

Handbook of
THE PHILOSOPHY OF SCIENCE

General Editors: DOV M. GABBAY, PAUL THAGARD, AND JOHN WOODS

PHILOSOPHY *of* STATISTICS



Edited by **Prasanta S. Bandyopadhyay**
and **Malcolm R. Forster**



Philosophy of Statistics

Handbook of the Philosophy of Science

General Editors

Dov M. Gabbay
Paul Thagard
John Woods



AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
North Holland is an imprint of Elsevier



Handbook of Philosophy of Science

Volume 7

Philosophy of Statistics

Edited by

Prasanta S. Bandyopadhyay
Montana State University, USA

Malcolm R. Forster
Tsinghua University, China and
University of Wisconsin-Madison, USA



AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
North Holland is an imprint of Elsevier



North Holland is an imprint of Elsevier
The Boulevard, Langford lane, Kidlington, Oxford, OX5 1GB, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA

First edition 2011

Copyright © 2011 Elsevier B.V. All rights reserved

No part of this publication may be reproduced, stored in a retrieval system
or transmitted in any form or by any means electronic, mechanical, photocopying,
recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights
Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333;
email: permissions@elsevier.com. Alternatively you can submit your request online by
visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting
Obtaining permission to use Elsevier material

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

ISBN: 978-0-444-51862-0

ISSN: 0031-8019

For information on all North Holland publications visit our web site at elsevierdirect.com

Printed and bound in Great Britain

11 12 13 14 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabrc.org

ELSEVIER

BOOK AID
International

Sabre Foundation

In memory of M. A. Sattar
(PSB)

In loving memory of my parents Lyn Forster and Ray Forster
(MRF)

This page intentionally left blank

GENERAL PREFACE

Dov Gabbay, Paul Thagard, and John Woods

Whenever science operates at the cutting edge of what is known, it invariably runs into philosophical issues about the nature of knowledge and reality. Scientific controversies raise such questions as the relation of theory and experiment, the nature of explanation, and the extent to which science can approximate to the truth. Within particular sciences, special concerns arise about what exists and how it can be known, for example in physics about the nature of space and time, and in psychology about the nature of consciousness. Hence the philosophy of science is an essential part of the scientific investigation of the world.

In recent decades, philosophy of science has become an increasingly central part of philosophy in general. Although there are still philosophers who think that theories of knowledge and reality can be developed by pure reflection, much current philosophical work finds it necessary and valuable to take into account relevant scientific findings. For example, the philosophy of mind is now closely tied to empirical psychology, and political theory often intersects with economics. Thus philosophy of science provides a valuable bridge between philosophical and scientific inquiry.

More and more, the philosophy of science concerns itself not just with general issues about the nature and validity of science, but especially with particular issues that arise in specific sciences. Accordingly, we have organized this Handbook into many volumes reflecting the full range of current research in the philosophy of science. We invited volume editors who are fully involved in the specific sciences, and are delighted that they have solicited contributions by scientifically-informed philosophers and (in a few cases) philosophically-informed scientists. The result is the most comprehensive review ever provided of the philosophy of science.

Here are the volumes in the Handbook:

Philosophy of Science: Focal Issues, edited by Theo Kuipers.

Philosophy of Physics, edited by Jeremy Butterfield and John Earman.

Philosophy of Biology, edited by Mohan Matthen and Christopher Stephens.

Philosophy of Mathematics, edited by Andrew Irvine.

Philosophy of Logic, edited by Dale Jacquette.

Philosophy of Chemistry and Pharmacology, edited by Andrea Woody, Robin Hendry and Paul Needham.

Philosophy of Statistics, edited by Prasanta S. Bandyopadhyay and Malcolm Forster.

Philosophy of Information, edited by Pieter Adriaans and Johan van Benthem.

Philosophy of Technology and Engineering Sciences, edited by Anthonie Meijers.

Philosophy of Complex Systems, edited by Cliff Hooker.

Philosophy of Ecology, edited by Bryson Brown, Kent A. Peacock and Kevin deLaplante.

Philosophy of Psychology and Cognitive Science, edited by Paul Thagard.

Philosophy of Economics, edited by Uskali Mäki.

Philosophy of Linguistics, edited by Ruth Kempson, Tim Fernando and Nicholas Asher.

Philosophy of Anthropology and Sociology, edited by Stephen Turner and Mark Risjord.

Philosophy of Medicine, edited by Fred Gifford.

Details about the contents and publishing schedule of the volumes can be found at http://www.elsevier.com/wps/find/bookdescription.cws_home/BS_HPHS/description#description

As general editors, we are extremely grateful to the volume editors for arranging such a distinguished array of contributors and for managing their contributions. Production of these volumes has been a huge enterprise, and our warmest thanks go to Jane Spurr and Carol Woods for putting them together. Thanks also to Lauren Schultz and Derek Coleman at Elsevier for their support and direction.

PREFACE

Prasanta S. Bandyopadhyay and Malcolm R. Forster

“Nowadays,” complained the American statistician William Kurskal more than four decades ago, “writers calling themselves statisticians and those calling themselves philosophers of science often refer to each [other], but [their] communication is restricted and piecemeal.”¹ This volume aims to remedy these shortcomings which, unfortunately, have continued to plague the disciplines. It provides state-of-the-art research in the area of Philosophy of Statistics by encouraging numerous experts to communicate with one another without feeling “restricted” by their disciplines or thinking “piecemeal” in their treatment of issues. A second goal of this book is to present work in the field without bias toward any particular statistical paradigm.

Broadly speaking, the essays in this *Handbook* are concerned with problems of induction, statistics and probability. For centuries, foundational problems like induction have been among philosophers’ favorite topics; recently, however, non-philosophers have increasingly taken a keen interest in these issues. This volume accordingly contains papers by both philosophers and non-philosophers, including scholars from nine academic disciplines. In addition, while the *Handbook* deals primarily with Anglo-American analyses and data, it also includes various approaches by many non-Western authors. The contributors hail from more than ten countries spread over four continents, providing the volume with a valuable global dimension. Statistics, of course, affect virtually every aspect of human life everywhere, from Spain to China to Mauritius. Appreciating the power of statistics, many of the authors have utilized data far from their places of origin, and, to the book’s benefit, they are not hesitant to exploit this information to present their own interpretations of complex issues. Moreover, just as statistics has a theoretical side, so it also has an applied side. Several chapters in this volume have combined those two aspects. Additionally, we provide essays that are straight-forwardly historical and address both western and non-western views on probability, thus making available ramifications of statistics and probability in different cultures.

A few words are in order regarding our volume and the kind of assistance of other scholars. We relied on an international team of experts in preparing the papers offered here. The referees were extremely generous with their time, advice and suggestions. James Hawthorne did far more than his share of reading and offering valuable suggestions. Clark Glymour, Theodore Porter, Jeff Paris, Antony Eagle, David Hitchcock, Prabal K. Sen, Jose Miguel Ponciano, Mylavaram Deekshithulu,

¹W. Kruskal and J. Tanur, eds. *International Encyclopedia of Statistics*, Vol. 2, Free Press, Collier Macmillan Publishers, London, 1968, p.1082.

Jay Herson, Bijoy Mukherjee, and Oliver Schulte also deserve high praise. Without the help of these scholars, the *Handbook* would not have taken its present shape with such variety, complexity, depth and scope.

While writing and revising the Introduction, we received valuable help from the essays' authors regarding whether we correctly represented their views. Our drafts were shuttled back and forth between the authors and the editors until both were satisfied. The process enabled us to clarify our comments on their essays. We believe this makes the Introduction more representative of how both the authors and we the editors understand the contributors' arguments and contributions.

Although our *Handbook* is substantial in terms of its variety, scope, complexity, and depth, it is not designed to provide an all-encompassing Encyclopedia of Statistics. We were unable to include some significant topics like experimental designs, the identifiability problem in statistics, bootstrapping, and cross-validation. We hope that the readers will still find it useful and significant in terms of its contribution to numerous fields which apply and use statistics in a reflective manner.

The question remains whether scholars have been able to communicate more effectively across disciplinary divides since the publication of Kruskal's paper four decades ago. Some scholars have mixed views about the merit of this sort of interaction. One foremost philosophers of science noted that compared to twenty-five years ago, today's philosopher of science are considerably more dependent on knowing much more of the actual science. By implication, modern philosophy of science should be interdisciplinary, and that is good for the well-being of the subject. A statistician of great eminence, however, holds a different opinion about the interaction between statisticians and philosophers. Based on his experience, he notes that when statisticians and philosophers meet, they diverge like two logs floating in a river,² touching one another solely for the purpose of taking two different routes.³ What are we to make of these conflicting perspectives among experts? Ultimately, we leave it to the reader to evaluate whether this volume is able to make some headway in the direction of multi-disciplinary dialogue, and whether this dialogue is worthwhile for the well-being of these overlapping disciplines. However, we firmly believe that the effort at cross-disciplinary communication is a worthwhile endeavor.

It would be inappropriate to conclude the preface without recording our thanks to the people who were involved from its very inception. It would be more than an understatement if we claimed to be able to produce a volume of this size had we not leaned on many individuals for guidance and help. We thank John Woods, the series editor, for his guidance and suggestions. We are especially grateful to Jane Spurr who worked tirelessly toward the completion of this volume and responded patiently to our daily emails. At times, we began to feel the despair noted by

²This log analogy has been borrowed from the *Ramayana* (the "Ayodhya Section"), one of the epics of the Indian sub-continent.

³However, he does not like to sound so pessimistic. This came out more as his reaction to his experience of the present state of interactions between these two disciplines rather than what he believes theoretically possible about a better interaction between philosophy of science and statistics.

Woody Allen: “eternity is very long, especially towards the end.”⁴ During those trying times, Jane reassured us that the task was almost completed. We are truly indebted to her for both her ceaseless work and her dedication to the production of this volume.

We thank both A. Philip Dawid, Jayanta K. Ghosh, and Colin Howson for their suggestions regarding the volume, as well as our colleagues and friends for their support, including James Allard, Anindya Bandyopadhyay, Hironmay Banerjee, John G. Bennett, Gordon Brittan, Debiprasad Chattopadhyay, David Cherry, Steve Cherry, Abhijit Dasgupta, Simon Dixon, Arthur Falk, Dan Flory, Jack Gilchrist, Sanford Levy, Michelle Maskiell, Sue Monahan, Prodyot K. Mukhopadhyay, Mujib Rahman, Michael Reidy, Abdur Razzaque, Tasneem Sattar, Billy Smith and Mark Taper. We are also thankful to Derek Coleman and his team in Amsterdam and Mohana Natarajan and her team in Chennai for their constant assistance and suggestion with regard to printing style and publication quality. PSB acknowledges the help of his department’s three superb administrative assistants, Diane Cattrell, Deidre Manry, and Jessica Marks, for their smiling faces whenever he needed long printouts for the volume. PSB is also thankful to the Montana State University’s legal team for its expert opinion in spelling out some features of the contract. Last but not least, the generous support received by PSB from the Astrobiology Biogeocatalysis Research Centre (NASA grant #4w1781) and a 2007 summer grant from Montana State University’s Vice-Provost Office assisted him in carrying out this research. MRF thanks the University of Wisconsin-Madison and the University of Pittsburgh Centre for the Philosophy of Science for their support during the spring semester of 2006

PSB

MRF

25 November, 2010

⁴Quoted in Martin Rees’ *Just Six Numbers: The Deep Forces that Shape the Universe*. Basic Books, Great Britain, 2000, p.71.

This page intentionally left blank

CONTRIBUTORS

Davis Baird

Clark University, USA.

dbaird@clarku.edu

Prasanta S. Bandyopadhyay

Montana State University, USA.

psb@montana.edu

Deborah Bennett

New Jersey City University, USA.

dbennett@njcu.edu

José M. Bernardo

Universitat de València, Spain.

jose.m.bernardo@uv.es

Jeffrey D. Blume

Vanderbilt University School of Medicine, USA.

j.blume@vanderbilt.edu

Robert J. Boik

Montana State University, USA.

rjboik@math.montana.edu

Arijit Chakrabarti

Indian Statistical Institute, India.

arc@isical.ac.in

Richard Charnigo

University of Kentucky, USA.

rjcharn2@aol.com

Steve Cherry

Montana State University, USA.

cherry@math.montana.edu

Abhijit Dasgupta

University of Detroit Mercy, USA.

abhijit.dasgupta@udmercy.edu

A. Philip Dawid

Cambridge University, UK.

A.P.Dawid@statslab.cam.ac.uk

Michael Dickson

University of South Carolina, USA.

dickson@sc.edu

David L. Dowe

Monash University, Australia.

david~dot~dowe~at~infotech.monash.edu.au

Kenny Easwaran

University of Southern California, USA.

easwaran@usc.edu

Roberto Festa

University of Trieste, Italy.

festa@units.it

Malcolm R. Forster

Tsinghua University, China and University of Wisconsin-Madison, USA.

mforster@wisc.edu

Jayanta K. Ghosh

Indian Statistical Institute, India and Purdue University, USA.

ghosh@stat.purdue.edu

Sander Greenland

University of California, Los Angeles, USA.

lesdomes@ucla.edu

Mark C. Greenwood

Montana State University, USA.

greenwood@math.montana.edu

Jason Grossman

Australian National University, Australia.

jason.grossman@xenyn.net

Peter D. Grünwald

University of Leiden, The Netherlands.

peter.grunwald@cw.nl

Alan Hájek

Australian National University, Australia.

alanh@coombs.anu.edu.au

Gilbert Harman

Princeton University, USA.

harman@princeton.edu

Joel Harper

University of Montana, USA.

joel@mso.umt.edu

James Hawthorne

University of Oklahoma, USA.

hawthorne@ou.edu

Colin Howson

University of Toronto, Canada.
c.howson@lse.ac.uk

Kevin T. Kelly

Carnegie Mellon University, USA.
kk3n@andrew.cmu.edu

Sanjeev Kulkarni

Princeton University, USA.
kulkarni@princeton.edu

Subhash R. Lele

University of Alberta, Canada.
slele@ualberta.ca

Deborah G. Mayo

Virginia Tech, USA.
mayod@vt.edu

Johnnie Moore

University of Montana, USA.
johnnie.moore@umontana.edu

John D. Norton

University of Pittsburgh, USA.
jdnorton@pitt.edu

C. K. Raju

Universiti Sains Malaysia, Malaysia.
c_k_raju@hotmail.com

Jan-Willem Romeijn

University of Groningen, The Netherlands.
j.w.romeijn@rug.nl

Steven de Rooij

University of Cambridge, UK.
s.deRooij@statslab.cam.ac.uk

Elliott Sober

University of Wisconsin, USA.
ersoer@wisc.edu

Aris Spanos

Virginia Tech, USA.
aris@vt.edu

Peter Spirtes

Carnegie Mellon University, USA.
ps7z@andrew.cmu.edu

Cidambi Srinivasan

University of Kentucky, USA.
srini@ms.uky.edu

Daniel Steel

Michigan State University, USA.
steel@msu.edu

Mark L. Taper

Montana State University, USA.
MarkLTaper@gmail.com

Choh Man Teng

Institute for Human and Machine Cognition, USA.
cmteng@ihmc.us

C. Andy Tsao

Natioal Dong Hwa University, Taiwan.
chtsao@mail.ndhu.edu.tw

Susan Vineberg

Wayne State University, USA.
susan.vineberg@wayne.edu

Paul Weirich

University of Missouri, USA.
weirichp@missouri.edu

Gregory Wheeler

New University of Lisbon, Porugal.
grw@fct.unl.pt

Jon Williamson

University of Kent, UK.
j.williamson@kent.ac.uk

Sandy L. Zabell

Northwestern University, USA.
zabell@math.northwestern.edu

CONTENTS

General Preface	vii
Dov Gabbay, Paul Thagard, and John Woods	
Preface	ix
Prasanta S. Bandyopadhyay and Malcolm R. Forster	
List of Contributors	xiii
Introduction	
Philosophy of Statistics: An Introduction	1
Prasanta S. Bandyopadhyay and Malcolm R. Forster	
Part I. Probability & Statistics	
Elementary Probability and Statistics: A Primer	53
Prasanta S. Bandyopadhyay and Steve Cherry	
Part II. Philosophical Controversies about Conditional Probability	
Conditional Probability	99
Alan Hájek	
The Varieties of Conditional Probability	137
Kenny Easwaran	
Part III. Four Paradigms of Statistics	
Classical Statistics Paradigm	
Error Statistics	153
Deborah G. Mayo and Aris Spanos	
Significance Testing	199
Michael Dickson and Davis Baird	
Bayesian Paradigm	
The Bayesian Decision-Theoretic Approach to Statistics	233
Paul Weirich	

Modern Bayesian Inference: Foundations and Objective Methods José M. Bernardo	263
Evidential Probability and Objective Bayesian Epistemology Gregory Wheeler and Jon Williamson	307
Confirmation Theory James Hawthorne	333
Challenges to Bayesian Confirmation Theory John D. Norton	391
Bayesianism as a Pure Logic of Inference Colin Howson	441
Bayesian Inductive Logic, Verisimilitude, and Statistics Roberto Festa	473
Likelihood Paradigm	
Likelihood and its Evidential Framework Jeffrey D. Blume	493
Evidence, Evidence Functions, and Error Probabilities Mark L. Taper and Subhash R. Lele	513
Akaikean Paradigm	
AIC Scores as Evidence — a Bayesian Interpretation Malcolm Forster and Elliott Sober	535
Part IV: The Likelihood Principle	
The Likelihood Principle Jason Grossman	553
Part V: Recent Advances in Model Selection	
AIC, BIC and Recent Advances in Model Selection Arijit Chakrabarti and Jayanta K. Ghosh	583
Posterior Model Probabilities A. Philip Dawid	607

Part VI: Attempts to Understand Different Aspects of “Randomness”

Defining Randomness 633
Deborah Bennett

Mathematical Foundations of Randomness 641
Abhijit Dasgupta

Part VII: Probabilistic and Statistical Paradoxes

Paradoxes of Probability 713
Susan Vineberg

Statistical Paradoxes: Take It to The Limit 737
C. Andy Tsao

Part VIII: Statistics and Inductive Inference

Statistics as Inductive Inference 751
Jan-Willem Romeijn

Part IX: Various Issues about Causal Inference

Common Cause in Causal Inference 777
Peter Spirtes

The Logic and Philosophy of Causal Inference: A Statistical Perspective 813
Sander Greenland

Part X: Some Philosophical Issues Concerning Statistical Learning Theory

Statistical Learning Theory as a Framework for the Philosophy of Induction 833
Gilbert Harman and Sanjeev Kulkarni

Testability and Statistical Learning Theory 849
Daniel Steel

Part XI: Different Approaches to Simplicity Related to Inference and Truth

Luckiness and Regret in Minimum Description Length Inference 865
Steven de Rooij and Peter D. Grünwald

MML, Hybrid Bayesian Network Graphical Models, Statistical 901
Consistency, Invariance and Uniqueness
David L. Dowe

Simplicity, Truth and Probability 983
Kevin T. Kelly

Part XII: Special Problems in Statistics/Computer Science

Normal Approximations 1027
Robert J. Boik

Stein's Phenomenon 1073
Richard Charnigo and Cidambi Srinivasan

Data, Data, Everywhere: Statistical Issues in Data Mining 1099
Choh Man Teng

Part XIII: An Application of Statistics to Climate Change

An Application of Statistics in Climate Change: Detection of 1121
Nonlinear Changes in a Streamflow Timing Measure in the
Columbia and Missouri Headwaters
Mark C. Greenwood, Joel Harper and Johnnie Moore

Part XIV: Historical Approaches to Probability/Statistics

The Subjective and the Objective 1149
Sandy L. Zabell

Probability in Ancient India 1175
C. K. Raju

Index 1197

PHILOSOPHY OF STATISTICS: AN INTRODUCTION

Prasanta S. Bandyopadhyay and Malcolm R. Forster

1 PHILOSOPHY, STATISTICS, AND PHILOSOPHY OF STATISTICS

The expression “philosophy of statistics” contains two key terms: “philosophy” and “statistics.” Although it is hard to define those terms precisely, they convey some intuitive meanings. For our present purpose, those intuitive meanings are a good place to embark on our journey. Philosophy has a broader scope than the specific sciences. It is concerned with general principles and issues. In contrast, “statistics” is a specific branch of knowledge that, among many other activities, includes addressing reliable ways of gathering data and making inferences based on them. Perhaps the single most important topic in statistics is how to make reliable inferences. As a result, statisticians are interested in knowing which tools to use and what mechanisms to employ in making and correcting our inferences. In this sense, the general problem of statistics is very much like the problem of induction in which philosophers have long been interested. In fact, statisticians as diverse as Ronald Fisher [1973], Jerzy Neyman [1967] and Bruno de Finetti [1964] characterized the approaches they originated as methods for inductive inferences.^{1,2}

Before we begin our discussion, it is worthwhile mentioning a couple of salient features of this volume. It contains thirty-seven new papers written by forty-eight authors coming from several fields of expertise. They include philosophy, statistics, mathematics, computer science, economics, ecology, electrical engineering, epidemiology, and geo-science. In the introduction, we will provide an outline of each paper without trying to offer expert commentary on all of them. Our emphasis on some topics rather than others in the following discussion reflects our own interest and focus without downplaying the significance of those topics less discussed. We encourage readers to start with the paper(s) that kindles their interest and lie within their own research areas.

In the western world, David Hume [1739] has been credited with formulating the problem of induction in a particularly compelling fashion. The problem of induction arises when one makes an inference about an unobserved body of data

¹See [Seidenfeld, 1979] for a close look at Fisher’s views on statistical inference.

²For a clear discussion of de Finetti’s view on the connection between the subjective degree of belief and inductive learning, see [Skyrms, 1984].

based on an observed body of data.³ However, there is no assurance that the inference in question will be valid because the next datum we observe may differ from those already gathered. Furthermore, to assume that they will be the same, according to Hume, is to assume what we need to prove. The problem of induction in this sense has very possibly turned out to be an irresolvable problem. Instead of addressing the problem of induction in the way Hume has described it, in terms of certainty, we are more often interested in knowing how or whether we would be able to make better inductive inferences in the sense that they are likely to be true most of the time, that is, in terms of reliability.

However, to be able to make reliable inferences we still require substantial assumptions about the relationship between the observed data and the unobserved data. One such assumption, which is sometimes called “the uniformity of nature” assumption and was questioned by Hume, is that the future data will be like the past data. In addition, we sometime also make some empirical assumptions about the world. One such assumption is that the world is simple, such as in the sense of being isotropic that there are laws that apply across all points in space and time, or at least across the domain of interest. For philosophers, this assumption is, in some sense, reminiscent of the assumption involved in the contrast between the green hypothesis (i.e., all emeralds are green) and the grue-hypothesis (i.e., all emeralds are grue.) X is defined as grue if and only if x is green and was observed before time t or x is blue and was not observed before t . The hypotheses are equally consistent with current data, even though the hypotheses are different. So, which hypothesis should we choose given they are equally supported by the data? The grue-green problem teaches us that we do commonly make (unusually unexamined) assumptions about the concepts we can use and the way we can use them in constructing hypotheses. That is, we make (often unarticulated) assumptions about the ways in which the world is uniform or simple.

As there is a quandary over whether simplicity is a legitimate factor in scientific inference, so there is another heated discussion regarding what kinds of assumptions other than the uniformity of nature assumption, and perhaps also simplicity, are warranted. These debates over the correct assumptions and approaches to inductive inference are as rampant in the choice of one’s statistical paradigm (for example, classical/error statistics, Bayesianism, likelihoodism or the Akaikean framework, to mention only the most prominent) as they are in the applied approaches to automated inductive inference in computer science. In philosophy of statistics, we are interested in the foundational questions including the debates about statistical paradigms regarding which one provides the right direction and method for carrying out statistical inference, if indeed any of them do in general. On one end of the spectrum, we will discuss several major statistical paradigms with their assumptions and viewpoints. On the other end, we also consider broader issues — like the issue of inductive inference — after understanding these statistical paradigms in more detail. We are likewise interested in

³Taleb has adopted a distinctively unique approach to the issues concerning induction especially when they involve the 9/11 event and the Wall-Street crash of 2008 [Taleb, 2010].

specific questions between these two extremes, as well as more modern viewpoints that adopt a “tool-kit” perspective. In that tool-kit perspective, each paradigm is merely a collection of inferential tools with limitations and appropriate (as well as inappropriate) domain of application. We will thus consider issues including the following: the causal inference in observational studies; the recent advances in model selection criteria; such foundational questions as “whether one should accept the likelihood principle (LP)” and “what is conditional probability”; the nature of statistical/probabilistic paradoxes, the problems associated with understanding the notion of randomness, the Stein phenomenon, general problems in data mining, and a number of applied and historical issues in probability and statistics.

2 FOUR STATISTICAL PARADIGMS

Sometimes different approaches to scientific inference among philosophers, computer scientists, and statisticians stem from their adherence to competing statistical paradigms. Four statistical paradigms will be considered in the following discussion: (1) classical statistics or error statistics, (ii) Bayesian statistics, (iii) likelihood-based statistics, and (iv) the Akaikean-Information Criterion-based statistics. How do they differ in their approaches to statistical inference?

2.1 *Four statistical paradigms and four types of questions*

To address this question, consider two hypotheses: H , representing that a patient suffers from tuberculosis, and $\sim H$, its denial. Assume that an Xray, which is administered as a routine test, comes out positive for the patient. Based on this simple scenario, following the work of Richard Royall, one could pose three questions that underline the epistemological issue at stake within three competing statistical paradigms [Royall, 1997]:

- (i) Given the datum, what should we *believe* and to what *degree*?
- (ii) What does the datum say regarding *evidence* for H against its alternative?
- (iii) Given the datum what should we *do*?

The first question we call the *belief question*, the second the *evidence question*, and the third the *decision question*. Royall thinks that Bayesians address the belief question, that classical/error statistics address the decision question, and that only the likelihood program addresses the evidence question. Sharpening Royall’s evidence question, the AIC framework can be taken to be addressing what we call the *prediction question*:

(iv) What does the datum tell us about the *predictive* accuracy of the hypothesis?

We will discuss how four statistical paradigms revolve round these four types of questions: (i) the belief question, (ii) the evidence question, (iii) what to do question, and (iv) the prediction question.

2.2 *Classical statistics/error statistics paradigm*

Deborah Mayo and Aris Spanos have put forward an error statistical approach to statistical inference. Unlike Royall, however, they think that error statistics is successful in addressing the evidence question as well as the decision question. Error statistics provides both a tool-kit for doing statistics as well as advancing a philosophy of science in which probability plays a key role in arriving at reliable inferences and severe tests. Mayo and Spanos have proposed a detailed account of the severity of testing within error statistical framework. Suppose George measures his weight on a scale on two dates, and considers the hypothesis H that he has gained no more than 3 pounds during that time. If the measured difference is one pound, and the scale is known to be sensitive to the addition of a 0.1 pound potato, then we can say that the hypothesis has survived a severe test because the measured difference would have been greater if the hypothesis were false, and he had gained more than 3 pounds. The correct justification for such an inference is not that it would rarely be wrong in the long run of repetitions of weighing as on the strict behavioristic interpretation of tests. Instead, they argue, it is because the test had a very high capacity to have detected the falsity of H , and did not. Focusing on a familiar one sided Normal test of the mean, Mayo and Spanos show how a severity interpretation of tests addresses each of the criticisms often raised against the use of error probabilities in significance tests. Pre-data, error probabilities ensure that a rejection indicates with severity some discrepancy from the null, and that failing to reject the null rules out with severity those alternatives against which the test has high power. Post-data, one can go much further in determining the discrepancies from the null warranted by the actual data in hand. This is the linchpin of their error statistical philosophy of statistics.⁴

Taking frequency statistics as crucial for understanding significance tests, Michael Dickson and Davis Baird have discussed how, on numerous occasions in social science literature, significance tests have been used, misused and abused without implying that a well-designed significance test may not have any value. Both authors have explored the history of the use of significance tests, including the controversy between Mendelians and Darwinists in examining Mendel's work from a statistical perspective. In this regard, they discuss how Ronald Fisher, while attempting to reconcile the debates between Mendelians and Darwinists, came to realize that Mendel's report on 600 plants is questionable since the data

⁴We have lumped classical statistics with Fisher's significance testing following some of the authors of this volume. Historically, however, classical statistics is distinguished from the theory of significance testing. We owe this point of clarification to Sander Greenland.

that report exploited “were too good to be true”. One of the issues both Dickson and Baird wonder about is how auxiliaries along with the hypotheses are tested within this significance test framework. In a theory testing framework, philosophers are usually concerned about how or when a theory is confirmed or disconfirmed. According to P. Duhem and W. V. Quine and, a theory is confirmed or disconfirmed as a whole along with its auxiliaries and background information. Their view is known not surprisingly as the Duhem-Quine thesis. In the original Duhem’s statement: “[a]n experiment in physics can never condemn an isolated hypothesis but only a whole theoretical group.” Dickson and Baird wonder (without offering a solution) how or whether significance testing could contribute to our understanding of drawing inferences from a body of evidence within the context of the Duhem-Quine thesis.

2.3 Bayesian statistics paradigm

It is generally agreed by its supporters and critics alike that Bayesianism⁵ currently is the dominant view in the philosophy of science. Some statisticians have gone further, conjecturing years ago that Bayesian statistics will be *the* dominant statistics for the twenty-first century. Whether this claim can be substantiated is beyond the scope of this introduction. However, it is uncontested that the Bayesian paradigm has been playing a central role in such disciplines as philosophy, statistics, computer science, and even jurisprudence.

Bayesians are broadly divided into subjective and objective categories. According to all Bayesians, an agent’s belief must satisfy the rules of the probability calculus. Otherwise, in accordance with the familiar “Dutch Book” argument, the agent’s degree of belief is incoherent. Subjective Bayesians take this (probabilistic) coherence to be both a necessary and a sufficient condition for the rationality of an agent’s beliefs, and then (typically) argue that the beliefs of rational agents will converge over time. The point of scientific inference, and the source of its “objectivity,” is to guarantee coherence and ensure convergence. Objective Bayesians, on the other hand, *typically* insist that while the coherence condition is necessary, it is not also sufficient for the kind of objectivity which scientific methodologies are intended to make possible.

Paul Weirich’s paper in this volume focuses primarily on subjective probability. Weirich has developed a Bayesian decision theoretic approach where he considers how an agent’s beliefs can be revised in light of data. Probabilities represent an agent’s degree of belief. Weirich evaluates several charges against Bayesians. According to one objection he has considered, Bayesianism allows an agent’s degrees of belief to be anything as long as they satisfy the probability calculus. Weirich takes the objection to be implying that Bayesian subjective probabilities must represent an agent’s idiosyncratic beliefs. He has, however, rejected permissive Bayesianism in favor of his version of Bayesianism. The notion of conditional

⁵Howson and Urbach’s [2006] book, which is now a classic, has provided a clear account of Bayesian view.

probability on which the principle of conditionalization rests is central for him. According to this principle, an agent should update her degree of belief in a hypothesis (H) in light of data (D) in accordance with the principle of conditionalization, which says that her degree of belief in H after the data is known is given by the conditional probability $P(H|D) = P(H \& D)/P(D)$, assuming that $P(D)$ is not zero. Weirich also evaluates charges brought against the use of the principle of conditionalization. Finally, he compares Bayesian statistical decision theory with classical statistics, concluding his paper with an evaluation of the latter.

One central area of research in the philosophy of science is Bayesian confirmation theory. James Hawthorne takes Bayesian confirmation theory to provide a logic of how evidence distinguishes among competing hypotheses or theories. He argues that it is misleading to identify Bayesian confirmation theory with the subjective account of probability. Rather, any account that represents the degree to which a hypothesis are supported by evidence as a conditional probability of the hypothesis on the evidence, where the probability function involved satisfies the usual probabilistic axioms, will be a Bayesian confirmation theory, regardless of the interpretation of the notion of probability it employs. For, on any such account Bayes' theorem will express how *what hypotheses say about evidence* (via the likelihoods) influences *the degree to which hypotheses are supported by evidence* (via posterior probabilities). Hawthorne argues that the usual subjective interpretation of the probabilistic confirmation function is severely challenged by extended versions of the *problem of old evidence*. He shows that on the usual subjectivist interpretation even trivial information an agent may learn about an evidence claim may completely undermine the objectivity of the likelihoods. Thus, insofar as the likelihoods are supposed to be objective (or intersubjectively agreed), the confirmation function cannot bear the usual subjectivist reading. Hawthorne does take prior probabilities to depend on plausibility assessments, but argues that such assessments are not *merely subjective*, and that Bayesian confirmation theory is not severely handicapped by the sort of subjectivity involved in such assessments. He bases the latter claim on a powerful Bayesian convergence result, which he calls the *likelihood ratio convergence theorem*. This theorem depends only on likelihoods, not on prior probabilities; and it's a *weak law of large numbers* result that supplies explicit bounds on the rate of convergence. It shows that as evidence increases, it becomes highly likely that the evidential outcomes will be such as to make the likelihood ratios come to strongly favor a true hypothesis over each evidentially distinguishable competitor. Thus, any two confirmation functions (employed by different agents) that agree on likelihoods but differ on prior probabilities for hypotheses (provided the prior for the true hypothesis is not too near 0) will tend to produce likelihood ratios that bring posterior probabilities to converge towards 0 for false hypotheses and towards 1 for the true alternative.⁶

⁶Our readers might wonder if Hawthorne's Likelihood Ratio Convergence Theorem (LRCT) is different from de Finetti's theorem since in both theorems, in some sense, the swamping of prior probabilities can occur. According to de Finetti's theorem, if two agents are not pigheaded — meaning that neither of them assign extreme (0 or 1) but divergent probabilities to a hypothesis

John D. Norton seeks to provide a counterbalance to the now dominant view that Bayesian confirmation theory has succeeded in finding the universal logic that

(event) and that they accept the property of exchangeability for experiments — then as the data become larger and larger, they tend to assign the same probability value to the hypothesis (event) in question. In fact, Hawthorne's LRCT is different from de Finetti's theorem. There are several reasons for regarding why they are different.

(i) de Finetti's theorem assumes exchangeability, which is weaker than probabilistic independence, but exchangeability also entails that the evidential events are "identically distributed" — i.e. each possible experiment (or observation) in the data stream has the same number of "possible outcomes," and for each such possible outcome of one experiment (or observation) there is a "possible outcome" for each other experiment (or observation) that has the same probability. The version of the LRCT assumes the probabilistic independence of the outcomes relative to each alternative hypothesis (which is stronger than exchangeability in one respect), but does not assume "identically distributed experiments (or observations)." That is, each hypothesis or theory under consideration may contain within it any number of different statistical hypotheses about entirely different kinds of events, and the data stream may draw on such events, to which the various statistical hypotheses within each theory apply.

de Finetti's theorem would only apply (for example) to repeated tosses of a single coin (or of multiple coins, but where each coin is weighted in the same way). The theorem implies that regardless of an agent's prior probabilities about how the coin is weighted, the agent's posterior probabilities would "come to agree" with the posterior probabilities of those who started with different priors than the agent in question. By contrast, the LRCT could apply to alternative theories about how "weighting a coin" will influence its chances of coming up heads. Each alternative hypothesis (or theory) about how distributing the weight influences the chances of "heads" will give an alternative (competing) formula for determining chances from weightings. Thus, each alternative hypothesis implies a whole collection of different statistical hypotheses in which each statistical hypothesis will represent each different way of weighting the coin. So testing two alternative hypotheses of this sort against each other would involve testing one collection of statistical hypotheses (due to various weightings) against an alternative collection of statistical hypotheses (that propose different chances of heads on those same weightings). de Finetti's theorem cannot apply to such hypotheses, because testing hypotheses about how weightings influence chances (using various different weightings of the coins) will not involve events that are all exchangeable with one another. (ii) de Finetti's theorem need not be about testing one scientific hypothesis or theory against another, but the LRCT is about that. The LRCT shows that sufficient evidence is very likely to make the likelihood ratio that compares a false hypothesis (or theory) to a true hypothesis (or theory) very small — thus, favoring the true hypothesis. The LRCT itself only depends on likelihoods, not on prior probabilities. But the LRCT does imply that if the prior probability of the true hypothesis isn't too near 0, then the posterior probability of false competitors will be driven to 0 by the diminishing likelihood ratios, and as that happens the posterior probability of the true hypothesis goes towards 1. So the LRCT could be regarded as a "convergence to truth" result in the sense de Finetti's theorem is. The latter shows that a sufficient amount of evidence will yield posterior probabilities that effectively act as though there is an underlying simple statistical hypothesis governing the experiments; where, with regard to that underlying statistical hypothesis, the experiments will be independent and identically distributed. Regardless of the various "prior probability distributions" over the alternative possible simple statistical hypotheses, there will be "converge to agreement" result on the best statistical model (the best simple statistical hypothesis) that models the exchangeable (identically distributed) events as though they were independent (identically distributed) events. But de Finetti didn't think of this as "convergence to a true statistical hypothesis". He seemed to think of it as converging to the best instrumental model.

Summing up, the LRCT is more general than de Finetti's theorem in that the former applies to all statistical theories, not just to those that consist of a single simple statistical hypothesis that only accounts for "repetitions of the same kinds of experiments" that have the same (but unknown) statistical distribution. We owe this point to Hawthorne (in an email communication).

governs evidence and its inductive bearing in science. He allows that Bayesians have good reasons for optimism. Where many others have failed, their system succeeds in specifying a precise calculus, in explicating the inductive principles of other accounts and in combining them into a single consistent theory. However, he urges, its dominance arose only recently in the centuries of Bayesian theorizing and may not last given the persistence of the problems it faces.

Many of the problems Norton identifies for Bayesian confirmation theory concern technicalities that our readers may find more or less troubling. In his view, the most serious challenge stems from the Bayesian aspiration to provide a complete account of inductive inference that traces our inductive reasoning back to an initial, neutral state, prior to the incorporation of any evidence. What defeats this aspiration, according to Norton, is the well-known, recalcitrant problem of the priors, recounted in two forms in his chapter. In one form, the problem is that the posterior $P(H|D\&B)$, which expresses the inductive support of data D for hypothesis H in conjunction with background information B , is fixed completely by the two “prior” probabilities, $P(H\&D|B)$ and $P(D|B)$. If one is a subjectivist and holds that the prior probabilities can be selected at whim, subject only to the axioms of the probability calculus, then, according to Norton, the posterior $P(H|D\&B)$ can never be freed of those whims. Or if one is an objectivist and holds that there can be only one correct prior in each specific situation, then, as explained in his chapter, the additivity of a probability measure precludes one assigning truly “informationless priors.” That is for the better, according to Norton, since a truly informationless prior would assign the same value to every contingent proposition in the algebra. The functional dependence of a posterior on the priors would then force all non-trivial posteriors to a single, informationless value. Hence, a Bayesian account can be non-trivial, Norton contends, only if it begins with a rich prior probability distribution whose inductive content is provided by other, non-Bayesian means.

Three papers in the volume explore the possibility that Bayesian account could be shown as a form of logic. Colin Howson contends that Bayesianism is a form of deductive logic of inference, while Roberto Festa and Jan-Willem Romeijn contend that Bayesian theory can be cast in the form of inductive inference. To investigate whether Bayesian account can be regarded as a form of deductive inference, Howson looks briefly at the last three hundred’s years of scientific inference and then focuses on why he thinks that Bayesian inference should be considered a form of pure logic of inference. Taking into account the debate over whether probabilistic inference can be regarded as logic of consistency or coherence, he discusses de Finetti’s theory of probability where de Finetti took the theory of probability to say nothing about the world, but takes it as a “logic of uncertainty.” One motivating reason to consider why Bayesian inference should be taken as a logic of pure logic is to note his disagreement with Kyburg’s distinction between the expression “consistency” to be applicable to a system which contains no two inconsistent beliefs and the expression “coherence” to be applicable to degrees of belief. For Howson, the analogy with deductive logic is between the latter imposing con-

sistency constraints on truth-evaluations and the rules of the probability theory imposing constraints in degree of belief. The remainder of his paper is devoted to developing and interpreting Bayesian inference as a form of pure logic of inference.

Both Festa and Romeijn regret that in the past century statistics and inductive inference have developed and flourished more or less independently of one another, without clear signs of symbiosis. Festa zooms in on Bayesian statistics and the Carnap's theory of inductive probabilities, and shows that in spite of their different conceptual bases, the methods worked out within the latter are essentially identical to those used within the former. He argues that some concepts and methods of inductive logic may be applied in the rational reconstruction of several statistical notions and procedures. According to him, inductive logic suggests some new methods which can be used for different kinds of statistical inference involving analogical considerations. Finally, Festa shows how a Bayesian version of truth approximation can be developed and integrated into a statistical framework.⁷

Romeijn also investigates the relationship between statistics and inductive logic. Although inductive logic and statistics have developed separately, Romeijn thinks, like Festa, that it is time to explore the interrelationship between the two. In his paper, he investigates whether it is possible to represent various modes of statistical inference in terms of inductive logic. Romeijn considers three key ideas in statistics to forge the link. They are (i) Neyman-Pearson hypothesis testing (NPTH), (ii) maximum-likelihood estimation, and (iii) Bayesian statistics. Romeijn shows, using both Carnapian and Bayesian inductive logic, that the last of two of these ideas (i.e., maximum-likelihood estimation and Bayesian statistics) can be represented naturally in terms of a non-ampliative inductive logic. In the final section of his chapter, NPTH is joined to Bayesian inductive logic by means of interval-based probabilities over the statistical hypotheses.

As there are subjective Bayesians so there are objective Bayesians. José Bernardo is one of them. Since many philosophers are not generally aware of Bernardo's work, we will devote a relatively longer discussion to it. Bernardo writes that "[i]t has become standard practice, . . . , to describe as 'objective' any statistical analysis which only depends on the [statistical] model assumed. In this precise sense (and only in this sense) reference analysis is a method to produce 'objective' Bayesian inference" [Bernardo, 2005].

For Bernardo, the reference analysis that he has advocated to promote his brand of objective Bayesianism should be understood in terms of some parametric model of the form $M \equiv \{P(x|w), x \in X, w \in \Omega\}$, which describes the conditions under which data have been generated. Here, the data x are assumed to consist of one observation of the random process $x \in X$ with probability distribution $P(x|w)$ for some $w \in \Omega$. A parametric model is an instance of a statistical model. Bernardo defines $\theta = \theta(w) \in \Theta$ to be some vector of interest. All legitimate Bayesian inferences about the value θ are captured in its posterior distribution $P(\theta|x) \propto \int_{\Lambda} P(x|\theta, \lambda)P(\theta, \lambda)d\lambda$ provided these inferences are made under an assumed model.

⁷For an excellent discussion on the history of inductive probability, see Zabell's series of papers [2005].

Here, λ is some vector of nuisance parameters and is often referred to as “model” $P(x|\lambda)$.

The attraction of this kind of objectivism is its emphasis on “reference analysis,” which with the help of statistical tools has made further headway in turning its theme of objectivity into a respectable statistical school within Bayesianism. As Bernardo writes, “[r]eference analysis may be described as a method to derive model-based, non-subjective *posteriors*, based on the information-theoretical ideas, and intended to describe the inferential content of the data for scientific communication” [Bernardo, 1997]. Here by the “inferential content of the data” he means that the former provides “the basis for a method to derive non-subjective posteriors” (*Ibid*). Bernardo’s objective Bayesianism consists of the following claims.

First, he thinks that the agent’s background information should help the investigator build a statistical model, hence ultimately influence which prior the latter should assign to the model. Therefore, although Bernardo might endorse arriving at a unique probability value as a goal, he does not require that we need to have the unique probability assignment in all issues at our disposal. He writes, “[t]he analyst is supposed to have a *unique* (often subjective) prior $p(w)$, independently of the design of the experiment, but the scientific community will presumably be interested in comparing the corresponding analyst’s personal posterior with the *reference* (consensus) posterior associated to the published experimental design.” [Bernardo, 2005, p. 29, the first emphasis is ours]. Second, for Bernardo, statistical inference is nothing but a case of deciding among various models/theories, where decision includes, among other things, the utility of acting on the assumption of the model/theory being empirically adequate. Here, the utility of acting on the empirical adequacy of the model/theory in question might involve some loss function [Bernardo and Smith, 1994, p.69]. In his chapter for this volume, he has developed his version of objective Bayesianism and has addressed several charges raised against his account.

In their joint chapter, Gregory Wheeler and Jon Williamson have combined objective Bayesianism with Kyburg’s evidential theory of probability. This position of Bayesianism or any form of Bayesianism *seems* at odds at Kyburg’s approach to statistical inference that rests on his evidential theory of probability. We will consider Kyburg’s one argument against Bayesianism. Kyburg thinks that we should not regard partial beliefs as “degrees of belief” because (strict) Bayesians (like Savage) are associated with the assumption of a unique probability of a proposition. He discussed the interval-based probability as capturing our partial beliefs about uncertainty. Since the interval-based probability is not Bayesian, it follows that we are not allowed to treat partial beliefs as degrees of belief. Given this opposition between Kyburg’s view on probability and objective Bayesian view, Wheeler and Williamson have tried to show how core ideas of both these two views could be fruitfully accommodated within a single account of scientific inference.

To conclude our discussion on Bayesian position while keeping in mind Royall’s attribution of the belief question to Bayesians, many Bayesians would have mixed feelings about this attribution. To some extent some of them might consider it

to be inappropriately simple-minded. Howson would agree with this attribution with the observation that this would miss some of the nuances and subtleties of Bayesian theory. He broadly follows de Finetti's line in taking subjective evaluations of probability. These evaluations are usually called "degrees of belief." So to that extent he certainly thinks that there is a central role for degrees of belief, since after all they are what is referred to directly by the probability function. Therefore, according to him, the attribution of the belief question to Bayesians makes some sense. However, he thinks that the main body of the Bayesian theory consists in identifying the constraints which should be imposed on these to ensure their consistency/coherence. His paper has provided that framework for Bayesianism. Hawthorne might partially disagree with Royall since his Likelihood Ratio Convergence Theorem shows that how different agents could agree in the end even though they could very well start with varying degrees of belief in a theory. Both Weirich and Norton, although they belong to opposing camps insofar as their stances toward Bayesianism are concerned, might agree that Royall's attribution to Bayesians is after all justified. With regard to the prediction question, many Bayesians, including those who work within the confines of confirmation theory, would argue that an account of confirmation that responds to the belief question is able to handle the prediction question as, for Bayesians, the latter is a sub-class of the belief question.

2.4 *Likelihood-based statistics paradigm*

Another key statistical paradigm is the "likelihood framework," broadly construed. It stands between Bayesianism and error statistics (i.e., Frequentist, or hypothesis testing), because it focuses on what is common to both frameworks: the likelihood function. The data influence Bayesian calculations only by the way of the likelihood function, and this is also true of most Frequentist procedures. So Likelihoodists argue that one should simply interpret this function alone to understand what the data themselves say. In contrast, most Bayesians combine the likelihood function with a subjective prior probability about the hypotheses of interest and Frequentists combine the likelihood function with a decision theory understood in terms of two types of errors (See Mayo and Spanos's chapter for a discussion of those types of errors.) The Likelihood function is the same no matter which approach is taken and Likelihoodists have been quick to point out that historic disagreements between Bayesian and Frequentists can be traced back to these additions to the likelihood function.

Jeffrey Blume, however, argues that the disagreement between different statistical paradigms is really due to something more fundamental: the lack of an adequate conceptual framework for characterizing statistical evidence. He argues that any statistical paradigm purporting to measure the strength of statistical evidence in data must distinguish between the following three distinct concepts: (i) a measure of the strength of evidence ("How strong is the evidence for one hypothesis over another?"), (ii) the probability that a particular study will generate

misleading evidence (“what is the chance that the study will yield evidence that is misleading?”) and (iii) the probability that observed evidence — the collected data — is misleading (What is the chance that this final result is wrong or misleading?”) Blume uses the likelihood paradigm to show that these quantities are indeed conceptually and mathematically distinct. However, this framework clearly transcends any single paradigm. All three concepts, he contends, are essential to understanding statistics and its use in science for evaluating the strength of evidence in data for competing theories. One need to take note of the fact is that here “evidence” means “support for one hypothesis over its rival” without implying “support for a single hypothesis being true” or some other kind of truth-related virtues.

Before collecting any data, Blume contends, we should identify the mathematical tool that we will use to measure the evidence at the end of the study and we should report the tendency for that tool to be mistaken (i.e., to favor a false hypothesis over the true hypothesis). Once data are collected, we use that tool to calculate the strength of observed evidence. Then, we should report the tendency for those observed data to be mistaken (i.e., the chance that the observed data are favoring a false hypothesis over the true one). The subtle message is that the statistical properties of the study design are not the statistical properties of the observed data. Blume contends that this common mistake — attributing the properties of the data collection procedure to the data themselves — is partly to blame for those historic disagreements.

Let us continue with our earlier tuberculosis (TB) example to explore how Blume has dealt with these three distinct concepts, especially when they provide philosophers with an opportunity of investigating the notion of evidence from a different angle. Suppose it is known that the individuals with TB are positive 73.33% of the time and individuals without TB test positive only 2.85% of the time. As before, let H represent the simple hypothesis that an individual has tuberculosis and $\sim H$ the hypothesis that she does not. These two hypotheses are mutually exclusive and jointly exhaustive. Suppose our patient tests positive on an X-ray. Now, let D , for data, represents the positive X-ray result. Blume offers the likelihood ratio (LR) as a measure of the strength of evidence for one hypothesis over another. The LR is an example of the first concept; the first evidential quantity (EQ1).

$$LR = \left[\frac{P(D|H1)}{P(D|H2)} \right] \quad (E1)$$

According to the law of the likelihood, the data D , a positive X-ray in our example provide support for $H1$ over $H2$ if and only if their ratio is greater than one. An immediate corollary of equation 1 is that there is equal evidential support for both hypotheses only when $LR = 1$ (the likelihood ratios are always positive). Note that in (E1) if $1 < LR \leq 8$, then D is often said to provide weak evidence for $H1$ against $H2$, while when $LR > 8$, D provides fairly strong evidence. This benchmark, discussed in [Royall, 1997], holds the same meaning regardless of the context. In the tuberculosis example, the LR for $H1$ (presence of the disease)

over it $H2$ (absence of the disease) is $= (0.7333/0.0285) \approx 26$. So, the strength of evidence is strong and the hypothesis that the disease is present is supported by a factor of 26 over its alternative.

Now that we know how we will measure evidence, we turn our attention to the study design. Blume's second evidential concept (EQ2), the probability that a particular study design will generate misleading evidence, involves the *probability of a future event*. It is the probability that an investigator will collect a set of data that will yield an evidence measure in the wrong direction (i.e., that it will support the wrong hypothesis over the true one). To address the probability that a particular study design will generate misleading evidence, we examine the diagnostic properties of an X-ray for detecting TB (table 1 below).

Table 1. X-ray results

	Positive	Negative
Disease present	73.33%	26.67%
Disease is absent	2.85%	97.15%

The inference that a positive X-ray is evidence that the disease is present is correct regardless of a subject's true disease status. This is because the $LR = P(D|H1)/P(D|H2) = 0.733/0.0285 \approx 26$ regardless of the subject's true disease status. However, the positive X-rays can yield evidence that is misleading. That is, it is possible for a non-diseased patient to have a positive X-ray. Table 1 shows that the probability of the positive test being misleading given that the disease is absent is only 2.85%. So this test would be considered a good test in the sense that this happens only 2.85% of times given the hypothesis. This probability called "the probability of misleading evidence" is an example of the second evidential concept (EQ2). Of course, we never know if an observed positive test result is *truly* misleading because we never know the underlying disease status of the patient. However, we can determine *the tendency for* observed positive test results to be misleading and this may be helpful in judging the reliability of the observed data.

The third key concept (EQ3) involves the tendency for our final observed likelihood ratio to favor the false hypothesis. Unlike our discussion of EQ2 — the probability of observing misleading evidence — in the last paragraph, EQ3 conditions in the observed data set. At this stage, both the data and the likelihood ratio are fixed; it is the true hypothesis that is unknown. Therefore, we are interested in the probability that a hypothesis is true given the observed data. Back in our example, an observed positive result is misleading if and only if the subject in question does not have the disease. That quantity is $P(\sim H|D)$. Similarly, one could construe an observed negative test result to be misleading if and only if the subject does have the disease. That quantity is $P(H|\sim D)$. Both quantities, $P(\sim H|D)$, and $P(H|\sim D)$, can be computed using Bayes theorem as long we know the prevalence of tuberculosis in our testing population, $Pr(H)$, and the study design probabilities, $P(D|H)$ and $P(D|\sim H)$.

$P(H)$ is a prior probability and typically the specification of this probability would be subjective. But in our example it is easy to imagine that this probability is known and would be agreed upon by many experts. Here we will use a 1987 survey where there were 9.3 cases of tuberculosis per 100,000 population [Pagano and Gauvrau, 2000]. Consequently, $P(H) = 0.0093\%$ and $P(\sim H) = 99.9907\%$. From an application of Bayes Theorem we obtain $P(\sim H|D) = (0.028/0.0280) \approx 1$ and $P(H|\sim D) = 0.0025\%$. Thus the probability that an observed positive test result is misleading is nearly 100% because the disease is so rare. Nevertheless, our interpretation of the LR as strong evidence the disease is present is correct. It is wrong to argue that a positive test for TB is evidence that disease is absent! Blume and Royall [1997] independently provide less extreme examples, but this scenario shows why it is important to distinguish between the first and third evidential quantities (See Blume's chapter for the full reference to this paper.) A positive test result is evidence that the disease is present even though in this population a positive test is not all that reliable (this is not the case for all populations of course, just those with very rare diseases).

One of the problems of invoking prior probabilities, as we know, is the problem of subjectivity. The prior probabilities we have discussed in the diagnostic case study are frequency-based prior probabilities; as a result, the problem of subjectivity does not arise. However, there are often cases when the invocation of prior probabilities might be required to handle the third concept within the likelihood framework, thus potentially leading to subjectivity. At first this seems to pose a problem. However Blume shows that, for any prior, the probability that observed data are misleading is driven to zero as the likelihood grows. Large likelihood ratios are misleading less often. Therefore, the prior plays an important role in the starting point for this probability, but the data — acting through the likelihood ratio — will eventually drive this probability to zero. Thus, Blume argues, it is not critical to know the prior, since a large enough likelihood ratio will render the effect of any reasonable prior moot. In any case, we need to be aware of this dependence on which the third key concept might rely in some cases. Here, readers are invited to explore how Hawthorne's Bayesian Likelihood Ratio Convergence Theorem could be compared and contrasted with the result Blume has been discussing in regard to likelihood framework.

This third concept, the probability that observed evidence is misleading, helps bridge the gap between the likelihood and Bayesian frameworks: The Bayesian posterior probability is the probability that the observed evidence is misleading. What is missing in the Bayesian framework, Blume implies, are the first two concepts. Exactly what is the Bayesian measure of the strength of evidence and how often is that measure misleading? Likewise, Blume argues, the frequentists also must provide similar clarity. They cannot use a tail area probability for both the first and second quantities and then misinterpret that tail area probability as the third quantity once data are observed.

Taking a cue from Royall's work on likelihood framework and in some sense similar to Blume's paper, Mark Taper and Subhash Lele have proposed a more

general version of the likelihood framework that they call “evidentialism.” To give a motivation for why they think that likelihood framework is a special case of the latter here is some justification one of the authors provides elsewhere [Lele, 2004]. Lele defines a class of functions called “the evidence functions” to quantify the strength of evidence for one hypothesis over the other. He imposes some desiderata on this class of evidence functions, which could be regarded as epistemological conditions. Some of the conditions satisfied by the evidence function are:

1. The translation invariance condition: If one translates an evidence function by adding a constant to it to change the strength of the evidence, then the evidence function should remain unaffected by that addition of a constant.
2. The scale invariance condition: If one multiplies an evidence function by a constant to change the strength of evidence, then the evidence function should remain unaffected by that constant multiplier.
3. The reparameterization invariance: The evidence function must be invariant under reparameterization. It means that if there is an evidence function, $Ev1$ and the latter is reparameterized to $Ev2$, then both $Ev1$ and $Ev2$ must provide the identical quantification of the strength of evidence.
4. The invariance under transformation of the data: The evidence function should remain unaffected insofar as the quantification of the strength is concerned if one uses different measuring units.

Lele adds two more conditions on the evidence function, which he calls “regularity conditions,” so that the probability of strong evidence for the true hypothesis should converge to 1 as the sample size increases. He demonstrates that the likelihood ratio becomes an optimal measure of evidence under those epistemological and regularity conditions, providing a justification for the use of likelihood ratio as a measure of evidence. Lele believes that showing the optimality of the likelihood ratio amounts to providing necessary and sufficient conditions for the optimality of the law of likelihood. According to the law of likelihood, observation O favors $H1$ over $H2$ if and only if $P(O|H1) > P(O|H2)$. Taper and Lele also mention that information criteria, or at least order-consistent information criteria, are also evidence functions. By “order-consistent information criteria”, they mean that if the true model is in the model set, then sufficient amount of data should be able to find the true model.

Having established the relationship between both the likelihood and information criteria paradigms and evidentialism, Taper and Lele explore the likelihood and information criteria, another pillar of frequentist statistics-error statistics. They distinguish between two concepts: global reliability and local reliability. They seem to be sympathetic with some sort of reliable account of justification which, roughly speaking, states that strength of evidence is justified if and only if it has been produced by a reliable evidence producing mechanism. Given their opinion that scientific epistemology is, or at least should be, a public and not a private epistemology, Taper and Lele are no friends of Bayesians because they think Bayesian

subjectivity infects the objective enterprise of scientific knowledge⁸ accumulation. They contend that “evidentialism” of their variety will track the truth in the long run.⁹ According to them, global reliability describes the truth-tracking or error avoidance behavior of an inference procedure over all of its potential applications in the long run. They argue that global reliability measures commonly used in statistics are provided by Neyman/Pearson test sizes (α and β) and confidence interval levels. These measures, according to them, describe the reliability of the test procedures (and not individual inferences) as a global reliability that is a property of a long-run relative frequency of repeatable events. Royall’s probability of misleading evidence is a similar measure of the long run reliability of evidence procedures to produce evidence. Taper and Lele distinguish this global reliability measure from its local version, which is the acquisition of or arriving at the truth in a specific scenario. They think that Fisher’s p-values and Mayo and Spanos’ concept of severity provide a local reliability measure. The probability of obtaining a value for the test statistic that is as extreme, or more extreme, is called the p-value. According Mayo and Spanos the test passes with severity with respect to specific observed outcome relative to a specific test (see section 2.2 for Mayo and Spanos’s view on severity). Taper and Lele define the notion of local reliability of the evidence that they call “ M_L ”, as the probability of obtaining misleading evidence for one model over the alternative at least as strong as the observed evidence. The smaller “ M_L ” is, the greater one’s confidence that one’s evidence is not misleading. What is significant in their paper is that their concept of evidence and reliability are distinct, and that both may be used in making inference. This differs from an error statistical analysis such as given by Mayo and Spanos.

According to them, “ M_L ”, is clearly different from their notion of global reliability. It is also different from the third key concept (discussed in Blume’s chapter) that is interested in knowing the probability of the observed evidence to be misleading. To compute the probability of the observed evidence to be misleading, one needs to fall back on the posterior probability value; as a result, the measure associated with the third concept is open to the charge of subjectivity. However, M_L does not depend on any posterior probability calculation. In the remainder of the paper, they connect philosophers’ work on “reliability” with their evidentialism, although they think that the latter is a research program which is evolving and thriving. As a result, more research needs to be carried out before we would have a fully developed account of evidentialism of their variety.

⁸By “knowledge,” they must be meaning something different than what would be called “knowledge” in most epistemological literature in philosophy. They reject the truth of any comprehensible proposition of scientific interest and don’t believe that belief is a necessary component in any scientific enterprise concerning “knowledge.”

⁹Any reader familiar with epistemological literature in philosophy might wonder about the apparent problem of reconciling “reliability” with “evidentialism” since evidentialists have raised objections to any account of justification that rests on reliabilism. However, Taper and Lele’s senses of “reliabilism” and “evidentialism” may not be mapped exactly onto philosophers’ locution. Hence, there is less chance of worrying about an inconsistency in their likelihood framework.

2.5 *The Akaikean information-criterion-based statistics paradigm*

The last paradigm to be considered is the Akaikean paradigm. In philosophy of science, the Akaikean paradigm has emerged as a prominent research program due primarily to the work of Malcolm Forster and Elliott Sober on the curve-fitting problem [Forster and Sober, 1994]. An investigator comes across “noisy data” from which she would like to make a reliable inference about the underlying mechanism that has generated the data. The problem is to draw inferences about the “signal” behind the noise. Suppose that the signal is described in terms of some mathematical formula. If the formula has too many adjustable parameters, then it will begin to fit mainly to the noise. On the other hand, if the formula has too few adjustable parameters, then the formula provides a small family of curves, thereby reducing the flexibility needed to capture the signal itself. The problem in curve-fitting is to find a principled way of navigating between these undesirable extremes. In other words, how should one trade off the conflicting considerations of simplicity and goodness-of-fit. This is a problem that applies to any situation in which sets of equations with adjustable parameters, commonly called models, are fitted to data, and subsequently used for prediction.

In general, with regard to the curve-fitting problem, the goal of Forster and Sober is to measure the degree to which a model is able to capture the signal behind the noise, or equivalently, to maximizing the accuracy of predictions, since only the signal can be used for prediction, given that noise is unpredictable by its very nature. The Akaikean Information Criterion (AIC) is one possible way of achieving this goal. AIC assumes that a true probability distribution exists that generates independent data points. Even though the true probability distribution is unknown, AIC provides an unbiased estimate of the predictive accuracy of a family of curves that are fitted to a data set of that size under surprisingly weak assumptions about the nature of the true distribution.

Forster and Sober want to explain how we can predict future data from past data. An agent uses the available data to obtain the maximum likelihood estimates (MLE) of the parameters of the model under consideration, which yields a single “best-fitting curve” that is capable of making predictions. The question is how well this curve will perform in predicting unseen data. A model might fare well on one occasion, but fail to do so on another. The predictive accuracy of a model depends on how well it would do on average, were these processes repeated again and again.

AIC can be considered to be an approximately unbiased estimator for the predictive accuracy of a model. By an “unbiased estimator,” we mean that the estimator will equal the population parameter on average, with the average being relative to repeated random sampling of data of the same size as the actual data. Using AIC, the investigator attempts to select the model with minimum expected Kullback-Leibler (KL) distance across these repeated random samples, based on a single observed random sample. For each model, the AIC score is

$$\text{AIC} = -2\log(\hat{L}) + 2k \tag{E2}$$

Here “ k ” represents the number of adjustable parameters. The maximum likelihood, represented by $\hat{L}$, is simply the probability of the actual data given by the model fitted to the same data. The fact that the maximum likelihood term uses the same data twice (once to fit the model and once to determine the likelihood of that the fitted model) introduces a bias; the same curve will not fit unseen data quite as well. The penalty for complexity is introduced in order to correct for that bias.

Another goal of AIC is to minimize the average (expected) distance between the true distribution and the estimated distribution. It is equivalent to the goal of maximizing predictive accuracy. AIC was designed to minimize the KL distance between a fitted MLE model and the distribution actually generating the data. Specifically, AIC can be considered as an approximately unbiased estimator for the expected KL divergence which uses a MLE to create estimates for the parameters of the model.

With this background, it is worthwhile to consider their contribution to the volume where Forster and Sober analyze the Akaikean framework from a different perspective. AIC is a frequentist construct in the sense that AIC provides a criterion, or a rule of inference, that is evaluated according to the characteristics of its long-run performance in repeated instances. Bayesians find the use of the frequency-based criteria to be problematic. Since AIC is a frequentist construct, Bayesians worry about the foundation of AIC. What Forster and Sober have done in their chapter is to show that Bayesians can regard AIC scores as providing evidence for hypotheses about the predictive accuracies of models. A key point in their paper is to point out that this difference of evidential strength between hypotheses about predictive accuracy can be interpreted in terms of the law of likelihood (LL), which is something that Bayesians can accept. According to the LL, observation O favors $H1$ over $H2$ if and only if $P(O|H1) > P(O|H2)$. The secret is to take O to be the “observation” that AIC has a certain value and $H1$ and $H2$ to be competing “meta”-hypotheses about the predictive accuracy of a model (or of the difference in predictive accuracies between two models). According to Forster and Sober, Bayesians’ worry about AIC has turned out to be untenable because now AIC scores are evidence for hypotheses about predictive accuracy according to the LL which is one of the cornerstones of Bayesianism.

Revisiting the four types of questions, we find that the Akaike framework is capable of responding to the prediction question. Forster and Sober maintain that, in fact, it is able to do this in terms of the evidence question because the prediction question can be viewed a sub-class of the evidence question.

3 THE LIKELIHOOD PRINCIPLE

So far, we have considered four paradigms in statistics along with their varying and often conflicting approaches to statistical inference and evidence. However, we also want to discuss some of the most significant principles whose acceptance and rejection pinpoint the central disagreement among the four schools. The likelihood

principle (LP) is definitely one of the most important. In his chapter, Jason Grossman analyzes the nature and controversies surrounding the LP. Bayesians are fond of this principle. To know whether a theory is supported by data, according to Bayesians, LP says that the only part of data that is relevant is the likelihood of the actual data given the theory.

The LP is derivable from two principles; the first is the weak sufficiency principle (WSP) and the second is the weak conditionality principle (WCP). The WSP claims that if T is sufficient statistic and if $T(x_1) = T(x_2)$, then both x_1 and x_2 will provide equal evidential support. A statistic is sufficient, if given it, the distribution of any other statistic does not involve the unknown parameter θ . Both Bayesians and likelihoodists like both Royall and Blume accept the LP, whereas classical/error statisticians do not. Grossman clarifies the reason that classical/error statisticians question the principle. What data investigators might have obtained, but do not actually have, according to error statisticians, should play a significant role in evaluating statistical inference. Significance testing, hypothesis testing and confidence interval approach, which are tool-kits for classical/error statistics, violate the LP since they incorporate both actual and possible data in their calculation. In this sense, Grossman's chapter overlaps with many papers in the section on "Four Paradigms," especially Mayo and Spanos' paper in this volume.

Grossman also discusses the relationship between the law of likelihood (LL) and LP. They are different concepts; hence, supporting or denying one does not necessarily lead to supporting the other, and conversely. The LP only says where the information about the parameter θ is and suggests that the data summarization can be done through the likelihood function, in some way. It does not clearly indicate how to compare the evidential strength of two competing hypotheses. In contrast, the LL compares the evidential strength of two contending hypotheses.

4 THE CURVE-FITTING PROBLEM, PROBLEM OF INDUCTION, AND ROLE OF SIMPLICITY IN INFERENCE

The previous discussion of four statistical paradigms and how they address four types of questions provides a framework for evaluating several approaches to inductive inference. Two related problems that often confront any approach to inductive inference are the pattern recognition problem and the "the curve-fitting problem." In both cases one is confronted with two conflicting desiderata, "simplicity" and "goodness-of-fit." Numerous accounts within statistics and outside statistics have been proposed about how to understand these desiderata and to reconcile them in an optimal way.

Statistical learning theory (SLT) is one such approach that looks at these problems while being motivated by certain sets of themes. Learning from a pattern is a fundamental aspect of inductive inference. It becomes all the more significant if a theory is able to capture our learning via *pattern recognition* in our day to day life as this type of learning is not easily suited to systematic computer programming.

Suppose for example that we want to develop a system for recognizing whether a given visual image is an image of a cat. We would like to come up with a function from a specification of an image to a verdict, a function that maximizes the probability of a correct verdict. To achieve this goal, the system is given several examples of cases in which an “expert” has classified an image as of a cat or not of a cat. We have noted before that there is no assumption-free inference possible in the sense that we have to assume that data at hand must provide some clues about the future data. Note that this assumption is a version of the uniformity of nature assumption. In order for the investigator to generate examples, she assumes that there is an unknown probability distribution characterizing when particular images will be encountered and relating images and their correct classification. We assume that the new cases of the examples that we will come across are also randomly sampled from that probability distribution. This is similar to what is assumed in the Akaikean framework, discussed earlier.

We assume that the probability of the occurrence of an item with a certain characterization and classification is independent of the occurrence of other items and that the same probability distribution governs the occurrence of each item. The reason for having the probabilistic independence assumption is to imply that each new observation provides maximum information. The identical probability distribution assumption implies that each observation gives exactly the same information about the underlying probability distribution as any other. (These assumptions can be relaxed in various ways.) One central notion in SLT is the notion of VC-Dimension, *which is defined in terms of shattering*. A set of hypothesis S shatters certain data if and only if S is compatible with every way of classifying the data. That is, S shatters a given feature vectors if for every labeling of the feature vectors (e.g., as “cat” or “not cat”) the hypothesis in S generates this labeling. The finite VC-dimension of a set of rules C is the largest finite number N for which some set of N points is shattered by rules in C ; otherwise the VC-dimension is infinite. The VC-dimension of C provided a measure of the “complexity” of C . Various learning methods aim to choose a hypothesis in such a way as to minimize the expected error of prediction about the next batch of observations. In developing their account about inductive inference, Gilbert Harman and Sanjeev Kulkarni in their joint paper have contended that SLT has a lot to offer to philosophers by way of better understanding the problem of induction and finding a reliable method to arrive at the truth. They note similarities between Popper’s notion of falsifiability and VC-dimension and distinguish low VC-dimension from simplicity in Popper’s or any ordinary sense. Both Harman and Kulkarni address those similarities between the two views and argue how SLT could improve Popper’s account of simplicity. Daniel Steel, while bringing in Popper’s notion of falsifiability, has gone further to argue that the aim of Popper’s account seems different from the aim of SLT. According to Popper, the scientific process of conjectures and refutations generates ever more testable theories that more closely approximate the truth. This staunch realism toward scientific theories, according to Steel, is absent in SLT which aims at minimizing the expected

error of prediction. Even though there is an apparent difference between these two approaches, Steel conjectures, there might be some underlying connection between predictive accuracy and efficient convergence to truth.

Like SLT, both the Minimum Description Length (MDL) principle and the earlier Minimum Message Length (MML) principle aim to balance model complexity and goodness-of-fit in order to make reliable inferences. Like SLT, the MDL and MML, are also motivated by a similar consideration regarding how to make reliable inferences from data.¹⁰ In both these approaches, the optimal tradeoff is considered to be the one that allows for the best compression of the data, in the sense that the same information is described in terms of a shorter representation. In MML, it is important that the compression be in two parts: hypothesis ($H1$) followed by data given the hypothesis (D given $H1$).

According to the MDL principle, the more one could compress a given set of data, the more one has learned about the data. The MDL inference requires that all hypotheses ought to be specified in terms of codes. A code is a function that maps all possible outcomes to binary sequences such that the length of the encoded representation can be expressed in terms of bits. The MDL principle provides a recipe regarding how to select the hypothesis: choose the hypothesis H for which the length of the hypothesis $L(H)$ along with the length of the description of the data using the hypothesis $LH(D)$ is the shortest. The heart of the matter is, of course, how these codes $L(H)$ and $LH(D)$ should be defined. In their introduction to MDL learning, Steven de Rooij and Peter Grünwald explain why these codes should be defined to minimize the worst-case regret (roughly, the coding overhead compared to the best of the considered codes), while at the same time achieving especially short code-lengths if one is lucky, in the sense that the data turn out to be easy to compress.

By minimizing regret in the worst case over all possible data sequences, it is not necessary to make any assumptions as to what the data will be like. If one performs reasonably for the worst possible data set, then one will perform reasonably well for any data set. However, in general it may not even be possible to learn from data. Rather than assuming the truth to be simple, De Rooij and Grünwald introduce the alternative concept of luckiness: codes are designed in such a way that if the data turn out to be simple, we are lucky and we will learn especially well.

The Minimum Message Length (MML) principle is similar to the MDL principle in that it is also interested in proposing a resolution for what we have called the curve-fitting problem. Like MDL, data compression plays a key role in MML.

¹⁰There are considerable disagreements among experts on the SLT, MML, and MDL regarding which one is a more general theory than the other in the sense whether, for example, the SLT is able to include discussion of the properties of the rest. The SLT adherents argue that the SLT is the only general theory, whereas the MML adherents contend that the MML should be credited with the only approach which is the most general between the two. One MDL author, however, thinks that it is misleading to take any of the three approaches as any more fundamental than the rest. We are thankful to Gilbert Harman, David Dowe and especially Peter Grünwald for their comments on this debate. For a discussion of these entangled issues, see [Grünwald, 2007, chapter 17]. We, as editors, however, would like to report those disagreements among these experts without taking sides in the debate.

The more one could compress the data, the more we are able to get information from the data; moreover, the shorter the length of the code for presenting that information, the better will it be in terms of MML. One way to motivate either the MDL or the MML approach is to think of the length of codes in terms of Kolmogorov complexity in which the shortest input to a Turing Machine will generate the original data string (For Kolmogorov's complexity and its relation to random sequences see section 7.3). This approach uses two-part codes. The first part always represents the information one is trying to learn, that is, of encoding the hypothesis H and then making the Turing Machine prepare to read and generate data, assuming that the data were generated by the hypothesis H encoded in the first part. In the first part of the message, the codes do not cause the Turing Machine to write. The second part of the message encodes the data assuming the (hypothesis or) model given in the first part, and then makes the Turing Machine write the data. In the use of two-part codes there is very little difference between MML and MDL. However, there is a fundamental difference between the two. MML is a subjective Bayesian approach in its interpretation of the used codes, whereas MDL eschews any subjectivism in favor of the concept of luckiness. The MML can exploit an agent's prior (degree of) beliefs about the data generating process, but it can also attempt to make our priors as objective as possible in MML by using a simplest Universal Turing Machine.

In his paper on the Bayesian information-theoretic MML principle, David Dowe surveys a variety of statistical and philosophical applications of MML, including relating MML to hybrid Bayesian nets. The relationship underlying MML is the idea of information theory where information is taken to be the negative logarithm of probability. This view has also led him to his two recent results: (i) the only scoring system which remains invariant under-reframing of questions is the logarithm of probability score, and (ii) a related new uniqueness result about the Kullback-Leibler divergence between probability distributions. Dowe re-states his conjecture that, for problems where the amount of data per parameter is bounded above (e.g., Neyman-Scott problem, latent factor analysis, etc.), to guarantee both statistical invariance and statistical consistency in general, it appears that one needs either MML or a closely-related Bayesian approach. Using the statistical consistency of MML and its relation to Kolmogorov complexity, Dowe independently re-discovers Scriven's human unpredictability as the "elusive model-paradox" and then, resolves the paradox (independently of Lewis and Shelby Richardson [1966]) using the undecidability of the Halting problem. He also provides an outline of the differences between MML and the various variations of the later MDL principle that have appeared over the years (for references of the papers in the last paragraph, see Dowe's paper in the volume.)

The notion of simplicity in statistical inference is a recurring theme in several papers in this volume. De Rooij and Grünwald have addressed the role of simplicity, which they call "the principle of parsimony" in learning. Those who think that simplicity has an epistemological role to play in statistical inference contend that simpler theories more likely to be true. There could be two opposing camps in

statistical inference regarding the epistemological role of simplicity. One could be Bayesians. The other could be non-Bayesian [Forster and Sober, 1994]. De Rooij and Grünwald have, however, identified the epistemological interpretation of simplicity with Bayesians. A likely natural extension of the epistemological construal of simplicity in statistical inference is to *believe* that simpler theories are more likely to be true. Subjective Bayesians subscribe to this epistemological account of simplicity. The hypothesis with a maximum posteriori probability is believed most likely to be true. De Rooij and Grünwald distance themselves from this interpretation, because the philosophy behind MDL aims to find *useful* hypotheses without making any assertions as to their truth.

De Rooij and Grünwald assert that one fundamental difference between MDL, on the one hand, and MML and SLT, on the other, is that the former does not seem to have any form of the uniformity of nature assumption inbuilt in its philosophy that we find in the latter two. They would hesitate to assume that the data at hand necessarily provide clues about the future data. They prefer not to discount the possibility that it may not. Instead, they design methods so that we learn from the data if we are in the lucky scenario, where such is possible. According to them, this is a key distinction between the MDL approach and any other approaches including MML and SLT.

In his paper, Kevin Kelly agrees with non-Bayesians like De Rooij and Grünwald about the Bayesian explanation of the role of simplicity in scientific theory choice. The standard Bayesian argument for simplicity, as already stated, is that simpler theories are more likely to be true. Bayesians use some form of Bayes' theorem to defend their stance toward the role of simplicity in theory choice. This could take the form of comparing posterior probabilities of two competing theories in terms of the posterior ratio:

$$\frac{P(S|D)}{P(C|D)} = \frac{P(S)}{P(C)} \times \frac{P(D|S)}{P(D|C)}, \quad (\text{E3})$$

where theory S is simple (in the sense of having no free parameters) and theory C is more complex (in the sense of having free parameter θ that ranges, say, over k discrete values). The first-quotient on the right hand side of (E3) is the ratio of the prior probabilities. According to Kelly, setting $P(S) > P(C)$ clearly begs the question in favor of simplicity. So he supposes, out of “fairness”, that $P(S)$ is roughly equal to $P(C)$, so that the comparison depends on the second quotient on the right hand side of (E3), which is called the *Bayes factor*. According to him, the Bayes factor appears “objective”, but when expanded out by the rule of total probability, it assumes the form:

$$\frac{P(S|D)}{P(C|D)} = \frac{P(S)}{P(C)} \times \frac{P(D|S)}{\sum_{\theta} P(D|C_{\theta})P(C_{\theta}|C)},$$

which involves the subjective prior probabilities $P(C_{\theta}|C)$. Kelly's point is that typically there is some value of θ such that $P(D|S) = P(D|C_{\theta})$. If $P(C_0|C) = 1$,

then the posterior ratio evaluates to 1 (the complex theory is as credible as the simple theory). But, in that case the parameter θ is not “free”, since one has strong *a priori* views about how it would be set if C were true. To say that is “free” is to adopt a more or less uniform distribution over k values of θ . In that case, the posterior ratio evaluates to k — a strong advantage for the simple theory that becomes arbitrarily large as the number of possible values of θ goes to infinity. But, objectively speaking, C_0 predicts D just as accurately as S does. The only reason C is “disconfirmed” compared to S in light of D is that the *subjective* prior probability $P(C_0|C) = 1/k$ gets passed through Bayes’ theorem. Kelly concludes, therefore, that the Bayesian argument for simplicity based on Bayes’ factor is still circular, since it amounts to a prior bias in favor of the simple world S in comparison to each of the complex possible C_0 .

Kelly proposes a new, alternative explanation of Ockham’s razor that is supposed to connect simplicity with truth in a non-circular way. The explanation is based on the Ockham Efficiency theorem, according to which Ockham’s razor is the unique strategy that keeps on the most direct path to truth, where directness is measured in terms of jointly minimizing course reversals en route to the truth and the times at which these courses reversals occur. Since no prior probabilities are involved in the theorem, Kelly maintains that it does not beg the question the way simplicity-biased prior probabilities do. Furthermore, since Kelly views directness of pursuit of the truth as a concept of truth conduciveness, he views the Ockham Efficiency Theorem, as a foundation for scientific inference rather than instrumentalistic model selection. In that respect, he parts company with the anti-realism of De Rooij and Grünwald, who maintain, in light of the MDL approach, that simplicity plays primarily a heuristic role in making a theory useful. Kelly argues that in the case of causal theory choice from non-experimental data (see section 6 below), getting the causal arrows backwards yields extremely inaccurate policy predictions, so the Ockham Efficiency Theorem, according to Kelly, is the only available, non-circular foundational viewpoint for causal discovery from non-experimental data.

5 RECENT ADVANCES IN MODEL SELECTION

We touched on the model selection problems when we discussed the curve-fitting problem, especially with regard to the AIC paradigm in statistics. Model selection is such a hotly debated topic in statistical literature that it deserves a special place in our volume. However, model selection has a narrowly focused area in statistics in which investigators are confronted with choosing the model that provides the best possible account for the underlying mechanism generating the data and the role simplicity contributes to their choice. It is a delicate question whether there is a characteristic difference between theory choice and model selection. In revolutions in physical theories, Einstein’s theory replaced the Newtonian theory or a new discipline like bio-chemistry emerged as the theory of double helical structure of the DNA was discovered. However, model selection cannot or should not de-

cide between two contending theories, as we see in the case of physical theories. In contrast, model selection addresses a narrow topic with specific problems, like whether the data in question are coming from a normal distribution or from other non-normal distributions.

Setting the model selection issues in this context of theory choice in advanced physical theories, Arijit Chakrabarti and Jayanta Ghosh address two widely discussed model selection criteria: AIC and Bayesian Information Criterion (BIC). BIC says that $\text{Prob}(H_k | \text{data})$ is proportional to the log-likelihood of the sample n multiplied by $n^{-k/2}$. Surveying the literature in statistics and computer science where model selection criteria are extensively applied, Chakrabarti and Ghosh distinguish between the purposes for which AIC and BIC are introduced. They argue that the BIC is more useful when the true model is included within the model space. In contrast, the AIC is more effective in predicting the future observation. This depiction of the difference between the BIC and AIC sets the tone for a peaceful co-existence of both model selection criteria; too often, statisticians and philosophers are involved in “statistics war” pleading for the superiority of one criterion over the other. This also provides an elegant connection between the theme echoed in Chakrabarti and Ghosh’s paper on the one hand, and Norton’s and Kelly’s papers on the other. The common theme that has been shared by all four authors is that there should not be any such account of inductive inference that could defend “one-size-fits-all” philosophy. Although both Norton and Kelly are non-Bayesians and their targets of criticism are Bayesians, Chakrabarti and Ghosh are themselves Bayesians and their targets could be both Bayesians and non-Bayesians.

A.P. Dawid in his chapter has also focused on Bayesian model selection issues. Within a Bayesian framework, the conventional wisdom about the problem of model choice is that model choice is more sensitive to the prior than is standard parametric inference. Working with a Bayesian framework, he, however, rejects this conventional wisdom based on the study of the asymptotic dependency of posterior model probabilities on the prior specifications and the data. Moreover, the paper suggests a possible solution to the problematic Bayes factor with respect to improper priors by specifying a single overall prior and then focusing on the posterior model distributions. The general theory has been illustrated by constructing reference posterior probabilities for both normal regression models and analyses of an ESP experiment. However, in a sense, Dawid’s interest is much broader than just model selection issues. He is interested in Bayesian inference. Consequently, he has developed a general Bayesian theory of inference with a specific type of posterior probability construction.

6 CAUSAL INFERENCE IN OBSERVATIONAL STUDIES

Whether we engage in statistical inference or are involved with statistical evidence or model selection problem, data are enormously important for statistics and making reliable inference. Often, data or observations are randomized. Frequently, we

do not have the luxury of randomly sampling our observations so that our inference would be reliable in the sense of minimizing biases or systematic errors. Those cases arise in observational studies when we have no randomization option, but we still need to make inferences. Sometimes, we also make causal inferences based on observational studies. Peter Spirtes and Sander Greenland's two papers address these issues regarding how one could successfully make a reliable causal inference when topics in question involve observational studies.

In building a causal model from observational studies, Spirtes distinguishes between two kinds of questions, (i) qualitative and (ii) quantitative. A qualitative question asks "will manipulating a barometer reading affect subsequent rainfall?" A quantitative question asks, "how much does manipulating a barometer reading affect subsequent rainfall?" With observational studies, one might be confronted with several issues about how to make a reliable qualitative or quantitative inference. Spirtes is interested in making reliable causal inference in observational studies when two variables may be connected by an unmeasured common cause. For example, the correlation between two observed variables *barometer reading* and *rainfall* could be due to an unmeasured (latent) common cause such as *atmospheric pressure*; without further background knowledge the correlation could also be due to the *barometer reading* causing *rainfall*. Without some substantive causal assumptions of this entire causal setup, one cannot hope to make any reliable inference about which of these causal relationships between *barometer reading* and *rainfall* is correct simply from observing the correlation between them. Sometimes attempts are made to make reliable inferences about the causal relationship between *barometer reading* and *rainfall* by using temporal information and conditioning on as many potential causes as possible prior to the occurrence of the putative cause. However, in this scenario there is still no way to make reliable causal inferences without substantive causal assumptions regarding whether all the relevant variables are conditioned on; in addition standard causal inference methods can be misled by conditioning on too many variables, as well as too few. Spirtes discusses some general assumptions relating causal relationships to statistical relationships, which causal inferences cannot be reliably made under these assumptions when unmeasured common causes may be present, and alternative reliable methodologies for reliable causal inferences when it is possible.

Like Spirtes, Greenland is also interested in causal inference in observational studies. However, he is focused on a manipulative account of causation where an investigator wants to observe a particular outcome in a subject after she is given a treatment. This model is called the *Potential outcome* or *Counterfactual model*. In the potential outcome model, the investigator approaches the issue by defining the effects of a cause or intervention and then defining the effects of treatment. Suppose, we would like to observe the effects of different doses of AZT on patients affected with HIV virus. We would compare the one year mortality rate that would occur in a given population after they are administered certain doses of AZT or none at all. Here, the range of intervention in terms of patient's doses of AZT could very well correspond to different levels of their physical discomfitures, in

which the potential outcomes could be their length of survival, e.g., if the patients survive another year or not. Or it could be that their survival time range from one year to another five year from the day of the introduction of the drug to them.

In the deterministic situations like the case of the AZT doses and the survival rate of the patients, the purpose of observational studies, according to Greenland, is to find a causal link connecting treatments of the patients to their recovery/survival rate. In the deterministic world, we use probability to capture uncertainty, although ambiguity remains if the intervention is not completely specified. The situation becomes more complex in a stochastic world, such as in quantum world, where it is not even theoretically possible to know the future fully. The problem can be dealt with by allowing the potential outcomes to refer to distributions (or parameters) instead of realizations [Greenland *et al.*, 1999].

In the present chapter, however, Greenland confines himself to the deterministic world. In articulating the potential outcome model for such a deterministic world, he discusses the role of causal laws (expressed in terms of some structural equations) in which one shifts from the understanding of the counterfactual (if you would have administered treatment x instead of y to the patient, then what could have happened to her?) to the understanding of the law connecting the outcome variable to the antecedent variable. The law in question could be “if one drops both a feather and a coin at the same time from a height to the ground where there is no air-resistance, then they would reach the ground simultaneously.” Greenland thinks that if we find that the predictions of the theory have turned out to be false more often than not, then they would provide reasons to question its tenability. As students of philosophy of science we know, however, that the falsity of predictions should not necessarily lead to the rejection of that theory as its various auxiliaries could be called into question rather than the theory itself because of its wrong predictions. Similar tangled issues could crop up in studying causal inference in observational studies

To circumvent this possible objection that the falsity of predictions should not necessarily lead to the wholesale rejection of the theory, Greenland has demonstrated the benefits of the use of causal diagrams in similar and other relevant situations. A “causal system” as he calls it consists of the output w of a function $f(u, v)$ which might become an input to a later function $g(u, w)$ with output x . The arrows in the diagram representing this system are to be taken as causal arrows connecting input variables to output variables. The advantage of this kind of diagrams lies in its ability to do “local surgeries” if and when required. They are devised in such a way that they are able to isolate the effect of that rejection mentioned above in that causal network of relations among variables, and thus are able to pinpoint which specific part of theory or its auxiliaries have to be called into question when predictions of the theory don’t result in what is expected.

While developing the potential outcome models along with their structural equation generalizations, he makes the reader aware of the possible theoretical problems of this model. However, he concludes with a pragmatic note. What matters most, according to him, is the practical significance of these causal diagrams applied to

varied situations, although many foundational questions about the model are yet to be satisfactorily resolved.

7 SPECIFIC TOPICS OF INTEREST

We began this “Introduction” with the types of questions/problems with which philosophers are often confronted. We have seen that they are more often interested in general epistemic and causal issues, including issues about the foundations of statistical inference. Philosophers who are trained primarily in analytic tradition are sometimes interested in simple/specific problems. Nonetheless, it would be a mistake to think that only philosophers are interested in specific problems. This section examines how statisticians, mathematicians, and computer scientists along with philosophers are busy working out the details of some of the specific problems in the topics of their interest.

7.1 *Conditional probability*

One of these specific problems in which various philosophers have lately been getting interested is “what is conditional probability?” The conditional probability of $P(H|D)$ is traditionally defined as $P(H \& D)/P(D)$ when $P(D)$ is greater than zero. Here, as before, “ H ” stands for a *hypothesis* and “ D ” stands for *data*. Alan Hájek in his contribution suggests that although this way of understanding conditional probability is common and has been proposed by Kolmogorov, it is actually a *conceptual analysis* of conditional probability. According to Hájek, one is free to stipulate that a particular technical expression is a short-hand for a particular concept, but one is not free to assign that concept any meaning he/she chooses. Conditional probability is not simply an abbreviation; rather, it is answerable to certain pre-theoretical intuitions. Thus, although we could choose to make “ $P(H|D)$ ” a shorthand for this ratio, he argues, we do not have any good reason to identify the expression “the probability of H given D ” with this ratio. He evaluates several arguments for the ratio analysis, and ultimately rejects them. Instead, he argues, conditional probability should be taken as primitive.

One of the arguments he offers for dispensing with the ratio analysis is by way of illustrated by an example. What is the probability that a coin turns up heads given that I toss it fairly? It should be $\frac{1}{2}$. The problem, Hájek contends, is that according to the ratio analysis, the conditional probability is the ratio:

$$P(\text{the coin lands heads} \& \text{I toss the coin fairly})/P(\text{I toss the coin fairly})$$

and both unconditional probabilities need not be defined. After all, he argues, “you may simply not assign them values. After some thought, you may start to assign them values, but the damage has already been done; and then again, you may still not do so.” The damage that has been done is that there is a time at which an agent assigns a conditional probability ($\frac{1}{2}$) in the absence of the corresponding unconditional probabilities required by the ratio formula. As is evident, this is an immediate counterexample to the ratio formula. It does not save the

formula from this counterexample if later on the agent in question happens to assign the requisite unconditional probabilities; the counterexample concerned the earlier time, and its existence provides a refutation of the ratio analysis of the conditional probability. However, in this example, Hájek thinks that there is a clear-cut conception of conditional probability according to which the answer is $\frac{1}{2}$. He explores the fundamental nature of conditional probability in the various interpretations of probability. He canvases other arguments against the ratio analysis: cases in which the condition (D) has probability 0 (and thus can't appear in the denominator of the ratio), or unsharp probability, or vague probability. (These notions are explained in sections 4.2 and 4.3. of his chapter) He also shows how conditional probability plays a key role in various paradoxes, including Simpson's paradox (see section 7.2 for an example and exposition.)

One highlight of Hájek's paper is to investigate the debate over whether the notion of conditional probability is a primitive or a defined concept. In the usual way that we learn probability theory, conditional probability is defined in terms of two unconditional probabilities. Since he thinks that we have an intuitive understanding of conditional probability, he recommends reversing the order of analysis. Like Popper, Hájek argues that one should take conditional probability, $P(\cdot)$ to be fundamental and unconditional probability to be derivative: the unconditional probability of a is $P(a, \mathbf{T})$, where $\mathbf{T}$ is a logical truth.¹¹ Hájek presents Popper's axioms and postulates of primitive conditional probability functions, known as the *Popper Functions*. This mathematical set-up offers a rival approach to Kolmogorov's. Hájek also points out that Popper's approach has no difficulty in handling the objections raised against the ratio analysis.

Building on Hájek's paper, Kenny Easwaran has both addressed and evaluated many of the issues raised by Hájek. He agrees with Hájek that the ratio analysis of the conditional probability is faulty. However, he disagrees sharply with Hájek regarding his claim that conditional probability rather than unconditional probability should be regarded as primitive. Easwaran thinks that Hájek's argument in this connection needs to be separated for each interpretation of probability rather than taking it as one argument that cuts across all interpretations together. He thinks that, especially with regard to subjective interpretation of probability, a distinction needs to be made between $P(H|D)$, where " D " are themselves some possible event later in time than t , and $P(H|D)$, where " D " encompasses a complete description of the history of the universe up to time t . Easwaran contends that although it is evident that all non-logical interpretations (for example, the subjective or propensity interpretation) of probability depend on some *information* in order for some probabilities to be assigned to some particular cases, the role this information plays could very well vary. Thus, he concludes that Hájek has not provided a good argument for the claim that the notion of conditional probability should be counted as a primitive notion rather than the notion of the unconditional probability.

¹¹Hájek prefers to reserve the notations, " $P(|)$ " for the ratio analysis and " $P(\cdot)$ " for the conditional probability being primitive.

7.2 Probabilistic and statistical paradoxes

Paradoxes, albeit entertaining, challenge our intuitions in a fundamental way. There is no exception to this whether we are concerned with probabilistic or statistical paradoxes. It is often hard to distinguish between probabilistic paradoxes and statistical paradoxes. One could, however, propose that all statistical paradoxes are probabilistic paradoxes, whereas probabilistic paradoxes are not exactly statistical paradoxes. This leaves a more basic question unanswered: “what is a statistical paradox?” A statistical paradox is a paradox that can be understood in terms of notions that need not be probabilistic, for example, “confounding,” “intervening,” and so on. It still might not settle the dispute about the differences between statistical and probabilistic paradoxes, because probability theorists could contend that notions like “confounding” are subject to probabilistic reduction. It is, however, sufficient for our purpose to appreciate the difficulty associated with giving a clear-cut definition of either of the terms as we proceed to discuss probability paradoxes. One way to understand the difference between the two is to consider the “Monty Hall problem” (named after a television game show host), which, among many other paradoxes, Susan Vineberg discusses in her chapter.

Suppose you are a contestant on a quiz show. The host, Monty Hall, shows you the three doors (A , B , and C). Behind one door is an expensive new car and behind others are goats. You are to choose one door. If you choose the door with the car, you get it as a prize. If you choose a door with a goat, you get nothing. You announce your choice, and Monty Hall opens one of the unchosen doors, showing you a goat, and offers to let you change your choice. Should you change? Three crucial points need to be clearly stated which are usually overlooked in a popularized version of the Monty Hall problem. They are (i) the expensive car has an equal chance of being distributed behind any of the doors. (ii) If you choose one door without the prize behind it, Monty will open the door that does not have the prize behind it. And (iii) if you choose the door behind which there is a prize behind it, then Monty will open the door randomly which does not have the prize behind it. Given this setup, let us calculate the probability of opening the door B (*Monty B*) when the prize is behind A (A). Suppose the contestant has chosen door A .

The probability of Monty’s opening door B given that the prize is behind door A equals $P(B|A) = \frac{1}{2}$. (Recall that once you have chosen the door A Monty is not going to open it. So Monty is left with two choices.)

Compute the probability of Monty’s opening the door (*Monty B*) when there is the prize behind the door B .

$P(\text{Monty } B|B) = 0$ (Recall that he is not going to open the door behind which there is the prize).

Compute the probability of opening the door B when there is the prize behind door C .

$P(\text{Monty } B|C) = 1$ (because the contestant has chosen door A and the prize is behind door C . He is left with no other option than to open door B .)

We would like to know whether you should stay where you are (A) conditional (given) on the information that Monty has opened door B . That is, $P(A| \text{Monty } B)$. According to the Bayes' theorem,

$$P(A| \text{Monty } B) = \frac{(P(A) \times P(\text{Monty } B|A))}{\{(P(A) \times P(\text{Monty } B|A)) + (P(B) \times P(\text{Monty } B|B)) + (P(C) \times P(\text{Monty } B|C))\}}$$

Consider $P(A) \times P(\text{Monty } B|A) = 1/3 \times \frac{1}{2} = 1/6$

$$\begin{aligned} P(B) \times P(\text{Monty } B|B) &= 1/3 \times 0 = 0. \\ P(C) \times P(\text{Monty } B|C) &= 1/3 \times 1 = 1/3 \end{aligned}$$

Now $P(A| \text{Monty } B) = \frac{1/6}{1/6+0+1/3} = 1/6 \div \frac{1}{2} = 1/3$.

Thus, if you switch, you will get $2/3$ of a chance of winning.

Vineberg has considered the Monty Hall problem to be a probability paradox. One needs to be careful as to her reason for classifying this problem to be a paradox of probability. She takes it to be a paradox of probability because probabilities are appealed to in the reasoning that gives us the paradox, and not because of its solution which rests on using Bayes' theorem. This paradox might be said to be essentially probabilistic, because it is resolved by highlighting the correct use of Bayes' theorem, which yields $1/3$ along with the assumptions noted before. In contrast, Simpson's paradox, another paradox that Vineberg considers, can be considered to be both a statistical and a probability paradox at the same time. Simpson's Paradox involves the reversal of the direction of a comparison or the cessation of an association when data from several groups are combined to form a single whole. Suppose you are considering the acceptance rates for males and females in a graduate program that includes two departments. Consider an example of Simpson's paradox.

Table 2. Simpson's Paradox

CV	Dept. 1		Dept. 2		Acceptance Rates		Overall Acceptance Rates
	Accept	Reject	Accept	Reject	Dept. 1	Dept. 2	
F	180	20	100	200	90%	33%	56%
M	480	120	10	90	80%	10%	70%

Here, "CV" includes two categorical variables, "F" for "females" and "M" for "men." "A" and "R" represent "the rates of acceptance/rejection" for two departments, D_1 , and D_2 . Here is a formulation of the paradox, in which the association in the subpopulations is reversed in the combined population. Although the acceptance rates for females are higher than for males in each department, in the combined population, the rates have reversed. Vineberg has explained why it could be explained within the confines of the probability theory. However, as we know, it could also be regarded as a statistical paradox because notions like "confounding,"

which many statisticians consider to be statistical notions, could be used to explain it. In the above example of the paradox, for example, the effect on acceptance (A) of the explanatory variable, sex (S), is hopelessly mixed up (or “confounded”) with the effects on A of the other variable, department (D). According to some statisticians, we are interested in the direct effect of sex on acceptance and not an indirect effect by way of another variable such as department. The effect of S on A is confounded with the effect on A of a third variable D .

Although Vineberg is fully aware of this difficulty of characterizing a paradox completely in terms of the probability theory or statistics, she confines herself to the probability paradoxes, investigating possible reasons for regarding them as probability paradoxes. She counts all of the paradoxes that she has considered paradoxes of probability because the reasoning (and/or premise) involved in them is probabilistic, and not because of the form of their solution (for a clear understanding of the relationship between probabilistic reasoning and probabilistic/inductive inference, see the concluding section of Bandyopadhyay and Cherry’s chapter.) On her account, Simpson’s paradox is a paradox of probability is due to this specific reason. She argues that some of the paradoxes arise from misapplications of the rules of the probability theory. Some of the resolutions of the paradoxes (like the Monty Hall problem) are less controversial, whereas others are more controversial; as a result, they are hotly debated. She cites Newcomb’s problem as one such paradox in the latter category since it is liable to several competing solutions requiring radical rethinking about the foundations of Bayesian decision theory (see [Savage, 1972] for classical Bayesian decision theory).¹² The rationale she offers for not regarding it as a paradox of probability is that the reasoning leading to this problem requires non-probabilistic decision theoretic considerations.

C. Andy Tsao reviews two statistical paradoxes: (i) Lindley’s paradox and (ii) the Fieller-Creasy problem. There are already some discussions on those paradoxes in the literature, since both arose in the early 1960s. We will confine ourselves to Tsao’s discussion of the Lindley’s paradox. To set the paradox in a broader perspective, consider the formulation of the hypothesis testing problem where two hypotheses are H_0 = the null hypothesis and H_1 = the alternative hypothesis. There are infinitely many tests possible regarding the tenability of a hypothesis; some are conservative and some are liberal. In this connection, the term “conservative” (“liberal”) refers to a specific property of the tests that have a smaller (larger) probability of rejecting the null when the alternative is typically a new, or previously uncontested statement of a hypothesis/theory. Let us assume that H_0 = the suspect is innocent vs. H_1 = the suspect is guilty. A conservative judge will presume the innocence of the suspect even at the cost of letting some guilty walk free, while controlling the rate of error associated with judging the innocent to be guilty. “Is this statistical procedure justified” is one basic question

¹² In the last ten to fifteen years, in several fields a great deal of research has been done on Bayesian decision theory. However, for Bayesian decision theory there are very few books which are better than Berger [1985]’s classic book written from a statistician’s point of view. There are also a couple of recent good books on the same issue in philosophy. Joyce [1999] has addressed Bayesian decision theory from a casual decision-theoretic perspective. See also Weirich [2004] for a realistic non-ideal agents’ decision theory.

of statistical decision theory. Depending on the criterion, one could consider this procedure to be problematic, whereas under other criterion, one could deem it to be justified. However, we also have a sense of what counts as “justified,” and statisticians often fall back on it for supporting their stance toward evaluating a theory.

Lindley’s paradox demonstrates that the common belief about classical statistics being conservative is mistaken. Consider why conservatism has been usually associated with classical statistics. The confidence coefficient on which the idea of a confidence interval rests is the smallest coverage probability among all possible θ rather than the average of them or the size of a test is the largest of probability of Type I error for all possible θ in the null parameter space. Classical statistics is taken to be more conservative in rejecting the null. It is harder to reject the null if one’s tests are conservative. Consider two values, “a” and “b”, denoted by $a(x), b(x)$ as functions of the observed data. These two values are such that for any x , we have $a(x) < b(x)$. Thus, in this hypothesis testing setup, we will reject H_0 if $a(x) < \alpha$ (similarly $b(x) < \alpha$). Therefore, under this scenario, using $b(x)$ will be harder to reject H_0 than using $a(x)$.

According to Tsao’s analysis, classical statistics, however, has turned out to be less conservative than its Bayesian counterpart. In the literature, many classical/frequentist statistics are shown to be minimax procedure-based, in a way that they work best in the worst scenarios, therefore, they are conservative. Lindley’s paradox is a paradox in the sense that given the same data and setting, the “frequentist” p-value has become substantially smaller (less than α) than the minimum of “reasonable” Bayes estimates (greater than $1-\alpha$).

We need to mention a couple of especially worthwhile points about the paper. Lindley’s paradox, according to Tsao, is mathematically correct. This, in fact, points out that frequentist and Bayesian procedures may lead to different conclusions. Given the same data, a frequentist using p-value may reject H_0 , whereas, according to Bayesian use of the posterior probability argument, there is a very high probability for H_0 to be true. What his paper brings out is that those statistical procedures derived from frequentist and Bayesian principles have different types of guarantee. Frequentist hypothesis testing theory (Neyman-Pearson formulation) guarantees that the long run frequency of the Type I error will be less than α while Bayesian procedure will minimize the posterior expected loss function. In decision theory, a Bayesian estimator is a decision rule that minimizes the posterior expected value of a loss function. These two criteria are foundationally different and not necessarily lead to same procedure nor the corresponding conclusion.

One needs to be careful that it not because of the goals of the two paradigms, (i) classical statistics, and (ii) Bayesian statistics, are different that the conclusions are different. Tsao contends in fact that the goal of both paradigms in this specific case is the same: that is, to assess which hypothesis H_0 or H_1 is more likely to be the correct hypothesis. Typically, the frequentist procedures perform well when used repeatedly in the long run while the Bayesian procedure work well for the

current experiment if the prior agrees well with the parameter uncertainty. The readers will be able to find that some of the issues raised in this paper overlap with some of the issues covered both in Mayo and Spanos's paper on error statistics, and in Weirich's paper on Bayesian theory when Weirich contrasts the latter with classical statistical theory.

7.3 *On randomness*

Like probability and statistical paradoxes, the concept of randomness is a favorite topic for many probability theorists, philosophers, and other scholars. This concept initially plays a crucial role in understanding the frequency interpretation of probability. The frequency account of probability requires that the probability of an event is the long-run relative frequency in a sequence of repeatable events. The long-run relative frequency in a sequence of repeatable events is an objective property that the frequency theorists could measure and sample. To estimate long-run relative frequencies, we would like to measure the relative frequency in a random sub-sequence. We have two papers, one by Deborah Bennett and the other one by Abhijit Dasgupta, which discuss this elusive notion. Bennett provides an accessible introduction to the topic detailing some of the problems associated with defining "randomness." Dasgupta has provided rigorous mathematical foundations for randomness; he traces both its history and the mathematics leading to our present-day understanding of the topic. Since the topic itself along with its mathematical history could be of interest to our various readers, we will provide a background for both articles. We address the topic from three interrelated ways, borrowing insights from Dasgupta's paper: (i) the statement and scope of the problem, (ii) an historical outline, and (iii) the three paradigms of algorithmic randomness.

(i) Statement and scope of the problem:

We begin with two intuitive views about randomness and a problem regarding which/when sequences can be considered random. We also contrast this view with what Dasgupta calls the extensional "black-box view." Consider some examples that seem "random" from our intuitive understanding of what counts as random. Among the examples of these processes are the following: flipping a coin, turning a gambling wheel, taking a snapshot of weather data, measuring the time between two successive clicks of a Geiger counter detecting radioactive decay, or collecting the stock market index value. The outcomes of any such process is recorded as a list of symbols or numbers, technically called a *sequence*. Here, we assume that the list of outcomes is available to us only as a sequence of observations. The inner workings of the process is assumed to be unknown and will not concern us in any way. This is why the view is called the extensional "blackbox" view. This way, we completely "decouple" the generated sequence of outcomes from the process generating it.

The main problem can now be stated in the following:

Find a precise criterion to distinguish, among the collection of all possible sequences, the random ones from the non-random ones. In other words, which sequences are random and which are not?

Here is a further simplification: Consider the example of repeatedly flipping a fair coin, where heads and tails are equally likely and those flips are probabilistically independent of one another. Denoting heads by 1 and tails by 0, the outcomes of this process can be represented as the collection of all binary sequences like 1010011010... Let us call this probability space the *binary sequences with uniform probability*. Restricting our attention to this special case, our main problem becomes this: *Find a precise criterion to distinguish the random binary sequences from the non-random ones (for the uniform probability case)*. It is a remarkable but technical mathematical fact that this special case is able to simulate and model other complex processes in a faithful way, hence the restriction is not as severe as it may first appear. From now on, we only work with binary sequences with uniform probability.

(ii) An historical outline:

Our outline consists of seven short paragraphs showing the historical development of the problem about randomness leading to the present state of research on this.

1. *Absolute lawlessness is impossible!* The first intuitive and imprecise approach to the problem is to classify the random binary sequences as the ones whose bit-patterns obey no “law of regularity” whatsoever. Unfortunately, this intuitive approach is incorrect, since standard mathematical results (e.g., van der Waerden’s theorem) show that all binary sequences satisfy certain laws of regularity. In fact, as Borel showed in 1909, random sequences, instead of being completely lawless, will satisfy a law of *frequency stability*.
2. *1909: Borel’s strong law of large numbers.* In 1909, Borel proved his strong law of large numbers, which is a law of frequency stability, and is the first significant law of randomness. It states (recall our restriction to the uniform probability case) that the probability is one that in a binary sequence the proportion of 1’s among the first n terms approaches the value $\frac{1}{2}$ in the limit as n approaches infinity (the *frequency* of 1’s *stabilizes* to the value $\frac{1}{2}$). While Borel’s law of randomness is satisfied by all random binary sequences (see footnote 9 of Dasgupta’s article for an explanation), it fails to capture the notion of randomness since there are also many non-random sequences satisfying it. Thus, Borel’s strong law is a necessary but not sufficient condition for randomness, and so, in particular, it is not a criterion for randomness.
3. *1919: Von Mises defines randomness using frequency stability.* In 1919, Richard von Mises turned things around and gave a definition of, i.e. a criterion for, randomness in terms of frequency stability. He realized the

fundamental importance of frequency stability for defining randomness, and his definition of randomness is highly intuitive and appealing. A binary sequence, according to him, is random if and only if after erasing any “admissible” part of the sequence, the remaining sequence still satisfies Borel’s condition of frequency stability. This was considered to be the first mathematical definition of randomness. The idea of von Mises can be stated as follows: The randomness of a sequence is equivalent to its inherent *unpredictability*, which amounts to unbiasedness (frequency stability) in all its admissible parts. For von Mises, this also served as the foundation for the frequentist theory of probability that he was developing during this period.

4. *1940: Church brings in algorithms, forever.* The definition of von Mises, while being highly intuitive, was not precise mathematically, since he did not specify what the term “admissible” in his definition really means. In 1940, Church rectified the situation by bringing in the notion of algorithms to rigorously and precisely interpret the word “admissible” in von Mises’ definition. This gave the first mathematically precise definition of randomness. In addition, Church’s use of algorithms turned the subject of randomness permanently algorithmic, and the use of the notion of algorithm became increasingly important and relevant in the study of randomness. However, as will see, Church’s modification of von Mises’ definition has turned out to be inadequate.
5. *1965: Martin-Löf finds the first satisfactory definition of randomness.* Even after Church made von Mises’ definition of randomness mathematically precise, the definition remained inadequate. Ville showed that Church’s modification of von Mises’s definition was again only a necessary but not sufficient condition for randomness, and therefore is still not a criterion for randomness. To provide a rough understanding about Martin-Löf’s definition, call a property of sequences *special* if and only if the probability that a sequence has this property is zero. Some examples of special properties are “eventually constant”, “every third term is 0”, “there is no run of zeros of length five”, etc. If such a special property S is effectively specified, we observe first that the terms of any sequence satisfying S will be, roughly speaking, so contained that the sequence can’t be random. Second, by definition, the probability is zero that a “random” sequence will have this property S . Thus, a special property makes a sequence highly non-random, and, it is impossible for a random sequence to satisfy an effectively specified special property. We can now classify all sequences into two types. A sequence is *special* if and only if it satisfies at least one effectively specified special property. Otherwise, it is *typical*. If a sequence is to be random, it can’t be special. Hence, it must be typical. Thus, like the von Mises-Church definition, typicality is a necessary condition for randomness. Martin-Löf [Martin-Löf, 1965] turned this around to make typicality also a sufficient condition for randomness. A binary sequence is *random according to Martin-Löf* if and only if it is typi-

cal, i.e., if and only if it does not satisfy any all effectively specifiable typical property. Martin-Löf's work showed that this indeed constitutes a desirable definition of randomness, and does not exhibit the problems that plagued the von Mises-Church definition. In fact, Martin-Löf's definition has been the most satisfactory definition of randomness till to date.

6. *1960s–1970s: Solomonoff, Kolmogorov, Chaitin, etc, introduce the idea of complexity of finite strings.* This approach, now known as *Kolmogorov Complexity*, solved the problem of degrees of randomness in finite binary strings. The main idea is that a finite binary string is non-random if it has algorithmic descriptions that are shorter than the string itself, and it is random otherwise. More precisely, given a string x , the length $K(x)$ of its shortest description can be used to measure its degree of randomness: The smaller the value of $K(x)$ compared to the length of x , the less random x is. According to this idea, the randomness of a binary sequence is equivalent to its *incompressibility*, which means that none of its initial segments has much shorter descriptions.
7. *Martingales and an “unpredictability definition” of randomness.* It is also possible to give a stronger version of the von Mises-Church definition using capital betting strategies, or gambling strategies which take into account the amount of bet, technically known as *martingales*. This has resulted in a more satisfactory definition of randomness, equivalent to Martin-Löf's definition of randomness. In the following, when we mention randomness as unpredictability, we will mean this martingale definition.

(iii) Three paradigms of algorithmic randomness:

We discuss how apparently three distinct approaches to algorithmic randomness have merged in the end to an equivalent definition of randomness. As indicated in the brief historical outline above, there are three “paradigms” for defining randomness:

1. *Unpredictability*,
2. *Typicality*, and
3. *Incompressibility*.

The three paradigms appear to be very different approaches for defining randomness. It is therefore a remarkable mathematical theorem that the three apparently distinct notions in fact coincide! Roughly speaking, this means that a sequence is unpredictable if and only if it is typical if and only if it is incompressible. This remarkable coincidence has led to the formulation of the so called *Martin-Löf-Chaitin thesis*, which says that each (and so all) of these definitions gives the true definition of randomness. Like the Church-Turing thesis for the definition of algorithm, the Martin-Löf-Chaitin thesis is not a mathematical theorem or a

conjecture subject to proof, but rather a proposal arising from strong evidence. Algorithmic randomness thus provides so far an unrivaled mathematical foundation for randomness. Furthermore, since the works of von Mises, Church, and Martin-Löf, algorithmic randomness has been a very active and lively area of research providing deep philosophical insights into the notion of randomness and its finer ramifications. Dasgupta's article provides a survey of such research up to the present day.

In this subsection, so far we have followed Dasgupta's both mathematical and historical development of the topic. The difference between Dasgupta's paper and Bennett's paper is that the former is more mathematically oriented without being oblivious of the history of the development of the randomness of a sequence, whereas Bennett's paper is more intuitively accessible. Bennett discusses why randomness is hard to define. She makes a distinction between "relative randomness" and "absolute randomness" when the notion of relative randomness is defined relative to a set of properties. She thinks that both "relative randomness" and "absolute randomness" seem "fickle and unfair" in the short run, but they must appear predictable in the long run. She argues that we need the relative randomness to satisfy our notion of uncertainty. In fact, she also thinks that we need both notions of randomness, because we want the relative randomness to measure up to ideal randomness as much as possible.

Some special topics like "normal approximations", the "Stein phenomenon" (also known as the "Stein paradox"), and "data mining" are also crucial for understanding the state of the art research that has been done in statistics, computer science and other fields.

7.4 *Normal approximations*

The technique of approximating the distribution of a random variable by a normal (Gaussian) distribution is known as a normal approximation. It is the central limit theorem that justifies the use of normal approximations in commonly encountered settings. Boik describes the theory and application of the central limit theorem (CLT). According to the CLT, if certain mild regularity conditions are satisfied, then the distribution of suitably standardized sum or mean of a sequence of a random variable approaches a normal distribution as the number of random variables in the sequence increases. For example, if $Y_1, Y_2, \dots, Y_n$ is a random sample from an infinite sized population, whose mean and standard deviation are μ and σ respectively, then the distribution $\sqrt{n}(\bar{Y} - \mu)/\sigma$ converges to a normal distribution as the value of n increases to infinity, where $\bar{Y}$ is the mean of the random sample. Boik has considered how the scope of the CLT can be expanded to include non-linear functions of means and sums by the delta method and Slutsky's theorem. He also describes and illustrates several applications of the CLT. These applications include approximating the distributions of sums or means of discrete random variables; approximating the sampling distributions of sums or means of random samples from finite sized populations; approximating the sampling distributions

of test statistics in permutation tests; and many more. One highlight of his paper is a discussion of a result that justifies multivariate normal approximations for Bayesian posterior distributions of parameters. In addition, he describes how the accuracy of normal approximations can be improved by making small adjustments to the approximations.

7.5 *Stein phenomenon*

Richard Charnigo and Cidambi Srinivasan have written on the topic of the Stein phenomenon. Since among philosophers there is an intense interest in the problem, we will discuss this topic to make it more accessible to our general readers [Sober, 2008]. We first consider some elementary ideas from statistics to see how “Stein’s phenomenon” is startling from the perspective of the relation between the sample mean and the population mean. The sample mean is widely regarded as a good estimate of the population mean, since its expected value is the same as the population mean (it is an unbiased estimator) and it has the least variation around the population value compared to any other unbiased estimator under normal conditions. If we assume the data to be normally distributed with some unknown mean and variance, then the sample mean is, in fact, the maximum likelihood estimate (MLE) of the population mean. For a given data set and an underlying probability model, MLE picks the values of model parameter that makes the data “most likely.” In sum, the sample mean provides a natural way of “extracting information” about the population mean from the data. The sample mean is assumed to be optimal with regard to estimating the population mean because no other estimator has all these properties. Several optimality properties of the sample mean have been subsequently proved, providing a sound foundation for statistical methodology. These features of the sample mean led to the belief that when one would like to estimate several population means simultaneously, one should use sample means as they are likely to be optimal. Charles Stein showed that this belief is unfounded. In fact, when three or more parameters are estimated simultaneously, their combined estimator is more accurate (in the sense of minimizing the expected squared error) than any other statistic that estimates the parameters separately. This is known as the “Stein Phenomenon.”

Consider that we are running doughnut shops in the state of New York. Our intention is to know how many doughnuts we should expect to sell this year (2010) based on how many products we sold last year in each of our eight stores ($S_1, S_2, S_3, \dots, S_8$). Let us assume further that stores are positioned geographically so that sales in each store may be regarded as statistically independent of the sales of the other stores.¹³ We would like to estimate the parameter, μ_1 = the expected

¹³To make this example more intuitive, we could assume that we sell doughnuts in one store and completely unrelated items in the other stores. In the second store, we sell shoes, in the third store, we sell like insurance. In accordance with this setup of the example, our discussion in this subsection will change correspondingly. For example, we have to write now, similarly, μ = the expected sale of shoes in store two in 2010 based on the sample, X_2 = the observed sale of shoes of per 10,000 people in 2009 and so on for both another six parameters and observed sale

sale of doughnuts in store one in 2010 assuming that each month consists of 30 days, based on the sample X_1 = the observed sale of doughnuts in 2009. Similarly, μ_2 = the expected sale of doughnuts in store two in 2010 based on the sample, X_2 = the observed sale of doughnuts of per 10,000 people in 2009 and so on for both another six parameters and observed sale of doughnuts in 2009. “Stein’s Phenomenon” shows that if we are interested in estimating μ_1 through μ_8 , simultaneously, the best guesses are not X_1 through X_8 respectively. Rather we get a better estimate for each of the eight parameters by a formula that makes use of the other data points than any measure that estimates the parameters separately.

Consider taking the MLE of each population parameter based on each sample mean separately. Let x be the n -tuple of the observed values of $X_1, X_2, \dots, X_n$. Whereas x is the MLE estimate of the n -tuple of population means, the Stein estimator is

$$\left(1 - \frac{(n-2)\sigma^2}{|x|^2}\right)x,$$

where σ^2 is the common variance of each random variable, which we assume to be known. The Stein “correction” has the effect of shifting all the values towards zero. Compare these two estimates. In the Stein case, we get a better estimate of the eight parameters than what we can get in the first case, where “better” means that the *sum* of the expected squared errors is lower. In fact, Stein proved that the Stein estimator dominates the MLE estimator when n is three or more although for any single estimate MLE could perform better. In addition, for a specific estimate of a single parameter, the MLE and Stein estimate could be comparable. The fact that the Stein estimator dominates MLE and its cognates shows that the latter are suboptimal. The MLE and the like are not optimal in correctly estimating three or more unrelated parameters simultaneously.

Stein estimation has a seemingly paradoxical nature. As a result, it is also called “Stein’s Paradox.” When estimating the expected doughnut sale in store 1 in 2010 (i.e., μ_1), we should use the observations of the other stores even though their sales are statistically independent of doughnut sales. It seems paradoxical that the estimate for μ_1 should depend on X_2 or X_3 , since they are statistically independent. We need to be careful about the exact import of the Stein’s estimator. If we are only interested in minimizing the expected squared error for the sales in store 1, then there is no advantage of using the other variables. It is the sum of the expected squared errors that is made reliably smaller by Stein’s estimator, and not the expected squared errors individually.

One way to provide an intuitive understanding behind the Stein phenomenon and thereby to take away some paradoxical air from “Stein’s paradox” is to think of estimating eight parameters simultaneously as being randomly generated from a single distribution with one common mean, π , in terms of the expected annual sale of in 2010 per 10,000 people over eight stores. Then, even though every

for other unrelated items in 2009. Since this will make the discussion more complex, we have worked with the doughnut store example as stated above with some oversimplification.

observation $X_1, X_2, X_3 \dots X_8$ contains information about the common mean, π , and X_1 contains the most information about μ_1 , $X_2, X_3, X_4, \dots X_8$ also contains some indirect information about μ_1 . Whether this intuitive clarification really takes away the paradoxical nature of Stein phenomenon, one should be careful to remember that Stein's result is a mathematical result that holds independently of whether an intuitive explanation makes sense. Both Charnigo and Srinivas have made this point very clear in their chapter.

To return to the theme of the paradoxical nature of Stein's phenomenon, Robbins [1951] wrote " X_1 could be an observation on a butterfly in Ecuador, X_2 on an oyster in Maryland, X_3 the temperature of a star, and so on." The seeming weirdness of Stein's phenomenon is to ask, "to estimate the number of butterflies in Ecuador, should one jointly estimate the number of butterflies in Ecuador and oysters in Maryland"? If what we want to estimate is the number of butterflies, then we should estimate it. But, if we are concerned with both butterflies in Ecuador and oysters in Maryland, then we should get an estimate for both. For the latter estimating these two quantities jointly will be quite reasonable since Stein's phenomenon would allow for reduction in the expected error.

7.6 Data-mining

In the twenty-first century, we cannot afford to live without the use of data in some form or the other, whether the data involve our credit card information, information about car theft, or information about our genotype. How investigators could make use of data effectively to extract the right sort of information from them is a daunting task. Although there are traditional statistical tools available, today we tend to rely increasingly on computer applications for processing information from the data. Choh Man Teng in her chapter gives an overview of the area where data mining is at work. Knowledge discovery from data is one area that extracts information from a huge body of data. The information that can be discovered from several such areas need not be mutually exclusive. These areas could be understood as performing different tasks, namely (i) description, (ii) prediction, and (iii) explanation. Descriptive tasks consist of describing a large amount of data succinctly. This task does not involve uncertainty. The prediction task involves finding a mechanism that will reliably foretell the value of a target feature of an unseen instance based on information about some known features. Unlike the descriptive task, it makes an inference that involves uncertainty. However, in the prediction task, predictive accuracy is a central concept. The explanation task aims at discovering the underlying mechanism that generates the data thus providing an explanation for the generation of this set of data rather than another. To retrieve or construct a model from a given data set, investigators sometimes face both under-fitting and over-fitting problems and thus construction of a model results in a trade-off between bias and variance. Here, many statistical tools and analysis are applied to resolve or minimize the problem.

Interestingly, the reference class problem, well known to philosophers of probability, arises in data-mining. Consider an example of the reference class problem to see how the latter could arise in data-mining. In one dorm, there are students except Mary who watched a rental movie, which 90% of them dislike. In another dorm, all of them are French majors; 95% of them liked the film very much. Mary is a French major, but Mary lives in the first dorm. Mary is yet to watch the movie. Mary has no access to any other information on which to base an estimate of how much she will like the movie. Which reference class should Mary belong to? Teng discusses some measures of handling the reference class problem. Besides the reference class problem, she also discusses how sometimes we fall back on experts both to ensure correct analysis and to check whether procedures are properly executed. She also considers the advantages and disadvantages of replacing experts or statisticians by automatic computer guided data analysis.

8 AN APPLICATION OF STATISTICS TO CLIMATE SCIENCE

A great deal of statistical research has been used to analyze issues that confront modern society. The topics range from determining the causes of the cholera epidemic in London in 1854, to finding the cause of arsenic poisoning in Bangladesh, to the current hotly debated topic of climate change. Mark Greenwood, a statistician, teamed with Joel Harper and Johnnie Moore, two geo-scientists, to write a paper on the application of statistics to climate change focusing on finding evidence of climate change in a specific area. They provide an application of statistical modeling to climate change research. They assess the evidence of climate change in the timing of northern Rocky Mountains' stream-flows, which is one way of measuring potentially earlier snowmelt caused by warming climate. A nonparametric spatial-temporal model is used to assess evidence for change and then to estimate the amount of change by using that model. The methods illustrate estimation of nonlinear components in complicated models that account for spatial-temporal correlations in measurements. They found evidence for the presence of climate change by using a model selection criterion, the AIC (see section 2.5 for more on the AIC framework). What they have done in this paper is to use AIC with the assumption that data have a linear trend, and then applied AIC with the assumption that data have a non-linear trend. Strikingly, AIC with the non-linear trend agrees with the historical data much better, although the non-linear model uses many more degrees of freedom than the model that assumes the change has been linear. They compare results for models that include the non-linear trend, and then they constrain it to be linear; later, they remove it entirely from the model. Using AIC to assess evidence of climate change is a new idea because much of the literature on climate change has focused on hypothesis testing methods for evidence of climate change.

In contrast to much of the other research on stream-flow timing measures in the Western US, their methods provide a regional level estimate of the common climate change signal. They find the trend to be more complicated than a linear trend.

Thus, they conclude, the impacts in this region on the timing of stream-flow are *not linear* over the time period. In fact, their findings imply that the magnitude of the change is lower than has been previously suggested in other research. They propose methods that have the potential for application in other areas of climate change research, where *a priori* assumptions of linear change in systems over time may be suspect and the data sets are collected over space as well as time.

9 HISTORICAL TOPICS IN PROBABILITY AND STATISTICS

The final two papers of the volume are of historical nature. One chapter examines how the subjective and objective notions in probability evolved and developed in the western world over a period of three centuries, culminating in the middle of nineteenth century. The other paper looks at how the notion of probability has been used in Ancient India before probability made inroads in the western world. Let us begin with Sandy Zabell's paper on the first topic. The use of subjective and objective probabilities has been a great source of debate currently among different statistical paradigms. Bayesians often are pitted against non-Bayesians when the former are taken to invoke subjective priors, whereas the latter eschew them. Zabell looks closely at the historical development of the subjective and objective probabilities in the western world dating back to seventeenth century. Although there is no clear-cut distinction between subjective and objective probability reasoning as early as in seventeenth and eighteenth centuries, he contends that a clear-cut distinction between the two, subjective and objective probabilities, has emerged much later in the middle of the nineteenth century.

Looking at the original texts of the giants of the probability theory like Bernoulli brothers, Cournot, Poisson, Laplace, De Morgan, D'Alambert including Mill, Venn and others- he has set the debate in the context of some well-known philosophers of that era like Voltaire, Leibniz, Kant and Hume. Thus, Zabell is able to create a perspective where one could see the historical development of the distinction between the subjective and objective. In this paper, there are several historical and philosophical comments that many modern probability theorists might find fascinating. He discusses how meanings of some words like "subjectivity" underwent a sea-change as the expression "subjectivity" was used to mean "objectivity" during sixteenth century. Descartes' *Meditation III* is a case in point. He also discusses the influence of Kant on the thinking about probability, although the probability statement is not included in Kant's list of judgments. His interest does not overlook little known figures like Ellis. He discusses how Venn's *The Logic of Chance*, although borrowing most his ideas from Ellis' work, makes only one passing reference to it. Some of the most exciting comments of Zabell occur when he distinguishes three senses of epistemic probability while he discusses Bertrand's work. These three senses are

1. an epistemic probability could be subjective in the first sense when it is an attribute of the subject perceiving rather than the object perceived.

2. an epistemic probability could be subjective in the second sense because it can vary from one individual to another as different individuals might possess different information.
3. an epistemic probability could be subjective in the third sense because it could vary from one individual to another even if they possess the same information.

He attributes the first sense to Cournot and Venn and the second one to Bertrand. The third sense he thinks goes at the heart of the debate between the objective and subjective probability theorists of the 20th century (see section 2.3 for more on this debate).

C.K. Raju investigates the notion of probability and to some extent mathematics in Ancient India. He discusses how the notion of permutation and combination were connected with poetic meters and allied themes before the 3rd century CE in the *Chandahsutra*. He explains how the ideas of permutation and combination were actually applied to the Vedic meter, the earliest known written accounts relating to permutations and combinations. This dates back even earlier than the third century CE. He discusses how the game of dice is central to understanding probabilistic reasoning in Ancient India, because there are numerous stories in Ancient India where many of the protagonists were addicted gamblers. Raju goes back to some of the claims made by Ian Hacking who attributes it to an Indian statistician, Godambe, and claims that the notion of sampling is found first in the *Mahabharata*, one of two epics of Indian culture written more than 2000 years ago [Hacking, 1975].

Raju discusses the story behind Hacking's attribution of the idea of sampling to the *Mahabharata*. In the Nala-Damayanti episode, the King Rituparna told Nala, his charioteer, exactly how many nuts the entire tree had by sampling just one branch of the tree. However, Raju suggests why we will not find numerous references to the art of probabilistic reasoning in Ancient India by referring to the Nala-Rituparna dialogue. In this dialogue, Nala asked the king "how he could tell the exact number of nuts in the tree without counting them individually." Rituparna, the king, said "this knowledge is secret," implying that it should seldom be shared with others. This observation is as old as India herself. This is one of the reasons for the West tends to call much of ancient Indian philosophy "mystical" and "irrational." Raju shows that a philosophy of mathematics, which he calls "zeroism" can resolve a long-standing difficulty with the frequentist interpretation of probability, namely that relative frequency converges to probability only in a probabilistic sense. Further, he explores the relation of both Buddhist and Jaina logic to probabilities in quantum mechanics as, all of them, both Buddhist and Jaina logic, and probabilities in quantum mechanics, are non-truth-functional. A logic is considered to be non-truth-functional if and only if the truth value of a compound sentence is not function of its constituent sentences. Raju discusses how the distribution law between two truth-functional operators "*and*" and "*or*" fails in quantum mechanics. In a double-slit experiment, to say, that 'the electron reached

the screen and passed through slit *A* or slit *B*’ is *not* the same thing as saying that ‘the electron reached the screen and passed through slit *A* or the electron reached the screen and passed through slit *B*.’ In one case, the author contends, we get a diffraction pattern, whereas in the other, we get a superimposition of two normal distributions. Exploiting this result in quantum mechanics, Raju explains in which sense the Buddhist and Jaina logic are quantum mechanical.

10 PLANS BEHIND ARRANGING MAJOR SECTIONS/CHAPTERS

Having discussed all the papers briefly in our volume, it is worthwhile to give the reader some idea why we divided the entire volume into fourteen major divisions. Since many of the chapters have addressed key issues in philosophy of statistics, those issues and topics lend themselves to technicalities and intricacies in probability and statistics. To get our readers up to the speed, so to speak, regarding various papers in the volume, Prasanta Bandyopadhyay and Steve Cherry have written a primer on probability and statistics without presupposing any technical knowledge. The notion of conditional probability is a central concept for probability and statistics. However, it has raised several philosophical problems regarding its correct interpretation. As already noted, we have a section heading on “Philosophical Controversies about Conditional Probability” which contains two papers: (i) Hájek’s paper on “Conditional Probability” followed by Easwaran’s paper on “The Varieties of Conditional Probability,” which explain the concept in detail.

Given this preparation on the fundamentals of probability and statistics, we have then introduced four influential statistical paradigms: (1) Classical/error-statistics, (2) Bayesianism, (3) Likelihoodism, and finally (4) the Akaikean framework. Each section dealing with the four paradigms usually consists of more than one paper. The classical/error statistics section consists of two papers: (i) Mayo and Spanos’ paper on “error statistics” followed by Dickson and Baird’s paper on “Significance Testing.” Our largest section is on Bayesian paradigm. We divide that section under several sub-headings. Under the subsection on “Subjectivism”, we have Weirich’s paper on “Bayesian Decision Theoretic Approach.” The subsection on “Objective Bayesianism” consists of two papers: (i) Bernardo’s paper on “Modern Bayesian Inference: Foundations and Objective Methods” and (ii) Wheeler and Williamson’s paper on “Evidential Probability and Objective Bayesian Epistemology.” The next subsection on “Confirmation Theory and its Challenges,” consists of two papers: (i) Hawthorne’s “Confirmation Theory” followed by (ii) Norton’s paper “Challenges to Bayesian Confirmation Theory.” The subsection on “Bayesianism as a form of “logic” consists of (i) Howson’s paper “Bayesianism as a Pure Logic of Inference”, and (ii) Festa’s paper on “Bayesian Inductive Logic, Verisimilitude and Statistics.” The next paradigm is the likelihood framework. Two papers, (i) Blume’s paper on “Likelihood and its Evidential Framework”, and (ii) Taper and Lele’s paper on “Evidence, Evidence Functions, and Error Probabilities” belong to this section. The section on the Akaikean framework has just only one paper which is, Forster and Sober’s paper on “AIC Scores

as Evidence: A Bayesian Interpretation.” The next section consists of a single paper by Grossman on “The Likelihood Principle.” This is followed by the section called “Recent Advances in Model Selection.” This section contains two papers: (i) Chakrabarti and Ghosh’s paper on “The AIC, BIC and Recent Advances in Model Selection” and Dawid’s paper on “Posterior Model Probabilities.” Section six titled “Attempts to Understand Aspects of Randomness” has two papers. The first one is by Deborah Bennett on “Defining Randomness” and the second one is by Abhijit Dasgupta on “Mathematical Foundations of Randomness.”

Our section on “Probabilistic and Statistical Paradoxes” consists two papers. One is by Vineberg on “Paradoxes of Probability” and the other one is by Tsao on “Statistical Paradoxes: Take it to the Limit.” We have a single article by Romeijn under the section heading “Statistics as inductive inference”. This section on “Various Issues about Causal Inference” has two papers. One is by Spirtes on “Common Cause in Causal Inference” and the other one is by Greenland on “The Logic and Philosophy of Causal Inference: A Statistical Perspective.” We have two papers on the section on “Some Philosophical Issues Concerning Statistical Learning Theory.” The first paper is by Harman and Kulkarni on “Statistical Learning Theory as a Framework for Philosophy of Induction” and the second one is by Steel on “Testability and Statistical Learning Theory.” In the next section we have brought three papers under the heading “Different Approaches to Simplicity Related to Inference and Truth.” They are (i) De Rooij and Grünwald’s paper on “Luckiness, and Regret in Minimum Description Length” and (ii) Dowe’s paper on “MML, Hybrid Bayesian Network Graphical Models, Statistical Consistency, Invariance, and Uniqueness” and (iii) Kelly’s paper on “Simplicity, Truth, and Probability.” The section on “Special Problems in Statistics/Computer Science” includes three papers. They are (i) Boik’s paper on “Normal Approximations,” (ii) Charnigo and Srinivasan’s paper on “Stein Phenomenon”, and thirdly and finally (iii) Teng’s paper on “Data, Data, Everywhere: Statistical issues in Data Mining.” Greenwood, Harper, and Moore’s paper on “Applications of Statistics in Climate Change: Detection of Non-Linear Changes in a Stream-flow Timing Measure in the Columbia and Missouri Headwaters” constitutes a single contribution in the section on “A Statistical Application to Climate Change”. Our last section on the “Historical Approaches to Probability and Statistics” consists of two papers: (i) Zabell’s paper on “The Subjective and the Objective”, and Raju’s paper on “Probability in Ancient India”.

11 CODA¹⁴

We began our journey with the claim that both philosophers and statisticians are in a sense interested in the same problem, i.e., “the problem of induction”. The papers discussed in this Introduction explain at some length how they approach the problems and issues pertaining to statistical inference, broadly construed. How-

¹⁴The section has been promoted by Jayanta Ghosh’s observation.

ever, the time has come when we need to take pause and enumerate what or whether we have learned about the “philosophy of statistics” from these papers.

Readers have no doubt noticed disagreements between various authors regarding the proper approach to statistical problems. Observing such a disagreement, the British philosopher, George Berkeley might have commented, “[they] have first raised a dust and then complain that [they] cannot see [Berkeley, 1970, 1710].” However, this time “they” include both philosophers and non-philosophers.

We mentioned in the beginning that the philosophy of statistics is concerned with foundational questions in statistics. In fact, the debate in philosophy of statistics is, in some sense, comparable to the debate in the foundations of mathematics that began in the early part of 20th century. It is well-known that Frege, working on the foundations of arithmetic, assumed the consistency of set-theoretical axioms, that Russell showed later to be inconsistent. This was followed by Gödel’s incompleteness results shaking the very foundation of mathematics. It is true that there has not been any comparable result in the foundations of statistics as in the foundations of mathematics. Yet there is a similarity between the two in terms of the debates about their foundations and their subsequent developments toward real-world applications. Despite the fact that there are foundational questions that remain unresolved in mathematics, the latter has been used and applied extensively including in the development of the Newtonian physics, relativistic theories, quantum mechanics, the construction of the double helical structure of the DNA and the like that make stunning predictions. The point of this is that foundational debates have no longer stopped mathematics to be exploited, and then later on used to solve real world problems. In the same vein, even though there are debates about the correct foundation of statistics, statistics has an applied side in which different statistical tools and soft-wares have been widely used to make diagnostics studies, weather forecasting, estimating the possibility of the distribution of life in and beyond our solar system and so on.

The question remains, “what then is the philosophy of statistics that is taught by our esteemed authors?” Without waiting to see whether the dust has fully settled on the foundational issues, one could venture to address this pressing question. For a sympathetic reader, philosophy of statistics will appear to be more like a mosaic of themes consisting of several seemingly unrelated features. Different statistical schools contribute different aspects of our understanding of its underlying themes. Error statistics provides us with tools for handling errors in testing hypotheses while making us aware of the key role “errors” play in scientific investigations. The likelihood framework shows why the concept of evidence is so crucial in science and consequently needs to be separated from the concept of belief. The Akaikean framework is interested in the concept of prediction and provides a way of capturing this concept to some extent. In contrast, Bayesianism is an over-arching school which covers many of the concepts just mentioned if one is allowed to have prior information about the topics in question. In this mosaic of themes, both computer scientists and mathematicians have not lagged behind. They also help us understand the uncertainty involved in carrying out inductive

inference. Similarly, applied statisticians have helped us to appreciate another side of this problem. This applied side of statistics, in the form of making reliable inference, causal and non-causal in which new algorithms and causal diagrams, has begun to make breakthroughs in wider contexts, like diagnostic studies. Two historical papers have also their place in this mosaic. One traces back the debate about the subject/object notions of probability as early as to the seventeenth century western intellectual history. The other paper contends that we can't afford to be ethno-centric about some of the probabilistic/statistical notions as the latter can be found in other tradition(s) much before the emergence of probability and statistics in the western tradition. Much more research needs to be conducted in the emergence of probability in non-western traditions.

The philosophy of statistics has this huge canvass that includes, among other topics, these theoretical and practical sides that have so far been treated as two different aspects, with literally no hope of converging. Hopefully, in the next quarter of a century, we might be able to see much more interaction between them, so that the next generation of this series will keep us abreast of this symbiosis in addition to the raging foundational debates that will remain a staple of philosophers for many years to come.

ACKNOWLEDGEMENTS

We have received numerous help from several people while writing this introduction. Individuals who have offered their help in terms of their expertise are Deborah Bennett, Jeffrey Blume, Robert Boik, Richard Carnigo, Arijit Chakrabarti, Abhijit Dasgupta, Jason Davey, Philip Dawid, Steven de Rooij, David Dowe, Roberto Festa, Mark Greenwood, Peter Grünwald, Alan Hájek, Gilbert Harman, James Hawthorne, Colin Howson, Kevin Kelly, Deborah Mayo, Sanjeev Kulkarni, John Norton, C. K. Raju, Jan-Willem Romeijn, Peter Spirtes, Cidambi Srinivasan, Mark Taper, Choh Man Teng, C. Andy Tsao, Susan Vineberg, Gregory Wheeler, Paul Weirich, Jon Williamson, and John Woods. We are thankful to Jane Spurr for her willingness to incorporate even very minor changes numerous times when the introduction has been written. We are especially indebted to Gordon Brittan, Jayanta Ghosh, John G. Bennett, Sander Greenland and Billy Smith for their substantive suggestions regarding the entire manuscript either in terms of helping us sharpen various philosophical/statistical arguments or improving our writing style. Without the help from all these individuals, it would have been a much worse paper. We are, however, solely responsible for any error that it might still contain. PSB's research has been supported by the NASA's Astrobiology Research Center grant (#4w1781.)

BIBLIOGRAPHY

- [Berger, 1985] J. Berger. *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. Berlin: Springer-Verlag, 1985.
- [Berkeley, 1710] G. Berkeley. *A Treatise Concerning the Principles of Human Knowledge*. Turbayne, C. M., ed., Indianapolis: Bobbs-Merrill, 1710; Edition used is by Turbayne's 1970 edition.
- [Bernardo, 2005] J. Bernardo. Reference Analysis. In *Handbook of Statistics*. 25. P.17-90. (D.K. Dey and C.R. Rao, eds). Elsevier, North-Holland, 2005.
- [Bernardo, 1997] J. Bernardo. Noninformative Priors Do not Exist: A Dialogue with Jose Bernardo. *Journal of Statistical Planning and Inference*. 65, p.157-189, 1997.
- [Bernardo and Smith, 1994] J. Bernardo and A. Smith. *Bayesian Theory*. New York. John Wiley, 1994.
- [De Finetti, 1937] B. De Finetti. La prevision: ses lois logiques, ses sources subjectives. Translated in H.E. Kyburg, Jr. and H.E. Smokler (eds.) *Studies in Subjective Probability*, New York: Wiley, 1964, pp. 93-158, 1937.
- [Fisher, 1973] R. Fisher. *Statistical Methods and Scientific Inference*. 3rd edition. New York: Wiley, 1973.
- [Forster and Sober, 1994] M. Forster and E. Sober. How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions. *British Journal for the Philosophy of Science* 45. Pp.1-36, 1994.
- [Greenland *et al.*, 1999] S. Greenland, J. M. Robins, and J. Pearl. Confounding and Collapsibility in Causal Inference. *Statistical Science*, Vol. 19, pp. 29-46, 1999.
- [Grünwald, 2007] P. Grünwald. *The Minimum Description Length Principle*, MIT Press, 2007.
- [Hacking, 1975] I. Hacking. *The Emergence of Probability*. Cambridge University Press. UK, 1975.
- [Howson and Urbach, 2006] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*, 3rd edn. Open Court, Illinois, 2006.
- [Hume, 1739] D. Hume. *A Treatise of Human Nature*. London, 1739. Edition referred is by P.H. Nidditch, Oxford, The Clarendon Press, 1978 edition.
- [Joyce, 1999] J. Joyce. *The Foundations of Causal Decision Theory*, Cambridge University Press, UK, 1999.
- [Lele, 2004] S. Lele. Evidence Functions and the Optimality of the Law of Likelihood. In M. Taper, and L. S (eds.) *The Nature of Scientific Evidence*. University of Chicago Press. Chicago, 2004.
- [Lewis and Shelby-Richardson, 1966] D. Lewis and J. Shelby-Richardson. Scriven on Human Unpredictability. *Philosophical Studies*, vol 17, number, 5, Pp.69-74, 1966.
- [Martin-Löf, 1965] P. Martin-Löf. The Definition of Random Sequences. *Inform. Control* 9. 1966. pp.602-619, 1965.
- [Neyman, 1967] J. Neyman. *A Selection of Early Papers by J. Neyman*. Berkeley: University of California Press, 1967.
- [Pagano and Gauvreau, 2000] M. Pagano and K. Gauvreau. *Principles of Biostatistics*. (Second edition). Duxbury, Australia, 2000.
- [Robins, 1951] H. Robins. Asymptotically Subminimax Solutions of Compound Statistical Decision Problems. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. L. LeCam and J. Neyman (eds.) pp.355-372. University of California Press, 1951.
- [Royall, 1997] R. Royall. *Statistical Evidence. A Likelihood Paradigm*. Chapman & Hall. London, 1997.
- [Savage, 1972] L. Savage. *Foundations of Statistics*, Dover, New York, 1972.
- [Scriven, 1965] M. Scriven. An essential unpredictability in human behavior. In *Scientific Psychology: Principles and Approaches*, B. B. Wolman and E. Nagel, eds, pp. 411-425. Basic Books (Perseus Books), 1965.
- [Seidenfeld, 1979] T. Seidenfeld. *Philosophical Problems of Statistical Inference: Learning from R. A. Fisher*, Dordrecht, Boston. D. Reidal Publishing Company, 1979.
- [Skyrms, 1984] B. Skyrms. Learning from Experience. In *Pragmatics and Empiricism*. Yale University Press: New Haven; pp. 37-82, 1984.
- [Sober, 2008] E. Sober. *Evolution and Evidence*. Cambridge University Press. United Kingdom, 2008.

- [Taleb, 2010] N. Taleb. *The Black Swan: The Impact of the Highly Improbable*, 2nd edn. Random House. New York, 2010.
- [Vapnik, 1998] V. Vapnik. *Statistical Learning Theory*. New York. John Wiley, 1998.
- [Weirich, 2004] P. Weirich. *Realistic Decision Theory*. Oxford, 2004.
- [Zabell, 2005] S. Zabell. *Symmetry and its Discontents: Essays on the History of Inductive Probability*. Cambridge University Press, United Kingdom, 1998.

Part I

Probability & Statistics

This page intentionally left blank

ELEMENTARY PROBABILITY AND STATISTICS: A PRIMER

Prasanta S. Bandyopadhyay and Steve Cherry

1 INTRODUCTION

Some of the chapters of this volume, albeit thorough, are technical and require some familiarity with probabilistic and statistical reasoning. We believe that an introduction to some of the basic concepts of probability and statistics will be helpful for our general reader. Probability and statistics provide necessary tools to *capture* the *uncertain state* of our *knowledge*. In addition, probability/statistics will also provide necessary tool-kits to *quantify* our uncertainties. Our discussion will be mostly quantitative. Patient readers who work through some of those technicalities will be rewarded in the end by seeing how the technicalities contribute to a better appreciation of several of the papers which on many occasions lend themselves to philosophical issues of great importance.

Below we provide a brief introduction the theory of probability and statistics. The introduction is at the level found in introductory statistics classes and discussed in many statistical textbooks. Typical texts used over the past few years at Montana State University include Devore [2008], Moore and McCabe [2006], and Deveaux *et al.*, [2008]. Much of the material and many of the examples presented below are taken from Devore.

We begin our discussion on the nature of probabilistic/statistical reasoning by contrasting probabilistic/statistical reasoning/inference with deductive reasoning/inference. The property of deductive validity of an argument is central to understanding the distinction between deductive inference and inductive inference, when deductive validity could be understood in terms of the monotonic property of reasoning. First, we define “deductive validity” followed by an understanding of the property of monotonicity. An argument is deductively valid if and only if it is logically impossible for its premises to be true, but its conclusion to be false. A sentence is a *deductive consequence* of others when it is logically impossible that they should be true but that sentence is false. Consider the set S_1 which consists of the sentence p , and the sentence $p \rightarrow q$, where “ $\rightarrow$ ” captures any if p then q sentences and the symbol “ $\rightarrow$ ” is known as “material conditional.” Here, q is a deductive consequence of S_1 , which we write as “ $S_1 \models q$.” The symbol “ $\models$ ” denotes deductive consequence relation, which is a relation between a set of sentences and

a sentence. Monotonicity is a property of certain types of inferences and is appreciated in terms of the deductive consequence relation. A relation between sets of sentences and a sentence is *monotonic* if and only if when it holds between a set and a sentence, it also holds between any superset of the set and that sentence. S_2 can be taken to be the superset of S_1 consisting of S_1 and $\sim p$. The symbol “ $\sim$ ” means “it is false that.” Any deductive consequence relation represented by “ $\models$ ” is by definition monotonic. For example, if S_2 is any superset of the set S_1 above (for which we have $S_1 \models q$) then $S_2 \models q$ must hold too.

To appreciate this point, consider the rule of *modus ponens*, which is a well-known rule of deductive logic.

1. If it rains, then the ground will be wet.
2. It has rained. Therefore, the ground will be wet.

“ p ” which is called “the antecedent” of the if-then sentence, represents the proposition “it rains.” “ q ”, which is called “the consequent” of the if-then sentence, represents the proposition “the ground will be wet.” The standard rules of deductive logic including *modus ponens* imply that adding new premises to a store of information can only *increase* the class of conclusions that can be deduced from them. If one adds, for example, $\sim p$ to the above set comprising of premise 1 and premise 2, then one would be able to deduce q along with other additional conclusions from the set, $\{\sim p, p, p \rightarrow q\}$, although the entire set would turn out to be inconsistent. Non-monotonic reasoning, which underlies much of our probabilistic/statistical reasoning, by contrast, allows the possibility that adding new information can actually result in *dispensing* with conclusions previously held. Many of our daily experiences are characteristic of non-monotonic reasoning in which we lose information in light of new evidence.

Consider the following example.

1. All *observed* crows are black.
2. Therefore, all crows are black.

Suppose we have observed so far that all crows are black. Seeing a picture of an albino crow (and assuming we have no reason to be skeptical about that picture) in the local newspaper, our belief that all crows are black is undermined. So an addition of information to our belief set could force us to lose some of our cherished beliefs. Here, the new premise about the presence of an albino crow has led to the rejection of its conclusion that all crows are black. This loss of information is solely a feature of non-monotonic reasoning that underlies much of probabilistic/statistical reasoning. Although we have introduced the idea of non-monotonic and monotonic reasoning intuitively in terms of whether we lose our existing information in one case and not in another, we need to be aware that it is the deductive validity that goes at the heart of monotonic reasoning which can’t be undermined by adding a new premise. The purpose of our examples

is just to illustrate this theme behind monotonic reasoning. Probability theory provides a better tool for handling inductive arguments via which non-monotonic reasoning has primarily been expressed. We will begin with probability theory. At the final section of this chapter, after learning about probability theory and statistics, we will, however, return to the theme of non-monotonic reasoning to evaluate whether arguments used in statistical/probabilistic inference *necessarily* involve non-monotonic reasoning.

2.1 Basic and derived rules of the probability calculus

Probability theory provides a tool for handling inductive arguments. Typically the first step is to define some needed terms. We define a probabilistic experiment (broadly) to be any process that produces outcomes *which are not predictable in advance*. The toss of a single six-sided die is a simple example. We know that one of six numbers between one and six will occur on any given toss but prior to a toss we cannot state with certainty which outcome will be observed. A list of all the possible outcomes of an experiment is called the sample space. We may be interested in the individual outcomes themselves or in sets of outcomes, i.e. we will be interested in subsets of the sample space and will refer to such subsets as events. In tossing a single six-sided die once the sample space is

$$S = \{1, 2, 3, 4, 5, 6\}.$$

We may be interested in the event of observing an even number. This is the subset $\{2, 4, 6\}$. If any of the individual outcomes in this event occur then the event itself is said to have occurred. This is true in general. For example, if we toss the die and observe a 3 then the events $\{1, 3, 5\}$ and $\{1, 2, 3\}$ have both occurred as has any other event containing 3 as an outcome. For any sample space the empty or null set is considered to be an event along with the entire sample space itself.

The union of two events $A \cup B$ is the event that occurs if either A or B or both occur. The intersection of two events $A \cap B$ is the event that occurs if both A and B occur. The complement of any event A is denoted A^c is the event comprised of all outcomes that are not in A . A collection of events $A_1, A_2, \dots, A_n$ is said to be mutually exclusive (or disjoint) if none of the pairs have any outcomes in common.

The probability of an event A , denoted $P(A)$ is a quantitative measure of how likely we are to observe A in a single trial of an experiment. Initially, we define or interpret probability as is typically done in introductory statistics text books. The probability of event is the long run relative frequency with which the event occurs over very many independent trials of the experiment. This definition is controversial in part because it does not always make sense (What is the probability the Denver Broncos will win the Super Bowl in 2010?). For further discussion and additional interpretations of probability see [Hájek, 2007]. We note that the mathematical rules and properties of probability described below do not depend on the specific interpretation of probability.

The mathematical development of probability starts with three basic rules or axioms:

1. For any event A ; $0 \leq P(A) \leq 1$. If A has probability 0 then it is impossible and if its probability is 1 then it is certain to occur. Any event of practical interest will never have probabilities of 0 or 1.
2. Denoting the sample space by S , $P(S) = 1$ (something has to happen). S could be considered a tautology.
3. For a sequence of mutually exclusive events $A_1, A_2, A_3, \dots$,

$$P(A_1 \cup A_2 \cup A_3 \dots) = \sum_{i=1}^{\infty} P(A_i)$$

All other rules and properties of probability can be derived from these three axioms.

Two simple rules that are usually established quickly are that the probability of the empty set $\emptyset$ is 0 and the so-called Complement Rule: for any event A

$$P(A^c) = 1 - P(A).$$

Another rule that can be derived quickly is a modification of axiom 3. For a finite sequence of mutually exclusive events $A_1, A_2, \dots, A_n$

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n P(A_i).$$

Most events are not mutually exclusive. It can be shown that for any two events A and B

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Intuitively this makes sense. The probability that event A or event B (or both) occur is equal to the probability that A occurs plus the probability that B occurs minus the probability that both occur (which has been double counted by the addition of the probabilities of A and B). Note that if A and B are mutually exclusive then their intersection is the empty set with probability 0 and we are back to the addition rule for mutually exclusive events given above.

This rule can be extended to more than two events: for any three events A, B , and C ,

$$\begin{aligned} P(A \cup B \cup C) = & P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - \\ & P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

In any given experiment we need to be able to assign probability to outcomes in some way. There are potentially many different ways of doing so as long as they do

not violate the axioms and other rules of probability. We have seen, for example, that the sample space for the experiment of tossing a six-sided die once is

$$S = \{1, 2, 3, 4, 5, 6\}.$$

We note that the outcomes are six mutually exclusive events and so

$$P(S) = P(1) + P(2) + \cdots + P(6) = 1.$$

As long as the probabilities we assign to the individual outcomes are all between 0 and 1 and sum to 1, we will have a mathematically valid probability model for the experiment. There are an infinity of such possibilities. An empirical method of assigning probabilities, based on our frequentist interpretation of probability, is to toss the die a large number of times and record the outcomes. The assignment of probabilities would be based on the proportion of times each outcome occurred. Or we might reason that the die is fair in that we do not expect to see any one outcome more often than any other, i.e. each outcome is equally likely and taken together must sum to 1. Thus, each of the outcomes would be assigned a probability of $1/6$.

2.2 Conditional probability and marginal probability

The sample space of an experiment contains all the possible outcomes. Sometimes additional information becomes available that constrains the set of possible outcomes. The probability of an event then is computed based on the outcomes in the constrained sample space and such a probability is called a conditional probability. Before giving a mathematical definition we look at a simple example.

A geneticist studying two genes has access to a population of 1000 individuals. The table below shows a categorization of the individuals with respect to the genes.

Gene 1	Gene 2 Dominant	Gene 2 Recessive
Dominant	560	240
Recessive	140	60

The experiment will consist of drawing an individual at random from the population. By drawing “at random” we mean that all individuals have the same chance of being drawn, $1/1000$ based on the information given in the table. If the person drawn has two dominant genes we label her $D1D2$. If the person has a dominant Gene 1 and a recessive Gene 2 we label her $D1R2$ and so on. The sample space is

$$S = \{D1D2, D1R2, R1D2, R1R2\}.$$

Based on what we expect to see over the long run a reasonable assignment of probabilities is

$$P(D1D2) = 0.56, P(D1R2) = 0.24, P(R1D2) = 0.14, P(R1R2) = 0.06.$$

This assignment was made by simply counting the number of individuals in each category and dividing by the total population size. Each of the numbers is between 0 and 1 and sum to 1 so the assignment is mathematically valid. Suppose we are interested in the probability that the person selected has a dominant Gene 2. We can directly apply the addition rule for mutually exclusive events and compute this as

$$P(D2) = P(D1D2 \cup D1R2) = P(D1D2) + P(D1R2) = 0.56 + 0.24 = 0.80.$$

In this simple case we can also note that 800 of the 1000 individuals have a dominant Gene 1 and conclude that the probability is $800/1000=0.80$. This is the unconditional (or marginal) probability of selecting an individual with a dominant Gene 2.

Now suppose we are told that we will be selecting our individual from the group with a recessive Gene 1. We now know that we are only dealing with the subpopulation of 200 individuals with such a gene. The sample space is now

$$S = \{R1D2, R1R2\}$$

Of the 200 individuals in this subpopulation 140 are $R1D2$ and 60 are $R1R2$ and so the probability of a dominant Gene 2 is $140/200 = 0.70$. We say that the conditional probability of a dominant Gene 2 given a recessive Gene 1 is 0.70. Mathematically we write

$$P(D2|R1) = 140/200 = 0.70.$$

In general, for any two events A and B with $P(B) \neq 0$, we define the conditional probability of A given B to be

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

Considering the genetics example we see that

$$P(D2|R1) = \frac{P(D2 \cap R1)}{P(R1)} = \frac{140/1000}{200/1000} = \frac{140}{200} = 0.70.$$

Failure to recognize the appropriate restriction of the sample space leads to common errors in calculating probability. Suppose we are told that a family has two children and that at least one is a girl. What is the probability that both are girls? The most common intuitive answer to this question is, assuming boys and girls occur with equal probability, 0.5. This answer is wrong. Let GG denote the outcome of a girl born first and a girl born second, GB denote the outcome of a girl born first and a boy born second and so on. The sample space is

$$S = \{GG, GB, BG, BB\}$$

and, assuming equal chances for a boy or girl, a reasonable assignment of probabilities to the outcomes is 0.25, i.e. each is equally likely. We are being asked to find the conditional probability of GG given or conditional on the information that at least one of the children is a girl. Denoting this event by B , we have $B = \{GG, GB, BG\}$. The probability that GG occurs is 0.25 and the probability that B occurs is 0.75. Note also that $(GG \cap B) = GG$. Thus,

$$P(GG|B) = \frac{P(GG \cap B)}{P(B)} = \frac{P(GG)}{P(B)} = \frac{0.25}{0.75} = 1/3.$$

Suppose we had phrased the question as: What is the probability of two girls given that the oldest is a girl. Denoting the latter event by B , we have $B = \{GG, GB\}$, and

$$P(GG|B) = \frac{P(GG \cap B)}{P(B)} = \frac{P(GG)}{P(B)} = \frac{0.25}{0.50} = 1/2.$$

The definition of conditional probability leads immediately to the so-called Multiplication Rule:

$$P(A \cap B) = P(A|B)P(B)$$

The Law of Total Probability is an important result. Let $A_1, \dots, A_k$ be a sequence of mutually exclusive and exhaustive sets. By exhaustive we mean that the union of these sets is the sample space. For any other event B , the unconditional probability of B is

$$P(B) = P(B|A_1)P(A_1) + \dots + P(B|A_k)P(A_k).$$

This rule can be used to prove Bayes's Rule (or Theorem):

Let $A_1, \dots, A_k$ be a sequence of mutually exclusive and exhaustive sets, and B be event such that $P(B) \neq 0$. Then

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_i^k P(B|A_i)P(A_i)}.$$

The middle term is just an application of the definition of conditional probability. The numerator of the last term follows from the Multiplication Rule and the denominator is the Law of Total Probability.

Bayes' Rule provides insight into another type of problem for which our intuition leads to incorrect probability calculations. Suppose a test exists for a rare (only 1 in 1000 adults has the disease) but serious disease and that the test is quite accurate in that when the disease is present it will return a positive result 99% of the time and when the disease is absent a positive result will occur only 2% of the time. You are given the test and receive a positive result. How worried should you be? Your first inclination would be to focus on how likely the test is to return a positive result when the disease is present and many people would choose that figure (0.99) as the probability of having the disease. But the event of interest is not observing a positive result given the disease but the event of having the disease

given a positive test result. Let D denote the event of having the disease and $+$ denote the event of testing positive. Based on the above information we know the following:

$$P(D) = 0.001, P(+|D) = 0.99, P(+|D^c).$$

By Bayes' Rule

$$\begin{aligned} P(D|+) &= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)} \\ &= \frac{0.99(0.001)}{0.99(0.001) + 0.02(0.999)} \\ &= \frac{0.00099}{0.02097} = 0.047. \end{aligned}$$

2.3 Probabilistic independence and generalized probability conjunction rule

Two events are said to be probabilistically independent if and only if the probability of the occurrence or non-occurrence of one event in no way affects the probability of the occurrence or non-occurrence of the other. Mathematically, we define two events to be independent if

$$P(A|B) = P(A).$$

It can be shown that this equality also implies $P(B|A) = P(B)$. Note that under an assumption of probabilistic independence the Multiplication Rule becomes

$$P(A \cap B) = P(A|B)P(B) = P(A)P(B).$$

Some textbooks define independence in this way and then show that independence implies

$$P(A|B) = P(A).$$

Typically in practical problems independence is an assumption that is made. It may be reasonable in some cases such as in assuming that repeated tosses of a six-sided die yield outcomes that are independent of one another (getting a 6 on one toss does not affect the chances of getting a 6 on the next, or any other toss). But in other situations it may not be a reasonable assumption. Subsequent to the explosion of the space shuttle Challenger in 1988 it was determined that the computed risk of catastrophic failure had been seriously underestimated due, in part, to unwarranted assumptions of independence of failure of related parts.

Mutual independence of more than two events is more complex. Mathematically it is defined as follow. A sequence of events $A_1, \dots, A_k$ are said to be mutually independent if for every $k(k = 2, 3, \dots, n)$ and every set of indices $i_1, \dots, i_k$

$$P(A_{i_1} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \dots P(A_{i_k}).$$

For example, three events A , B , and C are mutually independent if and only if all the following conditions are met

$$\begin{aligned} P(A \cap B) &= P(A)P(B) \\ P(A \cap C) &= P(A)P(C) \\ P(B \cap C) &= P(B)P(C) \\ P(A \cap B \cap C) &= P(A)P(B)P(C) \end{aligned}$$

Pairwise independence does not imply mutual independence. Consider tossing two coins, a quarter and a dime. Let $H1$ denote the event of observing a heads on the quarter, $T2$ denote the event of observing a tails on the dime, and C denote the event of observing either two heads or two tails. We assume the tosses are independent of one another. The sample space of the experiment consists of four equally likely outcomes of

$$S = \{H1H2, H1T2, T1H2, T1T2\}.$$

We have

$$\begin{aligned} P(H1 \cap T2) &= P(H1T2) = 1/4 = P(H1)P(T2) \\ P(H1 \cap C) &= P(H1H2) = 1/4 = P(H1)P(C) \\ P(T2 \cap C) &= P(T1T2) = 1/4 = P(T2)P(C) \\ P(H1 \cap T2 \cap C) &= P(\emptyset) = 0 \neq P(H1)P(T2)P(C) \end{aligned}$$

Thus, although the events are pairwise independent they are not mutually independent.

2.4 Probabilistic/logical independence and mutual exclusiveness of propositions

Two events may be logically independent without being probabilistically independent. Let A be the event that a randomly chosen individual drinks coffee and B be the event that a randomly chosen individual smokes. These two events are logically independent of one another because one does not imply the other nor does one imply the negation of the other. However, social research has shown that smokers are more likely to drink coffee than non-smokers and thus these two events are probabilistically dependent.

Students in introductory probability courses frequently struggle with the relationship between two events being mutually exclusive of one another and being probabilistically dependent/independent of one another. It seems that it is a natural first inclination to consider two mutually exclusive events to be independent,

but in fact if two events are independent of one another they must overlap in some way. If two events A and B are mutually exclusive (and both have non-zero probability of occurrence.) then knowing one has occurred automatically rules the other out because

$$P(A|B) = 0 \neq P(A).$$

For example, if we let H denote getting a heads on a single toss of a fair coin and T denote getting a tails, then these two events are clearly mutually exclusive and dependent;

$$P(H|T) = 0 \neq P(H).$$

Formally, we say that mutual exclusivity implies dependence but the converse is not true. Thus, dependent events may or may not be mutually exclusive. The negation (independence implies two events are not mutually exclusive) is, of course, also true.

For an example of two overlapping independent events, consider the experiment of tossing a fair coin twice. Let $H1$ denote the event of getting a heads on the first toss and let B denote the event of two heads or two tails. These two events overlap but

$$P(H1 \cap B) = 1/2 = P(H1).$$

3 FROM PROBABILITY TO STATISTICS AND A ROAD-MAP FOR THE REST OF THE CHAPTER

3.1 *The fundamental difference between probability and statistics*

One way to appreciate the difference between statistics and probability is to see how these two disciplines distinctively make inference based on data/sample. By “inference”, we mean the procedure by which one goes from the known data to the unknown value of the parameter which we are interested in knowing/estimating. The style of inference applied in statistics is an inference that goes from a sample to the population. It is the paradigm example of uncertain inference. Uncertainty of statistical inference stems from the fact that the inference is made from the known sample to the unknown population. In contrast, there are other kinds of inference where we make inferences about the unknown samples based on our information about the population. The latter kind of inference is known as non-statistical inference or inference involving probability. In short, drawing an inference from a sample to a population is statistics and drawing an inference from a population to a sample is mathematics that rests on using the theory of probability.

Consider an urn consisting of a large number of balls of the same size, but of different colors. Suppose we know that 25% of the balls are red. We are going to draw a sample of 100 balls randomly with replacement from the urn. What is the probability that the proportion of red balls in the sample will lie between 0.20 and 0.30? Using the laws of probability we can answer this question exactly. Suppose

we know nothing about the proportion of colors in the urn and we sample balls as described above. We wish to use the information in the sample to estimate the proportion of red balls in the population. What is the best way of doing this? If the observed proportion is 0.25 what conclusions can be drawn about the true but unknown proportion of red balls in the population? Is the observed proportion in the sample consistent with a true proportion of say 0.45, i.e. how likely are the observed data if the true proportion is 0.45? These are the types of questions of interest to statisticians (and to scientists using data to draw inferences to populations or processes from which the data were drawn).

3.2 Five Types of questions in statistics

Although we have already touched some of the central issues of statistics above, we will elaborate how we could understand the enterprise of statistics in terms of five types of questions. Since data/sample are at the heart of statistics and often statistics is regarded as the science of data analysis, our questions will begin with data/sample first. These five questions are as follows:

1. How could one describe the sample effectively?
2. From the sample, how one could make an inference about the total population?
3. How reliable will be the conclusion of our inference?
4. (a) Is there is a relation between two or more variables?
(b) If so, then what could we say about the relation? Is it a simple association, or is there a causal relation between them? And finally,
5. How should one collect the sample so that it will help to produce the most reliable estimate?

For the sake of discussion, we are going to outline how different sections revolve around one or two of these questions although in this short review we won't be able to cover all key questions about statistics. (1) pertains to what we ordinarily mean by "statistics" which we take to be a collection of data. We want to know what those data tell us about a single variable representative of some object of interest. Section 4, among other topics, responds to (1). Section 6 which discusses two distinct ways of doing statistical inference, estimation and hypothesis testing, are devoted to addressing (2). In most colleges and universities, these two approaches to doing statistical inference are an essential part of learning statistics. To be able to do statistical inference we need to develop probability models. In one sense, a probability model provides the tools by connecting data to how one should be doing inference in a systematic fashion. Section 5 discusses additional probability results needed to make reliable inferences based on the sample and a probability model. The ideas behind sampling distributions and the central limit theorem

will be expanded in section 5. Section 6 will address (3). Statistics is an ever growing discipline with multi-layered complexity. We cannot address all aspects of this complexity in this short section. As a result, we won't address (5) in detail although it is important.

4 DATA REPRESENTED AND DESCRIBED

Data/sample are the brick and mortar of statistics, but a data set does not have to be particularly large before a simple listing of data values becomes overwhelming and effective ways of summarizing the information in a data set are needed. The first chapter or two in most introductory textbooks deal with graphical and numerical methods of summarizing data. First, we discuss how the bridge between the observation and data we have, and the world they represent could be built in terms of some non-ambiguous terms, like objects, variables and scales which constitute some basics of mathematical statistics. Second, we introduce three measures of central tendency and point out their implications in critical reasoning. Third, we discuss the variance and the standard deviation, the two key units of measuring dispersion of data.

4.1 Understanding data in terms of objects, variables, and scales

In deductive logic, we attribute truth-values to propositions. In probability theory, we attribute probability values both to events and propositions. In statistics, data which stand for our observations about the world lie at its core. In order for data to be converted into some language which is free from any ambiguity, so that what the data could furnish us with reliable information about the world, we take recourse to language of mathematical statistics.

The discussion of data in most introductory statistics textbooks typically starts with the definition of a population as a collection of objects of interest to an investigator. The investigator wishes to learn something about selected properties of the population. Such properties are determined by the characteristics of the individuals who make up the population and these characteristics are referred to as variables because their values vary over the individuals in the population. These characteristics can be measured on selected members of the population. If an investigator has access to all members of a population then he has conducted a census. A census is rarely possible and an investigator will select instead a subset of the population called a sample. Obviously, the sample must be representative of the population if it is to be used to draw inferences to the population from which it was drawn.

An important concept in statistics is the idea of a data distribution which is a list of the values and the number of times (frequency) or proportion of the time (relative frequency) those values occur.

Variables can be classified into four basic types — nominal, ordinal, interval, and ratio. Nominal and ordinal variables are described as qualitative while interval

and ratio scale variables are quantitative.

Nominal variables differ in kind only. For example, political party identification is a nominal variable whose “values” are labels; e.g. Democrat, Republican, Green Party. These values do not differ in any quantitative sense. This remains true even if we represent Democrats by 1, Republicans by 2 and so on. The numbers remain just labels identifying group membership without implying that 1 is superior to 2. Because this scaling is not liable to quantification does not mean that it has no value. In fact, it helps us to summarize a large amount of information into a relatively small set of non-overlapping groups of individuals who share a common characteristic.

Sometimes the values of a qualitative variable can be placed in a rank order. The latter might stand for the quality of toys received in different overseas cargos. Each toy in a batch receives a quality rating (Low, Medium, and High). They could also be given numerical codes (e.g. 1 for high quality, 2 for medium quality, and 3 for low quality). This ordinal ranking implies a hierarchy of quality in a batch of toys received from overseas. This ranking must satisfy the law of transitivity implying that if 1 is better than 2 and 2 is better than 3 then 1 must be better than 3. Since both nominal and ordinal scales are designated as qualitative variables, they are regarded as non-metric scales.

Interval scale variables are quantitative variables with an arbitrarily defined zero value. Put another way, a value of 0 does not mean the absence of whatever is being measured. Temperature measured in degrees Celsius is an interval scale variable. This is a metric scale in which for example the difference between 2 and 5 is the same as the difference between 48 and 51.

In contrast to interval scale data, in “ratio” scale data, zero is actually a pointer of “nothing” scored on the scale just as we see zero on a speedometer which signifies no movement of a car. Temperature measured in degrees Kelvin is a ratio scale variable because a value of 0 implies the absence of all motion at the atomic level.

Mathematical operations make sense with quantitative data whereas this is not true in general of qualitative data. This should not be taken to mean that qualitative data cannot be analyzed using quantitative methods, however. For example, gender is a qualitative variable and it makes no sense to talk about the “average” gender in a population but it makes a lot of sense to talk about the proportions of men and women in a population of interest.

4.2 Measures of central tendency

The data distribution can be presented graphically (e.g. in a histogram) or tabularly. Graphical analyses are an important and often overlooked aspect of data analysis itself in part because it seems so simple. We do not go into much detail here because we are more interested in statistical inference but we emphasize the importance of graphical summaries as an overall part of an initial data analysis.

Graphical summaries can provide a quick overall impression of the data but we need more. One important property of a sample (and by extension of the

population from which it was drawn) is the location of its “center”.

We suppose we have a sample of size n from some population of interest. We denote the values of the observations in the data set by $x_1, x_2, \dots, x_n$. There are at least three distinct measures of central tendency: (i) the mode, (ii) the mean and (iii) the median. The mode is the most frequent value in the data. The mean is computed by adding up all the numbers and dividing by the size of the sample. We denote the mean by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

The median is the middle value in the sense that half the values lie at or above the median and half lie at or below it. Computation of the median starts with ordering the data, generally from lowest to highest. If the sample size is odd the median is the middle value. If the sample size is even, the median is the mean of the two middle values.

By way of example consider the number of home runs hit by Babe Ruth during his 15 years with the New York Yankees (1920 to 1934),

$$54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22$$

Forty six home runs appear most often (3 times) and this is the mode. The mean number of home runs hit by Ruth is

$$\bar{x} = (1/15)(54 + 59 + \dots + 22) = 659/15 = 43.9.$$

We compute the median by first ordering the data from low to high

$$22, 25, 34, 41, 41, 46, 46, 46, 47, 49, 54, 54, 59, 60.$$

Since n is odd, we choose the middle value which is 46 (the same as the mode). The median is not unique in this data set.

The question of which measure is most appropriate arises immediately. The answer depends on context. The mode is generally easy to identify if one has already carried out a graphical assessment and may be perfectly adequate in some cases. However, it is rarely used in practical statistics anymore. Indeed it is rare to see the mode given in research papers in the biological and ecological sciences and we do not discuss it further here.

The mean and median are commonly used as numerical summary statistics, however. Often both are provided. The mean has the advantage that it uses all the data; however it is not resistant to unusual values in the data set. A common example used by one of the authors when he teaches introductory statistics is to imagine the effect on the mean annual income of the people in the classroom if Bill Gates was to walk in the door. The median is resistant to unusual or atypical observations but it only requires one or two data points and thus ignores a lot of information in the data set.

We can illustrate these points by considering the graphical representation of idealized data distributions below. The graphs are histograms with smooth curves added. The data in Fig. 1 are skewed to the right (they have a long right tail)., the distribution in Fig. 2 is skewed to the left , and Fig. 3 is symmetric.

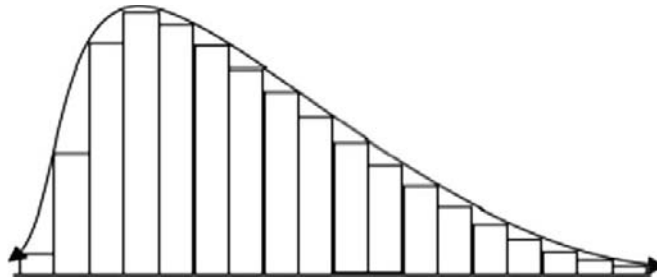


Figure 1. Right skewed

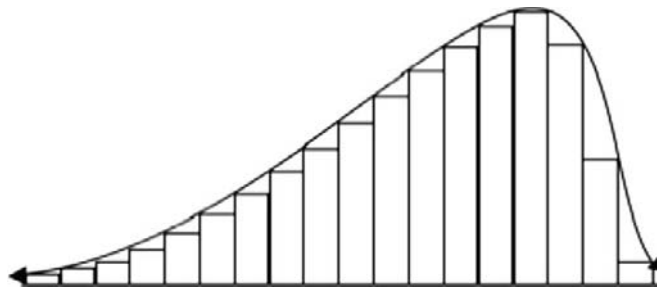


Figure 2. Left skewed

In each figure the median is the value that has half the area under the curve above it and half the area below it. The mean of the data set depicted in Fig. 1 will be greater than the median because the values in the long right tail “pull” the mean that direction. Similarly the mean in Fig. 2 will be less than the median. The mean and median will be equal to one another in Fig.3. The implication then is that the median is a better measure of central tendency (in the sense of identifying the most typical value) when data distributions are skewed and that it is comparable to the mean when the data distribution is symmetric. This would seem to imply that the median is then superior to the mean, and this argument may well be valid for descriptive data analysis. But the mean is more important for inferential analysis as will be discussed below. Note that the mode would generally not be interpretable as the most typical value when data are skewed. The mode, median, and mean are all equivalent for descriptive purposes when the data distribution is symmetric. However, the mean is more commonly used.

Which measure to use depends in large part on how it will be used. Generally, an

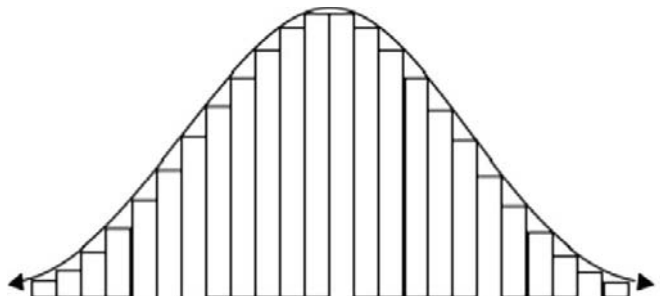


Figure 3. Symmetric

investigator is interested in drawing an inference from a sample to a population. An investigator may be interested in estimating an unknown population mean defined to be (for a finite population of size N)

$$\mu = \frac{1}{N} \sum_{j=1}^N x_j.$$

The sample mean $\bar{x}$ would seem to be the natural estimator to use in this case. However if the population distribution is symmetric and the sample is drawn in such a way that it is representative of the population (in which case the sample distribution is approximately symmetric) the population mean could be estimated by any of the central tendency measures described above. Which is the best? Answering this question is one reason why probability theory is so important to the theory of statistic and we will return to this topic below (see section 6).

Strictly speaking these measures of center are only appropriate for interval or ratio scale data. They are sometimes applied to ordinal data under the (strong) assumption that the rankings can be treated as numerical. Nominal data are sometimes summarized by providing the proportion of individuals in each category. Proportions can be viewed as means. For example, we have n observations of people in a population classified as male or female we could give a value of 1 to males and a value of 0 to females. We would have a list of n 0s and 1s and the sum of those divided by the sample size would be the proportion of males in the sample.

4.3 Measures of Dispersion

There is an old joke in statistics about a statistician who had his head in an oven and his feet in a freezer. When asked how he felt he replied, “On average I am quite comfortable”. How the values in a data set are spread out around the center is an important part of data description. Such variation is commonplace (indeed it is the rationale for the definition of the term variable) and in many cases it may be as interesting if not more interesting than the center.

The appropriate measure of dispersion to use is determined in part by the measure of center used. When the mean is chosen interest centers on quantifying the deviations of the observations about the mean. The sample variance is typically defined to be

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

This is a kind of average squared deviation about the mean. The units of measurement of the sample variance are awkward, (if the variable is cost of a home then the units of the variance are in squared dollars) and, especially for descriptive purposes, the positive square root of the variance is often given. This is referred to as the standard deviation. Both the variance and standard deviation are non-negative and will be equal to 0 if and only if all the values in a data set are the same. This will never or at least very rarely happen in practice.

The sample mean and variance are both sensitive to atypical observations because of their dependence on the sample mean.

Why do we divide by $n-1$? Not every text does and if the goal is to describe variability in a sample then either one can be used. However, the ultimate goal is to use the sample numerical summary statistics to estimate their unknown population counterparts. Consider a finite population of size N . The population variance is defined to be

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^N (x_j - \mu)^2.$$

If we are to use the sample variance to estimate the population variance and if we knew the population mean μ then dividing by n in the formula for the sample variance would be appropriate. But as a general rule we will not know the population mean and will estimate it using the sample mean. The variability of the n observations in the sample about the sample mean will tend to be less than their variability about the population mean. Thus, dividing by the sample size will tend to underestimate the population variance. Dividing by $n-1$ provides the necessary correction to this underestimation. We will return to this topic below when we discuss bias in statistical estimators.

As an example consider a simple data set: 10, 12, 14 and 16. The table below summarizes computation of the sample variance and standard deviation.

Observation	Value	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	10	10-13=-3	9
2	12	12-13=-1	1
3	14	14-13=1	1
4	16	16-13=3	9
Total	52	0	20

The sample mean is $52/4=13$. The sample variance is $20/3=6.66$ and the sample standard deviation is the square root of 6.66 or 2.58. Note that the sum of the deviations in column 3 is 0. This is true in general and is the reason why we cannot just “average” the deviations from the mean.

The variance and standard deviation are computed using the mean and are thus naturally paired with the mean. It is not correct, for example, to summarize location using the median and summarizing variability using the standard deviation. If the median is chosen as the measure of center another measure of variability is needed. One such is based on the sum of the absolute deviations from the median. Another cruder measure is the Interquartile Range which is the difference between the third quartile (75th percentile) and first quartile (25th percentile). We will not discuss such measures further here.

For most distributions the bulk of the possible values will lie within a couple of standard deviations of the mean. For symmetric distributions like the idealized example in Fig. 4, about 68% of the values will lie within one standard deviation of the mean, about 95% will lie within two standard deviations of the mean and over 99% will lie within three standard deviations of the mean. These figures are roughly accurate for skewed distributions as long as they are not badly skewed. The figures are exact for so-called normal distributions (also referred to as bell-shaped or Gaussian distributions). We will discuss this more below.

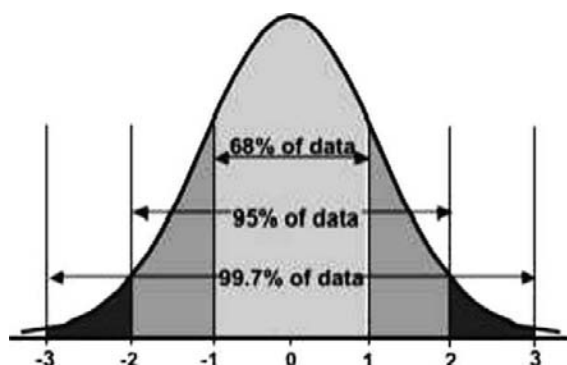


Fig. 4. 68-95-99.7 rule.

Figure 4. A graphical representation of the values in an idealized symmetric population with mean 0 and standard deviation 1.

4.4 From descriptive statistics to inference

The material presented in the preceding section is typically referred to as descriptive statistics and no probability is required to understand and to use the methods described above. However, at points in the discussion we alluded to the use of descriptive summary statistics as estimators of unknown population quantities,

e.g. using the sample mean to estimate a population mean. The sample mean is based on values observed in a subset of the population and will not in general be equal to the population mean. If we observe a sample mean of say 400 what can we say about the population mean? Is it exactly equal to 400? This is not very likely. Could the population mean be 425 or 350 or 675?

In determining how to quantify the uncertainty in the observed value of a statistic such as the sample mean, we ask ourselves the question, “What would we expect to see if the basic sampling process was repeated over and over again, independently, and under the same conditions”? That is, we conduct the thought experiment of repeating our sampling plan many times, computing the value of the statistic of interest each time and looking at the resulting distribution of those values. We are interested in the long run behavior of the statistic. Obviously we cannot do this in practice. To carry out this thought experiment we must make an assumption about the distribution of the values in the population and about how we draw values from the population. For the latter we must assume that we have a sample that is representative of the population in which we have an interest. Formally, we say that we have a simple random sample from the population, by which we mean that all possible samples of the same size have the same probability of being selected. Once we have a reasonable probability model for the process that generated our data we can use the results of probability theory to investigate the probabilistic behavior of sample statistics of interest. In the next section, we continue our discussion of probability theory with the goal of doing just that.

5 RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

There will be many experimental situations in which interest is primarily focused on numerical characteristics of various outcomes. We might be interested in questions such as “What is the probability of observing 6 or more heads in 10 tosses of a fair coin”?, or perhaps more practically, “What is the probability of surviving five years after being diagnosed with cancer”? We will need to be able to associate each outcome with a number. The rule which determines the association is called a random variable. In some ways this is an unfortunate terminology. Strictly speaking, a random variable is neither random nor a variable – it is a mathematical function mapping outcomes in a sample space to the real line. Typically however, introductory statistics textbooks steer clear of this technical definition and motivate the name as, for example, Devore [2009, p. 87] does when he writes that a random variable is “a variable because different numerical values are possible and random because the observed value depends on which of the possible experimental outcomes results”. We will need to assign probability to the values a random variable can assume. This assignment is done by means of a probability model or probability distribution.

We will divide this section into three short subsections. First, we will introduce the notion of a random variable, the expectation of a random variable and some basics of probability distribution. Once we become acquainted with random vari-

ables, we will discuss two types of popular probability distributions. They are, (i) normal distribution and (ii) binomial distribution. We spend some time on the nature of normal distribution and how we could convert it to its standard normal form.

5.1 Random variables and the basics of probability distributions

In simple language, a random variable is a rule for assigning a number to the outcome of a random process. By convention random variables are denoted by upper case letters from the Roman alphabet (X, Y, W , etc). Realizations (i.e. observed values of a random variable are denoted by the corresponding lower case letters (x, y, w , etc).

Consider tossing a fair coin three times and recording the outcomes in terms of heads or tails. The sample space is comprised of eight equally likely outcomes

$$S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}.$$

Let X = the number of heads observed on any single trial of this experiment. X is a random variable. It can take on one of four values; 0, 1, 2, or 3. The probability that X takes on these values is determined by the probabilities associated with the outcomes in the sample space. For example $X=2$ if we observe any one of three possible outcomes HHT , HTH , or TTH . Thus, the probability $X=2$ is

$$P(X = 2) = P(HHT \cup HTH \cup TTH) = 3/8$$

by the addition rule for mutually exclusive events. We summarize all the possibilities in the probability distribution for X :

X	0	1	2	3
$p_r(x) = P(X = x)$	1/8	3/8	3/8	1/8

Note that all values of $p_r(x)$ are between 0 and 1 and sum to 1 so this is a valid assignment of probabilities to the values of the random variable. The probability distribution $p_r(x)$ of the random variable X = the number of heads observed in three tosses of a fair coin describes how probability is assigned to the possible values X can assume.

Random variables can be discrete or continuous. A discrete random variable is either finite or countably infinite. The random variable X = the number of heads in three tosses of a fair coin is discrete and finite. As an example of an infinite discrete random variable, consider the experiment of tossing a coin until a head is observed. Let X be the number of failures recorded until the heads occurs. Of course, in practice we would never toss the coin forever but there is no well-defined upper limit and it may make sense to model the outcomes of this experiment by assigning X the possible values 0, 1, 2, 3,... Most commonly used discrete

random variables are integer valued but they do not need to be. Continuous random variables can take on uncountably many values in an interval on the real line. By uncountable we mean that the values of the random variable cannot be placed in a one-to-one correspondence with the natural numbers. Consider the random variable, X = the length of a randomly selected phone call. Although one could measure the call in terms of its minutes and seconds, the latter still fail to measure it in terms of its other shorter intervals. As a result, it is better to regard it as a continuous random variable capable of taking on any value on the positive real line.

In this chapter, we will mostly confine ourselves to discrete random variables. Each random variable is associated with a probability distribution (function), which must satisfy the rules of the probability theory. We saw a simple example above. In general a probability distribution for a discrete random variable (sometimes also called a probability mass function) is defined as follows. A discrete random variable X takes a finite or countably infinite number of values. The values it can assume are determined by the outcomes in the associated sample space, i.e. for every outcome o in the sample space there is an associated value for X . Mathematically then X is a function mapping outcomes o to numbers x , $X(o) = x$ although for convenience we usually suppress the functional notation. The probability distribution (or model) for X is then

$$p_r(x) = P(X = x) = P(o \in S, X(o) = x).$$

A function $p_r(x)$ is a valid probability distribution if

1. $0 \leq p_r(x) \leq 1$ for all x
2. $\sum_x p_r(x) = 1$ where the sum is taken over all possible values of x

The probability $P(X \in A)$ is found by summing the probabilities associated with the values x in the event A .

Consider tossing a fair die once and recording the number on the upturned face, X . Clearly there are six different values this random variable can take on and each one is equally likely, i.e.

$$P(X = 1) = \dots = P(X = 6) = 1/6.$$

The probability that we observe a value of X greater than or equal to 3 is

$$P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) = 2/3.$$

We will observe different values of a random variable over many repeated trials of its associated experiment. The expected value of a random variable is a weighted average of its possible values. Formally, for a discrete random variable X the expected value of X is defined to be

$$E(X) = \mu_X = \sum_x xp(x).$$

There are several ways to think about expected values. Given a long-run relative frequency interpretation of probability the expected value can be thought as the long-run average. That is, if we imagine repeating the experiment a large number of times and computing the average of the observed values of the random variable that average should be close to the expected value. Another way to think about it is to imagine a “perfect” sample, i.e. a sample that contains the values we expect it to contain given the probability distribution. The average of that perfect sample is equal to the expected values.

X	0	1	2	3
$p_r(x)$	0.48	0.39	0.12	0.01

Consider a random variable X with the following probability distribution. Now imagine the experiment is repeated 100 times so that we observe 100 values of X . Based on the probability distribution we expect to see 48 values of 0, 39 values of 1, 12 values of 2, and 1 value of 3. The mean of these 100 numbers will be

$$\frac{48(0) + 39(1) + 12(2) + 3(1)}{100} = 0.66$$

which we can rewrite as

$$0.48(0) + 0.39(1) + 0.12(2) + 3(1) = 0.66$$

which is the expected value of X

We need some way to quantify the variability in those possible values of a random variable. We define the variance of a random variable to be

$$V(X) = \sigma_X^2 = \sum_x p_r(x) (x - E(x))^2.$$

There is a short cut formula which is easier for computational purposes when computing variances by hand

$$\sigma_X^2 = \sum_x p_r(x) x^2 - \mu_x^2.$$

The variance depends on how often one expects each value of X to occur and how far away the X values are from the expected value of X . One could interpret the variance of X as the weighted average squared distance from $E(X)$, using the probabilities of each value of X as the weights. The standard deviation σ_X of the random variable is the positive square root of the variance of X .

Recall the example above involving tossing a six-sided die once. We let X denote the number of dots on the upturned face and noted that under an assumption that the die was fair we had the following probability distribution

$$P(X = 1) = \dots = P(X = 6) = 1/6.$$

The expected value is

$$\mu_X = (1/6)(1 + 2 + 3 + 4 + 5 + 6) = 3.5.$$

The variance, computed using the short cut formula is

$$\sigma_X^2 = (1/6)(1 + 4 + 9 + 16 + 25 + 36) - 3.5^2 = 2.92.$$

The standard deviation is 1.71.

Two discrete random variables X and Y are said to be independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y) = p_r(x)p_r(y)$$

for all possible values of x and y . Essentially what this says is that two discrete random variables are independent if the joint probability mass function factors into the product of the marginal mass functions.

Modifications of the above arguments are needed for continuous random variables. Whereas probability distributions for discrete random variables assign probability to single values probability distributions for continuous random variables assign probability to intervals of real numbers. The probability distribution for a continuous random variable is called a probability density function. A mathematically valid probability density function is any function f satisfying the following properties,

1. $f(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

The probability that X takes on a value in the interval (a, b) is

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

The expected value of a continuous random variable is defined to be

$$\mu_X = \int_{-\infty}^{\infty} xf(x)dx$$

The variance of a continuous random variable is defined to be

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)dx.$$

Unlike discrete random variables the probability that a continuous random variable will take on a single value is 0

$$P(X = a) = \int_a^a f(x)dx = 0.$$

Two continuous random variables X and Y are independent if their joint density factors into the product of the marginals

$$f(x, y) = f_X(x)f_Y(y).$$

There are an infinite number of both discrete and continuous probability distributions. In practice however a relative handful of families of distributions are used. We will briefly consider one discrete distribution and one continuous distribution important in the practice of statistics.

5.2 The normal distribution

The normal distribution is the most important in statistics. It is the distribution people have in mind when they referred to the “bell-shaped” curve and it is also often referred to as the Gaussian distribution after the mathematician Karl Friedrich Gauss.

A continuous random variable X is said to have a normal distribution (or be normally distributed) with mean μ and variance σ^2 if its probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}, -\infty < x < \infty.$$

The mean can be any real number and the variance any positive real number. We say that X is $N(\mu, \sigma^2)$.

There is a different normal distribution for each pair of mean and variance values and it is mathematically more appropriate to refer to the family of normal distributions but this distinction is generally not explicitly made in introductory courses.

The history of the normal distribution is fascinating [Stigler, 1986]. It seems to have first appeared in the work of Abraham DeMoivre in the mid-18th century and Gauss found it useful for work he was doing in the late 18th and early 19th centuries. It was imbued with semi-mystical significance initially. Some were impressed by the fact that this one distribution contained the three most famous irrational numbers, e , $\sqrt{2}$, and π . Normally distributed variables were considered to be a law of nature. Although not viewed quite as reverentially today it is still important for reasons which will be discussed in more detail below.

The graph of the density function is symmetric about μ . There is a different curve for every μ and σ^2 . In any normal distribution 68% of the data fall within σ (one sigma) of the mean μ , 95% of the data fall within 1.96σ of μ , and 99.7% of the data fall within 3σ of μ . These proportions are the same for any normally distributed population. For simplicity, we frequently convert the values from the units in which they were measured to unitless standard values. Figure 4 shows an example of the so-called standard normal distribution with mean 0 and variance equal to 1 along with an illustration of the 68-95-99.7 rule.

To transform a normally distributed random variable into a standard normal random variable we subtract the mean and divide by the standard deviation. The result is typically referred to as a Z score,

$$Z = \frac{X - \mu}{\sigma}$$

The random variable Z has a $N(0, 1)$ distribution.

As an example, suppose the heights of American young women are approximately normally distributed with $\mu = 65.5$ inches and $\sigma = 2.5$ inches. The standardized height

$$Z = \frac{\text{height} - 65.5}{2.5}$$

follows a standard normal distribution. A woman's standard height is the number of standard deviations by which her height differs from the mean height of all American young women. A woman who is 61 inches tall, for example, has a standard height of

$$Z = \frac{61 - 65.5}{2.5} = -1.8$$

or 1.8 standard deviations less than the mean height.

The standard normal distribution is important in introductory classes because it simplifies probability calculations involving normally distributed random variables. Because the normal distribution is a continuous distribution probabilities can be computed as areas under the density curve. But the probability density function does not have a closed form integral solution and those areas must be determined numerically. Further, many introductory courses in statistics do not require calculus as a prerequisite and so integration is not an assumed skill. Tables of probabilities (areas) associated with the standard normal distribution are provided in introductory statistics texts. Finding the probability that a normally distributed random variable X with mean μ and variance σ^2 falls in some interval (a, b) is solved by converting to standard units and using the tabled values.

Using the standard normal distribution to solve probability problems is no longer of much practical importance because probabilities can now be determined using computer software but the standard normal random variable still plays a major role in statistical inference as we will see.

The family of normal distributions has some nice mathematical properties not shared by other probability distributions. These mathematical properties explain,

in part, the importance of the normal distribution in statistics. Users of statistics are often interested in linear transformations of their data or in combining data through the use of linear combinations. For example, the sample mean is a linear combination of observed data. The application of probability theory to data analysis must take this into account. Linear functions and linear combinations of normally distributed random variables have a property that is not shared by other probability distributions that might serve as a model for a data generating process.

Suppose X is a normally distributed random variable with mean μ and variance σ^2 . Then any linear transformation of the form $Y = a + bX$ (with $b \neq 0$) will be normally distributed with mean $a + b\mu$ and variance $b^2\sigma^2$.

Suppose we have a sequence of independent random variable $X_1, X_2, \dots, X_n$ with means $\mu_1, \mu_2, \dots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$. Let $a_1, a_2, \dots, a_n$ be constants. What is the probability distribution of the linear combination

$$Y = a_1X_1 + a_2X_2 + \dots + a_nX_n?$$

It can be shown that if each of the X_i s is normally distributed then the linear combination Y is also normally distributed with mean $a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n$ and variance

$$a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2.$$

The important point of the above results is not the resulting means and variances. Those results hold for any probability distribution. What is important and, mathematically speaking, remarkable is that linear transformations and linear combinations of normally distributed random variables are themselves normally distributed.

Many commonly used statistical methods start with an assumption that observed data are a representative sample drawn from a population of individuals. A frequent goal is to use summary information in the sample to draw inferences to unknown corresponding quantities in the population. For example, the sample mean of a data set is commonly used to estimate an unknown population mean. Quantifying the uncertainty associated with the estimate requires a probability model. The data are viewed as realizations of a sequence of independent random variables $X_1, X_2, \dots, X_n$. The sample mean, viewed as a random variable is

$$\bar{X} = (1/n)(X_1 + X_2 + \dots + X_n).$$

Given the additional assumption that the values in the population can be approximated by a normal distribution with mean μ and variance σ^2 then $\bar{X}$ will be normally distributed with mean μ and variance σ^2/n . We discuss the implications of this result in more detail below.

The normal distribution is important in statistics for another reason, a truly remarkable and fascinating result: The Central Limit Theorem (CLT). There are different versions of the CLT but we will consider it as it pertains to the probability distribution of the particular linear combination of random variables called

the sample mean. Suppose we have a sequence of independent random variables $X_1, X_2, \dots, X_n$ all sharing the same finite mean μ and finite variance σ^2 . In the context of statistics we think of these random variables as constituting a random sample from a population with mean μ and variance σ^2 . We make no other distributional assumptions about the random variables, i.e. about the distribution of values in the population. The sample mean is

$$\bar{X} = (1/n)(X_1 + X_2 + \dots + X_n).$$

The Central Limit Theorem says that if the sample size n is large enough then $\bar{X}$ will be approximately normally distributed as $N(\mu, \sigma^2/n)$.

How large is “large enough”? A value frequently seen in introductory statistics texts is that $n \geq 30$. But, like all rules of thumb this one should not be applied indiscriminately. For some well behaved distributions (i.e. symmetric with small variances) sample sizes of 5 to 10 may suffice. For other distributions, especially those with high variance (known as fat or heavy-tailed distributions) the required sample size can be significantly greater than 30.

5.3 The binomial distribution

Consider an experiment that has only two possible outcomes which can be labeled a “success” or a “failure”. The probability of a success is denoted by p and the probability of a failure by $1-p$. The simplest such experiment is a coin toss with Heads counting as a “success”. If the coins is fair then $p = 0.5$. An experiment of this type is called a Bernoulli trial.

We will often be interested in counting the number of successes in a sequence of n such trials. The number of successes is a random variable with a probability distribution. We make the following two assumptions:

1. The n trials are independent, i.e. the outcome of any given trial does not affect the outcomes of other trials.
2. The probability of success p is the same on all the trials.

Then $X =$ the number of successes is said to have a binomial distribution with parameters n and p , or using a common form of mathematical shorthand we say that X is $B \text{ in}(n, p)$. The probability distribution (or probability mass function) is

$$p(x) = P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

for $x = 0, 1, \dots, n$. The $n!$ term is called “ n factorial” and is defined as

$$n! = n(n-1)(n-2) \dots (3)(2)(1)$$

for any positive integer n . We define $0! = 1$. This distribution is derived using the rules of probability. Any given sequence of x successes and $n-x$ failures in

has probability $p^x(1-p)^{n-x}$ from the assumption of independence and constant probability of success. There are

$$\frac{n!}{x!(n-x)!}$$

mutually exclusive sequences with x successes and $n-x$ failures. Thus to compute the probability of x successes we add up the $\frac{n!}{x!(n-x)!}$ probabilities $p^x(1-p)^{n-x}$. We have used the multiplication rule for mutually independent events and the addition rule for mutually exclusive events.

The mean and variance of a binomially distributed random variable X are $\mu_X = np$ and $\sigma_X^2 = np(1-p)$, respectively.

The binomial distribution is a common probability model chosen for data analysis problems involving proportions. Examples include estimating the proportion of voters who favor a particular candidate, the proportion of cancer patients cured by a new treatment, the proportion of individuals in a population who have a particular disease, and so on. The proportion of successes in n Bernoulli trials is

$$\hat{p} = X/n.$$

The sample proportion is a linear transformation of X and is also a random variable. It has mean $\mu_{\hat{p}} = p$ and variance $\sigma_{\hat{p}}^2 = p(1-p)/n$, respectively. Use of $\hat{p}$ as a statistical estimator requires knowledge of its probability distribution and unlike normally distributed random variables linear transformations of binomial random variables are not binomially distributed. However, if the sample size is large enough then the Central Limit Theorem implies that $\hat{p}$ will be approximately

$$N\left(p, \frac{p(1-p)}{n}\right).$$

This result follows from the fact that the sample proportion can be considered the mean of a random sample of Bernoulli random variables. In this case “large enough” is a value of n such that both np and $n(1-p)$ are greater than 10.

5.4 Sampling distributions

We know by now that in statistical inference, we make inferences about a population based on information in a sample which is a subset of the population. Mostly we do not know the details of the population and we will use the information in the sample to estimate the unknown population quantities of interest. Obviously, we require that the sample be representative of the population. We also need a probability model.

A population of interest has been identified and interest centers on a numerical characteristic of the population. We call such characteristics parameters and these are constants. A common parameter of interest is the population mean, μ . A sample will be drawn and the sample mean will be used to estimate the population

mean. The distribution of the values associated with the individuals in the population is called the population distribution. Prior to data collection we consider the sample to be a sequence of independent random variables. $X_1, X_2, \dots, X_n$. Each of these random variables has the same probability distribution with mean μ (equal to the population mean) and variance σ^2 (equal to the population variance). The sample mean is of course

$$\bar{X} = (1/n)(X_1 + X_2 + \dots + X_n).$$

A common assumption is that the population distribution is normal with mean μ and variance σ^2 . We know from the above that $\bar{X}$ is normally distributed with mean μ and variance σ^2/n . This is the probability distribution of the random variable $\bar{X}$. In the field of statistics this random variable will be put to a particular use and statisticians have their own terminology to describe this quantity. It is a statistic and its probability distribution is called a sampling distribution. In general the sampling distribution of a statistic is a description of how probability is assigned to the possible values the statistic can take on. What does the sampling distribution of $\bar{X}$ tell us? We see that the expected value or mean of this statistic is equal to the parameter it will be used to estimate. We say that $\bar{X}$ is an unbiased estimator of μ . We see that the variability in the values the sample mean can assume is a function of the population variance and the sample size. In particular as the sample size gets larger the variance of $\bar{X}$ decreases. This is a desirable characteristic; the larger the sample, i.e. the more information we have about the population, the more we can trust that $\bar{X}$ will produce a good estimate of μ . The sample mean then has two properties we would like to see in any statistic; we want statistics to be unbiased (at least approximately) and we want the variability in the possible values the statistic can assume to decrease with increasing information (larger sample size).

The rules of probability mean that an assumption of a normal population distribution results in a normal sampling distribution for the sample mean. For example, we know that prior to data collection there is a 95% probability that the calculated value of the sample mean will lie within $1.96 \sigma/\sqrt{n}$ units of the population mean.

If the assumption of normality for the population is not valid then the sampling distribution of the sample mean will no longer be normal although the sample mean still has expected value (mean) of μ and variance of σ^2/n . Thus, it is still unbiased and it still has a variance that is decreasing as the sample size increases. Further, if the sample size is large enough, the sampling distribution of the sample mean will still be approximately normal by the Central Limit Theorem and prior to data collection there is an approximate 95% probability that the calculated value of the sample mean will lie within $1.96 \sigma/\sqrt{n}$ units of the population mean.

A related and common application is estimation of a population proportion. The population of interest can be visualized as a population of two values, 0 or 1 with 1 denoting a "success". A simple random sample of size n will be taken and X = the number of successes will be counted.

The sample proportion

$$\hat{p} = X/n$$

will be computed and used to estimate the population proportion p . Given these assumptions X will have a binomial distribution with mean np and variance $np(1-p)$ and $\hat{p}$ will have mean p and variance $p(1-p)/n$. The statistic $\hat{p}$ is an unbiased estimator of the parameter p and its variance decreases with increasing sample size. Its sampling distribution is known but is awkward to work with. From the Central Limit Theorem we know that if n is large enough then the sampling distribution of $\hat{p}$ will be approximately normal with mean p and variance $p(1-p)/n$.

6 STATISTICAL INFERENCE

Statistical inference involves drawing conclusions about an entire population based on the information in a sample. For the sake of discussion, we divide this section into three subsections. First, we broach the two ways of doing inference, (i) estimation, and (ii) the test of significance. Second, we distinguish between two ways of conducting estimation, (i) point estimation and (ii) confidence interval estimation. Third and finally, we introduce the concept of the test of significance.

6.1 *Two types of ways of doing theory of inference: estimation & testing*

Consider first the estimation problem and the test of significance. In the problem of estimation, we determine the values of the parameters, and in the test of significance, we determine whether the result we observe is due to random variation of the sample or due to some other factor. Problems of estimation can be found almost everywhere, in science and in business. Consider an example in physics where Newcomb measured the speed of light. Between July and September 1882, he made 66 measurements of the speed of light. The speed of light is a constant but Newcomb's measurements varied. Newcomb was interested in determining the "true" value of the parameter; the speed of light. In business, a retailer might like to know something about the average income of families living within 3 miles of her store.

Note the difference in the two populations in the examples. The retailer's population of interest is finite and well-defined. Theoretically she could sample every household of interest and compute the true value of the parameter. Logistically, such a sampling plan, called a census, is often practically impossible and a sample is taken from that population. Newcomb's population is hypothetical. It is comprised of all possible measurements he could have theoretically made and estimation of the parameter of interest requires an assumption that the 66 observed measurements are a random sample from that population.

Scientists and other users of statistics often have a different question of interest. Imagine Newcomb in an argument with another scientist. Newcomb believes the

speed of light is equal to one value and the other scientist believes the value is less than that. The goal of the study would have been to assess the strength of evidence in the data for or against these hypotheses. The retailer could be considering an expansion of her store but will not do it unless the mean income of families in the nearby area exceeds a specified threshold. She takes a sample, not so much with the intent of estimating the income but of determining if there is sufficient evidence that it is high enough to justify expansion.

Statisticians have argued and continue to argue over the best ways to do these two types of inference. Testing in particular is controversial. Below we will continue with the approach we have taken above and present the basics as they are often presented in introductory courses in statistics. Our goal is to present the concepts, not practically useful methods.

6.2 Two types of estimation problems: point and interval estimation

An investigator is interested in a population, in particular, in the value of some constant numerical quantity θ , called a parameter. The value of θ is unknown. A representative sample of size n will be taken (denoted $X_1, \dots, X_n$) and information in the sample will be used to estimate the parameter. This is done by distilling the information in the n values in the data set into a single summary quantity called a point estimator or statistic, $\hat{\theta}$. Under an assumption of random sampling the estimator (statistic) is a random variable that is a function of the data,

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n).$$

For convenience the functional notation is generally suppressed but it is important to understand that point estimators are random variables that are functions of data.

In practice, a single sample is taken and a single value of the point estimator is obtained. That value, a realization of the estimator, is called an estimate. In practice, we only have a single sample and a single estimate but we know that if we took another sample from the same population that we would not get the same value. Recall the above thought experiment of taking realizations of the estimator. The distribution of those values is the sampling distribution of the estimator. Of course, we cannot take very many samples to determine the sampling distribution and this is why probability theory is so important — under some assumptions the sampling distribution of an estimator can be derived using the results of probability.

Simon Newcomb took 66 measurements of the time it took light to travel a known distance. The values are shown below. The values are the number of nanoseconds each measurement differed from 0.000024800 seconds. Thus, the first measurement was actually 0.000024828 seconds. Although the speed of light was believed to be constant (at least according to theory) there is variability in the measurements. Denote the true time it takes light to travel the distance in Newcomb's experiment by μ .

Table 1. Newcomb's measurements of the speed of light data

28	22	36	26	28	28
26	24	32	30	27	24
33	21	36	32	31	25
24	25	28	36	27	32
34	30	25	26	26	25
-44	23	21	30	33	29
27	29	28	22	26	27
16	31	29	36	32	28
40	19	37	23	32	29
-2	24	25	27	24	16
29	20	28	27	39	23

Using the information in the sample requires several assumptions. We assume that μ is the mean of a hypothetical population of all possible measurements. We assume the 66 observed measurements are representative of that population. The measurements are independent of one another in the sense that knowing the value of the i^{th} measurement provides no information about the $(i+1)^{st}$. Stated another way, we could permute the values of the observations in the table without impacting the analysis. We also assume that the population of possible measurements is normally distributed.

This latter assumption is questionable in light of the two outliers -44 and -2. Such atypical values are not unusual in data sets and the question of what to do with them is not easily answered. Here, for simplicity, we omit them and proceed with an analysis of the remaining 64 observations.

The sampling distribution of the sample mean $\bar{X}$ is normal with mean μ and standard deviation $\sigma/\sqrt{64}$. The point estimate, i.e. the realized value of the sample mean is $\bar{x} = 27.75$. In a sense this is our best guess, given the data, about the value of μ . However, there is uncertainty about this estimate because we know the value of the sample mean varies from sample to sample. That is, if Newcomb had taken another 64 observations it is highly unlikely that he would have observed the same value of the sample mean. The variability is quantified by the standard deviation, $\sigma/\sqrt{64}$, which in practice is itself an unknown parameter which must be estimated by the sample standard deviation which is 5.08 for this data set. We will assume, unrealistically, that this, in fact, is the population standard deviation and that $\sigma/\sqrt{64} = 5.08/8 = 0.64$.

We have a point estimate and we have quantified the uncertainty associated with this estimate. We go further however and compute an interval estimate of μ , called a confidence interval. Prior to data collection we know, from the properties

of the normal distribution, that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

follows a standard normal distribution and that

$$P(-1.96 < Z < 1.96) = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95.$$

A relatively simple set of algebraic manipulations leads to the conclusion that

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Thus, prior to data collection we know that the random interval $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$ will capture or contain μ with probability 0.95. Over the long run, 95% of all such intervals will contain the population mean but 5% will not. When we plug in the values from Newcomb's data we have the realized value of the interval,

$$\bar{x} \pm 1.96\sigma/\sqrt{n} = 27.75 \pm 1.96(5.08)/8 = (26.5, 29.0).$$

The probability that μ lies between 26.5 and 29.0 is either 0 or 1 because μ is a fixed constant. If $\mu=23$ then the probability it lies between 26.5 and 29 is 0. If $\mu = 28$ then the probability it lies between 26.5 and 29 is 1. However, the process that produced the interval is reliable in that over the long run 95% of all intervals produced in this way will contain μ so we say that we are 95% confident that μ lies in the interval (26.5, 29.0).

It is possible to compute intervals for other levels of confidence. Let $z_{\alpha/2}$ be the value from a standard normal distribution with $\alpha/2$ of the area under the normal curve lying above it, i.e.

$$P(Z > z_{\alpha/2}) = \alpha/2.$$

Then a $100(1 - \alpha)\%$ confidence interval for the mean μ of a normally distributed population when the value of σ is known is given by

$$\bar{x} \pm z_{\alpha/2} (\sigma/\sqrt{n}).$$

The most common confidence levels used in practice are 90% ($z_{0.05} = 1.645$), 95% ($z_{0.025} = 1.96$), and 99% ($z_{0.005} = 2.575$).

Three key assumptions are required for such intervals to be valid:

- (i) The data must represent a simple random sample from the population of interest.
- (ii) The population must be normally distributed.

(iii) The standard deviation of the population is known.

By far the most important assumption is (i). One implication of the assumption is that observations can be considered independent of one another. This assumption along with assumption (ii) is required for the sampling distribution of $\bar{X}$ to be normal with the indicated mean and variance. It is the normality of $\bar{X}$ that was necessary for the above derivation of the confidence interval procedure. As long as the sample size is large enough the procedure is fairly robust to violations of assumptions (ii) and (iii). If the population is not normal but the sample size is large enough then the Central Limit Theorem will result in an approximate normally distributed $\bar{X}$ and the confidence levels of intervals will be approximately correct. If the population standard deviation is not known then we estimate it with the sample standard deviation. Given a large enough sample size this estimate should be reasonably close to the true standard deviation, close enough that the intervals will have a true confidence level close to the advertised level.

One might wonder why 100% confidence intervals are not computed. The answer is that they are so wide as to be practically useless (a demographer can be 100% confident that the population of the earth is between 0 and one trillion people). Construction of confidence intervals is a balancing act between achieving a high enough level of confidence and a narrow enough interval to be useful. The half-width of a confidence interval is $z_{\alpha/2}\sigma/\sqrt{n}$. Clearly the larger the population standard deviation the wider an interval will be. Investigators generally have little control over the population standard deviation but do have control over the level of confidence and the sample size. However, specification of a higher level of confidence requires use of a larger $z_{\alpha/2}$ leading to a wider interval. Larger sample sizes will lead to narrower intervals. However, collecting samples can be expensive and there will often be limits on how large a sample can be selected. Also, because the half-width of the interval is inversely proportional to the square root of the sample size increasing the sample size by a factor of say k does have the same effect of decreasing width. For example, increasing the sample size by 4 times only cuts the interval width in half. It is possible in the simple setting described here to manipulate the confidence interval formula to determine a sample size needed to achieve a specified width and level of confidence. However, this is more difficult in general.

Confidence intervals are widely used but remain controversial among many statisticians. The interpretation of confidence intervals in particular is difficult for many students and even more sophisticated users to grasp. It is easy to say that one is 95% confident that an interval contains a parameter of interest but when pressed as to exactly what that means many cannot explain it satisfactorily. They want to say, indeed they almost surely believe, that the interval contains the parameter with 95% probability. Some statisticians empathize with such users. They object to the invention of another way of quantifying uncertainty, especially one difficult to understand. To these statisticians the natural language of uncertainty is probability, but they tend to interpret probability differently from frequentists. The approach taken by these statisticians is typically referred to

as Bayesian because of the prominent role by Bayes' Rule in their methodology. This is perhaps an unfortunate description because Bayes' Rule is a theorem, consequently it works in relevant cases quite well as it should regardless of one's interpretation of probability but it has been used too long to change now. We do not discuss this approach to interval estimation here as our goal is to present a primer of statistics as it is taught in most introductory courses. However, some of the chapters in this volume will discuss the Bayesian approach in more detail.

6.3 *The test of significance and hypothesis testing*

The question of interest in point and interval estimation is, "What is the value of an unknown parameter and how do we quantify our uncertainty about our estimate of the parameter?" Testing is motivated by another question; "How consistent are the data with a stated claim or hypothesis about the value of a parameter?"

The hypothesis takes the form of specifying a particular value or of specifying a range of values within which the true value lies. Implicit in the specification of a hypothesis about the value of a parameter is another hypothesis (the negation of the first) that the value of a parameter is something else. The two hypotheses are called the null hypothesis (H_0) and the alternative hypothesis (H_a). The null hypothesis specifies a value or range of values of a parameter and the alternative is the value or range of values deemed plausible if the null hypothesis is not true.

Confusingly there are two basic approaches to testing encountered in the introductory statistical literature; the test of significance and the hypothesis test. The first approach is attributed to the English statistician and population geneticist Sir Ronald A. Fisher. Hypothesis testing is attributed to the Polish mathematician Jerzy Neyman and the English statistician Egon Pearson. Different texts treat the topic in different ways. Devore [2008] discusses Neyman-Pearson hypothesis testing. DeVeaux, Velleman, and Bock [2008] place more emphasis on significance testing although they briefly describe the Neyman-Pearson approach. In general, the discussion of testing in such textbooks takes a middle ground between the two approaches that neither Fisher nor Neyman-Pearson would have approved of. We will start with significance testing and have a little to say about hypothesis testing at the end.

The following simple example illustrates the basic reasoning in a test of significance. An hypothesis is stated along with an alternative. Data are collected and steps are taken to determine if the data are consistent with the null hypothesis. In this example, it is possible that the data are consistent with the null hypothesis and the decision is to fail to reject the null. If the data are not consistent with the null hypothesis then it might be decided to reject the null. In practice, the data may be equivocal and no decision is warranted. Or it may be that different investigators make a different decision given the same data.

We now describe the procedure for tests of significance about population means. We will assume that we have a simple random sample of size n from a normally distributed population with unknown mean μ and variance σ^2 (or standard de-

viation σ). For definiteness, suppose that another physicist working at the same time as Newcomb believed that the speed of light was such that the mean time it took light to travel the distance in Newcomb's experiment was 29 (using the scaled units from above). We suppose (in our fictional account) that Newcomb did not believe this physicist was correct. Being a scientist he was willing to put his supposition to the test. Further, he was willing to start by assuming that he was wrong and then use the data to show that this assumption is not tenable. This is a classic argument by contradiction. Newcomb did not have the tools to carry out the test that we have but we will see what would have happened if he had. Denoting the true time it takes light to travel the distance in the experiment by μ the null and alternative hypotheses are

$$\begin{aligned} H_0 : \mu &= 29 \\ H_a : \mu &\neq 29 \end{aligned} .$$

We need a test statistic; a means of quantifying the evidence against the null hypothesis. It is natural to base the test statistic on the sample mean $\bar{X}$. Note that both large and small values of the sample mean would provide evidence against the null and for the alternative. We will use the 64 measurements given above. Recall that we assumed that the standard deviation was known, $\sigma = 5.08$ and that we computed the realized value of the sample mean to be $\bar{x} = 27.75$. Is this consistent with the null hypothesis, i.e. could a value of 27.75 plausibly be observed if the true mean is indeed 29? We know that the sampling distribution of the sample mean $\bar{X}$ is normal with mean μ and standard deviation $\sigma/\sqrt{n} = 5.08/8 = 0.635$. If the null hypothesis is true then the mean is 29. Thus, probability theory gives us the sampling distribution of the sample mean under an assumption that the null hypothesis is true. We can use this information to see how unusual the observed value of the sample mean is given an assumed population mean of 29. The test statistic commonly used is

$$Z = \frac{\bar{X} - 29}{0.635}.$$

Under the assumptions Z will have a standard normal distribution. We compute the value of the test statistic

$$z = \frac{27.75 - 29}{0.635} = -1.97.$$

If the null hypothesis is true then the observed value of the sample mean lies 1.97 standard deviations below the hypothesized mean of 29. In our example, Newcomb did not seem to have a direction in mind initially, i.e. he simply stated his belief that the true mean was not equal to 29. This is an example of a two-sided (or two-tailed) test because as noted above both large and small values of the sample mean would be unusual under the null hypothesis. Of course, any single observation is probabilistically uncertain (with our underlying continuous probability model any single observation has probability 0). We quantify the evidence against the null by computing the probability of observing a value of the test statistic as extreme

or more extreme than the one we actually observed COMPUTED UNDER AN ASSUMPTION THAT THE NULL HYPOTHESIS IS TRUE. This probability is referred to as a P -value. Given the two-sided nature of the test we compute

$$P(Z \leq -1.97) + P(Z \geq 1.97) = 0.049.$$

This calculation was carried out assuming that Z is a standard normal random variable which is true if the null hypothesis is true. Thus, if the null hypothesis is true we would expect to see a value of the sample mean of 27.75 or something more “extreme” (less than 27.75 or greater than 30.25) only about 4.9% of the time if the experiment were repeated over and over again, independently, and under the same conditions. Is this evidence strong enough to reject the null hypothesis? A frequently used cut-off value between “strong enough” and “not strong enough” is 0.05. If the P -value is less than or equal to 0.05 then reject the null hypothesis and if the P -value is greater than 0.05 then fail to reject the null hypothesis. If one adheres strictly to this criterion then the evidence would be strong enough to reject.

The selection of a cut-point between reject and fail-to-reject is called fixed-level testing. The cut-point itself is referred to as an α -level. Other commonly chosen levels are $\alpha = 0.01$ and $\alpha = 0.10$. These three values (0.01, 0.05, and 0.10) have become so ingrained in statistical practice that many people assume there is some scientific justification for them but that is not true.

Two different people could come to different conclusions about the results of a test based on their particular selection of a significance level. If Newcomb chose $\alpha = 0.05$ he would reject the null hypothesis but his imaginary opponent could have demanded stronger evidence say by choosing $\alpha = 0.01$ in which case he would fail to reject the null hypothesis.

We have seen a specific example of a two-sided test. More generally a two-sided test consists of the following steps.

1. State the null and alternative hypotheses:

$$\begin{array}{l} H_0: \mu = \mu_0 \\ H_a: \mu \neq \mu_0 \end{array} .$$

2. Compute the value of the test statistic;

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} .$$

3. Compute the P -value,

$$P - value = 2P(Z \geq |z_{obs}|) .$$

where Z is assumed to be standard normal and z_{obs} is the observed value of the test statistic.

There are two other one-sided versions of significance tests. The upper-tailed version has null and alternative hypotheses of

$$\begin{aligned} H_0: \mu &\leq \mu_0 \\ H_a: \mu &> \mu_0 \end{aligned}$$

And the lower-tailed version has null and alternative hypotheses of

$$\begin{aligned} H_0: \mu &\geq \mu_0 \\ H_a: \mu &< \mu_0 \end{aligned}$$

The test statistic is the same in all cases. The P -value for the upper-tailed version is the probability of getting a value of the test statistic equal to or greater than the one actually observed computed under an assumption that the true mean is μ_0 . If the evidence against the null is strong enough to reject the null when the true mean is μ_0 it will be even stronger when $\mu < \mu_0$. Similarly, the P -value for the lower-tailed version is the probability of getting a value of the test statistic equal to or less than the one actually observed computed under an assumption that the true mean is μ_0 .

If fixed-level testing is desired then an appropriate significance level is chosen and the decision of whether to reject or fail to reject the null hypothesis is based on a comparison of the P -value to the significance level.

Hypothesis testing is carried out in a manner that is mechanically identical to the above. A key difference is that interest centers on controlling the rates at which various errors are made. Hard and fast rules are specified for whether or not to reject the null hypothesis. The probability of rejecting the null hypothesis when $\mu = \mu_0$ is called a Type I error and the significance level is the rate at which that error will occur. The probability of failing to reject the null hypothesis when it is false is called a Type II error. The power of a test is $1 - P(\text{Type II error})$. In practice an investigator decides how much of a difference between the true mean and the hypothesized value μ_0 he is willing to accept, i.e. he decides how false the null has to be before he needs a high probability of detecting it.

Obviously the ideal hypothesis test will have low rates of both types of errors but specifying a small chance of a Type I error leads to a larger chance of a Type II error. This tension between the two types of errors is frequently misunderstood. An investigator will choose a low level of significance (a small chance of committing a Type I error), say 1%, failing to understand that the probability of a Type II error can be quite high.

Hypothesis testing is frequently characterized as a decision making process. As such a null hypothesis may be “accepted”, not because it is believed to be true but because it is believed to be true enough to justify acting as if it is true. It may make sense in applications in quality control, for example. A shipment of raw material arrives at a plant and it must be accepted or rejected. Type I

and Type II error rates are determined based on economic considerations and on purity considerations. It is known they over the long run a certain proportion of acceptable shipments will be rejected and a certain proportion of unacceptable shipments will be accepted but those error rates are economically justifiable.

As mentioned above Fisher is generally credited with inventing significance testing and Neyman and Pearson credited with hypothesis testing. Fisher, who was a scientist, was critical of the Neyman-Pearson approach. He did not concern himself with Type II errors. He did not see science as a decision making process. Neyman-Pearson did not use P -values in their original work. They were interested in controlling error rates. The test statistics in tests of population means were sample means themselves. They used the specified values of error rates to find so-called rejection and acceptance regions for values of the sample mean. If an observed value fell in a rejection region the null was rejected. If a value fell in an acceptance region the null was “accepted”. Later, it was noticed that the P -value could be used to determine if the sample mean was in or out of a rejection region. If the P -value was less than the significance level then the value of the sample mean fell in the rejection region and vice versa. But the size of the P -value meant little in Neyman-Pearson hypothesis testing. If the significance level was chosen to be 0.05 then a P -value of 0.049 meant the same as a P -value of 0.00000006, and a P -value of 0.0495 would lead to a decision to reject the null while one of 0.0505 would lead to a decision to fail to reject the null. Fisher believed the observed size of the P -value was important.

We will not go into the details of the distinction between the two methods of testing. Royall [1997] has a good discussion of the differences. We will also not concern ourselves with the controversies that still surround this topic. Some of those controversies are addressed in chapters to follow. Our goal here has been merely to provide an overview of the main results from probability theory and mathematical statistics to readers who may not have seen this material before or to those who saw it a long time ago.

Also, we have presented the basic concepts of estimation and testing hypothesis about population means and with the unreasonable assumption that the population standard deviation is known. There are many more parameters of interest (variables and standard deviations, correlation coefficients, differences in means, slopes and intercepts of regression lines, etc), and different assumptions are needed in many other cases (the population standard deviation is rarely if ever known in practice). The details differ but the broad concepts presented here remain the same and those are the concepts we wanted to address.

7 CONCLUSION

It is time to wrap up our primer on probability and statistics. Before we conclude we revisit where we began our journey. We defined a notion called “monotonic reasoning” that underlies deductive consequence relation which, in fact, captures the core of deductive reasoning. We have seen that a deductive consequence relation is

by definition monotonic. The way we have understood monotonic reasoning is this: A relation between a set of sentences and a sentence is monotonic if and only if when it also holds between a set and a sentence, it also holds between any superset of the set and the sentence. We mentioned that probability theory provides tools for both understanding and handling inductive arguments. However, there is a possibility of misunderstanding that whenever the expression “probability” occurs in statistical or philosophical literature, it necessarily means that the context is inductive, and therefore, cannot be monotonic. We will show that there are contexts in which although the expression “probability” occurs they do not have anything to do with inductive inferences. To drive this point home, we make a distinction between “probabilistic inference” and “probabilistic conclusion”. By “probabilistic inference,” we mean an inference engine that takes statements as inputs and spits out as its output a conclusion where the inference itself is probabilistic. All inductive inferences are probabilistic inferences in this sense. Consequently, they are non-monotonic. We mean by “probabilistic conclusion” an argument whose conclusion contains a probabilistic statement. An argument with a probabilistic conclusion need not be inductive. Hence, it need not involve non-monotonic reasoning.

In the case of an inference involving probability, its conclusion could be a categorical statement in which the conclusion does not follow necessarily from the premises of the argument. Since it is concerned with inference itself the inference is sometimes called “uncertain inference.” However, in other cases of inference, the conclusion of that inference could be a probabilistic statement, but it could follow deductively from its premises. It is a case of inference about uncertainty (See, [Hempel, 1965; Kyburg, 2000; Bandyopadhyay and Bennett, 2004]). We get four possible combinations between probabilistic conclusion and probabilistic inference. They are as follows:

- Probabilistic conclusion, but not probabilistic inference.
- Probabilistic inference, but no probabilistic conclusion.
- Probabilistic inference and probabilistic conclusion. And
- Neither probabilistic conclusion nor probabilistic inference.

- (i) Probabilistic conclusion, but not a case of probabilistic inference.

P1: This die has six faces, labeled 1, 2, 3, 4, 5, 6.

P2: Each face is equally probable.

C: The probability of rolling a 3 on this die is $1/6$

The conclusion makes a statement about the probability of a certain event and the conclusion follows deductively from the premises. Given the probability calculus, it is a deductively valid argument in which if the premises are true, then the conclusion must be true, too. If we add a new premise to this argument then we won’t be able to undermine its deductive validity.

Therefore, it involves monotonic reasoning. At this point, we want to mention that Simpson's paradox which is a well-known paradox in probability and statistics in fact falls under this category. Simpson's Paradox involves the reversal of the direction of a comparison or the cessation of an association when data from several groups are combined to form a single whole (see the "Introduction" together with Hájek's paper in the volume for this position). This shows that an argument with a probabilistic conclusion does not necessarily make it a probabilistic inference; therefore, the argument in question does not stand for an inductive inference.

- (ii) Probabilistic inference, but no probabilistic conclusion.

P1. 95% of the balls in the urn are red.

P2: A ball is randomly selected from the urn.

C: This ball drawn will be red.

The conclusion does not follow deductively from its premises, although it is a categorical statement about a particular ball, which may be false. Given a long-run relative frequency interpretation of probability we expect to see a red ball about 95% of the time with repeated draws. We might be willing to bet a large sum of money that a particular draw will be red, but we cannot state this unequivocally. It will be fair to say that even though the premises are true, the conclusion could be very well false. Hence, the inference is not deductive. As a result, the reasoning involved is non-monotonic and therefore, this is an example of inductive inference.

- (iii) Probabilistic inference and probabilistic conclusion

The following is an example in which the inference is probabilistic as well as its conclusion.

P1. The die has six faces, labeled 1, 2, 3, 4, 5, 6

P2. In a sequence of 250 rolls, a 3 was rolled 50 times.

C: The probability of rolling a 3 with this die is about $1/5$.

In this situation, we happen to observe 50 threes in 250 rolls of a fair die. The probability of obtaining a three on a single roll is $1/5$. It is an instance of inductive inference since even though the premises are true the conclusion could be false. Therefore, the reasoning the argument employs is non-monotonic. To understand the reasoning involved in (iii), compare (iii) with (i). In (iii) we have a specific outcome of a random process and we are reasoning from the sample to the population. In this sense, we are going from the specific to the general. In (i), however, we are reasoning from the population process to the behavior of samples over the long run. Here, we are making an inference about a specific outcome from a general claim.

Therefore, (i) is a case of deductive inference involving monotonic reasoning, whereas (iii) is a case of inductive inference involving non-monotonic reasoning.

- (iv) Neither probabilistic conclusion nor probabilistic inference.

Consider an example which does not involve probabilistic inference nor does it involve probabilistic conclusion.

P1: Wherever there is smoke, there is fire.

P2: The hill has smoke.

C: There is fire on the hill.

Neither its premises nor its conclusion contain a probabilistic claim. This example is a case of non-probabilistic inference (deductive argument) with a non-probabilistic (categorical) conclusion. Therefore, from the perspective of our probabilistic curiosity, it becomes uninteresting. Although this specific argument is deductively valid and therefore, involves monotonic reasoning, any example under this rubric need not be deductively valid. Since our present interest lies in probabilistic reasoning, an example of this specific argument is of no concern for us.

This ends our primer where we have discussed basic idea and theories behind probability and statistics. We have also discussed some of the fundamental differences between inductive and deductive arguments. We hope that this will provide a background for the general reader to read and appreciate many of the papers in the volume.

ACKNOWLEDGEMENTS

We would like to thank Abhijit Dasgupta, Malcolm Forster, John G. Bennett, and Monami Chakrabarti for their comments and suggestions regarding our chapter which have helped to improve the chapter considerably. PSB's research has been supported by Montana State University's *NASA* Astrobiology research center grant (#4w1781).

BIBLIOGRAPHY

- [Bandyopadhyay and Bennett, 2004] P. Bandyopadhyay and J. G. Bennett. Commentary in *The Nature of Evidence*, M. Taper and S. Lele, eds. University of Chicago Press, 2004.
- [DeVeaux *et al.*, 2008] R. DeVeaux, P. Vellman, and D. Bock. *Stats: Data and Models, Second Edition*. Pearson-Addison Wesley, New York, NY. USA, 2008.
- [Devore, 2008] J. Devore. *Probability and Statistics for Engineering and the Sciences, Seventh Edition*. Brooks-Cole, Belmont, CA. USA, 2008.
- [Hájek, 2007] A. Hájek. Interpretations of Probability in *Stanford Encyclopedia of Philosophy*, 2007.
- [Hempel, 1965] C. G. Hempel. *Aspects of Scientific Explanation*. New York: Free Press, 1965.

- [Kyburg, 2000] H. Kyburg. Probable Inference and Probable Conclusion. In *Science, Explanation, and Rationality*. J. Fetzer (ed.) Oxford University Press., New York: New York, 2000.
- [Moore and McCabe, 2006] D. Moore and G. McCabe. *An Introduction to the Practice of Statistics, Fifth Edition*. W.H. Freeman & Company, New York, NY. USA, 2006.
- [Royall, 1997] R. Royall. *Statistical Evidence: A likelihood paradigm*. Chapman & Hall: London, 1997.
- [Stigler, 1986] S. Stigler. *History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press: Cambridge: Mass, 1986.

This page intentionally left blank

Part II

**Philosophical Controversies
about
Conditional Probability**

This page intentionally left blank

CONDITIONAL PROBABILITY

Alan Hájek

1 INTRODUCTION

A fair die is about to be tossed. The probability that it lands with ‘5’ showing up is $1/6$; this is an *unconditional probability*. But the probability that it lands with ‘5’ showing up, *given* that it lands with an odd number showing up, is $1/3$; this is a *conditional probability*. In general, conditional probability is probability *given* some body of evidence or information, probability *relativised to* a specified set of outcomes, where typically this set does not exhaust all possible outcomes. Yet understood that way, it might seem that *all* probability is conditional probability — after all, whenever we model a situation probabilistically, we must initially delimit the set of outcomes that we are prepared to countenance. When our model says that the die may land with an outcome from the set $\{1, 2, 3, 4, 5, 6\}$, it has already ruled out its landing on an edge, or on a corner, or flying away, or disintegrating, or . . . , so there is a good sense in which it is taking the non-occurrence of such anomalous outcomes as “given”. Conditional probabilities, then, are supposed to earn their keep when the evidence or information that is “given” is more specific than what is captured by our initial set of outcomes. In this article we will explore various approaches to conditional probability, canvassing their associated mathematical and philosophical problems and numerous applications. Having done so, we will be in a better position to assess whether conditional probability can rightfully be regarded as the fundamental notion in probability theory after all.

Historically, a number of writers in the pantheon of probability took it to be so. Johnson [1921], Keynes [1921], Carnap [1952], Popper [1959b], Jeffreys [1961], Renyi [1970], and de Finetti [1974/1990] all regarded conditional probabilities as primitive. Indeed, de Finetti [1990, 134] went so far as to say that “every prevision, and, in particular, every evaluation of probability, is conditional; not only on the mentality or psychology of the individual involved, at the time in question, but also, and especially, on the state of information in which he finds himself at that moment”. On the other hand, orthodox probability theory, as axiomatized by Kolmogorov [1933], takes unconditional probabilities as primitive and later analyses conditional probabilities in terms of them. Whatever we make of the primacy, or otherwise, of conditional probability, there is no denying its importance, both in probability theory and in the myriad applications thereof — so much so that the author of an article such as this faces hard choices of prioritisation. My choices are targeted more towards a philosophical audience, although I hope that they will be of wider interest as well.

2 MATHEMATICAL THEORY

2.1 Kolmogorov's axiomatization, and the ratio formula

We begin by reviewing Kolmogorov's approach. Let Ω be a non-empty set. A *field (algebra)* on Ω is a set $\mathcal{F}$ of subsets of Ω that has Ω as a member, and that is closed under complementation (with respect to Ω) and union. Assume for now that $\mathcal{F}$ is finite. Let P be a function from $\mathcal{F}$ to the real numbers obeying:

1. $P(A) \geq 0$ for all $A \in \mathcal{F}$. (Non-negativity)
2. $P(\Omega) = 1$. (Normalization)
3. $P(A \cup B) = P(A) + P(B)$ for all $A, B \in \mathcal{F}$ such that $A \cap B = \emptyset$ (Finite additivity)

Call P a *probability function*, and $(\Omega, \mathcal{F}, P)$ a *probability space*.

One could instead attach probabilities to members of a collection of *sentences* of a formal language, closed under truth-functional combinations.

Kolmogorov extends his axiomatization to cover infinite probability spaces. Probabilities are now defined on a σ -*field* (σ -*algebra*) — a field that is further closed under *countable* unions — and the third axiom is correspondingly strengthened:

- 3'. If $A_1, A_2, \dots$ is a countable sequence of (pairwise) disjoint sets, each belonging to $\mathcal{F}$, then
- $$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) \quad (\text{Countable additivity})$$

So far, all probabilities have been unconditional. Kolmogorov then introduces the *conditional probability of A given B* as the ratio of unconditional probabilities:

$$(\text{RATIO}) \quad P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ provided } P(B) > 0$$

(On the sentential formulation this becomes:

$$P(A|B) = \frac{P(A \& B)}{P(B)}, \text{ provided } P(B) > 0.)$$

This is often called the “definition” of conditional probability, although I suggest that instead we call it a *conceptual analysis*¹ of conditional probability. For ‘conditional probability’ is not simply a technical term that one is free to introduce however one likes. Rather, it begins as a pre-theoretical notion for which we have

¹Or (prompted by Carnap [1950] and Maher [2007]), perhaps it is an *explication*. I don't want to fuss over the viability of the analytic/synthetic distinction, and the extent to which we should be refining a folk-concept that may not even be entirely coherent. Either way, my point stands that Kolmogorov's formula is not merely definitional.

associated intuitions, and Kolmogorov's ratio formula is answerable to those. So while we are free to stipulate that ' $P(A|B)$ ' is merely shorthand for this ratio, we are *not* free to stipulate that 'the conditional probability of A , given B ' should be identified with this ratio. Compare: while we are free to stipulate that ' $A \supset B$ ' is merely shorthand for a connective with a particular truth table, we are *not* free to stipulate that 'if A , then B ' in English should be identified with this connective.

And Kolmogorov's ratio formula apparently answers to most of our intuitions wonderfully well.

2.2 Support for the ratio formula

Firstly, it is apparently supported on a case-by-case basis. Consider the fair die example. Intuitively, the probability of '5', given 'odd', is $1/3$ because we imagine narrowing down the possible outcomes to the three odd ones, observing that '5' is one of them and that probability is shared equally among them. And the ratio formula delivers this verdict:

$$P(5|odd) = \frac{P(5 \cap odd)}{P(odd)} = \frac{1/6}{1/2} = 1/3.$$

And so it goes with countless other examples.

Secondly, a nice heuristic for Kolmogorov's axiomatization is given by van Fraassen's [1989] "muddy Venn diagram" approach, which suggests an informal argument in favour of the ratio formula. Think of the usual Venn-style representation of sets as regions inside a box (depicting Ω). Think of probability as mud spread over the diagram, so that the amount of mud sitting above a given region corresponds to its probability, with a total amount of 1 unit of mud. When we consider the conditional probability of A , given B , we restrict our attention to the mud that sits above the region representing B , and then ask what proportion of *that* mud sits above A . But that is simply the amount of mud sitting above $A \cap B$, divided by the amount of mud sitting above B .

Thirdly, the ratio formula can be given a frequentist justification. Suppose that we run a long sequence of n trials, on each of which B might occur or not. It is natural to identify the probability of B with the relative frequency of trials on which it occurs:

$$P(B) = \frac{\#(B)}{n}$$

Now consider among *those* trials the proportion of those on which A also occurs:

$$P(A|B) = \frac{\#(A \cap B)}{\#(B)}$$

But this is the same as

$$\frac{\#(A \cap B)/n}{\#(B)/n}$$

which on the frequentist interpretation is identified with

$$\frac{P(A \cap B)}{P(B)}.$$

Fourthly, the ratio formula for subjective conditional probability is supported by an elegant *Dutch Book* argument originally due to de Finetti [1937] (here simplified). Begin by identifying your subjective probabilities, or *credences*, with your corresponding betting prices. You assign probability p to X if and only if you regard pS as the value of a bet that pays S if X , and nothing otherwise. Symbolize this bet as:

S	if X
0	otherwise

For example, my credence that this coin toss results in heads is $1/2$, corresponding to my valuing the following bet at 50 cents:

\$ 1	if heads
0	otherwise

A *Dutch Book* is a set of bets bought or sold at such prices as to guarantee a net loss. An agent is susceptible to a Dutch Book if there exists such a set of bets, bought or sold at prices that she deems acceptable. Now introduce the notion of a *conditional bet* on A , given B , which

- pays \$1 if $A \cap B$
- pays 0 if $A^c \cap B$
- is called off if B^c (that is, the price you pay for the bet is refunded if B does not occur).

Identify your $P(A|B)$ with the value you attach to this conditional bet — that is, to:

\$1	if $A \cap B$
0	if $A^c \cap B$
$P(A B)$	if B^c

Now we can show that if your credences violate (RATIO), you are susceptible to a Dutch Book. For the conditional bet can be regarded as equivalent (giving the same pay-offs in every possible outcome) to the following pair of bets:

\$ 1	if $A \cap B$
0	if $A^c \cap B$
0	if B^c

0	if $A \cap B$
0	if $A^c \cap B$
$P(A B)$	if B^c

which you value at $P(A \cap B)$ and $P(A|B)P(B^c)$ respectively. So to avoid being Dutch Booked, you must value the conditional bet at $P(A \cap B) + P(A|B)P(B^c)$. That is:

$$P(A|B) = P(A \cap B) + P(A|B)P(B^c),$$

from which the ratio formula follows (since $P(B^c) = 1 - P(B)$).

2.3 Some basic theorems involving conditional probability

With Kolmogorov's axioms and ratio formula in place, we may go on to prove a number of theorems involving conditional probability. Especially important is the *law of total probability*, the simplest form of which is:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

This follows immediately from the additivity formula

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

by two uses of (RATIO). The law generalizes to the case in which we have a countable partition $B_1, B_2, \dots$:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots$$

This tells us that the unconditional probability $P(A)$ can be identified with a weighted average, or expectation, of probabilities conditional on each cell of a partition, the weights being the unconditional probabilities of the cells. We will see how the theorem underpins Kolmogorov's more sophisticated formulation of conditional probability (§5), and a rule for updating probabilities (Jeffrey conditionalization, §7.2). In the meantime, notice how it yields versions of the equally celebrated *Bayes' theorem*:

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} && \text{(by two uses of (RATIO))} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} && \text{(by the law of total probability)} \end{aligned}$$

More generally, suppose there is a partition of hypotheses $\{H_1, H_2, \dots\}$, and evidence E . Then for each i ,

$$P(H_i|E) = \frac{P(E|H_i)P(H_i)}{\sum_j P(E|H_j)P(H_j)}$$

The $P(E|H_i)$ terms are called *likelihoods*.

Bayes' theorem has achieved such a mythic status that an entire philosophical and statistical movement — "Bayesianism" — is named after it. This is meant

to honour the important role played by Bayes' theorem in calculating terms of the form ' $P(H_i|E)$ ', and (at least among philosophers) Bayesianism is particularly associated with a subjectivist interpretation of ' P ', and correspondingly with the thesis that rational degrees of belief are probabilities. This may seem somewhat curious, as Bayes' theorem is neutral vis-à-vis the interpretation of probability, being purely a theorem of the formal calculus, and just one of many theorems at that. In particular, it provides just one way to calculate a conditional probability, when various others are available, all ultimately deriving from (RATIO); and as we will see, often conditional probabilities can be ascertained directly, without any calculation at all. Moreover, a diachronic prescription for *revising* or *updating* probabilities, which we will later call 'conditionalization', is sometimes wrongly called 'updating by Bayes' theorem'. In fact the theorem is a 'static' rule relating probabilities synchronically, and being purely a piece of mathematics it cannot by itself have any normative force.

2.4 Independence

Kolmogorov's axioms assimilate probability theory to measure theory, the general theory of length, area, and volume. (Think of how these quantities are non-negative, additive, and can often be normalized.) Conditional probability is a further, distinctively probabilistic notion without any obvious counterpart in measure theory. Similarly, *independence* is distinctively probabilistic, and ultimately parasitic on conditional probability. Let $P(A)$ and $P(B)$ both be positive. According to Kolmogorov's theory, A and B are *independent*

$$\begin{aligned} \text{iff } P(A|B) &= P(A); \text{ equivalently,} \\ \text{iff } P(B|A) &= P(B). \end{aligned}$$

These equations are supposed to capture the idea of A and B being uninformative regarding each other: that one event occurs in no way affects the probability that the other does. To be sure, there is a further equivalent characterization of independence that is free of conditional probability:

$$P(A \cap B) = P(A)P(B).$$

But its rationale comes from the equalities of the corresponding conditional and unconditional probabilities. It has the putative advantage of applying even when A and/or B have probability 0; although it is questionable whether probability 0 events should automatically be independent of everything (including themselves, and their complements!).

When we say that ' A is independent of B ', we suppress the fact that such independence is really a *three*-place relation between an event, an event, and a *probability function*. This distinguishes probabilistic independence from such two-place relations as logical and counterfactual independence. Probabilistic independence is assumed in many of probability theory's classic limit theorems.

We may go on to give a Kolmogorovian analysis of *conditional independence*. We say that A is independent of B , given C , if

$$P(A \cap B|C) = P(A|C)P(B|C), \text{ provided } P(C) > 0.$$

2.5 Conditional expectation

Conditional probability underlies the concept of *conditional expectation*, also important in probability theory. A *random variable* X (on Ω) is a function from Ω to the set of real numbers, which takes the value $X(\omega)$ at each point $\omega \in \Omega$. If X is a random variable that takes the values $x_1, x_2, \dots$ with probabilities $p(x_1), p(x_2), \dots$, then the *expected value* of X is defined as

$$E(X) = \sum_i x_i p(x_i)$$

provided that the series converges absolutely. (For continuous random variables, we replace the $p(x_i)$ by values given by a density function, and replace the sum by an integral.) A *conditional expectation* is an expectation of a random variable with respect to a conditional probability distribution. Let X and Y be two random variables with joint distribution

$$P(X = x_j \cap Y = y_k) = p(x_j, y_k) (j, k = 1, 2, \dots)$$

The conditional expectation $E(Y|X = x_j)$ of Y given $X = x_j$ is given by:

$$\sum_k y_k P(Y = y_k|X = x_j) = \frac{\sum_k y_k p(x_j, y_k)}{p(x_j)}$$

(Again, this has a continuous version.) We may generalize to conditional expectations involving more random variables.

So the importance of conditional probability in probability theory is beyond dispute. The same can be said of its role in many philosophical applications.

3 PHILOSOPHICAL APPLICATIONS

Conditional probability is near-ubiquitous in both the methodology — in particular, the use of statistics and game theory — of the sciences and social sciences, and in their specific theories. Various central concepts in statistics are defined in terms of conditional probabilities: significance level, power, sufficient statistics, ancillarity, maximum likelihood estimation, Fisher information, and so on. Game theorists appeal to conditional probabilities for calculating the expected payoffs in correlated equilibrium; computing the Bayesian equilibrium in games of incomplete information; in certain Bayesian dynamic updating models of equilibrium selection in repeated games; and so on. Born's rule in quantum mechanics is often

understood as a method of calculating the conditional probability of a particular measurement outcome, given that a measurement of a certain kind is performed on a system in a certain state. In medical and clinical psychological testing, conditional probabilities of the form $P(\text{disorder} \mid \text{positive test})$ (“diagnosticity”) and $P(\text{positive test} \mid \text{disorder})$ (“sensitivity”) take centre stage. Mendelian genetics allows us to compute probabilities for an organism having various traits, given information about the traits of its parents; and population genetics allows us to compute the chance of a trait going to fixation, given information about population size, initial allele frequencies, and the fitness gradient. The regression equations of economics, and many of the results in time series analysis, are claims about conditional probabilities. And so it goes — this little sampler could be extended almost indefinitely.

Moreover, conditional probability is a staple of philosophy. The next section surveys a few of its philosophical applications.

3.1 *Conditional probability in the philosophy of probability*

A central issue in the philosophical foundations of probability is that of *interpreting* probability — that is, of analysing or explicating the ‘ P ’ that appears in its formal theory. Conditional probability finds an important place in all of the leading interpretations:

Frequentism: Probability is understood as relative frequency (perhaps in an infinite sequence of hypothetical trials) — e.g. the probability of heads for a coin is identified with the number of heads outcomes divided by the total number of trials in some suitable sequence of trials. Recalling our third justification for the ratio formula in §2.2, this seems to be naturally understood as a conditional probability, the condition being whatever determines the suitability of that sequence.

Propensity: Probability is a measure of the tendency for a certain kind of experimental set-up to produce a particular outcome, either in the single case [Giere, 1973], or in the long run [Popper, 1959a]. Either way, it is a conditional probability, the condition being a specification of the experimental set up.

Classical: Probability is assigned by one in an epistemically neutral position with respect to a set of “equally possible” cases — outcomes on which one’s evidence bears equally. Such an assignment must thus be relativised to such evidence.

Logical: Probability is a measure of inductive support or partial entailment, generalizing both deductive logic’s notion of entailment and the classical interpretation’s assignments to “equally possible” cases. In Carnap’s notation, $c(h, e)$ is a measure of the degree of support that evidence e confers on h . This is explicitly a conditional probability.

Subjective: Probability is understood as the degree of belief of some agent (typically assumed to be ideally rational). As we have seen, some subjectivists (e.g. Jeffreys, de Finetti) explicitly regarded subjective *conditional* probability to be basic. But even subjectivists who regard unconditional probability as basic find an important place for conditional probability. Subjectivists are unified in regarding

conformity to the probability calculus as a rational requirement on credences. They often add further constraints, couched in terms of conditional probabilities; a number of examples follow.

Gaifman [1988] coins the term “expert probability” for a probability assignment that a given agent with subjective probability function P strives to track. We may codify this idea as follows (simplifying his characterization at the expense of some generality):

$$(\text{Expert}) \quad P(A|pr(A) = x) = x, \text{ for all } x \text{ such that } P(pr(A) = x) > 0.$$

Here $pr(A)$ is the assignment that the agent regards as expert. For example, if you regard the local weather forecaster as an expert, and she assigns probability 0.1 to it raining tomorrow, then you may well follow suit:

$$P(\text{rain} | pr(\text{rain}) = 0.1) = 0.1.$$

More generally, we might speak of an entire probability function as being such a guide for an agent, over a specified set of propositions — so that (Expert) holds for any choice of A from that set. A *universal expert function* would guide *all* of the agent’s probability assignments in this way. van Fraassen [1984; 1995], following Goldstein [1983], argues that an agent’s *future* probability functions are universal expert functions for that agent — his *Reflection Principle*:

$$P_t(A|P_{t'}(A) = x) = x, \text{ for all } A \text{ and for all } x \text{ such that } P_t(P_{t'}(A) = x) > 0,$$

where P_t is the agent’s probability function at time t , and $P_{t'}$ her function at later time t' . The principle encapsulates a certain demand for ‘diachronic coherence’ imposed by rationality. van Fraassen defends it with a ‘diachronic’ Dutch Book argument (one that considers bets placed at different times), and by analogizing violations of it to the sort of pragmatic inconsistency that one finds in Moore’s paradox.

We may go still further. There may be universal expert functions for all rational agents. The *Principle of Direct Probability* regards the relative frequency function as a universal expert function (cf. [Hacking, 1965]). Let A be an event-type, and let $relfreq(A)$ be the relative frequency of A (in some suitable reference class). Then for any rational agent, we have

$$P(A|relfreq(A) = x) = x, \text{ for all } A \text{ and for all } x \text{ such that } P(relfreq(A) = x) > 0.$$

Related, but distinct according to those who do not identify objective chances with relative frequencies, is Lewis’s [1980]:

$$(\text{Principal Principle}) \quad P(A|ch_t(A) = x \ \& \ E) = x$$

Here ‘ P ’ is an ‘initial’ rational credence function (the prior probability function of a rational agent who has acquired no information), A is a proposition, $ch_t(A)$ is the chance of A at time t and E is further evidence that may be acquired. In

order for the Principal Principle to be applicable, E cannot be relevant to whether A is true or false, other than by bearing on the chance of A at t ; E is then said to be *admissible* (strictly speaking: with respect to P, A, t , and x). The literature ritually misstates the Principal Principle, regarding ' P ' as the credence function of a rational agent quite generally, rather than an 'initial' credence function as Lewis explicitly formulated it. Misstated this way, it is open to easy counterexamples in which the agent has information bearing on A that has been incorporated into ' P ', although not explicitly written into the condition (in the slot that ' E ' occupies). Interestingly, admissibility is surely just as much an issue for the other expert principles, yet for some reason hardly discussed outside the Principal Principle literature, where it is all the rage.

Finally, some authors impose the requirement of *strict coherence* on rational agents: such an agent assigns $P(H|E) = 1$ only if E entails H . See Shimony [1955].

3.2 *Some uses of conditional probability in other parts of philosophy*

The use of conditional probability in updating rules for credences, and in the semantics of conditionals, has been so important and fertile that I will devote entire sections to them later on (§7 and §9). In the meantime, here are just a few of the myriad applications of conditional probability in various other areas of philosophy.

Probabilistic causation

A major recent industry in philosophy has been that of providing analyses of causation compatible with indeterminism. At a first pass, we might analyze 'causation' as 'correlation' — that is, analyze ' A causes B ' as

$$P(B|A) > P(B|A^c).$$

This analysis cannot be right. It wrongly classifies spurious correlations and effects of common causes as instances of causation; moreover, it fails to capture the asymmetry of the causal relation. So a number of authors refine the analysis along the following lines (e.g. [Suppes, 1970; Cartwright, 1979; Salmon, 1980; Eells, 1991]):

A causes B iff $P(B|A \cap X) > P(B|A^c \cap X)$ for every member X of some 'suitable' partition.

The exact details vary from author to author; what they share is the fundamental appeal to inequalities among conditional probabilities.

Reichenbach's [1956] famous *common cause principle* is again couched in terms of inequalities among conditional probabilities. The principle asserts that if A and B are simultaneous events that are correlated, then there exists an earlier common

cause C of A and B , such that for every member X of some ‘suitable’ partition,

$$\begin{aligned} P(A|C) &> P(A|C^c), \\ P(B|C) &> P(B|C^c), \\ P(A \cap X|C) &= P(A|C)P(B|C), \text{ and} \\ P(A \cap X|C^c) &= P(A|C^c)P(B|C^c). \end{aligned}$$

That is, C is correlated with A and with B , and C screens off A from B (they are independent conditional on C).

Bayesian networks

We may model a causal network as a directed acyclic graph with nodes corresponding to variables. If one variable directly causes another, we join the corresponding nodes with a directed edge, its arrow pointing towards the ‘effect’ variable. We may naturally employ a genealogical nomenclature. We call the cause the ‘parent’ variable, the effect a ‘child’ variable, and call iterations of these relationships ‘ancestors’ and ‘descendants’ in the obvious way.

In a Bayesian network, a probability distribution is assigned across the nodes. The *Causal Markov condition* is a commonly held assumption about conditional independence relationships. Roughly, it states that any node in a given network is conditionally independent of its non-descendants, given its parents. More formally (with obvious notation): “Let G be a causal graph with vertex set V and P be a probability distribution over the vertices in V generated by the causal structure represented by G . G and P satisfy the Causal Markov Condition if and only if for every W in V , W is independent of $V \setminus (\text{Descendants}(W) \cup \text{Parents}(W))$ given $\text{Parents}(W)$ ” [Spirtes *et al.*, 2000, 29]. (“ $\setminus$ ” denotes set subtraction.) *Faithfulness* is the converse condition that the set of independence relations derived from the Causal Markov Condition is exactly the set of independence relations that hold for the network. (See [Spirtes *et al.*, 2000; Hausman and Woodward, 1999].)

Inductive-statistical explanation:

Hempel [1965] regards scientific explanation as a matter of subsuming an explanandum E under a law L , so that E can be derived from L in conjunction with particular facts. He also recognizes a distinctive kind of “inductive-statistical” (IS) explanation, in which E is subsumed under a *statistical* law, which will take the form of a statement of conditional probability; in this case, E cannot be validly derived from the law and particular facts, but rather is rendered probable in accordance with the conditional probability.

Confirmation

While ‘correlation is causation’ is an unpromising slogan, ‘correlation is confirmation’ has fared much better. Confirmation is a useful concept, because even

if Hume was right that there are no necessary connections between distinct existences, still it seems there are at least some non-trivial probabilistic relations between them. That's just what we mean by saying things like ' B supports A ', or ' B is evidence for A ', or ' B is counterevidence for A ', or ' B disconfirms A '. So, many Bayesians appropriate the unsuccessful first attempt above to analyze causation, and turn it into a far more successful attempt to analyze confirmation — confirmation is positive correlation, disconfirmation is negative correlation, and evidential irrelevance is independence:

Relative to probability function P ,

- E confirms H iff $P(H|E) > P(H)$
- E disconfirms H iff $P(H|E) < P(H)$
- E is evidentially irrelevant to H iff $P(H|E) = P(H)$

Curve-fitting, and the Akaike Information Criterion

Scientists are familiar with the problem of fitting a curve to a set of data. Forster and Sober [1994] argue that the real problem is one of trading off verisimilitude and simplicity: for a given set of data points, finding the curve that best balances the desiderata of predicting the points as accurately as possible using a function that has as few parameters as possible, so as not to 'overfit' the points. They argue that simplicity should be attributed to *families* of curves rather than to individual curves. They advocate selecting the family F with the best expected 'predictive accuracy', as measured by the Akaike Information Criterion:

$$AIC(F) = \frac{1}{N} \log(P(\text{Data}|L(F)) - k),$$

where $L(F)$ is the member of F that fits the data best, and k is the number of adjustable parameters of members of F .

Various other approaches to the curve-fitting problem similarly appeal to likelihoods (at least tacitly), and thus to conditional probabilities. They include the Bayesian Information Criterion [BIC; Schwarz, 1978], Minimum Message Length inference [MML; Wallace and Dowe, 1999; Dowe *et al.*, 2007] and Minimum Description Length inference [MDL; Grunwald *et al.*, 2005].

Decision theory

Decision theory purports to tell us how an agent's beliefs and desires in tandem determine what she should do. It combines her utility function and her probability function to give a figure of merit for each possible action, called the *expectation*, or *desirability* of that action (rather like the formula for the expectation of a random variable): a weighted average of the utilities associated with each action. In so-called 'evidential decision theory', as presented by Jeffrey [1983], the weights are

conditional probabilities for states, given actions. Let $S_1, S_2, \dots, S_n$ be a partition of possible states of the world. The choice-worthiness of action A is given by:

$$V(A) = \sum_i u(A \& S_i) P(S_i | A)$$

And so it goes again — this has just been another sampler. Given the obvious importance of conditional probability in philosophy, it will be worth investigating how secure are its foundations in (RATIO).

4 PROBLEMS WITH THE RATIO ANALYSIS OF CONDITIONAL PROBABILITY

So far we have looked at success stories for the usual understanding of conditional probability, given by (RATIO). We have seen that several different kinds of argument triangulate to it, and that it subserves a vast variety of applications of conditional probability — indeed, this latter fact itself provides a further pragmatic argument in favour of it. But we have not yet reached the end of the story. I turn now to four different kinds of problem for the ratio analysis, each mitigating the arguments in its favour.

4.1 *Conditions with probability zero*

$\frac{P(A \cap B)}{P(B)}$ is undefined when $P(B) = 0$; the ratio formula comes with the proviso that $P(B) > 0$. The proviso would be of little consequence if we could be assured that all probability-zero events of any interest are impossible, but as probability textbooks and even Kolmogorov himself caution us, this is not so. That is, we could arguably dismiss probability zero antecedents as ‘don’t cares’ if we could be assured that all probability functions of any interest are *regular* — that is, they assign probability 0 only to the empty set. But this is not so. Worse, there are many cases of such conditional probabilities in which intuition delivers a clear verdict as to the correct answer, but (RATIO) delivers no verdict at all.

Firstly, in uncountable probability spaces one cannot avoid probability zero events that are possible — indeed, we are saddled with uncountably many of them.² Consider probability spaces with points taken from a continuum. Here is an example originating with Borel: A point is chosen at random from the surface of the earth (thought of as a perfect sphere); what is the probability that it lies in the Western hemisphere, given that it lies on the equator? $1/2$, surely. Yet the probability of the condition is 0, since a uniform probability measure over a sphere must award probabilities to regions in proportion to their area, and the equator

²Here I continue to assume Kolmogorov’s axiomatization, according to which probabilities are *real-valued*. Regularity may be achieved by the use of *infinitesimal probabilities* (see e.g. [Skyrms, 1980]); but see [Hájek, 2003] for some concerns.

has area 0. The ratio analysis thus cannot deliver the intuitively correct answer. Obviously there are uncountably many problem cases of this form for the sphere.

Another class of problem cases arises from the fact that the power set of a denumerable set is uncountable. For example, the set of all possible infinite sequences of tosses of a coin is uncountable (the sets of positive integers that could index the heads outcomes form the power set of the positive integers). Any particular sequence has probability zero (assuming that the trials are independent and identically distributed with intermediate probability for heads). Yet surely various corresponding conditional probabilities are defined — e.g., the probability that a fair coin lands heads on every toss, given that it lands heads on tosses 3, 4, 5, ..., is $1/4$.

More generally, the various classic ‘almost sure’ convergence results — the strong law of large numbers, the law of the iterated logarithm, the martingale convergence theorem, etc. — assert that certain convergences take place, not with certainty, but ‘almost surely’. This is not merely coyness, since these convergences may fail to take place — again, genuine possibilities that receive probability 0, and interesting ones at that. The fair coin may land heads on every toss, and it would be no less fair for it.

Zero probability events also arise naturally in countable probability spaces if we impose certain symmetry constraints (such as the principle of indifference), and if we are prepared to settle for finite additivity. Following de Finetti [1990], imagine an infinite lottery whose outcomes are equiprobable. Each ticket has probability 0 of winning, although with probability 1 some ticket will win. Again, various conditional probabilities seem to be well-defined: for example, the probability that ticket 1 wins, given that either ticket 1, 2, or 3 wins, is surely $1/3$.

The problem of zero-probability conditions is not simply an artefact of the mathematics of infinite (and arguably idealized) probability spaces. For even in finite probability spaces, various possible events may receive probability zero. This is most obvious for subjective probabilities, and in fact it happens as soon as an agent updates on some non-trivial information, thus ruling out the complement of that information — e.g., when you learn that the die landed with an odd number showing up, thus ruling out that it landed with an even number showing up. But still it seems that various conditional probabilities with probability-zero conditions can be well-defined — e.g., the probability that the die landed 2, *given* that it landed 2, is 1. Indeed, it seems that there are some contingent propositions that one is rationally *required* to assign probability 0 — e.g., ‘I do not exist’. But various associated conditional probabilities may be well-defined nonetheless — e.g., the probability that I do not exist, *given* that I do not exist, is 1. Perhaps such cases are more controversial. If so, it matters little. As long as there is *some* case of a well-defined conditional probability with a probability-zero condition, then (RATIO) is refuted as an analysis of conditional probability. (It may nonetheless serve many of our purposes well enough as a *partial* — that is, incomplete — analysis. See [Hájek, 2003] for further discussion.)

4.2 Conditions with unsharp probability

A number of philosophers and statisticians eschew the usual assumption that probabilities are always real numbers, sharp to infinitely many decimal places. Instead, probabilities may for example be intervals, or convex sets, or sets of real numbers more generally. Such probabilities are given various names: “indeterminate” [Levi, 1974; 2000], “vague” [van Fraassen, 1990], “imprecise” [Walley, 1991], although these words have other philosophical associations that may not be intended here. Maybe it is best to mint a new word for this purpose. I will call them *unsharp*, marking the contrast to the usual sharp probabilities, while remaining neutral as to how unsharp probabilities should be modelled.

What is the probability that the Democrats win the next U.S. election? Plausibly, the answer is unsharp. This is perhaps clearest if the probability is subjective. If you say, for example, that your credence that they win is 0.6, it is doubtful that you really mean 0.60000 . . . , precise to infinitely many decimal places. Now, what is the probability that the Democrats win the next U.S. election, *given* that they win the next U.S. election? Here the answer is sharp: 1. Or what is the probability that this fair coin will land Heads when tossed, given the Democrats win the next U.S. election? Again, the answer seems to be sharp: $1/2$. In Hájek (2003) I argue that cases like these pose a challenge to the ratio analysis: it seems unable to yield such results. To be sure, perhaps that analysis coupled with suitable devices for handling unsharpness — e.g. supervaluation — can yield the results (although I argue that they risk being circular). Still, the point remains that the ratio analysis cannot be the complete story about conditional probability.

4.3 Conditions with vague probability

A superficially similar, but subtly different kind of case involves conditions with what I will call *vague* probability. Let us first be clear on the distinction between unsharpness and vagueness in general, before looking at probabilistic cases (this is a reason why I did not adopt the word “vagueness” in the previous section). The hallmark of vagueness is often thought to be the existence of *borderline* cases. A predicate is vague, we may say, if there are possible individuals that do not clearly belong either to the extension or the anti-extension of the predicate. For example, the predicate “fortyish” is vague, conveying a fuzzy region centered around 40 for which there are borderline cases (e.g. a person who is 43). By contrast, I will think of an unsharp predicate as admitting of a range of possible cases, but not borderline cases. “Forty-something” is unsharp: it covers the range of ages in the interval $[40, 50)$, but any particular person either clearly falls under the predicate or clearly does not. However we characterize the distinction, the phenomena of vagueness and unsharpness appear to be different. I now turn to the problem that *vague* probability causes for the ratio analysis.

Suppose that we run a million-ticket lottery. What is the probability that a *large-numbered* ticket wins? It is vague what counts as a ‘large number’ — 17 surely doesn’t, 999,996 surely does, but there are many numbers that are not so

easily classified. The probability assignment plausibly inherits this vagueness — it might be, for example, ‘0.3-ish’, again with borderline cases. Now, what is the probability that a large-numbered ticket wins, *given* that a large-numbered ticket wins? That is surely razor-sharp: 1. As before, the challenge to the ratio analysis is to do justice to these facts.

4.4 *Conditions with undefined probability*

Finally, we come to what I regard as the most important class of problem cases for (RATIO), for they are so widespread and often mundane. They arise when neither $P(A \cap B)$ nor $P(B)$ is defined, and yet the probability of A , given B , is defined. Here are two kinds of case, the first more intuitive, the second more mathematically rigorous, both taken from [Hájek, 2003].

The first involves a coin that you believe to be fair. What is the probability that it lands heads, given that I toss it fairly? $1/2$, of course. According to the ratio analysis, it is

$$\frac{P(\text{the coin lands heads} \mid \text{I toss the coin fairly})}{P(\text{I toss the coin fairly})},$$

However, these unconditional probabilities may not be defined — e.g. you may simply not assign them values. After some thought, you may start to assign them values, but the damage has already been done; and then again, you may still not do so. In [Hájek, 2003] I argue that this ratio may well remain undefined, and I rebut various proposals for how it may be defined after all.

The second kind of case involves non-measurable sets. Imagine choosing a point at random from the $[0, 1]$ interval. We would like to model this with a uniform probability distribution, one that assigns the same probability to a given set as it does to any translation (modulo 1) of that set. Assuming the axiom of choice and countable additivity, it can be shown that for any such distribution P there must be sets that receive no probability assignment at all from P — so called ‘non-measurable sets’. Let N be such a set. Then $P(N)$ is undefined. Nonetheless, it is plausible that the probability that the chosen point comes from N , given that it comes from N , is 1; the probability that it does not come from N , given that it comes from N , is 0; and so on. The ratio analysis cannot deliver these results.

The coin toss case may strike you as contentious, and the non-measurable case as pathological (although in [Hájek, 2003] I defend them against these charges). But notice that many of the paradigmatic applications of conditional probability canvassed in the previous section would seem to go the same way. For example, the Born rule surely should *not* be understood as assigning a value to a ratio of unconditional probabilities of the form

$$\frac{P(\text{measurement outcome } O_k \text{ is observed} \cap \text{measurement } M \text{ is performed})}{P(\text{measurement } M \text{ is performed})}.$$

Among other things, the terms in the ratio are clearly not given by quantum mechanics, and may plausibly not be defined at all, involving as they do a tacit quantification over the free actions of an open-ended set of experimenters.

To summarize: we have seen four kinds of case in which the ratio analysis appears to run aground: conditional probabilities with conditions whose probabilities are either zero, unsharp, vague, or undefined. Now there is a good sense in which these are problems with *unconditional* probability in its own right, which I am parlaying into problems for conditional probability. For example, the fact that Kolmogorov's theory of unconditional probability conflates zero-probability possibilities with genuine impossibilities may seem to be a defect of that theory, quite apart from its consequences for conditional probability. Still, since his theory of conditional probability is parasitic on his theory of unconditional probability, it should come as no surprise that defects in the latter can be exploited to reveal defects in the former. And notice how the problems in unconditional probability theory can be *amplified* when they become problems in conditional probability theory. For example, the conflation of zero-probability possibilities with genuine impossibilities might be thought of as a minor 'blurriness in vision' of probability theory; but it is rather more serious when it turns into problems of outright undefinedness in conditional probability, total blind spots.

Here are two ways that one might respond. First, one might preserve the conceptual priority that Kolmogorov gives to unconditional over conditional probability, but seek a more sophisticated account of conditional probability. Second, one might reverse the conceptual order, and regard conditional probability as the proper primitive of probability theory. The next two sections discuss versions of these responses, respectively.

5 KOLMOROGOV'S REFINEMENT: CONDITIONAL PROBABILITY AS A RANDOM VARIABLE

(This section is more advanced, and may be skipped by readers who are more interested in philosophical issues. Its exposition largely follows [Billingsley, 1995]; the ensuing critical discussion is my own.)

Kolmogorov went on to give a more sophisticated account of conditional probability as a random variable.

5.1 Exposition

Let the probability space $\langle \Omega, \mathcal{F}, P \rangle$ be given. We will interpret P as the credence function of an agent, which assumes the value $P(\omega)$ at each point $\omega \in \Omega$. Fixing $A \in \mathcal{F}$, we may define the random variable whose value is:

$$\begin{aligned} &P(A|B) \text{ if } \omega \in B, \\ &P(A|B^c) \text{ if } \omega \in B^c. \end{aligned}$$

Think of our agent as about to learn the result of the experiment regarding B , and she will update accordingly. (§7 discusses updating rules in greater detail.)

Now generalize from the 2-celled partition $\{B, B^c\}$ to any countable partition $\{B_1, B_2, \dots\}$ of Ω into $\mathcal{F}$ -sets. Let $\mathcal{G}$ consist of all of the unions of the B_i ; it is the smallest sigma field that contains all of the B_i . $\mathcal{G}$ can be thought of as an experiment. Our agent will learn which of the B_i obtains — that is, the outcome of the experiment — and is poised to update her beliefs accordingly. Fixing $A \in \mathcal{F}$, consider the function whose values are:

$$\begin{aligned} &P(A|B_1) \text{ if } \omega \in B_1, \\ &P(A|B_2) \text{ if } \omega \in B_2, \\ &\dots \end{aligned}$$

when these quantities are defined. If $P(B_i) = 0$, let the corresponding value of the function be chosen arbitrarily from $[0, 1]$, this value constant for all $\omega \in B_i$. Call this function the *conditional probability of A given $\mathcal{G}$* , and denote it $P[A|\mathcal{G}]$. Given the latitude in assigning a value to this function if $P(B_i) = 0$, $P[A|\mathcal{G}]$ stands for any one of a family of functions on Ω , differing on how this arbitrary choice is made. A specific such function is called a *version* of the conditional probability. Thus, any two versions agree except on a set of probability 0. Any version codifies all of the agent's updating dispositions in response to all of the possible results of the experiment.

Notice that since any $G \in \mathcal{G}$ is a disjoint union $\cup_k B_{i_k}$, the probability of any set of the form $A \cap G$ can be calculated by the law of total probability:

$$P(A \cap G) = \sum_k P(A|B_{i_k})P(B_{i_k}) \quad (1)$$

We may generalize further to the case where the sigma field $\mathcal{G}$ may not necessarily come from a countable partition, as was previously the case. Our agent will learn for each G in $\mathcal{G}$ whether $\omega \in G$ or $\omega \in G^c$. Generalizing (1), we would like to be assured of the existence of a function $P[A|\mathcal{G}]$ that satisfies the equation:

$$P(A \cap G) = \int_G P[A|\mathcal{G}] dP \text{ for all } G \in \mathcal{G}.$$

That assurance is provided by the *Radon-Nikodym theorem*, which for probability measures ν and P defined on $\mathcal{F}$ states:

If $P(X) = 0$ implies $\nu(X) = 0$ then there exists a function f such that

$$\nu(A) = \int_A f dP$$

for all $A \in \mathcal{F}$.

Let $\nu(G) = P(A \cap G)$ for all $G \in \mathcal{G}$. Notice that $P(G) = 0$ implies $\nu(G) = 0$ so the Radon-Nikodym theorem applies: the function $P[A|\mathcal{G}]$ that we sought does indeed

exist. As before, there may be many such functions, differing on their assignments to probability-zero sets; any such function is called a *version* of the conditional probability.

Stepping back for a moment: $\int_G P[A|\mathcal{G}]dP$ is the expectation of the random variable $P[A|\mathcal{G}]$, conditional on G , weighted according to the measure P . We have come back full circle to the remark made earlier about the law of total probability: an unconditional probability can be identified with an expectation of probabilities conditional on each cell of a partition, weighted according to the unconditional probabilities of the cells.

5.2 Critical discussion

Kolmogorov's more sophisticated formulation of conditional probability provides some relief from the problem of conditions with probability zero — there is no longer any obstacle to such conditional probabilities being defined. However, the other three problems for the ratio analysis — conditions with unsharp, vague, or undefined probability — would appear to remain. For the more sophisticated formulation equates a certain integral, in which the relevant conditional probability figures, to the probability of a conjunction; but when this latter probability is either unsharp, vague, or undefined, the analysis goes silent.

Moreover, there is further trouble that had no analogue for the ratio analysis, as shown by Seidenfeld, Schervish, and Kadane in their [2001] paper on “regular conditional distributions” — i.e. distributions of the form $P[\cdot|\mathcal{A}]$ that we have been discussing. Let $P[\cdot|\mathcal{A}](\omega)$ denote the regular conditional distribution for the probability space $(\Omega, \mathcal{B}, P)$ given the conditioning sub- σ -field $\mathcal{A}$, evaluated at the point ω . Following Blackwell and Dubins [1975], say that a regular conditional distribution is *proper at ω* if it is the case that whenever $\omega \in A \in \mathcal{A}$,

$$P(A|\mathcal{A})(\omega) = 1$$

The distribution is *improper* if it is not everywhere proper. Impropriety seems to be disastrous. We may hold this truth to be self-evident: the conditional probability of anything consistent, given *itself*, should be 1. Indeed, it seems to be about as fundamental fact about conditional probability as there could be, on a par with the fundamental fact in logic that any proposition implies itself. So the possibility of impropriety, however minimal and however localized it might be, is a serious defect in an account of conditional probability. But Seidenfeld *et al.* show just how striking the problem is. They give examples of regular conditional distributions that are *maximally* improper. They are cases in which

$$P[A|\mathcal{A}](\omega) = 0$$

(as far from the desired value of 1 as can be), and this impropriety holds *almost everywhere* according to P , so the impropriety is maximal both locally and glob-

ally.³ This is surely bad news for the more sophisticated analysis of conditional probability — arguably fatal.

6 CONDITIONAL PROBABILITY AS PRIMITIVE

A rival approach takes conditional probability $P(-, -)$ as primitive. If we like, we may then define the unconditional probability of a as $P(a, \mathbf{T})$, where $\mathbf{T}$ is a logical truth. (We use lower case letters and a comma separating them in keeping with Popper's formulation, which we will soon be presenting.) Various axiomatizations of primitive conditional probability have been defended in the literature. See Roeper and Leblanc [1999] for an encyclopedic discussion of competing theories of conditional probability, and Keynes [1921], Carnap [1950], Popper [1959b], and Hájek [2003] for arguments that probability is inherently a two-place function. As is so often the case, their work was foreshadowed by Jeffreys [1939/1961], who axiomatized a comparative conditional probability relation: p is more probable than q , given r .

In some ways, the most general of the proposed axiomatizations is Popper's [1959b], and his system is the one most familiar to philosophers. Renyi's [1970] axiomatization is undeservedly neglected by philosophers. It closely mimics Kolmogorov's axiomatization, replacing unconditional with conditional probabilities in natural ways. I regard it as rather more intuitive than Popper's system. But since the latter has the philosophical limelight, I will concentrate on it here.

Popper's primitives are: (i) Ω , the universal set; (ii) a binary numerical function $p(-, -)$ of the elements of Ω ; a binary operation ab defined for each pair (a, b) of elements of Ω ; a unary operation $\neg a$ defined for each element a of Ω . Each of these concepts is introduced by a postulate (although the first actually plays no role in his theory):

Postulate 1. The number of elements in Ω is countable.

Postulate 2. If a and b are in Ω , then $p(a, b)$ is a real number, and the following axioms hold:

- A1. (Existence) There are elements c and d in Ω such that $p(a, b) \neq p(c, d)$.
- A2. (Substitutivity) If $p(a, c) = p(b, c)$ for every c in Ω , then $p(d, a) = p(d, b)$ for every d in Ω .
- A3. (Reflexivity) $p(a, a) = p(b, b)$.

Postulate 3. If a and b are in Ω , then ab is in Ω ; and if c is also in Ω , then the following axioms hold:

³A necessary condition for this is that the conditioning sub-sigma algebra is not countably generated.

B2. (Monotony) $p(ab, c) \leq p(a, c)$

B2. (Multiplication) $p(ab, c) = p(a, bc)p(b, c)$

Postulate 4. If a is in Ω , then $\neg a$ is in Ω ; and if b is also in Ω , then the following axiom holds:

C. (Complementation) $p(a, b) + p(\neg a, b) = p(b, b)$, unless $p(b, b) = p(c, b)$ for every c in Ω .

Popper also adds a “fifth postulate”, which may be thought of as giving the definition of absolute (unconditional) probability:

Postulate AP. If a and b are in Ω , and if $p(b, c) \geq p(c, b)$ for every c in Ω , then $p(a) = p(a, b)$.

Popper’s axiomatization thus generalizes ordinary probability theory. Intuitively, b can be regarded as a logical truth. Unconditional probability, then, can be regarded as probability conditional on a logical truth. However, a striking fact about the axiomatization is that it is *autonomous* — it does not presuppose any set-theoretic or logical notions (such as “logical truth”). A function $p(-, -)$ that satisfies the above axioms is called a *Popper function*.

A well-known advantage of the Popper function approach is that it allows conditional probabilities of the form $p(a, b)$ to be defined, and to have intuitively correct values, even when the ‘condition’ b has absolute probability 0, thus rendering the usual conditional probability ratio formula inapplicable — we saw examples in §4.1. Moreover, Popper functions can bypass our concerns about conditions with unsharp, vague, or undefined probabilities — the conditional probabilities at issue are assigned directly, without any detour or constraint given by unconditional probabilities.

McGee [1994] shows that in an important sense, probability statements cast in terms of Popper functions and those cast in terms of nonstandard probability functions are inter-translatable. If r is a nonstandard real number, let $st(r)$ denote the *standard part of r* , that is, the unique real number that is infinitesimally close to r . McGee proves the following theorem: If P is a nonstandard-valued probability assignment on a language $\mathcal{L}$ for the classical sentential calculus, then the function $C : \mathcal{L} \times \mathcal{L} \rightarrow \mathbf{R}$ given by

$$\begin{aligned} C(a, b) &= st\left(\frac{P(ab)}{P(b)}\right), \text{ provided } P(b) > 0 \\ &= 1, \text{ otherwise} \end{aligned}$$

is a Popper function. Conversely, if C is a Popper function, there is a nonstandard-valued probability assignment P such that

$$P(b) = 1 \text{ iff } C(-, b) \text{ is the constant function } 1$$

and

$$C(c, b) = st\left(\frac{P(cb)}{P(b)}\right) \text{ whenever } P(b) > 0.$$

The arguments adduced in §4 against the ratio analysis of conditional probability indirectly support taking conditional probability as primitive, although they also leave open the viability of some *other* analysis of conditional probability in terms of unconditional probability. However, there are some considerations that seem to favour the primacy of conditional probability.

The conditional probability assignments that I gave in §4's examples are seemingly non-negotiable. They can, and in some cases must, stand without support from corresponding unconditional probabilities. Moreover, the examples of unsharp, vague, and undefined probabilities suggest that the problem with the ratio analysis is not so much that it is a *ratio* analysis, but rather that it is a *ratio analysis*. The problem lies in the very attempt to analyze conditional probabilities in terms of unconditional probabilities at all. It seems that any other putative analysis that treated unconditional probability as more basic than conditional probability would meet a similar fate — as Kolmogorov's elaboration did.

On the other hand, given an unconditional probability, there is always a corresponding conditional probability lurking in the background. Your assignment of $1/2$ to the coin landing heads superficially seems unconditional; but it can be regarded as conditional on tacit assumptions about the coin, the toss, the immediate environment, and so on. In fact, it can be regarded as conditional on your total evidence — recall the quotation from de Finetti in the second paragraph of this article. Now, perhaps in very special cases we can assign a probability free of all assumptions — an assignment of 1 to 'I exist' may be such a case. But even then, the probability is easily recovered as probability conditional on a logical truth or some other *a priori* truth. Furthermore, we can be sure that there can be no analogue of the argument that conditional probabilities can be defined even when the corresponding unconditional probabilities are not, that runs the other way. For whenever an unconditional probability $P(X)$ is defined, it trivially equals the conditional probability of X given a logical/*a priori* truth. Unconditional probabilities are special cases of conditional probabilities.

These considerations are supported further by our discussion in §3.1 of how according to the leading interpretations probability statements are always at least tacitly relativised — on the frequency interpretations, to a reference class; on the propensity interpretation, to a chance set-up; on the classical and logical interpretation, to a body of evidence; on the subjective interpretation, to a subject (who has certain background knowledge) at a time, and who may defer to some 'expert' (a person, a future self, a relative frequency, a chance).

Putting these facts together, we have a case for regarding conditional probability as conceptually prior to unconditional probability. So I suggest that we reverse the traditional direction of analysis: regard conditional probability to be the primitive notion, and unconditional probability as the derivative notion. But I also recommend Kenny Easwaran's contribution to this volume ("The Varieties

of Conditional Probability”) for a different perspective.

7 CONDITIONAL PROBABILITIES AND UPDATING RULES

7.1 *Conditionalization*

Suppose that your degrees of belief are initially represented by a probability function $P_{initial}(-)$, and that you become certain of E (where E is the strongest such proposition). What should be your new probability function P_{new} ? The favoured updating rule among Bayesians is *conditionalization*; P_{new} is related to $P_{initial}$ as follows:

$$(\text{Conditionalization}) \quad P_{new}(X) = P_{initial}(X|E) \quad (\text{provided } P_{initial}(E) > 0)$$

Conditionalization is supported by some arguments similar to those that supported the ratio analysis. Firstly, there is case-by-case evidence. Upon receiving the information that the die landed odd, intuition seems to judge that your probability that it landed 5 should be revised to $1/3$, just as conditionalization would have it. Similarly for countless other judgments. Secondly, the muddy Venn diagram can now be given a dynamic interpretation: learning that E corresponds to scraping all mud off $\neg E$. What to do with the mud that remains? It obviously must be rescaled, since it amounts to a total of only $P_{initial}(E)$, whereas probabilities must sum to 1. Moreover, since nothing stronger than E was learned, any movements of mud within E seem gratuitous, or even downright unjustified. So our desired updating rule should preserve the profile of mud within E but renormalize it by a factor of $1/P_{initial}(E)$; this is conditionalization.

Thirdly, conditionalization is supported by a ‘diachronic’ Dutch Book argument (see [Lewis, 1999]): on the assumption that your updating is rule-governed, you are subject to a Dutch book (with bets placed at different times) if you do not conditionalize. Equally important is the converse theorem [Skyrms, 1987]: if you do conditionalize, then you are immune to such a Dutch Book.

Then there are arguments for conditionalization for which there are currently no analogous arguments for (RATIO) — although I suggest that it would be fruitful to pursue such arguments. For example, Greaves and Wallace [2006] offer a “cognitive decision theory”, arguing that conditionalization is the unique updating rule that maximizes expected *epistemic* utility.

However, there are also some sources of suspicion and even downright dissatisfaction about conditionalization. There are apparently some kinds of belief revisions that should *not* be so modelled. Those involving *indexical* beliefs are a prime example. I am currently certain that my computer’s clock reads 8:33; and yet by the time I reach the end of this sentence, I find that I am certain that it does *not* read 8:33. Probability mud isn’t so much scraped away as *pushed sideways* in such cases. Levi [1980] insists that conditionalization is also not appropriate in cases where an agent “contracts” her “corpus” of beliefs — when her stock of settled assumptions is somehow challenged, forcing her to reduce it. See [Hild, 1998;

Bacchus *et al.*, 1990; Arntzenius, 2003] for further objections to conditionalization.

Much as the considerations supporting conditionalization are similar to those supporting the ratio analysis, the considerations *counter*-supporting the latter *counter*-support the former. In particular, the objections that I raised in §4 would seem to have force equally against the adequacy of conditionalization. Recall the problem of conditions with probability zero. A point has just been chosen at random from the surface of the earth, and you learn that it lies on the equator. Conditionalization cannot model this revision in your belief state, since you previously gave probability zero to what you learned. (The same would be true of any line of latitude on which you might learn the point to be.) Similarly for your learning that Democrats won the U.S. election; similarly for your learning that a large-numbered ticket was picked in the 1,000,000-ticket lottery; similarly for your learning that I tossed the fair coin; similarly for your learning that a randomly chosen point came from the non-measurable set N .

To be sure, the key idea behind conditionalization can be upheld while disavowing the ratio analysis for conditional probability. Upon receiving evidence E , one's new probability for X should be one's initial conditional probability for X , given E — this is neutral regarding how the conditional probability should be understood. My point is that the standard formulation of conditionalization, stated above, is *not* neutral: it presupposes the ratio analysis of conditional probability and inherits its problems. (Recall that $P(-| -)$ is shorthand for a ratio of unconditional probabilities.)

Popper functions allow a natural reformulation of updating by conditionalization, so that even items of evidence that were originally assigned such problematic unconditional probabilities by an agent can be learned. The result of conditionalizing a Popper function $P(-, -)$ on a piece of evidence encapsulated by e is $P(-, -e)$ — for example, $P(a, b)$ gets transformed to $P(a, be)$.

7.2 Jeffrey conditionalization

Jeffrey conditionalization allows for less decisive learning experiences in which your probabilities across a partition $\{E_1, E_2, \dots\}$ change to $\{P_{new}(E_1), P_{new}(E_2), \dots\}$, where none of these values need be 0 or 1:

$$P_{new}(X) = \sum_i P_{initial}(X|E_i)P_{new}(E_i)$$

[Jeffrey, 1965; 1983; 1990]. Notice that if we replace $P_{initial}(X|E_i)$ by $P_{new}(X|E_i)$, we simply have an instance of the law of total probability. This theorem of the probability calculus becomes a norm of belief revision, assuming that probabilities conditional on each cell of the partition should stay ‘rigid’, unchanged throughout such an experience. Diaconis and Zabell [1982] show, by reasonable criteria for determining a metric on the space of probability functions, that this rule corresponds to updating to the nearest function in that space, subject to the constraints. One might interpret this as capturing a kind of epistemic conservatism in the spirit

of a Quinean “minimal mutilation” principle: staying as ‘close’ to your original opinions as you can, while respecting your evidence.

Jeffrey conditionalization is again supported by a diachronic Dutch book argument [Armendt, 1980]. It should be noted, however, that diachronic Dutch Book arguments have found less favour than their synchronic counterparts. Levi [1991] and Maher [1992] insist that the agent who fails to conditionalize and who thereby appears to be susceptible to a Dutch Book will be able to ‘see it coming’, and thus avoid it; however, see also Skyrms’ [1993] rebuttal. Christensen [1991] denies that the alleged ‘inconsistency’ dramatized in such arguments has any normative force in the diachronic setting. van Fraassen [1989] denies that rationality requires one to follow a rule in the first place. Levi [1967] also criticizes Jeffrey conditionalization directly. For example, repeated operations of the rule may not commute, resulting in a path-dependence of one’s final epistemic state that might be found objectionable. However, Lange [2000] argues that this non-commutativity is a virtue rather than a vice.

8 SOME PARADOXES AND PUZZLES INVOLVING CONDITIONAL PROBABILITY AND CONDITIONALIZATION

8.1 *The Monty Hall problem*

Let’s begin with a problem that is surely *not* a paradox, even though it is often called that. You are on the game show *Let’s Make a Deal* hosted by Monty Hall. Before you are three doors; behind exactly one of them is a prize, which you will win if you choose its door correctly. First, you are to nominate a door. Monty, who knows where the prize is and will not reveal it, ostentatiously opens *another* door, revealing it to be empty. He then gives you the opportunity to switch to the remaining door. Should you do so? Many people intuit that it doesn’t matter either way: you’re as likely to win the prize by sticking with your original door as you are by switching. That’s wrong — indeed, you are twice as likely to win by switching than by sticking with your original door. An easy way to see this is to consider the probability of *failing* to win by switching. The only way you could fail would be if you had initially nominated the correct door — probability $1/3$ — and then, unluckily, switched away from it when given the chance. Thus, the probability of winning by switching is $2/3$.

The reasoning just given is surely too simple to count as paradoxical. But the problem does teach a salutary lesson regarding the importance on conditionalizing on one’s *total* evidence. The fallacious reasoning would have you conditionalize on the evidence that the prize is not behind the door that Monty actually opens (e.g. door 1) — that is, to assign a probability $1/2$ to each of the two remaining doors (e.g. doors 2 and 3). But your actual evidence was stronger than that: you also learned that Monty *opened the door that he did*. (If you initially chose the correct door, he had a genuine choice.) A relatively simple calculation shows that conditionalizing on your total evidence yields the correct answer: your updated

probability that the remaining door contains the prize is $2/3$, so you should switch to it.

8.2 *Simpson's paradox*

Again, it is questionable whether an observation due to Simpson deserves to be called a “paradox”; rather, it is a fairly straightforward fact about inequalities among conditional probabilities. But the observation is undoubtedly rather counterintuitive, and it has some significant ramifications for scientific inference.

The paradox was once famously instantiated by the U.C. Berkeley's admission statistics. Taken as a whole, admissions seemed to favour males, as suggested by the correlations inferred from the relative frequencies of admission of males and females:

$$P(\text{admission} \mid \text{male}) > P(\text{admission} \mid \text{female}).$$

Yet disaggregating the applications department by department, the correlations went the other way:

$$P(\text{admission} \mid \text{male \& department 1 applicant}) < P(\text{admission} \mid \text{female \& department 1 applicant})$$

$$P(\text{admission} \mid \text{male \& department 2 applicant}) < P(\text{admission} \mid \text{female \& department 2 applicant}),$$

and so on for every department.

How could this be? A simple explanation was that the females tended to apply to more competitive departments with lower admission rates. This lowered their university-wide admission rate compared to males, even though department by department their admission rate was superior.

More generally, Simpson's paradox is the phenomenon that correlations that appear at one level of partitioning may disappear or even reverse at another level of partitioning:

$$P(E|C) > P(E \mid \sim C) \text{ is consistent with}$$

$$P(E|C \& F_1) < P(E \mid \sim C \& F_1),$$

$$P(E|C \& F_2) < P(E \mid \sim C \& F_2),$$

...

$$P(E|C \& F_n) < P(E \mid \sim C \& F_n),$$

for some partition $\{F_1, F_2, \dots, F_n\}$.

Pearl [2000] argues that such a pattern of inequalities only seems paradoxical if we impose a causal interpretation on them. In our example, being male is presumably regarded as a (probabilistic) *cause* of being admitted, perhaps due to discrimination in favour of men and against women. We seem to be reasoning: “Surely *unanimity* in the departmental causal facts has to be preserved by the

university at large!” Pearl believes that if we rid ourselves of faulty intuitions about correlations revealing causal relations, the seeming paradoxicality will vanish.

I demur. I think that we are just as liable to recoil even if the data is presented as inequalities among ratios, with no causal interpretation whatsoever. Department by department, the *ratio* of admitted women is greater than the *ratio* of admitted men, yet university-wide the inequality among the *ratios* goes the other way. How could this be? “Surely *unanimity* in the departmental ratio inequalities has to be preserved by the university at large!” Not at all, as simple arithmetic proves. We simply have faulty arithmetical intuitions.

8.3 The Judy Benjamin problem

The general problem for probability kinematics is: given a prior probability function P , and the imposition of some constraint on the posterior probability function, what should this posterior be? This problem apparently has a unique solution for certain constraints, as we have seen — for example:

1. Assign probability 1 to some proposition E , while preserving the relative odds of all propositions that imply E . Solution: conditionalize P on E .
2. Assign probabilities $p_1, \dots, p_n$ to the cells of the partition $\{E_1, \dots, E_n\}$, while preserving the relative odds of all propositions within each cell. Solution: Jeffrey conditionalize P on this partition, according to the specification.

But consider the constraint:

3. Assign conditional probability p to B , given A .

The *Judy Benjamin problem* is that of finding a rule for transforming a prior, subject to this third constraint [van Fraassen, 1989].

van Fraassen provides arguments for three distinct such rules, and surmises that this raises the possibility that such uniqueness results “will not extend to more broadly applicable rules in general probability kinematics. In that case rationality will not dictate epistemic procedure even when we decide that it shall be rule governed” [1989, p. 343].

8.4 Non-conglomerability

Call P *conglomerable in the partition* $X = \{x_1, x_2, \dots\}$ if

$$k_1 \leq P(Y) \leq k_2 \text{ whenever } k_1 \leq P(Y|X = x_i) \leq k_2 \text{ for all } i = 1, 2, \dots$$

Here’s the intuitive idea. Suppose that you know now that you will learn which member of a particular partition is true. (A non-trivial partition might have as few as two members, such as {Heads} and {Tails}, or it might have countably many members.) Suppose further that you know now that *whatever* you learn, your probability for Q will lie in a certain interval. Then it seems that you should

now assign a probability for Q that lies in that interval. If you *know* that you are going to have a certain opinion in the future, why wait? — Make it your opinion *now!* More generally, if you *know* that a credence of yours will be bounded in a particular way in the future, why wait? — Bound that credence in this way *now!* ‘Conglomerability in a partition’ captures this desideratum.

Failures of conglomerability arise when P is finitely additive, but not countably additive. As Seidenfeld *et al.* [1998] show, in that case there exists some countable partition in which P is not conglomerable. If updating takes place by conditionalization, failures of conglomerability lead to curious commitments reminiscent of violations of the Reflection Principle: “My future self, who is ideally rational and better informed than I am, will definitely have a credence for Q in a particular interval, but my credence for Q is not in this interval.” (See [Jaynes, 2003, Ch. 15] for a critique of Seidenfeld *et al.* See also Kadane *et al.* [1986] for a non-conglomerability result even assuming countable additivity, in uncountable partitions.)

8.5 The two-envelope paradoxes

As an example of non-conglomerability, consider the following infinite version of the so-called ‘two envelope’ paradox: Two positive integers are selected at random and turned into dollar amounts, the first placed in one envelope, the second placed in another, whose contents are concealed from you. You get to choose one envelope, and its contents are yours. Suppose that following de Finetti [1972], and in violation of countable additivity, you assign probability 0 to all finite sets of positive integers, but (of course), probability 1 to the entire set of positive integers. Let X be the amount in your envelope and Y the amount in the other envelope. Then very reasonably you assign:

$$P(X < Y) = 1/2.$$

But suppose now that we let you open your envelope. You may see \$1, or \$2, or \$3, or ... Yet whatever you see, you will want to switch to holding the other envelope, for

$$P(X < Y | X = x) = 1 \text{ for } x = 1, 2, \dots$$

Why wait? Since you know that you will want to switch, you should switch now. That is absurd: you surely cannot settle from the armchair that you have made the wrong choice, however you choose.

A better-known version of the two-envelope paradox runs as follows. One positive integer is selected and that number of dollars is placed in an envelope. Twice as much is placed in another envelope. The contents of both envelopes are concealed from you. You get to choose one envelope, and its contents are yours. At first you think that you have no reason to prefer one envelope over another, so you choose one. But as soon as you do, you feel regret. You reason as follows: “I am holding some dollar amount — call it n . The other envelope contains either

$2n$ or $n/2$, each with probability $1/2$. So its expectation is $(2n)^{1/2} + (n/2)^{1/2} = 5n/4 > n$. So it is preferable to my envelope.” This is already absurd, as before. Worse, if we let you switch, your regret will immediately run the other way: “I am holding some dollar amount – call it $m \dots$ ” And similar reasoning seems to go through even if we let you open your envelope to check its contents!

Let X be the random variable ‘the amount in your envelope’, and let Y be ‘the amount in the other envelope’. Notice that a key step of the reasoning moves from

$$\text{for any } n, E(Y|X = n) > n = E(X|X = n) \quad (*)$$

to the conclusion that the other envelope is preferable. A missing premise is that

$$E(Y) > E(X).$$

This may seem to follow straightforwardly from $(*)$. But that presupposes conglomerability with respect to the partition of amounts in your envelope, which is exactly what should be questioned. See [Arntzenius and McCarthy, 1997; Chalmers, 2002] for further discussion.

9 PROBABILITIES OF CONDITIONALS AND CONDITIONAL PROBABILITIES

A number of authors have proposed that there are deep connections between conditional probabilities and conditionals. Ordinary English seems to allow us to shift effortlessly between the two kinds of locutions. ‘The probability of it raining, given that it is cloudy, is high’ seems to say the same thing as ‘the probability of it raining, if it is cloudy, is high’ — the former a conditional probability, the latter the probability of a conditional.

The Ramsey test and Adams’ thesis

Ramsey [1931/1990, p. 155] apparently generalized this observation in a pregnant remark in a footnote: “If two people are arguing ‘If p will q ?’ and are both in doubt as to p , they are adding p hypothetically to their stock of knowledge and arguing on that basis about q ; ... We can say they are fixing their degrees of belief in q given p .” Adams [1975] more explicitly generalized the observation in his celebrated thesis that the probability of the indicative conditional ‘if A , then B ’ is given by the corresponding conditional probability of B given A . He denied that such conditionals have truth conditions, so this probability is not to be thought of as the probability that ‘if A , then B ’ is true. Further, Adams’ ‘probabilities’ of conditionals do not conform to the usual probability calculus — in particular, Boolean compounds involving them do not receive ‘probabilities’, as the usual closure assumptions (given in §2.1) would require.

For this reason, Lewis [1976] suggests that they be called “assertabilities” instead, a practice that has been widely adopted subsequently. Note, however, that

“assertability” seems to bring in the norms of *assertion*. For example, Williamson [2002] argues that you should only assert what you *know*; but then it is hard to make sense of assertability coming in all the degrees that Adams requires of it. And conditionals can be unassertable for all sorts of reasons that seem beside the point here — they can be inappropriate, irrelevant, uninformative, undiplomatic, and so on. This is a matter of the pragmatics of conversation, which is another topic. Perhaps the locution “degrees of acceptability” better captures Adams’ idea.

Stalnaker’s Hypothesis

Stalnaker [1970], by contrast, insisted that conditionals have truth conditions, and he and Lewis were engaged in the late 60s and early 70s in a famous debate over what they were. In particular, they differed over the status of *conditional excluded middle* — on whether sentences of the following form are tautologies or not:

$$(CEM) \quad (A \rightarrow B) \vee (A \rightarrow \neg B)$$

Stalnaker thought so; Lewis thought not. Stalnaker upheld the equality of genuine *probabilities* of conditionals with the corresponding conditional probabilities, and used the attractiveness of this thesis as an argument for his preferred semantics. More precisely, the hypothesis is that some suitably quantified and qualified version of the following equation holds:

$$(PCCP) \quad P(A \rightarrow B) = P(B|A) \text{ for all } A, B \text{ in the domain of } P, \text{ with } P(A) > 0.$$

(“ $\rightarrow$ ” is a conditional connective.)

Stalnaker’s guiding idea was that a suitable version of the hypothesis would serve as a criterion of adequacy for a truth-conditional account of the conditional. He explored the conditions under which it would be reasonable for a rational agent, with subjective probability function P , to believe a conditional $A \rightarrow B$. By identifying the probability of $A \rightarrow B$ with $P(B|A)$, Stalnaker was able to put constraints on the truth conditions of the ‘ $\rightarrow$ ’. In particular, if this identification were sound, it would vindicate conditional excluded middle. For by the probability calculus,

$$\begin{aligned} & P[(A \rightarrow B) \vee (A \rightarrow \neg B)] \\ = & P(A \rightarrow B) + P(A \rightarrow \neg B) \\ & \text{(assuming that the disjuncts are incompatible, as both authors did)} \\ = & P(B|A) + P(\neg B|A) \\ & \text{(by the identification of probabilities of conditionals with conditional probabilities)} \\ = & 1. \end{aligned}$$

So all sentences of the CEM form have probability 1, as Stalnaker required.

Some of the probabilities-of-conditionals literature is rather unclear on exactly what claims are under discussion: what the relevant quantifiers are, and their

domains of quantification. With the above motivations kept in mind, and for their independent interest, we now consider four salient ways of rendering precise the hypothesis that probabilities of conditionals are conditional probabilities:

Universal version: There is some $\rightarrow$ such that for all P , (PCCP) holds.

Rational Probability Function version: There is some $\rightarrow$ such that for all P that could represent a rational agent's system of beliefs, (PCCP) holds.

Universal Tailoring version: For each P there is some $\rightarrow$ such that (PCCP) holds.

Rational Probability Function tailoring version: For each P that could represent a rational agent's system of beliefs, there is some $\rightarrow$ such that (PCCP) holds.'indexorthogonal

Can any of these versions be sustained? The situation is interesting however we answer this question. If the answer is 'no', then seemingly synonymous locutions are not in fact synonymous: surprisingly, 'the probability of B , given A ' does not mean the same thing as 'the probability of: B if A '. If the answer is 'yes', then important links between logic and probability theory will have been established, just as Stalnaker and Adams hoped. Probability theory would be a source of insight into the formal structure of conditionals. And probability theory in turn would be enriched, since we could characterize more fully what the usual conditional probability ratio means, and what its use is. de Finetti [1972] laments that (RATIO) gives the formula, but not the *meaning*, of conditional probability. A suitably quantified hypothesis involving (PCCP) could serve to characterize more fully what the ratio means, and what its use is.

There is now a host of results — mostly negative — concerning PCCP. We will give a sample of some of the most important ones. We will then be in a position to assess how the four versions of the hypothesis fare, and what the prospects are for other versions. Some preliminary definitions will assist in stating the results.

If (PCCP) holds, we will say that $\rightarrow$ is a *PCCP-conditional* for P , and that P is a *PCCP-function* for $\rightarrow$. If (PCCP) holds for a particular $\rightarrow$ for each member P of a class of probability functions $\mathcal{P}$, we will say that $\rightarrow$ is a *PCCP-conditional* for $\mathcal{P}$. A pair of probability functions P and P' are *orthogonal* if, for some A , $P(A) = 1$ but $P'(A) = 0$. (Intuitively, orthogonal probability functions concentrate their probability on entirely non-intersecting sets of propositions.) Call a proposition A a *P -atom* iff $P(A) > 0$ and, for all X , either $P(AX) = P(A)$ or $P(AX) = 0$. (Intuitively, a P -atom is a proposition that receives an indivisible 'blob' of probability from P .) Finally, we will call a probability function *trivial* if it has at most 4 different values.

Most of the negative results are 'triviality results': given certain assumptions, only trivial probability functions can sustain PCCP. Moreover, most of them make no assumptions about the logic of the ' $\rightarrow$ ' — it is simply a two-place connective. The earliest and most famous results are due to Lewis [1976]:

First triviality result: There is no PCCP-conditional for the class of all probability functions.

Second triviality result: There is no PCCP-conditional for any class of probability functions closed under conditionalizing, unless the class consists entirely of trivial functions.

Lewis [1986] strengthens these results:

Third triviality result: There is no PCCP-conditional for any class of probability functions closed under conditionalizing restricted to the propositions in a single finite partition, unless the class consists entirely of trivial functions.

Fourth triviality result: There is no PCCP-conditional for any class of probability functions closed under Jeffrey conditionalizing, unless the class consists entirely of trivial functions.

These results refute the Universal version of the hypothesis. They also spell bad news for the Rational Probability Function version, for even if rationality does not require updating by conditionalizing, or Jeffrey conditionalizing, it seems plausible that it at least *permits* such updating. This version receives its death blow from the following result by Hall [1994], that significantly strengthens Lewis' results:

Orthogonality result: Any two non-trivial PCCP-functions defined on the same algebra of propositions are orthogonal.

It follows from this that the Rational Probability Function version is true only if any two distinct rational agents' probability functions are orthogonal — which is absurd.

So far, the 'tailoring' versions remain unscathed. The Universal Tailoring version is refuted by the following result due to Hájek [1989; 1993], which concerns probability functions that assume only a finite number of distinct values:

Finite-ranged Functions Result: Any non-trivial probability function with finite range has no PCCP-conditional.

This result also severely casts doubt on the Rational Probability Tailoring version, for it is hard to see why rationality requires one to adopt a probability function with infinite range. The key idea behind this result can be understood by considering a very simple case. Consider a three-ticket lottery, and let L_i = 'ticket i wins', $i = 1, 2, 3$. Let P assign probability $1/3$ to each of the L_i . Clearly, some conditional probabilities take the value $1/2$ — for example, $P(L_1|L_1 \vee L_2)$. But no unconditional probability can take this value, being constrained to be a multiple of $1/3$; *a fortiori*, no (unconditional) probability of a conditional can take this value. The point generalizes to all finite-ranged probability functions: there will always be some value of the conditional probability function that finds no match among the unconditional probabilities, and *a fortiori* no match among the (unconditional) probabilities of conditionals. Picture a dance at which, for a given finite-ranged probability function, all of the probability-of-a-conditional values line up along one wall, and all of the conditional probability values line up along the opposite wall.

Now picture each conditional probability value attempting to partner up with a probability of a conditional with the same value on the other side. According to Stalnaker's hypothesis, the dance would always be a complete success, with all the values finding their matches; the Finite-ranged Functions Result shows that the dance can never be a complete success. There will always be at least one wallflower among the conditional probabilities, which will have to sit out the dance — for example, $1/2$ in our lottery case.

If we make a minimal assumption about the logic of the $\rightarrow$, matters are still worse thanks to another result of Hall's [1994]:

No Atoms Result: Let the probability space $\langle \Omega, \mathcal{F}, P \rangle$ be given, and suppose that PCCP holds for this P , and a ' $\rightarrow$ ' that obeys modus ponens. Then $\langle \Omega, \mathcal{F}, P \rangle$ does not contain a P -atom, unless P is trivial.

It follows from this, on pain of triviality, that the range of P , and hence Ω and $\mathcal{F}$, are non-denumerable. All the more, it is hard to see how rationality requires this of an agent's probability space.

It seems, then, that all four versions of the hypothesis so far considered are untenable. (See also [Hájek, 1994] for more negative results.) For all that has been said so far, though, some suitably restricted 'tailoring' version might still survive. A natural question, then, is whether even Hall's 'no atoms' result can be extended — whether even uncountable probability spaces cannot support PCCP, thus effectively refuting any 'tailoring' version of the hypothesis. The answer is 'no' — and here we have a positive result due to van Fraassen [1976]. Suppose that $\rightarrow$ has this much logical structure:

- (i) $[(A \rightarrow B) \cap (A \rightarrow C)] = [A \rightarrow (B \cap C)]$
- (ii) $[(A \rightarrow B) \cup (A \rightarrow C)] = [A \rightarrow (B \cup C)]$
- (iii) $[A \cap (A \rightarrow B)] = (A \cap B)$
- (iv) $[A \rightarrow A] = \Omega$.

Such an $\rightarrow$ conforms to the logic *CE*. van Fraassen shows:

CE tenability result: Any probability space can be extended to one for which PCCP holds, with an $\rightarrow$ that conforms to CE.

Of course, the larger space for which PCCP holds is uncountable. In the same paper, van Fraassen also shows that $\rightarrow$ can have still more logical structure, while supporting PCCP, provided we restrict the admissible iterations of $\rightarrow$ appropriately.

A similar strategy of restriction protects Adams' version of the hypothesis from the negative results. He applies PCCP to unembedded conditionals — 'simple' conditionals of the form $A \rightarrow B$, where A and B are themselves conditional-free. As mentioned before, Adams does not allow the assignment of probabilities to Boolean compounds involving conditionals; ' P ' is thus not strictly speaking

a probability function (and thus the negative results, which presuppose that it is, do not apply). McGee [1989] shows how Adams' theory can be extended to certain more complicated compounds of conditionals, while still falling short of full closure.

10 CONCLUSION

This survey has been long, and yet still I fear that some readers will be disappointed that I have not discussed adequately, or at all, their favourite application of or philosophical issue about conditional probability. They may find some solace in the lengthy bibliography that follows.

Along the way we have seen some reasons for questioning the orthodoxy enshrined in Kolmogorov's ratio analysis; moreover his more sophisticated formulation seems not entirely successful either. I have argued that we should take conditional probability as the primitive notion in probability theory, although this still remains a minority position. However we resolve this issue, we have something of a mathematical and philosophical balancing act: finding an entirely satisfactory mathematical and philosophical theory of conditional probability that as much as possible does justice to our intuitions and to its various applications. It is an act worth getting right: the foundations of probability theory depend on it, and thus any theory that employs probability theory depends on it also — which is to say, any serious empirical discipline, and much of philosophy.⁴

BIBLIOGRAPHY

- [Adams, 1975] E. Adams. *The Logic of Conditionals*, Reidel, 1975.
- [Armendt, 1980] B. Armendt. Is There a Dutch Book Argument for Probability Kinematics? *Philosophy of Science* 47, No. 4 (December), 583-88, 1980.
- [Arntzenius, 2003] F. Arntzenius. Some Problems for Conditionalization and Reflection, *Journal of Philosophy* Vol. C, No 7, 356-371, 2003.
- [Arntzenius and McCarthy, 1997] F. Arntzenius and D. McCarthy. The Two Envelope Paradox and Infinite Expectations, *Analysis* 57, No. 1, 28-34, 1997.
- [Bacchus *et al.*, 1990] F. Bacchus, H. E. Kyburg Jr., and M. Thalos. Against Conditionalization, *Synthese* 85, 475-506, 1990.
- [Billingsley, 1995] P. Billingsley. *Probability and Measure*, John Wiley and Sons, Third Edition, 1995.
- [Blackwell and Dubins, 1975] D. Blackwell and L. Dubins. On Existence and Non-existence of Proper, Regular, Conditional Distributions, *The Annals of Probability* v3 i5, 741-752, 1975.
- [Carnap, 1950] R. Carnap. *Logical Foundations of Probability*, Chicago: University of Chicago Press, 1950.
- [Carnap, 1952] R. Carnap. *The Continuum of Inductive Methods*, Chicago: The University of Chicago Press, 1952.
- [Cartwright, 1979] N. Cartwright. Causal Laws and Effective Strategies, *Noûs* 13, 419-437, 1979.

⁴I am grateful to Elle Benjamin, Darren Bradley, Lina Eriksson, Marcus Hutter, Aidan Lyon, John Matthewson, Ralph Miles, Nico Silins, Michael Smithson, Weng Hong Tang, Peter Vanderschraaf, Wen Xuefeng, and especially John Cusbert, Kenny Easwaran and Michael Titelbaum for helpful suggestions.

- [Chalmers, 2002] D. Chalmers. The St. Petersburg Two-Envelope Paradox, *Analysis* 62, 155-57, 2002.
- [Christensen, 1991] D. Christensen. Clever Bookies and Coherent Beliefs, *The Philosophical Review* C, No. 2, 229-247, 1991.
- [de Finetti, 1937] B. de Finetti. La Prévision: Ses Lois Logiques, Ses Sources Subjectives, *Annales de l'Institut Henri Poincaré*, 7: 1-68, 1937. Translated as Foresight. Its Logical Laws, Its Subjective Sources, in *Studies in Subjective Probability*, H. E. Kyburg, Jr. and H. E. Smokler (eds.), Robert E. Krieger Publishing Company, 1980.
- [de Finetti, 1972] B. de Finetti. *Probability, Induction and Statistics*, Wiley, 1972.
- [de Finetti, 1974/1990] B. de Finetti. *Theory of Probability*, Vol. 1. Chichester: Wiley Classics Library, John Wiley & Sons, 1974/1990.
- [Diaconis and Zabell, 1982] P. Diaconis and S. L. Zabell. Updating Subjective Probability, *Journal of the American Statistical Association* 77, 822-30, 1982.
- [Dowe et al., 2007] D. Dowe, S. Gardner, and G. Oppy. Bayes not Bust! Why Simplicity is No Problem for Bayesians, *British Journal for Philosophy of Science* 58, 4, 709-54, 2007.
- [Eells, 1991] E. Eells. *Probabilistic Causality*, Cambridge University Press, Cambridge, 1991.
- [Eells and Skyrms, 1994] E. Eells and B. Skyrms, eds. *Probability and Conditionals*, Cambridge University Press, 1994.
- [Forster and Sober, 1994] M. Forster and E. Sober. How to Tell when Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions, *British Journal for the Philosophy of Science* 45, 1-35, 1994.
- [Gaifman, 1988] H. Gaifman. A Theory of Higher Order Probabilities, in *Causation, Chance, and Credence*, Brian Skyrms and William L. Harper, eds., Dordrecht: Kluwer Academic Publishers, 1988.
- [Giere, 1973] R. Giere. Objective Single-Case Probabilities and the Foundations of Statistics, in *Logic, Methodology and Philosophy of Science IV – Proceedings of the Fourth International Congress for Logic*, P. Suppes, L. Henkin, A. Joja, and G. Moisil (eds.), New York: North-Holland, 1973.
- [Goldstein, 1983] M. Goldstein. The Prevision of a Prevision, *Journal of the American Statistical Association* 78, 817-819, 1983.
- [Greaves and Wallace, 2006] H. Greaves and D. Wallace. Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility, *Mind* 115 (459), 607-632, 2006.
- [Grunwald et al., 2005] P. Grunwald, M. A. Pitt, and I. J. Myung, eds. *Advances in Minimum Description Length: Theory and Applications*, Cambridge, MA: MIT Press, 2005.
- [Hacking, 1965] I. Hacking. *The Logic of Statistical Inference*, Cambridge: Cambridge University Press, 1965.
- [Hájek, 1989] A. Hájek. Probabilities of Conditionals — Revisited, *Journal of Philosophical Logic* 18, No. 4, 423-428, 1989.
- [Hájek, 1993] A. Hájek. *The Conditional Construal of Conditional Probability*, Ph.D. Dissertation, Princeton University. Available at http://fitelson.org/conditionals/hajek_dissertation.pdf
- [Hájek, 1994] A. Hájek. Triviality on the Cheap?, in Eells and Skyrms, 113-140, 1994.
- [Hájek, 2003] A. Hájek. What Conditional Probability Could Not Be, *Synthese* 137, No. 3, (December), 273-323, 2003.
- [Hall, 1994] N. Hall. Back in the CCCP, in Eells and Skyrms, 141-60, 1994.
- [Hausman and Woodward, 1999] D. Hausman and J. F. Woodward. Independence, Invariance and the Causal Markov Condition, *The British Journal for the Philosophy of Science* 50: 44, 521-583, 1999.
- [Hempel, 1965] C. Hempel. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press, 1965.
- [Hild, 1998] M. Hild. The Coherence Argument Against Conditionalization, *Synthese* 115, 229-258, 1998.
- [Jaynes, 2003] E. T. Jaynes. *Probability Theory: The Logic of Science*, Cambridge: Cambridge University Press, 2003.
- [Jeffrey, 1965/1983/1990] R. C. Jeffrey. *The Logic of Decision*, Chicago: Chicago University Press, 1965/1983/1990.
- [Jeffreys, 1961] H. Jeffreys. *Theory of Probability*, Oxford: Oxford University Press (originally published in 1939, and now in the Oxford Classic Texts in the Physical Sciences series), 1961.
- [Johnson, 1921] W. E. Johnson. *Logic*, Cambridge: Cambridge University Press, 1921.

- [Keynes, 1921] J. M. Keynes. *A Treatise on Probability*, London: Macmillan, 1921.
- [Kadane *et al.*, 1986] J. B. Kadane, M. J. Schervish and T. Seidenfeld. Statistical Implications of Finitely Additive Probability, in *Bayesian Inference and Decision Techniques*, Prem K. Goel and Arnold Zellner (eds.), Amsterdam: Elsevier Science Publishers, 59-76, 1986.
- [Kolmogorov, 1933/1950] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitrechnung, Ergebnisse Der Mathematik*, 1933. Translated as *Foundations of Probability*. New York: Chelsea Publishing Company, 1950.
- [Lange, 2000] M. Lange. Is Jeffrey Conditionalization Defective by Virtue of Being Non-Commutative? Remarks on the Sameness of Sensory Experiences, *Synthese* 123, 393-403, 2000.
- [Levi, 1967] I. Levi. Probability Kinematics, *British Journal for the Philosophy of Science* 18, 197-209, 1967.
- [Levi, 1974] I. Levi. On Indeterminate Probabilities, *Journal of Philosophy* 71, 391-418, 1974.
- [Levi, 1980] I. Levi. *The Enterprise of Knowledge*, Cambridge: MIT Press, 1980.
- [Levi, 1991] I. Levi. *Consequentialism and Sequential Choice*, in Michael Bacharach and Susan Hurley (eds.), *Foundations of Decision Theory*, Oxford: Basil Blackwell, 92-146, 1991.
- [Levi, 2000] I. Levi. Imprecise and Indeterminate Probabilities, *Risk, Decision and Policy* 5, 111-122, 2000.
- [Lewis, 1973] D. Lewis. *Counterfactuals*, Blackwell and Harvard University Press, 1973.
- [Lewis, 1976] D. Lewis. Probabilities of Conditionals and Conditional Probabilities, *Philosophical Review* 85, 297-315, 1976.
- [Lewis, 1980] D. Lewis. A Subjectivist's Guide to Objective Chance, in *Studies in Inductive Logic and Probability*, Vol II., University of California Press, 263-293, 1980; reprinted in *Philosophical Papers Volume II*, 1986, Oxford: Oxford University Press.
- [Lewis, 1986] D. Lewis. Probabilities of Conditionals and Conditional Probabilities II, *Philosophical Review* 95, 581-589, 1986.
- [Lewis, 1999] D. Lewis. *Papers in Metaphysics and Epistemology*, Cambridge: Cambridge University Press, 1999.
- [Maher, 1992] P. Maher. Diachronic Rationality, *Philosophy of Science* 59, 120-141, 1992.
- [Maher, 2007] P. Maher. Explication Defended, *Studia Logica* 86, 331-341, 2007.
- [McGee, 1989] V. McGee. Conditional Probabilities and Compounds of Conditionals, *The Philosophical Review* 98, 485-541 1989.
- [McGee, 1994] V. McGee. Learning the Impossible, in Eells and Skyrms, 179-99, 1994.
- [Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference*, Cambridge: Cambridge University Press, 2000.
- [Popper, 1959a] K. Popper. The Propensity Interpretation of Probability, *British Journal of the Philosophy of Science* 10, 25-42, 1959.
- [Popper, 1959b] K. Popper. *The Logic of Scientific Discovery*, Basic Books, 1959.
- [Ramsey, 1931] F. P. Ramsey. General Propositions and Causality, in R. B. Braithwaite, ed., *The Foundations of Mathematics and Other Logical Essays*, Routledge; also in *Philosophical Papers*, ed. D. H. Mellor, Cambridge: Cambridge University Press, 1931.
- [Reichenbach, 1956] H. Reichenbach. *The Direction of Time*, Berkeley, University of California Press, 1956.
- [Renyi, 1970] A. Renyi. *Foundations of Probability*, Holden-Day, Inc. 1970.
- [Roeper and Leblanc, 1999] P. Roeper and H. Leblanc. *Probability Theory and Probability Logic*, Toronto: University of Toronto Press, 1999.
- [Salmon, 1980] W. Salmon. Probabilistic Causality, *Pacific Philosophical Quarterly* 61, 50-74, 1980.
- [Schwarz, 1978] G. Schwarz. Estimating the dimension of a model, *Annals of Statistics* 6 (2), 461-464, 1978.
- [Seidenfeld *et al.*, 1998] T. Seidenfeld, M. J. Schervish, and J. B. Kadane. "Non-Conglomerability for Finite-Valued, Finitely Additive Probability", *Sankhya Series A*, Vol. 60, No. 3, 476-491. 1998.
- [Seidenfeld *et al.*, 2001] T. Seidenfeld, M. J. Schervish, and J. B. Kadane. Improper Regular Conditional Distributions, *The Annals of Probability* 29, No. 4, 1612-1624, 2001.
- [Shimony, 1955] A. Shimony. Coherence and the Axioms of Confirmation, *Journal of Symbolic Logic* 20, 1-28, 1955.
- [Skyrms, 1980] B. Skyrms. *Causal Necessity*, Yale University Press, 1980.

- [Skyrms, 1987] B. Skyrms. Dynamic Coherence and Probability Kinematics, *Philosophy of Science* 54, No. 1 (March), 1-20, 1987.
- [Skyrms, 1993] B. Skyrms. A Mistake in Dynamic Coherence Arguments? *Philosophy of Science* 60, 320-328, 1993.
- [Spirtes *et al.*, 2000] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*, 2nd ed. New York, N.Y.: MIT Press, 2000.
- [Stalnaker, 1968] R. Stalnaker. A Theory of Conditionals, *Studies in Logical Theory*, American Philosophical Quarterly Monograph Series, No. 2, Oxford: Blackwell, 1968.
- [Stalnaker, 1970] R. Stalnaker. Probability and Conditionals, *Philosophy of Science* 37, 64-80, 1970.
- [Suppes, 1970] P. Suppes. *A Probabilistic Theory of Causality*, Amsterdam: North Holland Publishing Company, 1970.
- [van Fraassen, 1976] B. van Fraassen. Probabilities of Conditionals, in Harper and Hooker (eds.), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, Vol. I, Reidel, 261-301, 1976.
- [van Fraassen, 1984] B. van Fraassen. Belief and the Will, *Journal of Philosophy* 81, 235-256, 1984.
- [van Fraassen, 1989] B. van Fraassen. *Laws and Symmetry*, Oxford: Clarendon Press, 1989.
- [van Fraassen, 1990] B. van Fraassen. Figures in a Probability Landscape, in J. M. Dunn and A. Gupta (eds.), *Truth or Consequences*, Kluwer, 1990.
- [van Fraassen, 1995] B. van Fraassen. Belief and the Problem of Ulysses and the Sirens, *Philosophical Studies* 77, 7-37, 1995.
- [Wallace and Dowe, 1999] C. S. Wallace and D. L. Dowe. Minimum Message Length and Kolmogorov Complexity, *The Computer Journal* 42, No. 4, (special issue on Kolmogorov complexity), 270-283, 1999.
- [Walley, 1991] P. Walley. *Statistical Reasoning with Imprecise Probabilities*, London: Chapman Hall, 1991.
- [Williamson, 2002] T. Williamson. *Knowledge and Its Limits*, Oxford: Oxford University Press, 2002.

This page intentionally left blank

THE VARIETIES OF CONDITIONAL PROBABILITY

Kenny Easwaran

Alan Hájek discusses many interesting features and applications of the notion of conditional probability. In addition to this discussion, he also gives arguments for some specific philosophical views on the nature of conditional probability. In general I agree with most of his points, and in particular with his arguments that Kolmogorov’s “ratio analysis” of conditional probability is not correct. However, he suggests three more points that I would like to contest — that there is a single correct analysis of conditional probability across interpretations, that conditional probability is in general a *more* fundamental notion than unconditional probability, and that Popper’s account of conditional probability is the correct one.. I discuss all of these issues (and especially the third) in my dissertation [Easwaran, 2008], but here I will focus more on the first two. I will argue in section 1 that although the different interpretations of probability have many similarities, this similarity is not exact, even at the purely formal level. In particular, I will argue that the bearers of probability in the propensity, subjective, and logical interpretations of probability are distinct classes of objects, and that this changes the mathematical relations that must hold between conditional and unconditional probabilities. As for Hájek’s argument that conditional probabilities are fundamental, in section 2 I will demonstrate a distinction that I think his argument collapses.

1 PLURALISM ABOUT CONDITIONAL PROBABILITY

The fact that a single word, “probability”, is applied to many diverse phenomena, and the existence of a branch of mathematics called “probability theory” suggests that these diverse phenomena can all be understood as applications of the same formal system. However, there are many different formal systems that all claim to describe probability — for instance, both [Kolmogorov, 1950] and [Rényi, 1970] take the objects of the probability function to be sets, although only the latter allows for conditioning on events of probability zero; the system described in [Hailperin, 1984; Hailperin, 1997; Hailperin, 2000; Hailperin, 2006] and that of [Popper, 1959] both allow for the objects not to be sets, but only the latter takes conditional probability as fundamental. Because of this variety of available systems, it seems that an argument is needed to show that different interpretations of probability should use the same one.¹

¹There are some terminological issues here — some suggest that what it *means* to be an interpretation of probability requires satisfying some specific set of mathematical axioms. [Eells,

At first it might seem that principles like the “Principal Principle” [Lewis, 1980] give the required link to show that the different interpretations should use the same formalism. However, such arguments are not definitive. A standard, simplified version of the Principal Principle states that $CR(A|CH(A) = x) = x$, where CR is an agent’s credence function, and CH is the objective chance function. As stated, this principle requires that the objects of the two functions are the same sort of thing (since A is an object of both). It also gives conditions (when the agent knows all the chances) under which the two functions are apparently identical, so there should be no general axioms for subjective probability that rule out any function that satisfies the axioms for objective chance. However, phrasing the principle more carefully, we will be able to see that these two points are not actually correct.

First, Paul Humphreys has argued that the objects of the chance function are events, rather than sentences or sets of worlds. [Humphreys, 2004, p. 669] Even if this is right, a version of the Principal Principle can be stated that is compatible with the arguments of the credence function being sets, as Kolmogorov suggested. Consider the set of all epistemic possibilities for an agent.² Let $[A]$ be the set of epistemic possibilities according to which the physical event A occurs. Then the Principal Principle can be reformulated by saying that $CR([A]|CH(A) = x) = x$, so that the different functions can still be defined over different objects. The principle only requires that there be some way of associating some objects of the chance function and some objects of the credence function, not that the objects actually be the same things.

Further, although the Principal Principle may entail that there be no constraints on *unconditional* credences that don’t also apply to chances, it doesn’t directly relate *conditional* chances to credences, so there is still room for the conditional notions to satisfy different mathematical constraints.³ Although in many cases the values of conditional probabilities can be derived from the values of unconditional probabilities, this is not true when unconditional probabilities take the value 0, which is a case that Hájek and I agree is often interesting. And as I will suggest later, conditional chances may not be directly constrained by the unconditional ones even in cases where they are all non-zero. Similar moves can be made for other principles that link different interpretations of probability together — these principles don’t guarantee that there is a single unified mathematical theory that all interpretations must follow.

1983] I just take the term “interpretation of probability” to apply to the notions that have historically been considered paradigmatic cases. Since my arguments suggest that there is no unified formalism that applies to all of them, this will mean that most of them won’t really count as interpretations of probability in this other sense.

²These may be possible worlds, impossible worlds, maximal consistent sets of sentences, or some other sort of entities.

³[Lewis, 1994] gives a “New Principal Principle” that suggests equating conditional probabilities in some circumstances as well. But on both principles it is only under special conditions that the functions must agree. This suggests that the functions may disagree and have different formalisms in other conditions, though I will not investigate this possibility here.

Thus, the primary objection to the possibility of different formal accounts of different interpretations of probability is incorrect — the principles connecting different interpretations of probability have natural modifications that allow the objects of the functions to be different, and allow different relations between conditional and unconditional probability to hold in each. In the rest of this section I will consider several different interpretations of probability and argue for each that it differs from the standard mathematical account of probability given in the first few chapters of [Kolmogorov, 1950], on which probability is a countably-additive, non-negative, normalized function on an algebra of sets, and the conditional probability $P(A|B)$ is given by $P(A \cap B)/P(B)$, whenever $P(B)$ has a precise non-zero value. Since they differ from this account in different ways, there can be no unified account of all of them that is complete for any one.

1.1 Propensity

The first interpretation of probability I will consider is Hájek’s “propensity interpretation”. (I will use the terms “objective chance” and “propensity” interchangeably.) I will argue that the standard account of conditional probability is incorrect here. Hájek gives some brief arguments in Section 2.2 that various interpretations of probability must satisfy $P(A|B) = P(A \cap B)/P(B)$, and then raises problems in Section 4 in which $P(B)$ is zero, imprecise, vague, or undefined. However, none of the arguments he presents in favor of the ratio account applies specifically to the propensity interpretation. I will suggest that there may be reason to think that there are further failures of the ratio account for the propensity interpretation, even when $P(B)$ is well-defined and non-zero.

The beginnings of this criticism come from [Humphreys, 1985], which argues that conditional chances don’t obey Bayes’ Theorem. This claim is known as “Humphreys’ Paradox.” The idea is that if B comes causally before A , then $P(A|B)$ makes sense in a standard way, while $P(B|A)$ should either be 0, $P(B)$, undefined, or otherwise different from the value $P(A|B)P(B)/P(A)$. Several objections to Humphreys’ argument are raised in [McCurdy, 1996] and [Gillies, 2000], but there are responses in [Humphreys, 2004].

Regardless of the status of that argument however, I think it brings up an interesting point about the interpretation of *conditional* probability, when probability is interpreted as objective chance, or propensity. Under most accounts of this interpretation, this means that the probability function measures some sort of disposition. If this is right, then we might think of the distinction between conditional and unconditional chances as being parallel to a distinction between dispositions that are triggered by some external conditions, and those that have some tendency to manifest regardless of the conditions. However, some responses to Humphreys’ Paradox rely on interpreting the conditional probability $P(A|B)$ differently, as a measure of the system’s unconditional disposition to produce both A and B simultaneously, out of the cases where it happens to produce B . This interpretation (which Humphreys calls the “co-production” interpretation) effec-

tively ends up *defining* conditional chance by the ratio formula, rather than taking conditional chance as a notion in need of its own analysis.

For a potential counterexample to the ratio formula, where $P(B)$ is well-defined and positive, but $P(A|B) \neq P(A \wedge B)/P(B)$, consider a fair coin-flipping machine whose trigger is inside a locked room. The easiest way to open the door is to turn on a huge magnet, which unlocks the door, but also prevents the coin from rotating in the air, so that while the magnet is activated, the coin almost always lands heads. Then on the co-production account, the probability of the coin coming up heads given that it is flipped is close to 1, because the propensity of the situation to give rise to a coin landing heads is almost as great as the propensity of the situation to give rise to a coin being flipped at all. However, while the magnet is off and the room is locked, it seems that we may want to say that the propensity for the device to produce a coin landing heads, given that it produces a flip at all, is close to $1/2$ — this is exactly what we mean to say that it is a fair coin-flipping machine. This example suggests that the co-production account of conditional propensities gives rise to a notion that is not actually dispositional — it is a parallel to examples of finks that are standard in the literature on dispositions. [Fara, 2006] I suspect that similar examples can be devised to correspond to other problematic cases in the literature on dispositions, like so-called “masks” and “mimics”. To get the behavior I have suggested may be intuitive here, we may need to observe the distinction between intervening and conditioning given in [Meek and Glymour, 1994].

It might be objected to this apparent counterexample that often this sort of non-dispositional behavior is exactly what we want for conditional propensities. For example, [Gillies, 2000, p. 828] includes an example of a barometer, in which it seems that we want a notion of conditional propensity for which the probability of a storm, given that the barometer level drops, is fairly high, even though there is no causal influence of the barometer on the storm. The co-production account gives a high conditional probability in this situation (since most cases in which the barometer level drops are cases in which there is a storm), but the more directly dispositional account Humphreys and I suggest gives a fairly low value (since lowering the barometer level has no tendency to cause storms).

But I think it may be that we want such a notion of conditional *probability*, but it's not clear that a notion of conditional *propensity* is needed here. I think all the work of this conditional probability can be done with a notion like conditional *degree of belief*, even though an important part of the work here will be done by the notion of *unconditional* propensity (or objective chance).

To see how this works, we can observe how the desired conditional probability comes out of the unconditional chances by means of Lewis' Principal Principle. A simplified version of this principle states that if an agent knows at some time t the chance at that time that some event A will occur, then her degree of belief in that event should equal the chance of that event. In a toy example, we might consider an agent who knows that the chance of a storm is .25, the chance of a drop in the barometer is .25 and the chance of both occurring is .2. Because

she knows these chances, these will also be her credences. Because conditional credences are generally related to unconditional credences by the ratio formula $P(A|B) = P(A \wedge B)/P(B)$, she will have credence .8 in a storm conditional on a drop in the barometer. The conditional chances never enter into her credence function here, so her reasoning can work just as Gillies wants, without having to accept his argument as saying anything about conditional chances. Now, most actual agents don't know the precise chances of any of these events, but they do have some ideas about them. All that is needed to make a modified version of this account work is that the agent be fairly confident that the ratio of chance of storm together with drop in barometer to the chance of a drop in the barometer is higher than the chance of a storm.

Given that an account of conditional chance doesn't need to accommodate these evidential correlations, it seems that we can get a more useful and interesting theory by avoiding the co-production interpretation — just as chances are dispositions on the propensity account, conditional chances are conditional dispositions. As Humphreys puts it in his reply to Gillies and McCurdy, “the conditional propensity constitutes an objective relationship between two events and any increase in our information about one when we learn of the other is a completely separate matter.” [Humphreys, 1985, p. 563] But this ends up leading to the conclusion I mentioned above — conditional propensity doesn't obey the mathematical relationships traditionally used for conditional probability.

To make the minimal modification to the standard framework, one natural suggestion (embraced by Fetzer and others) is that in cases like the ones Humphreys describes, conditional propensities are defined in one direction, but undefined in the other. That is, $P(A|B)$ makes sense when A is the event of a coin flip coming up heads, and B is the event of it being flipped fairly, but not when the two are reversed. (This is a difference from Humphreys' picture, on which the inverse conditional propensity has a value of 0, 1, or the value of a related unconditional probability, rather than being undefined.) Then, we just require that conditional propensities obey the standard probability axioms whenever they are defined.

However, the example of the coin flipping device in a magnetic room seems to suggest otherwise - in that case, the only way to get the intuitively correct conditional probabilities involves violating the ratio account of conditional probability, even though there is no division by zero or anything else mathematically strange going on. But I think that making this much of a departure from the standard Kolmogorov axioms for probability makes sense in this case, at least in part because of the metaphysics of the objects of the probability functions. Humphreys suggests that “the arguments of the propensity functions are names designating specific physical events. They do not pick out subsets of an outcome space as in the measure-theoretic approach.” [Humphreys, 2004, p. 669] This, if correct, would be a further departure from the standard mathematical theory. Since much of the motivation for the ratio formula comes from taking the objects of the probability function to be sets or sentences, this helps further undermine the argument that chance must satisfy this formula.

Other considerations on conditional chance come up when considering its uses. Humphreys' paradox (which motivates moving away from the standard mathematical formalism) proceeds from an intuition about what conditional chance means. Hájek mentions that evidential decision theory uses conditional probability — it may be that evidential decision theory makes use of conditional credences, while *causal* decision theory [Joyce, 1999] makes use of the agent's expectation of the conditional chances. If so, then the differences between the decision theories motivates a difference between the notions of conditional probability.

In [Lewis, 1980], Lewis suggests another use for conditional chance, which is to describe how *unconditional* chances change over time. The idea is that chances are explicitly indexed to times. Thus, at two different times t_1 and t_2 , there are generally two distinct probability functions P_1 and P_2 giving the chances for various events. Lewis argues that $P_2(A) = P_1(A|B)$, where B is “the complete history of the interval between” the two times. [Lewis, 1980, p. 280] In [Lange, 2006], there are some arguments against this claim — however, these depend on whether objective chances can change without any particular non-dispositional fact being the basis of such a change. At any rate, the modified claim is that instead of B being the history of all events occurring in the interval, B is specified to be the conjunction of outcomes of all chance events in the interval.

But in either case, only very specific conditional probabilities are needed — these are always probabilities of a later event conditional on an earlier one (so none of Humphreys' inverse conditional probabilities need to be used), and these are in fact always probabilities conditional on complete sets of occurrences between two times. It is plausible that such cases are never like the case I described with the coin-flipping device in a magnetized room, so that the question of whether those sorts of cases also violate the ratio account of conditional probability doesn't come up. Thus, this use of conditional chance doesn't seem to cut one way or the other in a dispute about whether conditional chances should behave like other conditional probabilities. If anything, this use suggests that Humphreys may be right about the order of the events being relevant to whether the conditional chance is even defined.

In conclusion, I have argued that chances are properties of events, rather than propositions, sets of worlds, or sentences. Because of the connection between chances and dispositions, and the role of chance in causal decision theory, it seems that the ratio analysis of conditional probability must fail, even in cases where the probabilities are non-zero. I don't have a substantive positive theory to offer in place of this, but this is a question for future research.

1.2 Subjective Probability

The second interpretation of probability is the “subjective” or “Bayesian” interpretation. The probability function is said to give an agent's “degrees of belief” or

“credences”.⁴ In this case, I think the standard account is correct as far as it goes — the difference only arises in the case where $P(B)$ doesn’t have a precise non-zero value. In these cases, I argue that conditional probability is best analyzed by the account described by Hájek in Section 5, as “Kolmogorov’s refinement”. The primary reason is that such an account is entailed by the principle of “conglomerability,” (which Hájek mentions in Section 8.4) which states that there cannot be a partition $\mathbf{G}$ (a set of propositions such that the agent is certain that exactly one proposition in the set is true) such that for every $G \in \mathbf{G}$, $P(A|G) > P(A)$. Basically, this means that no experiment can be such that every single conceivable outcome of the experiment would confirm A . (Since most interpretations of probability lack a notion corresponding to confirmation, this motivation for the principle only extends to the subjective interpretation.) A slightly weaker principle that also pushes in this direction is an identification of two notions of independence — conglomerability entails that if A is independent of partition $\mathbf{G}$ in the sense that $P(A|G)$ is constant for all $G \in \mathbf{G}$, then A is independent of $\mathbf{G}$ in the sense that $P(A|G) = P(A)$ for all $G \in \mathbf{G}$.

I also suggest that the examples of “impropriety” that Hájek attributes in Section 5.2 to Seidenfeld, Schervish, and Kadane, are irrelevant for probability as degree of belief. The particular probability spaces and partitions they use are so complicated that they can only be grasped by minds that are far more complicated than the ones that we normally attribute subjective probabilities to. (In particular, to actually construct a specific example of such a space, an agent would need to independently consider uncountably many distinct propositions.) However, other interpretations of probability don’t obviously require that every partition involved be graspable by a finite mind. For those interpretations, conglomerability may well fail, and thus a different mathematical account of conditional probability will be relevant.⁵

A further apparent problem for Kolmogorov’s refinement is that a single event G may be an element of two different partitions $\mathbf{G}$ and $\mathbf{G}'$, and $P(A|G)$ may be constrained to have two distinct values by these different partitions. ([Kadane *et al.*, 1986] shows that there are certain types of probability space in which this is unavoidable.) However, I suggest that this means we must view conditional probability as (in general) a *three*-place function, depending not only on A and G , but also the partition $\mathbf{G}$ defining the set of “relevant alternatives” to G . In particular cases, this partition will be specified by the experiment an agent is considering G as an outcome to, or the set of alternative hypotheses under consideration, or some other contextual factor. Thus, we must think of conditional degree of belief

⁴For much more on this topic, including more detailed versions of these arguments, and responses to some objections, see my dissertation. [Easwaran, 2008] I also explain in more detail what role I think conditional and unconditional probability play on this interpretation, which is an important part of arguing for any particular formalism.

⁵The failure of conglomerability Hájek mentions in Section 8.5, in the two-envelope paradox, depends essentially on the possibility of unboundedly large payoffs. I endorse only conglomerability for probabilities and not for expectations, and since probabilities are never greater than 1, these sorts of cases can’t arise.

as a function $P(A|G, \mathbf{G})$ rather than just $P(A|G)$.⁶ In my dissertation I argue that this relativization of the notion to a further parameter doesn't cause any problems for the applications of conditional degree of belief in its usual settings.

I do, however, think that the standard account is right to treat the objects of the probability function as something like sets of possible outcomes (in this case, sets of epistemic possibilities, whatever those are). Hájek claims in Section 4.4 that Kolmogorov's account conflates probability zero events with impossibilities — however, I think Kolmogorov's account actually preserves this distinction better than some of the alternatives. On his account, events are represented by sets, and their probability is assigned by a function. It's true that this function assigns zero both to impossible and non-impossible events, but this distinction shows up in the sets themselves — impossible events are represented by the empty set, while non-impossible events are represented by non-empty sets. In my dissertation, I use this very argument to motivate the claim that at least subjective probabilities are best understood by taking the bearers of probability values to be sets of some sort. This contrasts with accounts like Hailperin's, on which the events themselves have no internal structure to distinguish impossible from non-impossible ones.⁷

In summary, subjective probabilities are relations between an agent and a set of epistemic possibilities, rather than a sentence or an event. Although Kolmogorov's axioms are all appropriate constraints for the relation between conditional and unconditional degrees of belief, the need for conglomerability (to preserve some basic ideas about evidence and independence) means that conditional degrees of belief must actually be relativized to partitions, and must obey Kolmogorov's more sophisticated account of conditional probability, rather than the simple ratio analysis.

1.3 Logical probability

This interpretation is intended to give some sort of generalization of deductive logic's notion of entailment to a notion of "partial" or "inductive" entailment. When B entails A deductively, $P(A|B) = 1$; when B entails $\neg A$, $P(A|B) = 0$; in other cases, intermediate values are possible. In this particular case, I agree with Hájek's claim that conditional probability must be the fundamental notion — just as a notion of logical entailment must underlie any notion of logical truth, it looks like a notion of logical conditional probability must underlie any notion of logical probability. Thus, the proper account must again differ from the standard

⁶One might worry that this would cause further problems with our ability to state the Principal Principle. However, the Principal Principle only considers probabilities conditional on specific values of the chance function, which suggests that we should use the partition by values of chances to fill in the third spot in the conditional probability function. Although this might seem not to put many constraints on credences with respect to other partitions, it does (together with conglomerability) entail that an agent's unconditional credence in an event match her expected value of the unconditional chance, which is most of the work the principle tends to be used for.

⁷Popper's account does better here — although it, like Hailperin's, has no internal structure to the events, the impossible events can be distinguished as those events B such that for every A , $P(A|B) = 1$.

account mentioned above, in a way different from the previous two interpretations. Although this might suggest that Popper's account is the right one for this interpretation, I think this is not totally clear.

For one thing, the standard notion of logical consequence is not fundamentally a relation $A \vdash B$ between two sentences, but rather a relation $\Gamma \vdash B$ between a *set* of sentences Γ and a sentence. To be sure, there is a special case of this relation where Γ contains exactly one sentence. Additionally, when Γ is a finite set, the set-based relation holds between Γ and B iff the sentence-based relation holds between the conjunction of the elements of Γ and B . Thus, if the set is finite, one might try to reduce a function taking a set and a sentence as arguments to a function taking two sentences as arguments.⁸

In the case of complete entailment (that is, when the conditional probability is 1 or 0), classical first-order logic allows for the infinite case to be reduced to the finite case. The compactness theorem states that if Γ is an infinite set and $\Gamma \vdash B$, then there is some finite $\Gamma_0 \subset \Gamma$ such that $\Gamma_0 \vdash B$. However, although complete entailment by an infinite set can be reduced to complete entailment by a finite subset, there is no guarantee that this will be the case for partial entailment. The logic of complete entailment is monotonic — if a subset of Γ entails B , then Γ entails B . But if a set of sentences only *partially* entails a sentence, then surely adding more sentences to the collection of premises can make this partial entailment either better or worse. Thus, there seems to be no hope of reducing the infinite-premise case to the single-premise case in general. If the underlying notion of entailment to be generalized is not classical first-order logic, then the reduction may even be impossible in the case of full entailment.

Thus, if the logical interpretation of probability makes sense (in spite of the many objections that have been raised), it may need to be given by a sort of conditional probability function that allows for conditionalizing on a *set* of sentences, rather than just a single sentence, as all current formalisms for conditional probability do.

2 CONDITIONAL PROBABILITY VS. PROBABILITY GIVEN THE BACKGROUND

Hájek gives several different arguments for the claim that conditional probability ought to be taken as the basic notion of probability theory, rather than unconditional probability, as is standard. Given what I have said above, these arguments will need to be made separately for each interpretation of probability, rather than for all of them simultaneously. But at any rate, I think an important distinction to be made in these arguments is the distinction in degree of belief terms between probability *conditional* on some evidence E , and probability on a corpus of background knowledge B . This distinction may be more clear for objective

⁸Even in the finite case though this seem to blur some distinctions. After all, the fact that $\{A, B\} \vdash A \wedge B$ surely tells us something different than the fact that $A \wedge B \vdash A \wedge B$.

chances — if P is the probability function giving the chances at time t , then this is the distinction between $P(A|E)$ where E is itself some possible event *later in time* than t , and $P(A|B)$, where B is the complete description of the history of the universe *up to* time t . Similar distinctions may apply for some of the other interpretations of probability. To distinguish these, I will call the latter notion “background probability” and reserve the term “conditional probability” for the former. The distinction is approximately that the “background” consists of a collection of information that is necessary for the assignment of probability to even make sense, while a “condition” is some further information that could itself have had a probability given the background. If there really are such sets of background information that are necessary for the assignment of probabilities, then background probabilities seem to play the role that we originally assigned for *unconditional* probability, and not for conditional probability. There would be no room for a truly “unconditional” probability independent even of a background.

Hájek’s argument at the end of Section 6 that the fundamental notion of probability theory is conditional probability seems to be based on this notion of background probability, rather than the more specific notion of conditional probability that I distinguish from it. Thus, if these two notions require different formalisms, then he has only established that background probabilities are fundamental, and not conditional probabilities.

To see this distinction in the case of the propensity interpretation, consider Hájek’s example of a use for conditional probability: “Probability is a measure of the tendency for a certain kind of experimental set-up to produce a particular outcome ... [I]t is a conditional probability, the condition being a specification of the experimental set up.” (Section 3.1) This contrasts with the account suggested in [Lewis, 1980], on which the propensity function is indexed to a world and time. With the former point of view, unconditional probabilities don’t even make sense (what’s the probability that a coin lands heads, if nothing about the coin’s history or composition or manner of flipping is definite?) while on the latter point of view, both conditional and unconditional probabilities can be understood. If t, w is the time/world-pair at which a coin is about to be flipped, then we can make sense of both $P_{t,w}(H)$ and $P_{t,w}(H|A)$, where H is the event of the coin coming up heads, and A is (for example) the event of a substantial amount of matter quantum tunneling away from the coin in mid-air, thus disrupting the rotation of the coin as it flips. Indexing to the time obviates the need to conditionalize on the set-up, or the history, or anything else in determining a probability value. Thus, there is one picture of the propensity interpretation on which all propensities are conditional on a background (but this is not real conditional probability, in which we might conditionalize on more information than what is necessary to set up the experiment), and another on which there are truly unconditional propensities (though they are indexed to worlds and times, which suffice to determine the background conditions). In neither case do the truly conditional propensities play the fundamental role.

In the case of degree of belief, there are again two options for describing things.

We can index the probability function to an agent at a time, who happens to already have certain knowledge and belief. In that case, there are unconditional probabilities (her degrees of belief in particular propositions that she doesn't already know or fully believe) and conditional probabilities (her degrees of belief in particular propositions when she temporarily supposes further propositions). The other option is to always explicitly include all the agent's knowledge and information and beliefs in the proposition that is being conditioned on. This is the picture that seems to fit Hájek's account (citing de Finetti) of all probabilities really being conditional probabilities. But this conflates the things the agent *actually* knows or believes with things that she is merely supposing for the sake of a conditional probability. Additionally, it invites consideration of probabilities conditional on sets of information that *don't* include her full set of knowledge and beliefs. In some cases we can perhaps make sense of these alternate conditional probabilities as expressing counterfactual possibilities of what she imagines she would believe if she didn't have some of the information she does now. (This would make the conditional probability function conflate *three* types of attitudes an agent might have to a proposition.) But in cases where we consider a radically impoverished set of information in the condition, it seems that the conditional probability doesn't really tell us anything meaningful about the agent. At least, the only way around this seems to be to take seriously the notion of a hypothetical prior probability function that the agent *would* have had in the absence of any information whatsoever.

In the case of logical probability, the distinction I am making between background and truly conditional probabilities doesn't seem to hold up. There is supposed to be one correct logical probability function, and this function measures degree of entailment between sentences. No background is necessary for assigning these probabilities — all probabilities are really conditional. This is different from the other cases, where it seems that some sort of background is necessary to even begin to assign probabilities (there are no chances without a world and time, there are no degrees of belief absent some agent and time), but that other information might be hypothetically added in a way that gives an interesting (but distinct) notion of conditional probability. For logical probabilities it seems that only the latter exists, but that it really is the basic notion.

This distinction can be challenged in both the propensity and degree of belief cases. But in the propensity case this would involve arguing that there really is some objective chance function that specified the chances of various events even in the absence of an experimental set-up, including any of the specifications of how the universe works. And in the degree of belief case it would involve arguing that there really are hypothetical priors that have some real meaning for a particular agent. The difficulty of arguing for such unified functions that can assign probabilities in the absence of information is one of the serious challenges that has been raised against the logical interpretation of probability. Instead, it seems more promising to me to just index chances to worlds and times, and index degrees of belief to agents and times, so that the background doesn't have to be explicitly mentioned in

every probability statement. But this just means that there is sense in these cases to be made of a notion of unconditional probability that is not merely derivative from conditional probability, and that some arguments suggesting a priority for conditional probability identify the background information of the situation with the actual objects of the conditional probability function.

Thus, although it may be the case that all non-logical interpretations of probability depend on some set of information in order to assign probabilities in particular cases, the role this information plays can be very different from the role that the antecedent of a conditional probability plays in that interpretation. Thus, this background information is not a good argument for the claim that conditional probabilities are more fundamental than unconditional.

BIBLIOGRAPHY

- [Easwaran, 2008] K. Easwaran. *The Foundations of Conditional Probability*. PhD thesis, University of California, Berkeley, 2008.
- [Eells, 1983] E. Eells. Objective probability theory theory. *Synthese*, 57:387–442, 1983.
- [Fara, 2006] M. Fara. Dispositions. *Stanford Encyclopedia of Philosophy*, 1006.
- [Gillies, 2000] D. Gillies. Varieties of propensity. *British Journal for the Philosophy of Science*, 51:807–835, 2000.
- [Goel and Zellner, 1986] P. K. Goel and A. Zellner, eds. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. North-Holland, 1986.
- [Hailperin, 1984] T. Hailperin. Probability logic. *Notre Dame Journal of Formal Logic*, 25(3):198–212, 1984.
- [Hailperin, 1997] T. Hailperin. Ontologically neutral logic. *History and Philosophy of Logic*, 18:185–200, 1997.
- [Hailperin, 2000] T. Hailperin. Probability semantics for quantifier logic. *Journal of Philosophical Logic*, 29:207–239, 2000.
- [Hailperin, 2006] T. Hailperin. Probability logic and combining evidence. *History and Philosophy of Logic*, 27:249–269, 2006.
- [Humphreys, 1985] P. Humphreys. Why propensities cannot be probabilities. *The Philosophical Review*, 94(4):557–570, 1985.
- [Humphreys, 2004] P. Humphreys. Some considerations on conditional chances. *British Journal for the Philosophy of Science*, 55:667–680, 2004.
- [Joyce, 1999] J. Joyce. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.
- [Kadane et al., 1986] J. B. Kadane, M. J. Schervish, and T. Seidenfeld. Statistical implications of finitely additive probability. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, P. K. Goel and A. Zellner, eds. North-Holland, 1986.
- [Kolmogorov, 1950] A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea, 1950.
- [Lange, 2006] M. Lange. Do chances receive equal treatment under the laws? or: Must chances be probabilities? *British Journal for the Philosophy of Science*, 57:383–403, 2006.
- [Lewis, 1980] Lewis, D. (1980). A subjectivist’s guide to objective chance. In Jeffrey, R., editor, *Studies in Inductive Logic and Probability*, volume II. University of California Press.
- [Lewis, 1994] D. Lewis. Humean supervenience debugged. *Mind*, 109(412):473–490, 1994.
- [McCurdy, 1996] C. McCurdy. Humphreys’s paradox and the interpretation of inverse conditional probabilities. *Synthese*, 108:105–125, 1996.
- [Meek and Glymour, 1994] C. Meek and C. Glymour. Conditioning and intervening. *British Journal for the Philosophy of Science*, 45:1001–1021, 1994.
- [Popper, 1959] K. Popper. *The Logic of Scientific Discovery*, chapter iv*, pages 326–348. Harper & Row, 1959.
- [Rényi, 1970] A. Rényi. *Foundations of Probability*. Holden-Day, 1970.

Part III

Four Paradigms of Statistics

This page intentionally left blank

Classical Statistics Paradigm

This page intentionally left blank

ERROR STATISTICS

Deborah G. Mayo and Aris Spanos

1 WHAT IS ERROR STATISTICS?

Error statistics, as we are using that term, has a dual dimension involving philosophy and methodology. It refers to a standpoint regarding both:

1. a cluster of statistical tools, their interpretation and justification,
2. a general philosophy of science, and the roles probability plays in inductive inference.

To adequately appraise the error statistical approach, and compare it to other philosophies of statistics, requires understanding the complex interconnections between the methodological and philosophical dimensions in (1) and (2) respectively. To make this entry useful while keeping to a manageable length, we restrict our main focus to (1) the error statistical philosophy. We will however aim to bring out enough of the interplay between the philosophical, methodological, and statistical issues, to elucidate long-standing conceptual, technical, and epistemological debates surrounding both these dimensions. Even with this restriction, we are identifying a huge territory marked by generations of recurring controversy about how to specify and interpret statistical methods. Understandably, standard explications of statistical methods focus on the formal mathematical tools without considering a general philosophy of science or of induction in which these tools best fit. This is understandable for two main reasons: first, it is not the job of statisticians to delve into philosophical foundations, at least explicitly. The second and deeper reason is that the philosophy of science and induction for which these tools are most apt—we may call it *the error statistical philosophy*—will differ in central ways from traditional conceptions of the scientific method. These differences may be located in contrasting answers to a fundamental pair of questions that are of interest to statisticians and philosophers of science:

- *How do we obtain reliable knowledge about the world despite uncertainty and threats of error?*
- *What is the role of probability in making reliable inferences?*

To zero in on the main issues, we will adopt the following, somewhat unusual strategy: We shall first set out some of the main ideas within the broader error statistical philosophy of induction and inquiry, identifying the goals and requirements that serve both to direct and justify the use of formal statistical tools. Next we turn to explicating statistical methods, of testing and estimation, while simultaneously highlighting classic misunderstandings and fallacies. The error statistical account we advocate builds on Fisherian and Neyman-Pearsonian methods; see [Fisher, 1925; 1956; Neyman, 1952; Pearson, 1962]. While we wish to set out for the reader the basic elements of these tools, as more usually formulated, we will gradually transform them into something rather different from both Fisherian and Neyman-Pearsonian paradigms.

Our goals are twofold: to set the stage for developing a more adequate philosophy of inductive inquiry, and to illuminate crucial issues for making progress on the “statistical wars”, now in their seventh decade¹.

1.1 The Error Statistical Philosophy

Under the umbrella of error-statistical methods, one may include all standard methods using error probabilities based on the relative frequencies of errors in repeated sampling – often called *sampling theory* or *frequentist statistics*. Frequentist statistical methods are sometimes erroneously equated to other accounts that employ frequentist probability, for example, the “frequentism” in the logic of confirmation. The latter has to do with using relative frequencies of occurrences to infer probabilities of events, often by the straight rule, e.g., from an observed proportion of *As* that are *Bs* to infer the proportion of *As* that are *Bs* in a population.

1. One central difference, as Neyman [1957] chided Carnap, is that, unlike frequentist logics of confirmation, frequentist statistics always addresses questions or problems within a *statistical model*² (or family of models) $\mathcal{M}$ intended to provide an approximate (and idealized) representation of the process generating data $\mathbf{x}_0 := (x_1, x_2, \dots, x_n)$. $\mathcal{M}$ is defined in terms of $f(\mathbf{x}; \theta)$, the probability distribution of the sample $\mathbf{X} := (X_1, \dots, X_n)$, that assigns probabilities to all events of interest belonging to the sample space $\mathbb{R}_X^n$. Formal error statistical methods encompass the deductive assignments of probabilities to outcomes, given a statistical model $\mathcal{M}$ of the experiment, and inductive methods from the sample to claims about the model. Statistical inference focuses on the latter step: moving from the data to statistical hypotheses, typically couched in terms of unknown parameter(s) θ , which governs $f(\mathbf{x}; \theta)$.

¹The ‘first act’ might be traced to the papers by Fisher [1955], Pearson [1955], Neyman [1956].

²Overlooking the necessity of clearly specifying the statistical model — in terms of a complete set of probabilistic assumptions — is one of the cardinal sins still committed, especially by non-statisticians, in expositions of frequentist statistics.

2. The second key difference is how probability arises in induction. For the error statistician probability arises not to measure degrees of confirmation or belief (actual or rational) in hypotheses, but to quantify how frequently methods are capable of discriminating between alternative hypotheses and how reliably they facilitate the detection of error. These probabilistic properties of inductive procedures are *error frequencies* or *error probabilities*.

The statistical methods of significance tests and confidence-interval estimation are examples of formal error-statistical methods. A statistical inference might be an assertion about the value of θ , say that $\theta > 0$. Error probabilities attach, not directly to $\theta > 0$, but to the inference tools themselves, whether tests or estimators. The claims concerning θ are either correct or incorrect as regards the mechanism generating the data. Insofar as we are interested in using data to make inferences about this mechanism, in this world, it would make no sense to speak of the relative frequency of $\theta > 0$, as ‘if universes were as plenty as blackberries from which we randomly selected this one universe’, as Peirce would say (2.684). Nevertheless, error probabilities are the basis for determining whether and how well a statistical hypothesis such as $\theta > 0$ is warranted by data $\mathbf{x}_0$ at hand, and for setting bounds on how far off parameter values can be from 0 or other hypothetical values. Since it is the use of frequentist error probabilities, and not merely the use of frequentist probability that is central to this account, the term *error statistics* (an abbreviation of error probability statistics) seems an apt characterization.

Statistical Significance Test

Formally speaking, the inductive move in error statistics occurs by linking special functions of the data, $d(\mathbf{X})$, known as *statistics*, to hypotheses about parameter(s), θ of interest. For example, a test might be given as a rule: whenever $d(\mathbf{X})$ exceeds some constant c , infer $\theta > 0$, thereby rejecting $\theta = 0$:

Test Rule: whenever $\{d(\mathbf{x}_0) > c\}$, infer $\theta > 0$.

Any particular application of an inductive rule can be ‘in error’ in what it infers about the data generation mechanism, so long as data are limited. If we could calculate the probability of the event $\{d(\mathbf{X}) > c\}$ under the assumption that $\theta=0$, we could calculate the probability of erroneously inferring $\theta > 0$. Error probabilities are computed from the distribution of $d(\mathbf{X})$, the *sampling distribution*, evaluated under various hypothesized values of θ . The genius of formal error statistics is its ability to provide inferential tools where the error probabilities of interest may be calculated, despite unknowns.

Consider a simple and canonical example of a statistical test, often called a statistical significance test. Such a test, in the context of a statistical model $\mathcal{M}$, is a procedure with the following components:

1. a *null hypothesis* H_0 , couched in terms of unknown parameter θ , and

2. a function of the sample, $d(\mathbf{X})$, the *test statistic*, which reflects how well or poorly the data $\mathbf{x}_0$ accord with the null hypothesis H_0 — the larger the value of $d(\mathbf{x}_0)$ the further the outcome is from what is expected under H_0 — with respect to the particular question being asked. A crucial aspect of an error statistical test is its ability to ensure the sampling distribution of the test statistic can be computed under H_0 and under hypotheses discrepant from H_0 . In particular, this allows computing:
3. the *significance level* associated with $d(\mathbf{x}_0)$: the probability of a worse fit with H_0 than the observed $d(\mathbf{x}_0)$, under the assumption that H_0 is true:

$$p(\mathbf{x}_0) = P(d(\mathbf{X}) > d(\mathbf{x}_0); H_0).$$

This is known either as the *observed* significance level or the *p-value*. The larger the value of the test statistic the smaller the p-value. Identifying a relevant test statistic, together with the need to calculate the error probabilities, restricts the choices of test statistic so as to lead to a uniquely appropriate test, whether one begins with Neyman-Pearson (N-P) or Fisherian objectives [Cox, 1958].

Consider for example the case of a random sample $\mathbf{X}$ of size n from a Normal distribution with *unknown* mean μ and, for simplicity, *known* variance σ^2 (denoted by $N(\mu, \sigma^2)$). We want to test the hypotheses:

$$(1) \quad H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0.$$

The test statistic of this one-sided test is $d(\mathbf{X}) = \frac{(\bar{X} - \mu_0)}{\sigma_x}$, where $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$ denotes the sample mean and $\sigma_x = (\sigma/\sqrt{n})$. Given a particular outcome $\mathbf{x}_0$, we compute $d(\mathbf{x}_0)$. An inductive or ampliative inference only arises in moving from $d(\mathbf{x}_0)$ — a particular outcome — to a hypothesis about parameter μ . Consider the test rule: whenever $\bar{X}$ exceeds μ_0 by $1.96\sigma_x$ or more, infer $H_1: \mu > \mu_0$. Use of the statistic $d(\mathbf{X})$ lets us write this test rule more simply:

Test Rule T : whenever $\{d(\mathbf{x}_0) > 1.96\}$, infer $H_1: \mu > \mu_0$.

We deductively arrive at the probability of the event $\{d(\mathbf{X}) > 1.96\}$ under the assumption that H_0 correctly describes the data generating process, namely $P(d(\mathbf{X}) > 1.96; H_0) = .025$, giving the statistical significance level .025.

Tests, strictly speaking, are formal mapping rules. To construe them as inference tools requires an interpretation beyond the pure formalism, and this can be done in various ways. For instance, a Fisherian may simply output “the observed significance level is .025”; a Neyman-Pearson tester, might report “reject H_0 ” having decided in advance that any outcome reaching the .025 significance level will lead to this output. But these reports themselves require fleshing out, and a good deal of controversy revolves around some of the familiar ways of doing so.

Behavioristic and evidential philosophies

By a “statistical philosophy” we understand a general conception of the aims and epistemological foundations of a statistical methodology. Thus an important task for an error statistical philosophy is to articulate the interpretation and the rationale for outputs of inductive tests. For instance, continuing with our example, a Fisherian might declare that whenever $\{d(\mathbf{x}_0) > 1.96\}$:

“infer that H_0 is falsified at level .025”.

A strict Neyman-Pearsonian might identify ‘reject H_0 ’ with a specific *action*, such as:

“publish the result” or “reject the shipment of bolts”.

Consider a third construal: whenever $\{d(\mathbf{x}_0) > 1.96\}$ infer that the data $\mathbf{x}_0$ is evidence for a (positive) discrepancy γ from μ_0 :

“infer $\mathbf{x}_0$ indicates $\mu > \mu_0 + \gamma$ ”.

The weakest claim is to infer *some* discrepancy — as with the typical non-null hypothesis. More informatively, as we propose, one might specify a particular positive value for γ . For each we can obtain error probabilistic assertions:

Assuming that H_0 is true, the probability is .025 that:

- H_0 is falsified (at this level),
- the shipment of bolts is rejected,
- some positive discrepancy γ from μ_0 is inferred.

Each of these interpretations demands a justification or rationale, and it is a crucial part of the corresponding statistical philosophy to supply it. Error statistical methods are typically associated with two distinct justifications:

Behavioristic rationale. The first stresses the ability of tests to control error probabilities at some low level in the long run. This goal accords well with what is generally regarded as Neyman’s statistical philosophy wherein tests are interpreted as tools for deciding “how to behave” in relation to the phenomena under test, and are justified in terms of their ability to ensure low long-run errors. Neyman [1971] even called his tests tools for *inductive behavior*, to underscore the idea that the test output was an action, as well as to draw the contrast with Bayesian inductive inference in terms of degrees of belief.

Inferential rationale A non-behavioristic or inferential justification stresses the relevance of error probabilities for achieving inferential and learning goals. To succeed it must show how error probabilities can be used to characterize warranted inference in some sense. The key difference between behavioristic and inferential construals of tests is not whether one views an inference as a kind of decision (which one is free to do), but rather in the justificatory role given to error probabilities.

On the behavioristic philosophy, the goal is to adjust our behavior so that in the long-run we will not act erroneously too often: it regards low long-run error rates (ideally, optimal ones) alone as what justifies a method. This does not yield a satisfactory error statistical philosophy in the context of scientific inference. How to provide an inferential philosophy for error statistics has been the locus of the most philosophically interesting controversies. Although our main focus will be on developing an adequate inferential construal, there are contexts wherein the more behavioristic construal is entirely appropriate, and we propose to retain it within the error statistical umbrella³. When we speak of “the context of scientific inference” we refer to a setting where the goal is an inference about what is the case regarding a particular phenomenon.

Objectivity in error statistics

Underlying the error statistical philosophy, as we see it, is a conception of the objective underpinnings for uncertain inference: although knowledge gaps leave plenty of room for biases, arbitrariness, and wishful thinking, in fact we regularly come up against experiences that thwart our expectations, disagree with the predictions and theories we try to foist upon the world, and this affords objective constraints on which our critical capacity is built. Getting it (at least approximately) right, and not merely ensuring internal consistency or agreed-upon convention, is at the heart of objectively orienting ourselves toward the world. Our ability to recognize when data fail to match anticipations is what affords us the opportunity to systematically improve our orientation in direct response to such disharmony. Much as Popper [1959] takes the ability to falsify as the foundation of objective knowledge, R.A. Fisher [1935, p. 16] developed statistical significance tests based on his view that “every experiment may be said to exist only in order to give the facts the chance of disproving the null hypothesis”. Such failures could always be avoided by “immunizing” a hypothesis against criticism, but to do so would prevent learning what is the case about the phenomenon in question, and thus flies in the face of the aim of objective science. Such a strategy would have a very high probability of saving false claims. However, what we are calling the error-statistical philosophy goes beyond falsificationism, of both the Popperian and Fisherian varieties, most notably in its consideration of what positive inferences are licensed when data do not falsify but rather accord with a hypothesis or claim.

Failing to falsify hypotheses, while rarely allowing their acceptance as precisely true, may warrant excluding various discrepancies, errors or rivals. Which ones?

³Even in science there are tasks whose goal is avoiding too much noise in the network.

Those which, with high probability, would have led the test to issue a more discordant outcome, or a more statistically significant result. In those cases we may infer that the discrepancies, rivals, or errors are ruled out with severity.

Philosophy should direct methodology (not the other way around). To implement the error statistical philosophy requires methods that can accomplish the goals it sets for uncertain inference in science. This requires tools that pay explicit attention to the need to communicate results so as to set the stage for others to check, debate, scrutinize and extend the inferences reached. Thus, any adequate statistical methodology must provide the means to address legitimate critical questions, to give information as to which conclusions are likely to stand up to further probing, and where weaknesses remain. The much maligned, automatic, recipe-like uses of N-P tests wherein one accepts and rejects claims according to whether they fall into prespecified ‘rejections regions’ are uses we would also condemn. Rather than spend time legislating against such tests, we set out principles of interpretation that automatically scrutinize any inferences based on them. (Even silly tests can warrant certain claims.) This is an important source of objectivity that is open to the error statistician: choice of test may be a product of subjective whims, but the ability to critically evaluate which inferences are and are not warranted is not.

Background knowledge in the error statistical framework of ‘active’ inquiry

The error statistical philosophy conceives of statistics very broadly to include the conglomeration of systematic tools for collecting, modeling and drawing inferences from data, including purely ‘data analytic’ methods that are normally not deemed ‘inferential’. In order for formal error statistical tools to link data, or data models, to primary scientific hypotheses, several different statistical hypotheses may be called upon, each permitting an aspect of the primary problem to be expressed and probed. An auxiliary or ‘secondary’ set of hypotheses are needed to check the assumptions of other models in the complex network; see section 4. Its ability to check its own assumptions is another important ingredient to the objectivity of this approach.

There is often a peculiar allegation (criticism) that:

(#1) error statistical tools forbid using any background knowledge,

as if one must start each inquiry with a blank slate. This allegation overlooks the huge methodology of experimental design, data analysis, model specification and work directed at linking substantive and statistical parameters. A main reason for this charge is that prior probability assignments in hypotheses do not enter into the calculations (except in very special cases). But there is no reason to suppose the kind of background information we need in order to specify and interpret statistical methods can or should be captured by prior probabilities in the hypotheses being studied. (We return to this in section 3). But background knowledge must enter

in designing, interpreting, and combining statistical inferences in both informal and semi-formal ways. Far from wishing to inject our background opinions in the hypotheses being studied, we seek designs that help us avoid being misled or biased by initial beliefs. Although we cannot fully formalize, we can systematize the manifold steps and interrelated checks that, taken together, constitute a full-bodied experimental inquiry that is realistic.

The error statistician is concerned with the critical control of scientific inferences by means of stringent probes of conjectured flaws and sources of unreliability. Standard statistical hypotheses, while seeming oversimple in and of themselves, are highly flexible and effective for the piece-meal probes the error statistician seeks. Statistical hypotheses offer ways to couch conjectured flaws in inference, such as:

- mistaking spurious for genuine correlations,
- mistaken directions of effects,
- mistaken values of parameters,
- mistakes about causal factors,
- mistakes about assumptions of statistical models.

The qualities we look for to express and test hypotheses about such inference errors are generally quite distinct from those required of the substantive scientific claims about which we use statistical tests to learn. Unless the overall error statistical philosophy is recognized, the applicability and relevance of the formal methodology will be misunderstood, as it often is. Although the overarching goal of inquiry is to find out what is (truly) the case about aspects of phenomena, the hypotheses erected in the actual processes of finding things out are generally approximations (idealizations) and may even be deliberately false.

The picture corresponding to error statistics is one of an activist learner in the midst of an inquiry with the goal of finding something out. We want hypotheses that will allow for stringent testing so that if they pass we have evidence of a genuine experimental effect. The goal of attaining such well-probed hypotheses differs crucially from seeking highly probable ones (however probability is interpreted). We will say more about this in section 3.

1.2 *An Error Statistical Philosophy of Science*

The error statistical philosophy just sketched alludes to the general methodological principles and foundations associated with frequentist error statistical methods. By an error statistical philosophy of science, on the other hand, we have in mind the application of those tools and their interpretation to problems of philosophy of science: to model scientific inference (actual or rational), to scrutinize principles of inference (e.g., prefer novel results, varying data), and to frame and tackle philosophical problems about evidence and inference (how to warrant data, pinpoint blame for anomalies, test models and theories). Nevertheless, each of the points

in 1.1 about statistical methodology has direct outgrowths for the philosophy of science dimension. The outgrowths yield:

- (i) requirements for an adequate philosophy of evidence and inference, but also
- (ii) payoffs for using statistical science to make progress on philosophical problems.

- (i) In order to obtain a philosophical account of inference from the error statistical perspective, one would require forward-looking tools for finding things out, not for reconstructing inferences as ‘rational’ (in accordance with one or another view of rationality). An adequate philosophy of evidence would have to engage statistical methods for obtaining, debating, rejecting, and affirming data. From this perspective, an account of scientific method that begins its work only once well-defined evidence claims are available forfeits the ability to be relevant to understanding the actual processes behind the success of science.
- (ii) Conversely, it is precisely because the contexts in which statistical methods are most needed are ones that compel us to be most aware of strategies scientists use to cope with threats to reliability, that considering the nature of statistical method in the collection, modeling, and analysis of data is so effective a way to articulate and warrant principles of evidence.

In addition to paving the way for richer and more realistic philosophies of science, we claim, examining error statistical methods sets the stage for solving or making progress on long-standing philosophical problems about evidence and inductive inference.

- Where the recognition that data are always fallible presents a challenge to traditional empiricist foundations, the cornerstone of statistical induction is the ability to move from less to more accurate data.
- Where the best often thought feasible is getting it right in some asymptotic long-run, error statistical methods ensure specific precision in finite samples, and supply ways to calculate how large a sample size n needs to be.
- Where pinpointing blame for anomalies is thought to present insoluble *Duhemian problems* and *underdetermination*, a central feature of error statistical tests is their capacity to evaluate error probabilities that hold regardless of unknown background or nuisance parameters.
- Where appeals to statistics in conducting a meta-methodology too often boil down to reconstructing one’s intuition in probabilistic terms, statistical principles of inference do real work for us — in distinguishing when and why violations of novelty matter, when and why irrelevant conjuncts are poorly supported, and so on.

Although the extended discussion of an error statistical philosophy of science goes beyond the scope of this paper (but see [Mayo, 1996; Mayo and Spanos, 2010]), our discussion should show the relevance of problems in statistical philosophy for addressing the issues in philosophy of science — which is why philosophy of statistics is so rich a resource for epistemologists. In the next section we turn to the central error statistical principle that links (1.1) the error statistical philosophy and (1.2) an error statistical philosophy of science.

1.3 *The Severity Principle*

A method's error probabilities describe its performance characteristics in a hypothetical sequence of repetitions. How are we to use error probabilities in making particular inferences? This leads to the general question:

When do data $\mathbf{x}_0$ provide good evidence for, or a good test of, hypothesis H ?

Our standpoint begins with the situation in which we would intuitively *deny* $\mathbf{x}_0$ is evidence for H . Data $\mathbf{x}_0$ fail to provide good evidence for the truth of H if the inferential procedure had very little chance of providing evidence against H , even if H is false.

Severity Principle (weak). Data $\mathbf{x}_0$ (produced by process G) do *not* provide good evidence for hypothesis H if $\mathbf{x}_0$ results from a test procedure with a very low probability or capacity of having uncovered the falsity of H , even if H is incorrect.

Such a test we would say is insufficiently stringent or severe. The onus is on the person claiming to have evidence for H to show that they are *not* guilty of at least so egregious a lack of severity. Formal error statistical tools are regarded as providing systematic ways to foster this goal, as well as to determine how well it has been met in any specific case. Although one might stop with this negative conception (as perhaps Fisher and Popper did), we will go on to the further, positive one, which will comprise the full severity principle:

Severity Principle (full). Data $\mathbf{x}_0$ (produced by process G) provides good evidence for hypothesis H (just) to the extent that test T severely passes H with $\mathbf{x}_0$.

Severity rationale vs. low long-run error-rate rationale (evidential vs. behavioral rationale)

Let us begin with a very informal example. Suppose we are testing whether and how much weight George has gained between now and the time he left for Paris, and do so by checking if any difference shows up on a series of well-calibrated and stable weighing methods, both before his leaving and upon his return. If no change

on any of these scales is registered, even though, say, they easily detect a difference when he lifts a .1-pound potato, then this may be regarded as grounds for inferring that George's weight gain is negligible within limits set by the sensitivity of the scales. The hypothesis H here might be:

H : George's weight gain is no greater than δ ,

where δ is an amount easily detected by these scales. H , we would say, has passed a severe test: were George to have gained δ pounds or more (i.e., were H false), then this method would almost certainly have detected this.

A *behavioristic rationale* might go as follows: If one always follows the rule going from failure to detect a weight gain after stringent probing to inferring weight gain no greater than δ , then one would rarely be wrong in the long run of repetitions. While true, this is not the rationale we give in making inferences about George. It is rather that this particular weighing experiment indicates something about George's weight. The long run properties — at least when they are relevant for particular inferences — utilize error probabilities to characterize the capacity of our inferential tool for finding things out in the particular case. This is the *severity rationale*.

We wish to distinguish the severity rationale from a more prevalent idea for how procedures with low error probabilities become relevant to a particular application; namely, the procedure is rarely wrong, therefore, the probability it is wrong in this case is low. In this view we are justified in inferring H because it was the output of a method that rarely errs. This justification might be seen as intermediate between full-blown behavioristic justifications, and a genuine inferential justification. We may describe this as the notion that the long run error probability 'rubs off' on each application. This still does not get at the reasoning for the particular case at hand. The reliability of the rule used to infer H is at most a necessary and not a sufficient condition to warrant inferring H . What we wish to sustain is this kind of counterfactual statistical claim: that were George to have gained more than δ pounds, at least one of the scales would have registered an increase. This is an example of what philosophers often call an *argument from coincidence*: it would be a preposterous coincidence if all the scales easily registered even slight weight shifts when weighing objects of known weight, and yet were systematically misleading us when applied to an object of unknown weight. Are we to allow that tools read our minds just when we do not know the weight? To deny the warrant for H , in other words, is to follow a highly unreliable method: it would erroneously reject correct inferences with high or maximal probability (minimal severity), and thus would thwart learning. The stronger, positive side of the severity principle is tantamount to espousing the legitimacy of strong arguments from coincidence. What statistical tests enable us to do is determine when such arguments from coincidence are sustainable (e.g., by setting up null hypotheses). It requires being very specific about which inference is thereby warranted—we may, for example, argue from coincidence for a genuine, non-spurious, effect, but not be able to sustain an argument to the truth of a theory or even the reality of an entity.

Passing a Severe Test.

We can encapsulate this as follows:

A hypothesis H passes a severe test T with data $\mathbf{x}_0$ if,

(S-1) $\mathbf{x}_0$ accords with H , (for a suitable notion of accordance) and

(S-2) with very high probability, test T would have produced a result that accords less well with H than $\mathbf{x}_0$ does, if H were false or incorrect.

Equivalently, (S-2) can be stated:

(S-2)*: with very low probability, test T would have produced a result that accords as well as or better with H than $\mathbf{x}_0$ does, if H were false or incorrect.

Severity, in our conception, somewhat in contrast to how it is often used, is not a characteristic of a test in and of itself, but rather of the test T , a specific test result $\mathbf{x}_0$, and a specific inference H (not necessarily predesignated) being entertained. That is, the severity function has three arguments. We use the notation: $SEV(T, \mathbf{x}_0, H)$, or even $SEV(H)$, to abbreviate:

“The severity with which claim H passes test T with outcome $\mathbf{x}_0$ ”.

As we will see, the analyses may take different forms: one may provide a series of inferences that pass with high and low severity, serving essentially as benchmarks for interpretation, or one may fix the inference of interest and report the severity attained.

The formal statistical testing apparatus does not include severity assessments, but there are ways to use the error statistical properties of tests, together with the outcome $\mathbf{x}_0$, to evaluate a test’s severity in relation to an inference of interest. This is the key for the inferential interpretation of error statistical tests. While, at first blush, a test’s severity resembles the notion of a test’s power, the two notions are importantly different; see section 2.

The severity principle, we hold, makes sense of the underlying reasoning of tests, and addresses chronic problems and fallacies associated with frequentist testing. In developing this account, we draw upon other attempts to supply frequentist foundations, in particular by Bartlett, Barnard, Birnbaum, Cox, Efron, Fisher, Lehmann, Neyman, E. Pearson; the severity notion, or something like it, affords a rationale and unification of several threads that we have extracted and woven together. Although mixing aspects from N-P and Fisherian tests is often charged as being guilty of an inconsistent hybrid [Gigerenzer, 1993], the error statistical umbrella, linked by the notion of severity, allows for a coherent blending of elements from both approaches. The different methods can be understood as relevant for one or another type of question along the stages of a full-bodied inquiry. Within the error statistical umbrella, the different methods are part of the panoply of methods that may be used in the service of severely probing hypotheses.

A principle for interpreting statistical inference vs. the goal of science

We should emphasize at the outset that while severity is the principle on which interpretations of statistical inferences are based, we are *not* claiming it is the goal of science. While scientists seek to have hypotheses and theories pass severe tests, severity must be balanced with informativeness. So for example, trivially true claims would pass with maximal severity, but they would not yield informative inferences⁴. Moreover, one learns quite a lot from ascertaining which aspects of theories have *not* yet passed severely. It is the basis for constructing rival theories which existing tests cannot distinguish, and is the impetus for developing more probative tests to discriminate them (see [Mayo, 2010a]).

2 A PHILOSOPHY FOR ERROR STATISTICS

We review the key components of error statistical tests, set out the core ingredients of both Fisherian and N-P tests, and then consider how the severity principle directs the interpretation of frequentist tests. We are then in a position to swiftly deal with the specific criticisms lodged at tests.

2.1 Key Components of Error-Statistical Tests

While we focus, for simplicity, on inferences relating to the simple normal model defined in section 1.1, the discussion applies to any well-defined frequentist test.

A One-Sided Test T_α . We elaborate on the earlier example in order to make the severity interpretation of tests concrete.

EXAMPLE 1. Test T_α . Consider a sample $\mathbf{X} := (X_1, \dots, X_n)$ of size n , where each X_k is assumed to be Normal ($N(\mu, \sigma^2)$), Independent and Identically Distributed (NIID), denoted by:

$$\mathcal{M} : X_k \sim \text{NIID}(\mu, \sigma^2), \quad -\infty < \mu < \infty, k=1, 2, \dots, n, \dots$$

Let μ be the parameter of interest and assume at first that σ^2 is known; this will be relaxed later. Consider the following null and alternative hypotheses of interest:

$$(2) \quad H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0.$$

The *test statistic* or distance function is: $d(\mathbf{X}) = \frac{(\bar{X} - \mu_0)}{\sigma_x}$, where $\sigma_x = (\sigma/\sqrt{n})$. Note that $d(\mathbf{X})$ follows a common pattern for forming such a measure of discordance:

[estimated — expected (under H_0)], measured in standard deviation units.

⁴For an extreme case, demonstrating that not- H yields a logical contradiction shows, with maximal severity, that H is tautologous.

Under the null, $d(\mathbf{X})$ is distributed as standard Normal, denoted by:

$$d(\mathbf{X}) = \frac{(\bar{X} - \mu_0)}{\sigma_x} \sim \mathcal{N}(0, 1).$$

This is a very important result because it enables one to ensure that:

$$P(d(\mathbf{X}) > c_\alpha; H_0) = \alpha.$$

For instance, c_α for $\alpha = .025$ is 1.96; see figures 1(a)-(d). We also have:

$$P(\text{observing a } p\text{-value} \leq \alpha) \leq \alpha.$$

In the simple Fisherian test, the p -value indicates the level of inconsistency between what is expected and what is observed in the sense that the smaller the p -value the larger the discordance between $\mathbf{x}_0$ and H_0 [Cox, 1958]. If the p -value is not small, if it is larger than some threshold α (e.g., .01) then the disagreement is not considered strong enough to indicate evidence of departures from H_0 . Such a result is commonly said to be insignificantly different from H_0 , but, as we will see, it is fallacious to automatically view it as evidence *for* H_0 . If the p -value is small enough, the data are regarded as grounds to reject or find a discrepancy from the null. Evidence against the null suggests evidence for *some* discrepancy from the null, although it is not made explicit in a simple Fisherian test.

Reference to ‘discrepancies from the null hypothesis’ leads naturally into Neyman-Pearson [1933] territory. Here, the falsity of H_0 is defined as H_1 the complement of H_0 with respect to the parameter space Θ . In terms of the p -value, the Neyman-Pearson (N-P) test may be given as a rule:

if $p(\mathbf{x}_0) \leq \alpha$, reject H_0 (infer H_1); if $p(\mathbf{x}_0) > \alpha$, do not reject H_0 .

Equivalently, the test fixes c_α at the start as the cut-off point such that any outcome smaller than c_α is taken to “accept” H_0 . Critics often lampoon an automatic-recipe version of these tests. Here the tester is envisioned as simply declaring whether or not the result was statistically significant at a fixed level α , or equivalently, whether the data fell in the rejection region.

Attention to the manner in which tests are used, even by Neyman, however, reveals a much more nuanced and inferential interpretation to which these formal test rules are open. These uses (especially in the work of E. Pearson) provide a half-way house toward an adequate inferential interpretation of tests:

Accept H_0 : statistically *insignificant result* — “decide” (on the basis of the observed p value) that there is insufficient evidence to infer departure from H_0 , and

Reject H_0 : statistically *significant result* — “decide” (on the basis of the observed p -value) that there is some evidence of the falsity of H_0 in the direction of the alternative H_1 .

Although one could view these as decisions, we wish to interpret them as inferences. All the N-P results would continue to hold with either construal.

The N-P test rule: Reject H_0 iff $d(\mathbf{x}_0) > c_\alpha$,

ensures the probability of rejecting (i.e., declaring there is evidence against) H_0 when H_0 is true — a type I error — is α . Having fixed α , the key idea behind N-P tests is to minimize the probability of a *type II error* (failing to reject H_0 when H_1 is true), written as $\beta(\mu_1)$:

$$P(d(\mathbf{X}) \leq c_\alpha; \mu = \mu_1) = \beta(\mu_1), \text{ for any } \mu_1 \text{ greater than } \mu_0.$$

That is, the test *minimizes* the probability of finding “statistical agreement” with H_0 when some alternative hypothesis H_1 is true. Note that the set of alternatives in this case includes all $\mu_1 > \mu_0$, i.e. H_1 is a *composite* hypothesis, hence the notation $\beta(\mu_1)$. Equivalently, the goal is to maximize the *power of the test*, for a fixed c_α :

$$POW(T_\alpha; \mu_1) = P(d(\mathbf{X}) > c_\alpha; \mu_1), \text{ for any } \mu_1 \text{ greater than } \mu_0.$$

In the behavioristic construal of N-P tests, these goals are put in terms of wishing to avoid behaving erroneously too often. But, the tests that grow out of the requirement to satisfy the N-P pre-data, long-run desiderata often lead to a uniquely appropriate test, whose error probabilities simultaneously can be shown to satisfy severity desiderata. The severity construal of N-P tests underscores the role of error probabilities as measuring the ‘capacity’ of the test to detect different discrepancies $\gamma \geq 0$ from the null, where $\mu_1 = (\mu_0 + \gamma)$. The power of a ‘good’ test is expected to increase with the value of γ .

Pre-data, these desiderata allow us to ensure two things:

- (i) a rejection indicates with severity *some* discrepancy from the null, and
- (ii) failing to reject the null rules out with severity those alternatives against which the test has high power.

Post-data, one can go much further in determining the magnitude γ of discrepancies from the null warranted by the actual data in hand. That will be the linchpin of our error statistical construal. Still, even N-P practitioners often prefer to report the observed p-value rather than merely whether the predesignated cut-off for rejection has been reached, because it “enables others to reach a verdict based on the significance level of their choice” [Lehmann, 1993, p. 62]. What will be new in the severity construal is considering sensitivity in terms of the probability of $\{d(\mathbf{X}) > d(\mathbf{x}_0)\}$, under various alternatives to the null rather than the N-P focus on $\{d(\mathbf{X}) > c_\alpha\}$. That is, the error statistical construal of tests will require evaluating this ‘sensitivity’ post-data (relative to $d(\mathbf{x}_0)$, not c_α); see [Cox, 2006].

We now turn to the task of articulating the error-statistical construal of tests by considering, and responding to, classic misunderstandings and fallacies.

2.2 *How severity gives an inferential interpretation while scotching familiar fallacies*

Suppose the observed p -value is .01. This report might be taken to reject the null hypothesis H_0 and conclude H_1 . Why? An N-P *behavioristic rationale* might note that deciding to interpret the data this way would rarely be wrong. Were H_0 true, so large a $d(\mathbf{x}_0)$ would occur only 1% of the time. In our inferential interpretation, the fact that the p -value is small ($p(\mathbf{x}_0) = .01$) supplies evidence for H_1 because H_1 has passed a severe test: with high probability ($1 - p(\mathbf{x}_0)$) such an impressive departure from H_0 would not have occurred if H_0 correctly described the data generating procedure. The severity definition is instantiated because:

(S-1): $\mathbf{x}_0$ accords with H_1 , and (S-2): there is a high probability (.99) that a less statistically significant difference would have resulted, were H_0 true.

This is entirely analogous to the way we reasoned informally about George's weight. Granted, evidence from any one test might at most be taken as some evidence that the effect is genuine. But after frequent rejections of H_0 , H_1 passes a genuinely severe test because, were H_1 false and the null hypothesis H_0 true, we would very probably have obtained results that accord less well with H_1 than the ones we actually observed. So the p -value gives the kind of data-dependency that is missing from the coarse N-P tests, and it also lends itself to a *severity construal* — at least with respect to inferring the existence of *some* discrepancy from the null. We have an inferential interpretation, but there are still weaknesses we need to get around. A pair of criticisms relating to statistically significant results, are associated with what we may call “fallacies of rejection”.

2.3 *Fallacies of rejection (errors in interpreting statistically significant results)*

First there is the weakness that, at least on an oversimple construal of tests:

(#2) All statistically significant results are treated the same,

and second, that:

(#3) The p -value does not tell us how large a discrepancy is found.

We could avoid these criticisms if the construal of a statistically significant result were in terms of evidence for a particular discrepancy from H_0 (an effect size), that is, for inferring: $H: \mu > \mu_1 = (\mu_0 + \gamma)$, (there is evidence of a discrepancy γ).

The severity reasoning can be used to underwrite such inferences about particular discrepancies $\gamma \geq 0$ from the null hypothesis, i.e., $\mu > (\mu_0 + \gamma)$. For each

result we need to show: (a) the discrepancies that are not warranted, and (b) those which are well warranted. The basis for doing so is summarized in (a) and (b):

- (a) If there is a very high probability of obtaining so large a $d(\mathbf{x}_0)$ (even) if $\mu \leq \mu_1$, then $SEV(\mu > \mu_1)$ is low. By contrast:
- (b) If there is a very low probability of obtaining so large a $d(\mathbf{x}_0)$ if $\mu \leq \mu_1$, then $SEV(\mu > \mu_1)$ is high.

There are two key consequences. First, two different statistically significant results are distinguished by the inferences they severely warrant (criticism #2). Second, for any particular statistically significant result, the severity associated with $\mu > \mu_2$ will differ from (be less than) that associated with $\mu > \mu_1$, for any μ_2 greater than μ_1 (criticism #3).

Let us illustrate in detail with reference to our test T_α of hypotheses:

$$H_0 : \mu = 0 \text{ vs. } H_1 : \mu > 0.$$

For simplicity, let it be known that $\sigma=2$, and suppose $n=100$, i.e. $\sigma_x=.2$. Let us call a result “statistically significant” if it is statistically significant at the .025 level, i.e., $d(\mathbf{x}_0) > 1.96$. To address criticism #2, consider three different significant results: $d(\mathbf{x}_0)=2.0$ ($\bar{x}=0.4$), $d(\mathbf{x}_0)=3.0$ ($\bar{x}=0.6$), $d(\mathbf{x}_0)=5.0$ ($\bar{x}=1.0$).

Each statistically significant result “accords with” the alternative ($\mu > 0$). So (S-1) is satisfied. Condition (S-2) requires the evaluation of the probability that test T_α would have produced a result that accords *less well* with H_1 than $\mathbf{x}_0$ does (i.e. $d(\mathbf{X}) \leq d(\mathbf{x}_0)$), calculated under various discrepancies from 0. For illustration, imagine that we are interested in the inference $\mu > .2$. The three different statistically significant outcomes result in different severity assignments for the same inference $\mu > .2$.

Begin with $d(\mathbf{x}_0)=2.0$ ($\bar{x}=0.4$). We have:

$$SEV(\mu > .2) = P(\bar{X} < 0.4; \mu > .2 \text{ is false}) = P(\bar{X} < 0.4; \mu \leq .2 \text{ is true}).$$

Remember, we are calculating the probability of the event, $\{\bar{X} < .4\}$, and the claim to the right of the “;” should be read “calculated under the assumption that” one or another values of μ is correct. How do we calculate $P(\bar{X} < .4; \mu \leq .2 \text{ is true})$ when $\mu \leq .2$ is a composite claim? We need only to calculate it for the point $\mu=.2$ because μ values less than .2 would yield an even higher SEV value. The severity for inferring $\mu > .2$, when $\bar{x}=.4$ is $SEV(\mu > .2) = .841$. This follows from the fact that the observation $\bar{x}=.4$ is one standard deviation ($\sigma_x=.2$) in excess of .2. The probability of the event $(\bar{X} > .4)$ under the assumption that $\mu=.2$ is .16, so the corresponding SEV is .841. By standardizing the difference $(\bar{x}-\mu)$, i.e. define a *standardized Normal* random variable $Z = \frac{(\bar{x}-\mu)}{\sigma_x} \sim N(0, 1)$, one can read off the needed probabilities from the standard Normal tables. Figures 1(a)-(d) show the probabilities beyond 1 and 2 standard deviation, as well as the .025 and .05 thresholds, i.e. 1.645 and 1.96, respectively.

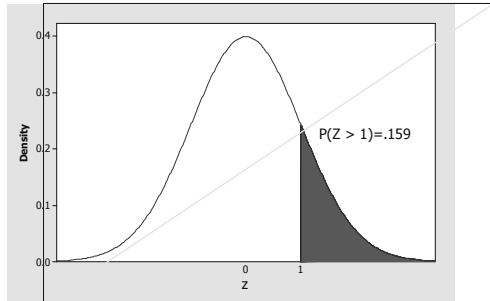


Fig 1a. $N(0, 1)$: Right tail probability beyond 1 (one) standard deviation (SD).

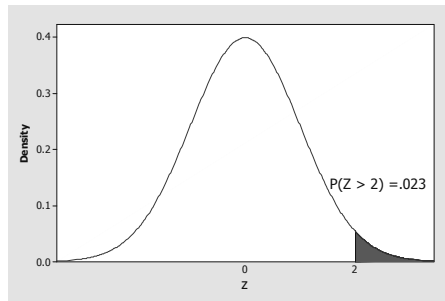


Fig 1b. $N(0, 1)$: Right tail probability beyond 2 (two) standard deviations.

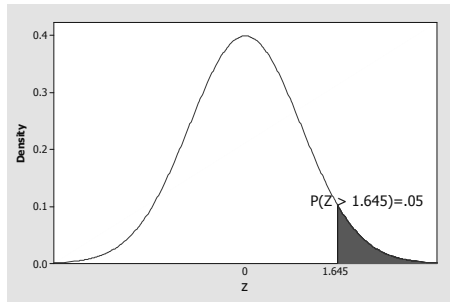


Fig 1c. $N(0, 1)$: 5% right tail probability.

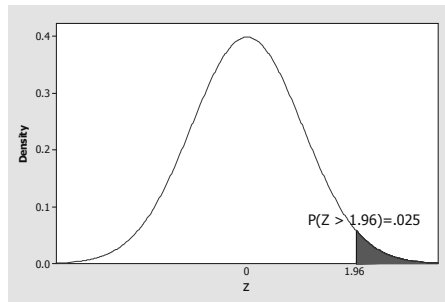


Fig 1d. $N(0, 1)$: 2.5% right tail probability.

Figure 1. Tail area probabilities of the standard normal ($N(0, 1)$) distribution

Now, let us consider the two other statistically significant outcomes, retaining this same inference of interest. When $\bar{x}=.6$, we have $SEV(\mu > .2)=.977$, since $\bar{x}=.6$ is 2 standard deviation in excess of the $\mu=.2$. When $\bar{x}=1$, $SEV(\mu > .2)=.999$, since $\bar{x}=1$ is 4 standard deviation in excess of $\mu=.2$. So inferring the discrepancy $\mu > .2$ is increasingly warranted, for increasingly significant observed values. Hence, criticisms #2 and #3 are scotched by employing the severity evaluation.

If pressed, critics often concede that one can avoid the fallacies of rejection, but seem to argue that the tests are illegitimate because they are open to fallacious construals. This seems to us an absurd and illogical critique of the foundations of tests. We agree that tests should be accompanied by interpretive tools that avoid fallacies by highlighting the correct logic of tests. That is what the error statistical philosophy supplies. We do not envision computing these assessments each time, nor is this necessary. The idea would be to report severity values corresponding to the inferences of interest in the given problem; several benchmarks for well warranted and poorly warranted inferences would suffice.

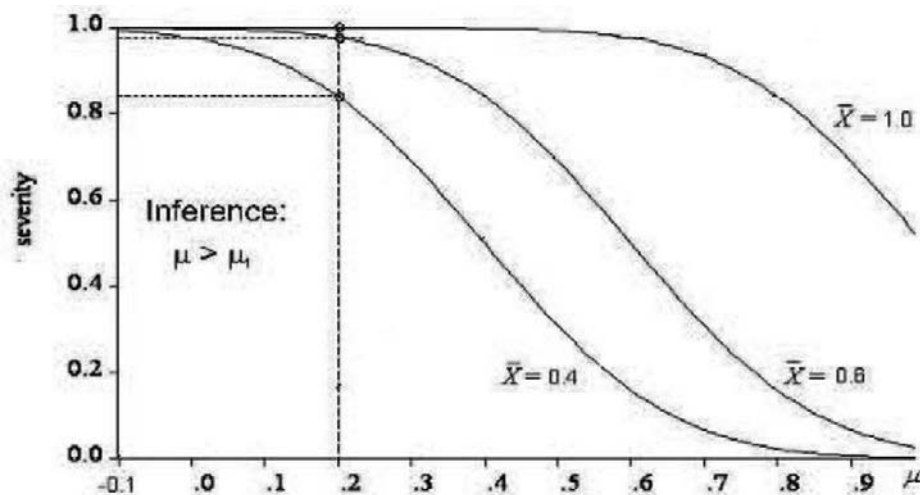


Figure 2. Significant result. Severity associated with inference $\mu > 0.2$, with different outcomes x_0 .

Figure 2 shows three *severity curves* for test T_α , associated with different outcomes $\mathbf{x}_0$, where, as before, $\sigma = 2, n = 100, d(\mathbf{x}_0) = \frac{(\bar{x} - \mu_0)}{\sigma_x}, \mu_0 = 0$:

$$\begin{aligned} \text{for } d(\mathbf{x}_0)=2.0 \ (\bar{x}=0.4) : \quad & SEV(\mu > 0.2)=.841, \\ \text{for } d(\mathbf{x}_0)=3.0 \ (\bar{x}=0.6) : \quad & SEV(\mu > 0.2)=.977, \\ \text{for } d(\mathbf{x}_0)=5.0 \ (\bar{x}=1.0) : \quad & SEV(\mu > 0.2)=.999. \end{aligned}$$

The vertical line at $\mu=.2$ pinpoints the inference in our illustration, but sliding it along the μ axis one sees how the same can be done for different inferences, e.g., $\mu>.3, \mu>.4, \dots$

Criticism (#3) is often phrased as “statistical significance is not substantive significance”. What counts as substantive significance is a matter of the context. What the severity construal of tests will do is tell us which discrepancies are and are not indicated, thereby avoiding the confusion between statistical and substantive significance.

To illustrate the notion of warranted discrepancies with data $\mathbf{x}_0$, consider figure 3 where we focus on just one particular statistically significant outcome, say $\bar{x}=0.4$, and consider different discrepancies γ from 0 one might wish to infer, each represented by a vertical line. To begin with, observe that $SEV(\mu > 0)=.977$, i.e. 1 minus the p-value corresponding to this test. On the other hand, as the discrepancy increases from 0 to .2 the $SEV(\mu > .2)$ is a bit lower, but not too bad: .841. We see that the SEV decreases as larger discrepancies from 0 are entertained (remembering the outcome is fixed at $\bar{x}=0.4$). An extremely useful benchmark is $\mu > .4$, since that is the inference which receives severity .5. So we know immediately that $SEV(\mu > .5)$ is less than .5, and in particular it is .3. So

$\bar{x}=0.4$ provides a very poor warrant for $\mu > .5$. More than half the time such a significant outcome would occur even if $\mu \leq .5$.

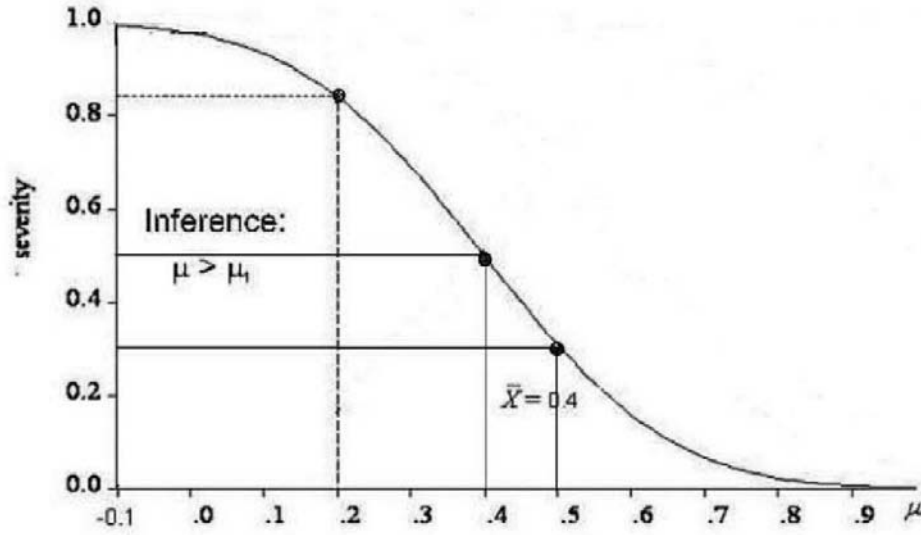


Figure 3. Significant result. The severity for inferring different discrepancies $\mu > \gamma$ with the same outcome $\bar{x}=0.4$

Many general relationships can be deduced. For example, since the assertions $\mu > \mu_1$ and $\mu \leq \mu_1$ constitute a partition of the parameter space of μ we have:

$$SEV(\mu > \mu_1) = 1 - SEV(\mu \leq \mu_1).$$

As before, severity is evaluated at a point μ_1 , i.e.

$$SEV(\mu > \mu_1) = P(d(\mathbf{X}) \leq d(\mathbf{x}_0); \mu=\mu_1).$$

Severity and Power with Significant Results: two key points

(i) It is important to note the relationship between our data-specific assessment of an α -level statistically significant result and the usual assessment of the power of test T_α at the alternative: $\mu_1=(\mu_0 + \gamma)$. Power, remember, is always defined in terms of a rejection rule indicating the threshold (c_α) beyond which the result is taken as statistically significant enough to reject the null; see section 2.1.

If $d(\mathbf{x}_0)$ is then significant at the α -level, $d(\mathbf{x}_0) > c_\alpha$, the severity with which the test has passed $\mu > \mu_1$ is:

$$P(d(\mathbf{X}) \leq c_\alpha; \mu=\mu_1) = 1 - POW(T_\alpha; \mu_1).$$

But the observed statistically significant $d(\mathbf{x}_0)$ could exceed the mere cut-off value for significance c_α . Should we take a result that barely makes it to the cut-off

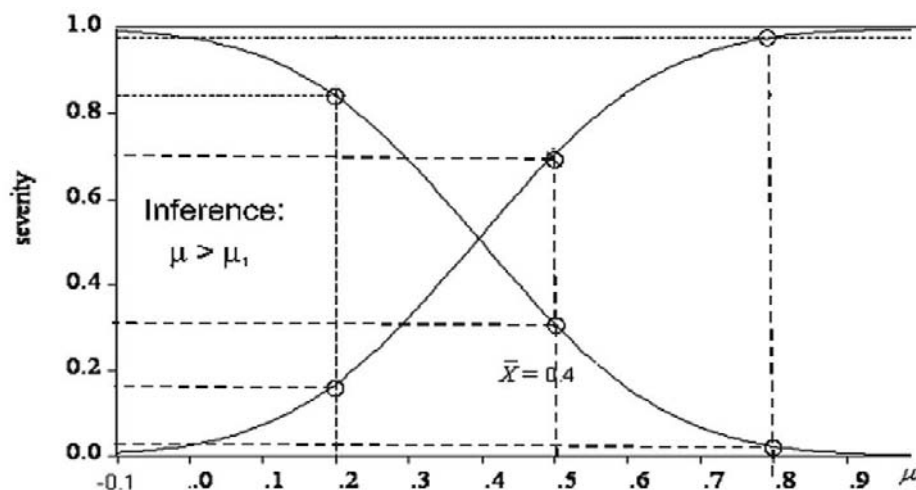


Figure 4. Juxtaposing the power curve with the severity curve for $\bar{x} = .4$.

just the same as one farther out into the rejection region? We think not, and the assessment of severity reflects this. As is plausible, the more significant result yields a higher the severity for the same inference $\mu > \mu_1$:

$$P(d(\mathbf{X}) \leq d(\mathbf{x}_0); \mu > \mu_1) \text{ exceeds } P(d(\mathbf{X}) \leq c_\alpha; \mu > \mu_1).$$

That is, one minus the power of the test at μ_1 provides a lower bound for the severity associated with the inference $\mu > \mu_1$.

The higher the power of the test to detect discrepancy γ , the lower the severity associated with the inference: $\mu > (\mu_0 + \gamma)$ when the test rejects H_0 .

Hence, the severity with which alternative $\mu > (\mu_0 + \gamma)$ passes a test is not given by, and is in fact inversely related to, the test's power at: $\mu_1 = (\mu_0 + \gamma)$.

This can be seen in figure 4 which juxtaposes the power curve with the severity curve for $\bar{x} = 0.4$. It is seen that the power curve slopes in the opposite direction from the severity curve. As we just saw, the statistically significant result, $\bar{x} = 0.4$, is good evidence for $\mu > .2$ (the severity was .841), but poor evidence for the discrepancy $\mu > .5$ (the severity was .3). If the result does not severely pass the hypothesis $\mu > .5$, it would be even less warranted to take it as evidence for a larger discrepancy, say $\mu > .8$. The relevant severity evaluation yields: $P(d(\mathbf{X}) \leq 2.0; \mu = .8) = .023$, which is very low, but the power of the test at $\mu = .8$ is very high, .977.

Putting numbers aside, an intuitive example makes the point clear. The smaller the mesh of a fishnet, the more capable it is of catching even small fish. So being given the report that (i) a fish is caught, and (ii) the net is highly capable of catching even 1 inch guppies, we would deny the report is good evidence of, say, a 9 inch fish! This takes us to our next concern.

2.4 Fallacies arising from overly sensitive tests

A common complaint concerning a statistically significant result is that for any discrepancy from the null, say $\gamma \geq 0$, however small, one can find a large enough sample size n such that a test, with high probability, will yield a statistically significant result (for any p-value one wishes).

(#4) With large enough sample size even a trivially small discrepancy from the null can be detected.

A test can be so sensitive that a statistically significant difference from H_0 only warrants inferring the presence of a relatively small discrepancy γ ; a large enough sample size n will render the power $POW(T_\alpha; \mu_1 = \mu_0 + \gamma)$ very high. To make things worse, many assume, fallaciously, that reaching statistical significance at a given level α is more evidence against the null the larger the sample size (n). (Early reports of this fallacy among psychology researchers are in Rosenthal and Gaito, 1963). Few fallacies more vividly show confusion about significance test reasoning. A correct understanding of testing logic would have nipped this fallacy in the bud 60 years ago. Utilizing the severity assessment one sees that an α -significant difference with n_1 passes $\mu > \mu_1$ less severely than with n_2 where $n_1 > n_2$.

For a fixed type I error probability α , increasing the sample size decreases the type II error probability (power increases). Some argue that to balance the two error probabilities, the required α level for rejection should be decreased as n increases. Such rules of thumb are too tied to the idea that tests are to be specified and then put on automatic pilot without a reflective interpretation. The error statistical philosophy recommends moving away from all such recipes. The reflective interpretation that is needed drops out from the severity requirement: increasing the sample size does increase the test's sensitivity and this shows up in the "effect size" γ that one is entitled to infer at an adequate severity level. To quickly see this, consider figure 5.

It portrays the severity curves for test T_α , $\sigma=2$, $n=100$, with the same outcome $d(\mathbf{x}_0)=1.96$, but based on different sample sizes ($n=50, n=100, n=1000$), indicating that: the *severity for inferring* $\mu > .2$ decreases as n increases:

for $n=50$:	$SEV(\mu > 0.2) = .895$,
for $n=100$:	$SEV(\mu > 0.2) = .831$,
for $n=1000$:	$SEV(\mu > 0.2) = .115$.

The facts underlying criticism #4 are also erroneously taken as grounding the claim:

"All nulls are false."

This confuses the true claim that with large enough sample size, a test has power to detect any discrepancy from the null however small, with the false claim that all nulls are false.

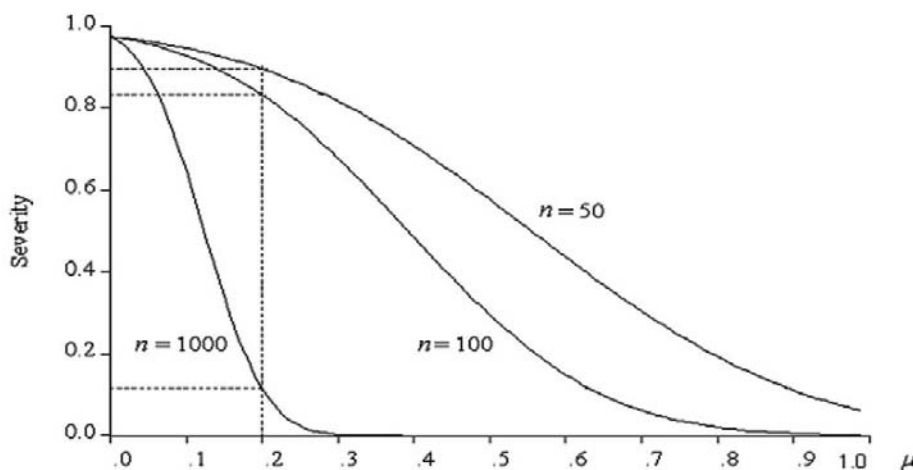


Figure 5. Severity associated with inference $\mu > 0.2$, $d(x_0)=1.96$, and different sample sizes n .

The tendency to view tests as automatic recipes for rejection gives rise to another well-known canard:

(#5) Whether there is a statistically significant difference from the null depends on which is the null and which is the alternative.

The charge is made by considering the highly artificial case of two point hypotheses such as: $\mu=0$ vs. $\mu=.8$. If the null is $\mu=0$ and the alternative is $\mu=.8$ then $\bar{x}=0.4$ (being $2\sigma_x$ from 0) “rejects” the null and declares there is evidence for .8. On the other hand if the null is $\mu=.8$ and the alternative is $\mu=0$, then observing $\bar{x}=0.4$ now rejects .8 and finds evidence for 0. It appears that we get a different inference depending on how we label our hypothesis! Now the hypotheses in a N-P test must exhaust the space of parameter values, but even entertaining the two point hypotheses, the fallacy is easily exposed. Let us label the two cases:

Case 1: $H_0: \mu=0$ vs. $H_1: \mu=.8$, **Case 2:** $H_0: \mu=.8$ vs. $H_1: \mu=0$.

In case 1, $\bar{x}=0.4$ is indeed evidence of *some* discrepancy from 0 in the positive direction, but it is exceedingly poor evidence for a discrepancy as large as .8 (see figure 2). Even without the calculation that shows $SEV(\mu > .8)=.023$, we know that $SEV(\mu > .4)$ is only .5, and so there are far less grounds for inferring an even larger discrepancy⁵.

⁵We obtain the standardized value by considering the sample mean ($\bar{x}=.4$) minus the hypothesize μ (.8), in standard deviation units ($\sigma_x=.2$), yielding $z = -2$, and thus $P(Z < -2)=.023$.

In case 2, the test is looking for discrepancies from the null (which is .8) in the negative direction. The outcome $\bar{x}=0.4$ ($d(\mathbf{x}_0)=-2.0$) is evidence that $\mu \leq .8$ (since $SEV(\mu \leq .8)=.977$), but there are terrible grounds for inferring the alternative $\mu=0$!

In short, case 1 asks if the true μ exceeds 0, and $\bar{x}=.4$ is good evidence of some such positive discrepancy (though poor evidence it is as large as .8); while case 2 asks if the true μ is less than .8, and again $\bar{x}=.4$ is good evidence that it is. Both these claims are true. In neither case does the outcome provide evidence for the point alternative, .8 and 0 respectively. So it does not matter which is the null and which is the alternative, and criticism #5 is completely scotched.

Note further that in a proper test, the null and alternative hypotheses must exhaust the parameter space, and thus, “point-against-point” hypotheses are at best highly artificial, at worst, illegitimate. What matters for the current issue is that the error statistical tester never falls into the alleged inconsistency of inferences depending on which is the null and which is the alternative.

We now turn our attention to cases of statistically insignificant results. Overly high power is problematic in dealing with significant results, but with insignificant results, the concern is the test is not powerful enough.

2.5 *Fallacies of acceptance: errors in interpreting statistically insignificant results*

(#6) Statistically insignificant results are taken as evidence that the null hypothesis is true.

We may call this the fallacy of interpreting insignificant results (or the fallacy of “acceptance”). The issue relates to a classic problem facing general hypothetical deductive accounts of confirmation: positive instances “confirm” or in some sense count for generalizations. Unlike logics of confirmation or hypothetico-deductive accounts, the significance test reasoning, and error statistical tests more generally, have a very clear basis for denying this. An observed accordance between data and a null hypothesis “passes” the null hypothesis, i.e., condition (S-1) is satisfied. But such a passing result is *not* automatically evidence *for* the null hypothesis, since the test might not have had much chance of detecting departures even if they existed. So what is called for to avoid the problem is precisely the second requirement for severity (S-2). This demands considering error probabilities, the distinguishing feature of an error statistical account.

Now the simple Fisherian significance test, where the result is either to falsify the null or not, leaves failure to reject in some kind of limbo. That is why Neyman and Pearson introduce the alternative hypothesis and the corresponding notion of power. Consider our familiar test T_α . Affirming the null is to rule out a discrepancy $\gamma > 0$. It is unwarranted to claim to have evidence for the null if the test had little capacity (probability) of producing a worse fit with the null even though the null is false, i.e. $\mu > 0$. In the same paper addressing Carnap, Neyman makes this

point (p. 41)⁶, although it must be conceded that it is absent from his expositions of tests. The severity account makes it an explicit part of interpreting tests (note that $d(x_0) = \frac{(\bar{x} - \mu_0)}{\sigma_x}$):

(a) If there is a very low probability that $d(\mathbf{x}_0)$ would have been larger than it is, even if μ exceeds μ_1 , then $\mu \leq \mu_1$ passes the test with low severity, i.e. $SEV(\mu \leq \mu_1)$ is low.

By contrast:

(b) If there is a very high probability that $d(\mathbf{x}_0)$ would have been larger than it is, were μ to exceed μ_1 , then $\mu \leq \mu_1$ passes the test with high severity, i.e. $SEV(\mu \leq \mu_1)$ is high.

To see how formal significance tests can encapsulate this, consider testing $H_0: \mu=0$ vs. $H_1: \mu > 0$, and obtaining a *statistically insignificant* result: $d(\mathbf{x}_0) \leq 1.96$. We have (S-1): $\mathbf{x}_0$ agrees with H_0 since $d(\mathbf{x}_0) \leq 1.96$.

(S-1): $\mathbf{x}_0$ agrees with H_0 since $d(\mathbf{x}_0) \leq 1.96$. We wish to determine if it is good evidence for $\mu \leq \mu_1$, where $\mu_1 = \mu_0 + \gamma$, by evaluating the probability that test T_α would have produced a more significant result (i.e. $d(\mathbf{X}) > d(\mathbf{x}_0)$), if $\mu > \mu_1$:

$$SEV(T_\alpha, \mathbf{x}_0, \mu \leq \mu_1) = P(d(\mathbf{X}) > d(\mathbf{x}_0); \mu > \mu_1).$$

It suffices to evaluate this at $\mu_1 = \mu_0 + \gamma$ because the probability increases for $\mu > \mu_1$. So, if we have good evidence that $\mu \leq \mu_1$ we have even better evidence that $\mu \leq \mu_2$, where μ_2 exceeds μ_1 (since the former entails the latter).

Rather than work through calculations, it is revealing to report several appraisals graphically. Figure 6 shows severity curves for test T_α , where $\sigma=2$, $n=100$, based on three different insignificant results:

$$d(\mathbf{x}_0)=1.95 \ (\bar{x}=.392), d(\mathbf{x}_0)=1.5 \ (\bar{x}=.3), d(\mathbf{x}_0)=.50 \ (\bar{x}=.1).$$

As before, let a statistically significant result require $d(\mathbf{x}_0) > 1.96$. None of the three insignificant outcomes provide strong evidence that the null is precisely true, but what we want to do is find the smallest discrepancy that each rules out with severity.

For illustration, we consider a particular fixed inference of the form $(\mu \leq \mu_1)$, and compare severity assessments for different outcomes. The low probabilities

⁶In the context where H_0 had not been “rejected”, Neyman insists, it would be “dangerous” to regard this as confirmation of H_0 if the test in fact had little chance of detecting an important discrepancy from H_0 , even if such a discrepancy were present. On the other hand if the test had appreciable power to detect the discrepancy, the situation would be “radically different.” Severity logic for insignificant results has the same pattern except that we consider the actual insignificant result, rather than the case where data just misses the cut-off for rejection.

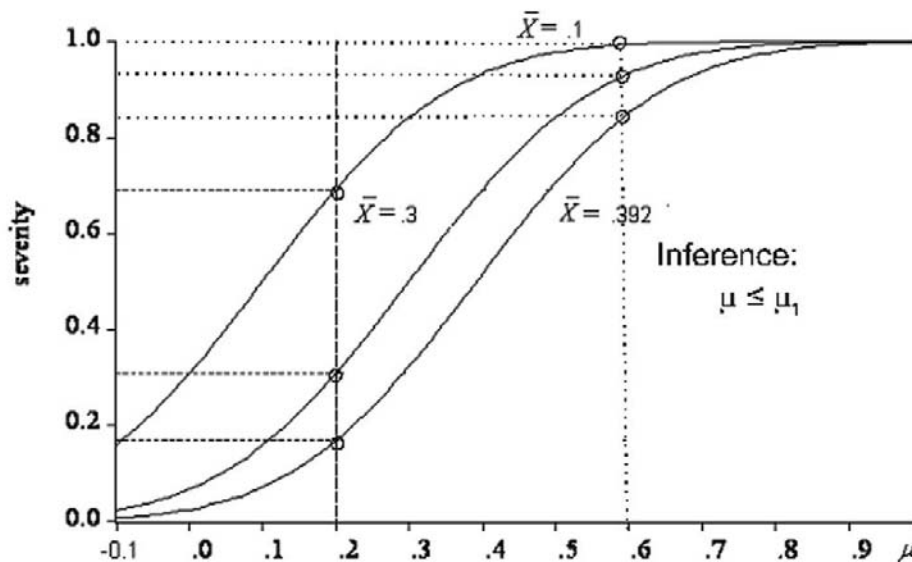


Figure 6. Insignificant result. Severity associated with inference $\mu \leq .2$ with different outcomes x_0 .

associated with the severity assessment of $\mu \leq .2$ indicates that, in all three cases, the claim that the discrepancy is $\mu \leq .2$ is unwarranted (to the degrees indicated):

$$\begin{array}{ll} \text{for } d(\mathbf{x}_0)=1.95 \ (\bar{x}=0.39) : & SEV(\mu \leq 0.2) = .171, \\ \text{for } d(\mathbf{x}_0)=1.5 \ (\bar{x}=0.3) : & SEV(\mu \leq 0.2) = .309, \\ \text{for } d(\mathbf{x}_0)=0.5 \ (\bar{x}=0.1) : & SEV(\mu \leq 0.2) = .691. \end{array}$$

So it would be fallacious (to different degrees) to regard these as warranting $\mu \leq 0.2$. To have a contrast, observe that inferring $\mu \leq .6$ is fairly warranted (to different degrees) for all three outcomes:

$$\begin{array}{ll} \text{for } d(\mathbf{x}_0)=1.95 \ (\bar{x}=0.39) : & SEV(\mu \leq 0.6) = .853, \\ \text{for } d(\mathbf{x}_0)=1.5 \ (\bar{x}=0.3) : & SEV(\mu \leq 0.6) = .933, \\ \text{for } d(\mathbf{x}_0)=0.5 \ (\bar{x}=0.1) : & SEV(\mu \leq 0.6) = .995. \end{array}$$

Working in the reverse direction, it is instructive to fix a high severity value, say, .95, and ascertain, for different outcomes, the discrepancy that may be ruled out with severity .95. For $\bar{x}=0.39$, $SEV(\mu \leq .72)=.95$, for $\bar{x}=0.3$, $SEV(\mu \leq .62)=.95$, and $\bar{x}=0.1$, $SEV(\mu \leq .43)=.95$. Although none of these outcomes warrants ruling out all positive discrepancies at severity level .95, we see that the smaller the observed outcome $\bar{x}$, the smaller is the μ_1 value such that $SEV(\mu \leq \mu_1) = .95$.

It is interesting to note that the severity curve associated with $d(\mathbf{x}_0) = 1.95$ virtually coincides with the power curve since $c_\alpha=1.96$ for $\alpha=.025$. The power of

the test to detect μ_1 gives the lower bound for the severity assessment for $(\mu \leq \mu_1)$; this is the lowest it could be when an insignificant result occurs. High power at μ_1 ensures that insignificance permits inferring with high severity that $\mu \leq \mu_1$. Thus severity gives an inferential justification for the predesignated power, but it goes further. Once the result is available, it directs us to give a more informative inference based on $d(\mathbf{x}_0)$.

P-values are Not Posterior Probabilities of H_0

The most well-known fallacy in interpreting significance tests is to equate the p-value with a posterior probability on the null hypothesis. In legal settings this is often called the *prosecutor's fallacy*. Clearly, however:

- (i) $P(d(\mathbf{X}) \geq d(\mathbf{x}_0); H_0)$ is *not* equal to (ii) $P(H_0 \mid d(\mathbf{X}) \geq d(\mathbf{x}_0))$.

The p-value assessment in (i) refers only to the sampling distribution of the test statistic $d(\mathbf{X})$; and there is no use of *prior probabilities*, as would be necessitated in (ii). In the frequentist context, $\{d(\mathbf{X}) > 1.96\}$ is an event and *not* a statistical hypothesis. The latter must assign probabilities to outcomes of an experiment of interest.

Could we not regard such events as types of hypotheses or at least predictions? Sure. But scientific hypotheses of the sort statistical methods have been developed to test are not like that. Moreover, no prior probabilities are involved in (i): it is just the usual computation of probabilities of events “calculated under the assumption” of a given statistical hypothesis and model. (It is not even correct to regard this as a conditional probability.) We are prepared to go further: it seems to us an odd way of talking to regard the null hypothesis as evidence for the event $\{d(\mathbf{X}) \leq 1.96\}$, or for its high probability. It is simply to state what is deductively entailed by the probability model and hypothesis. Most importantly, the statistical hypotheses we wish to make inferences about are not events; trying to construe them as such involves fallacies and inconsistencies (we return to this in Section 3).

Some critics go so far as to argue that despite it being fallacious (to construe error probabilities as posterior probabilities of hypotheses):

(#7) Error probabilities are invariably misinterpreted as posterior probabilities.

Our discussion challenges this allegation that significance tests (and confidence intervals) are invariably used in “bad-faith”. We have put forward a rival theory as to the meaning and rationale for the use of these methods in science: properly interpreted, they serve to control and provide assessments of the severity with which hypotheses pass tests based on evidence. The quantitative aspects arise in the form of degrees of severity and sizes of discrepancies detected or not. This rival theory seems to offer a better explanation of inductive reasoning in science.

2.6 Relevance for finite samples

Some critics charge that because of their reliance on frequentist probability:

(#8) **Error statistical tests are justified only in cases where there is a very long (if not infinite) series of repetitions of the same experiment.**

Ironically, while virtually all statistical accounts appeal to good asymptotic properties in their justifications, the major asset of error statistical methods is in being able to give assurances that we will not be too far off with specifiable finite samples. This is a crucial basis both for planning tests and in critically evaluating inferences post-data. Pre-data the power has an important role to play in ensuring that test T_α has enough ‘capacity’, say $(1-\alpha)$, to detect a discrepancy of interest γ for $\mu_1=\mu_0+\gamma$. To ensure the needed power one often has no other option but to have a large enough sample size n . How large is ‘large enough’ is given by solving the probabilistic equation:

$$\text{to ensure: } P(d(\mathbf{X}) > c_\alpha; \mu_1) = (1-\beta), \text{ set: } n = \{[(c_\alpha - c_\beta)\sigma]/\gamma\}^2,$$

where c_β is the threshold such that $P(Z \leq c_\beta) = \beta$ for $Z \sim \mathbf{N}(0, 1)$.

Numerical example. Consider test T_α with $\mu_0=0$, $\alpha=.025$, $\sigma=2$, and let the substantive discrepancy of interest be $\gamma=.4$. Applying the above formula one can determine that the sample size needed to ensure high enough power, say $(1-\beta)=.90$, to detect such a discrepancy is: $n = \{[(1.96+1.28)(2)]/(.4)\}^2 \approx 262$, i.e., the test needs 262 observations to have .9 power to detect discrepancies $\gamma \geq .4$.

If the sample size needed for informative testing is not feasible, then there are grounds for questioning the value of the inquiry, but not for questioning the foundational principles of tests.

This points to a central advantage of the error statistical approach in avoiding the limitations of those accounts whose reliability guarantees stem merely from asymptotic results, that is for n going to infinity. In particular, a test is consistent against some alternative μ_1 when its power $P(d(\mathbf{X}) > c_\alpha; \mu_1)$ goes to one as n goes to infinity. This result, however, is of no help in assessing, much less ensuring, the reliability of the test in question for a given n .

Considering discrepancies of interest restricts the latitude for test specification, not only in choosing sample sizes, but in selecting test statistics that permit error probabilities to be ‘controlled’ despite unknowns. We now turn to this.

2.7 Dealing with “Nuisance” Parameters

In practice, a more realistic situation arises when, in the above simple Normal model ($\mathcal{M}$), both parameters, μ and σ^2 , are *unknown*. Since the primary inference concerns μ and yet σ^2 is needed to complete the distribution, it is often called a “nuisance” parameter. Given the way nuisance parameters are handled in this

approach, changes to the testing procedure are rather minor, and the reasoning is unchanged. That is why we were able to keep to the simpler case for exposition. To illustrate, consider

EXAMPLE 2. Test T_α^* , when σ^2 is *unknown*.

First, the test statistic now takes the form: $d^*(\mathbf{X}) = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s}$, where $s^2 = \frac{1}{n-2} \sum_{k=1}^n (X_k - \bar{X})^2$ is the sample variance, that provides an unbiased estimator of σ^2 .

Second, the sampling distributions of $d^*(\mathbf{X})$, under both the null (H_0) and the alternative (H_1) hypotheses, are no longer Normal, but *Student's t*; see [Lehmann, 1986]. What matters is that the distribution of $d^*(\mathbf{X})$ under H_0 does not involve the nuisance parameter σ^2 ; the only difference is that one needs to use the Student's t instead of the Normal tables to evaluate c_α corresponding to a particular α . The distribution of $d^*(\mathbf{X})$ under H_1 *does* involve σ^2 but only through the non-centrality parameter affecting the power: $\delta = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}$. In practice one replaces σ with its estimate s when evaluating the power of the test at μ_1 — likewise for severity. All the other elements of test T_α remain the same for T_α^* , including the form of the test rule.

Ensuring error statistical calculations free of a nuisance parameter is essential for attaining objectivity: the resulting inferences are not threatened by unknowns. This important desideratum is typically overlooked in foundational discussions and yet the error statistical way of satisfying it goes a long way toward answering the common charge that:

(#9) Specifying statistical tests is too arbitrary.

In a wide class of problems, the error statistician attains freedom from a nuisance parameter by conditioning on a sufficient statistic for it; see [Cox and Hinkley, 1974], leading to a uniquely appropriate test. This ingenious way of dealing with nuisance parameters stands in contrast with Bayesian accounts that require prior probability distributions for each unknown quantity. (Nuisance parameters also pose serious problems for pure likelihood accounts; see [Cox, 2006]). Once this is coupled with the requirement that the test statistics provide plausible measures of “agreement”, the uniquely appropriate test is typically overdetermined: one can take one’s pick for the rationale (appealing to Fisherian, Neyman-Pearsonian, or severity principles).

2.8 Severe Testing and Confidence Interval (CI) Estimation

In CI estimation procedures, a statistic is used to set upper or lower (1-sided) or both (2-sided) bounds. For a parameter, say μ , a $(1-\alpha)$ CI estimation procedure leads to estimates of form: $\mu = \bar{x} \pm e$. Critics of significance tests often allege:

(#10) We should be doing confidence interval estimation rather than significance tests.

Although critics of significance tests often favor CI's, it is important to realize that CI's are still squarely within the error statistical paradigm. In fact there is a precise duality relationship between $(1-\alpha)$ CI's and significance tests: the CI contains the parameter values that would not be rejected by the given test at the specified level of significance [Neyman, 1935]. It follows that the $(1-\alpha)$ one sided interval corresponding to test T_α is:

$$\mu > \bar{X} - c_\alpha(\sigma/\sqrt{n}).$$

In particular, the 97.5% CI estimator corresponding to test T_α is:

$$\mu > \bar{X} - 1.96(\sigma/\sqrt{n}).$$

Now it is true that the confidence interval gives a data-dependent result, but so does our post-data interpretation of tests based on severity. Moreover, confidence intervals have their own misinterpretations and shortcomings that we need to get around.

A well known fallacy is to construe $(1-\alpha)$ as the degree of probability to be assigned the particular interval estimate formed, once $\bar{X}$ is instantiated with $\bar{x}$. Once the estimate is formed, either the parameter is or is not contained in it. One can say only that the particular estimate arose from a procedure which, with high probability, $(1-\alpha)$, would contain the true value of the parameter, whatever it is. This affords an analogous "behavioristic" rationale for confidence intervals as we saw with tests: Different sample realizations $\mathbf{x}$ lead to different estimates, but one can ensure that $(1-\alpha)100\%$ of the time the true parameter value μ , whatever it may be, will be included in the interval formed. Just as we replace the behavioristic rationale of tests with the inferential one based on severity, we do the same with confidence intervals.

The assertion $\mu > \bar{x} - c_\alpha(\sigma/\sqrt{n})$ is the one-sided $(1-\alpha)$ interval corresponding to the test T_α and indeed, for the particular value $\mu_1 = \bar{x} - c_\alpha(\sigma/\sqrt{n})$, the severity with which the inference $\mu > \mu_1$ passes test T_α is $(1-\alpha)$. The severity rationale for applying the rule and inferring $\mu > \bar{x} - c_\alpha(\sigma/\sqrt{n})$ might go as follows:

Suppose this assertion is false, e.g., suppose $\mu_1 = \bar{x} - 1.96(\sigma/\sqrt{n})$. Then the observed mean is 1.96 standard deviations in excess of μ_1 . Were μ_1 the mean of the mechanism generating the observed mean, then with high probability (.975) a result less discordant from μ_1 would have occurred. (For even smaller values of μ_1 this probability is increased.)

However, our severity construal also demands breaking out of the limitations of confidence interval estimation. In particular, in the theory of confidence intervals, a single confidence level is prespecified and the one interval estimate corresponding to this level is formed as the inferential report. The resulting interval is sometimes used to perform statistical tests: Hypothesized values of the parameter are accepted (or rejected) according to whether they are contained within (or outside) the resulting interval estimate. The same problems with automatic uses of tests with a single prespecified choice of significance level α reappear in

the corresponding confidence interval treatment. Notably, predesignating a single choice of confidence level, $(1-\alpha)$, is not enough.

Here is why: A $(1-\alpha)$ CI corresponds to the set of null hypotheses that an observed outcome would not be able to reject with the corresponding α -level test. In our illustrative example of tests, the null value is fixed (we chose 0), and then the sample mean is observed. But we could start with the observed sample mean, and consider the values of μ that would not be rejected, were they (rather than 0) the null value. This would yield the corresponding CI. That is, the observed mean is not sufficiently greater than any of the values in the CI to reject them at the α -level. But as we saw in discussing severity for insignificant results, this does not imply that there is good evidence for each of the values in the interval: many values in the interval pass test T_α with very low severity with $\mathbf{x}_0$. Yet a report of the CI estimate is tantamount to treating each of the values of the parameter in the CI on a par, as it were. That some values are well, and others poorly, warranted is not expressed. By contrast, for each value of μ in the CI, there would be a different answer to the question: How severely does $\mu > \mu_1$ pass with $\mathbf{x}_0$? The severity analysis, therefore, naturally leads to a sequence of inferences, or series of CIs, that are and are not warranted at different severity levels⁷.

3 ERROR STATISTICS VS. THE LIKELIHOOD PRINCIPLE

A cluster of debates surrounding error statistical methods, both in philosophy and in statistical practice, reflects contrasting answers to the question: what information is relevant for evidence and inference? Answers, in turn, depend on assumptions about the nature of inductive inference and the roles of probabilistic concepts in inductive inference. We now turn to this.

Consider a conception of evidence based just on likelihoods: data $\mathbf{x}_0$ is evidence for H so long as:

- (a) $P(\mathbf{x}_0; H) = \text{high (or maximal)}$
- (b) $P(\mathbf{x}_0; H \text{ is false}) = \text{low.}$

Although at first blush these may look like the two conditions for severity ((S-1) and (S-2)), conditions (a) and (b) together are tantamount to a much weaker requirement: $\mathbf{x}_0$ is evidence for H so long as H is more likely on data $\mathbf{x}_0$ than on the denial of H — referring to the mathematical notion of likelihood. To see that this is scarcely sufficient for severity, consider a familiar example.

⁷A discussion of various attempts to consider a series of CI's at different levels, confidence curves [Birnbaum, 1961], p-value functions [Poole, 1987], consonance intervals [Kempthorne and Folks, 1971] and their relation to the severity evaluation is beyond the scope of this paper; see [Mayo and Cox, 2006].

Maximally Likely alternatives. H_0 might be that a coin is fair, and $\mathbf{x}_0$ the result of n flips of the coin. For each of the 2^n possible outcomes there is a hypothesis H_i^* that makes the data $\mathbf{x}_i$ maximally likely. For an extreme case, H_i^* can assert that the probability of heads is 1 just on those tosses that yield heads, 0 otherwise. For any $\mathbf{x}_i$, $P(\mathbf{x}_i; H_0)$ is very low and $P(\mathbf{x}_i; H_i^*)$ is high — one need only choose for (a) the statistical hypothesis that renders the data maximally likely, i.e., H_i^* . So the fair coin hypothesis is always rejected in favor of H_i^* , even when the coin is fair. This violates the severity requirement since it is guaranteed to infer evidence of discrepancy from the null hypothesis even if it is true. The severity of ‘passing’ H_i^* is minimal or 0. (Analogous examples are the “point hypotheses” in [Cox and Hinkley, 1974, p. 51], Hacking’s [1965] “tram car” example, examples in [Mayo, 1996; 2008].)

This takes us to the key difference between the error statistical perspective and contrasting statistical philosophies; namely that to evaluate and control error probabilities requires going beyond relative likelihoods.

3.1 *There is Often Confusion About Likelihoods*

The distribution of the sample $\mathbf{X}$ assigns probability (or density) to each possible realization $\mathbf{x}$, under some fixed value of the parameter θ , i.e. $f(\mathbf{x}; \theta)$. In contrast, the likelihood assigns probability (or density) to a particular realization $\mathbf{x}$, under different values of the unknown parameter θ . Since the data $\mathbf{x}$ are fixed at $\mathbf{x}_0$ and the parameter varies, the likelihood is defined as proportional to $f(\mathbf{x}; \theta)$ but viewed as a function of the parameter θ :

$$L(\theta; \mathbf{x}_0) \propto f(\mathbf{x}_0; \theta) \text{ for all } \theta \in \Theta.$$

Likelihoods do not obey the probability axioms, for example, the sum of the likelihoods of a hypothesis and its denial is not one.

Hacking [1965] is known for having championed an account of comparative support based on what he called the “law of likelihood”: data $\mathbf{x}_0$ support hypothesis H_1 less than H_2 if the latter is more likely than the former, i.e., $P(\mathbf{x}_0; H_2) > P(\mathbf{x}_0; H_1)$; when H_2 is composite, one takes the maximum of the likelihood over the different values of θ admitted by H_2 . From a *theory of support* Hacking gets his *theory of testing* whereby, “an hypothesis should be rejected if and only if there is some rival hypothesis much better supported than it is...” [Hacking, 1965, p. 89].

Hacking [1980] distanced himself from this account because examples such as the one above illustrate that “there always is such a rival hypothesis, *viz.* that things just had to turn out the way they actually did” [Barnard, 1972, p. 129]. Few philosophers or statisticians still advocate a pure likelihood account of evidence (exceptions might be [Rosenkrantz, 1977; Sober, 2008], among philosophers, and [Royall, 1997] among statisticians). However, many who would deny that relative likelihoods are all that is needed for inference still regard likelihoods as all that is needed to capture the import of the data. For example, a Bayesian may hold

that inference requires likelihoods plus prior probabilities while still maintaining that the evidential import of the data is exhausted by the likelihoods. This is the gist of a general principle of evidence known as the *Likelihood Principle* (LP). Disagreement about the LP is a pivot point around which much of the philosophical debate between error statisticians and Bayesians has long turned. Holding the LP runs counter to distinguishing data on grounds of error probabilities of procedures.⁸

“According to Bayes’s theorem, $P(x|\mu)$...constitutes the entire evidence of the experiment, that is, it tells all that the experiment has to tell. More fully and more precisely, *if y is the datum of some other experiment, and if it happens that $P(x|\mu)$ and $P(y|\mu)$ are proportional functions of μ (that is, constant multiples of each other), then each of the two data x and y have exactly the same thing to say about the values of μ ...*” [Savage, 1962, p. 17]

The italicized portion defines the LP. If a methodology allows data to enter only through the likelihoods, then clearly likelihoods contain all the import of the data — for that methodology. The philosophical question is whether relevant information is not thereby being overlooked. The holder of the LP considers the likelihood of the actual outcome, i.e., just $d(\mathbf{x}_0)$, whereas the error statistician needs to consider, in addition, the sampling distribution of $d(\mathbf{X})$ or other statistic being used in inference. In other words, an error statistician could use likelihoods in arriving at (S-1) the condition of accordance or fit with the data, but (S-2) additionally requires considering the probability of outcomes $\mathbf{x}$ that accord less well with a hypotheses of interest H , were H false. In the error statistical account, drawing valid inferences from the data $\mathbf{x}_0$ that happened to be observed is crucially dependent on the relative frequency of outcomes other than the one observed, as given by the appropriate sampling distribution of the test statistic.

3.2 Paradox of Optional Stopping

The conflict we are after is often illustrated by a two-sided version of our test T . We have a random sample from a Normal distribution with mean μ and standard deviation 1, i.e.,

$$X_k \sim N(\mu, 1), \quad k = 1, 2, \dots, n,$$

and wish to test the hypotheses:

$$H_0: \mu = 0, \text{ vs. } H_1: \mu \neq 0.$$

To ensure an overall significance level of .05, one rejects the null whenever $|\bar{x}| > (1.96/\sqrt{n})$. However, instead of fixing the sample size in advance, we are to let n

⁸A weaker variation on the LP holds that likelihoods contain all the information within a given experiment, whereas the “strong” LP refers to distinct experiments. Here LP will always allude to the strong likelihood principle.

be determined by a *stopping rule*:

keep sampling until $|\bar{x}| > (1.96/\sqrt{n})$.

The probability that this rule will stop in a finite number of trials is 1, regardless of the true value of μ ; it is a *proper* stopping rule. Whereas with n fixed in advance, such a test has a type 1 error probability of .05, with this stopping rule the test would lead to an *actual* significance level that would differ from, and be greater than .05. This is captured by saying that significance levels are sensitive to the stopping rule; and there is considerable literature as to how to adjust the error probabilities in the case of ‘optional stopping’, also called *sequential tests* [e.g., Armitage, 1975]. By contrast, since likelihoods are unaffected by this stopping rule, the proponent of the LP denies there really is an evidential difference between the cases where n is fixed and where n is determined by the stopping rule⁹. To someone who holds a statistical methodology that satisfies the LP, it appears that:

(#11) Error statistical methods take into account the intentions of the scientists analyzing the data.

In particular, the inference depends on whether or not the scientist intended to stop at n or intended to keep going until a statistically significant difference from the null was found. The charge in #11 would seem to beg the question against the error statistical methodology which has perfectly objective ways to pick up on the effect of stopping rules: far from intentions “locked up in the scientist’s head” (as critics allege), the manner of generating the data alter error probabilities, and hence severity assessments. As is famously remarked in [Edwards *et al.*, 1963]: “The likelihood principle emphasized in Bayesian statistics implies, ... that the rules governing when data collection stops are irrelevant to data interpretation. This irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels — in the sense of [Neyman and Pearson, 1933, p. 239]. While it may restore “simplicity and freedom” it does so at the cost of being unable to adequately control error probabilities [Berger and Wolpert, 1988; Cox and Hinkley, 1974; Kadane *et al.*, 1999; Mayo and Kruse, 2001; Cox and Mayo, 2010].

3.3 The Reference Bayesians and the Renunciation of the LP

All error probabilistic notions are based on the sampling distribution of a statistic, and thus for an error statistician reasoning from data $\mathbf{x}_0$ always depends on considering how a procedure would handle outcomes other than $\mathbf{x}_0$; it is necessary to consider how often the result would occur in hypothetical repetitions. This conflicts with the likelihood principle (LP). Therefore, objectivity for the error

⁹Birnbaum [1962] argued that the LP follows from apparently plausible principles of conditionality and sufficiency. A considerable literature exists, see [Barnett, 1999]. Mayo [2010b] has recently argued that this “proof” is fallacious.

statistician entails violating the LP — long held as the core of Bayesianism [Mayo, 1983; 1985; 1996].

In fact Bayesians have long argued for foundational superiority over frequentist error statisticians on grounds that they uphold, while frequentists violate, the likelihood principle (LP), leading the latter into Bayesian *incoherency*. Frequentists have long responded that having a good chance of getting close to the truth and avoiding error is what matters [Cox and Hinkley, 1974]. However, many Bayesian statisticians these days, seem to favor the use of conventionally chosen or “reference” Bayesian priors, both because of the difficulty of eliciting subjective priors, and the reluctance of many scientists to allow subjective beliefs to overshadow the information provided by data. These reference priors however, violate the LP.

Over the past few years, leading developers of reference Bayesian methods [Bernardo, 2005; Berger, 2004] concede that desirable reference priors force them to consider the statistical model leading to violations of basic principles, such as the likelihood principle and the stopping rule principle; see [Berger and Wolpert, 1988]. Remarkably, they are now ready to admit that “violation of principles such as the likelihood principle is the price that has to be paid for objectivity” [Berger, 2004]. Now that the reference-Bayesian concedes that violating the LP is necessary for objectivity there may seem to be an odd sort of agreement between the reference Bayesian and the error statistician.

Do the concessions of reference Bayesians bring them closer to the error statistical philosophy? To even consider this possibility one would need to deal with a crucial point of conflict as to the basic role of probability in induction. Although Bayesians disagree among themselves about both the interpretation of posterior probabilities, and their numerical values, they concur that:

“what you ‘really’ want are posterior probabilities for different hypotheses.”

It is well known that error probabilities differ from posteriors. In a variation on the charge of misinterpretation in (#6), critics seek examples where:

“p-values conflict with Bayesian posteriors,”

leading to results apparently counterintuitive even from the frequentist perspective. We consider the classic example from statistics.

(Two-sided) Test of a Mean of a Normal Distribution

The conflict between p-values and Bayesian posteriors often considers the familiar example of the two sided $T_{2\alpha}$ test for the hypotheses:

$$H_0:\mu = 0, \text{ vs. } H_1:\mu \neq 0.$$

The difference between p-values and posteriors are far less marked with one-sided tests, e.g., [Pratt, 1977; Cassella and Berger, 1987]. Critics observe:

“If $n = 50$ one can classically ‘reject H_0 at significance level $p = .05$,’ although $P(H_0|x) = .52$ (which would actually indicate that the evidence favors H_0).” ([Berger and Sellke, 1987, p. 113], we replace Pr with P for consistency.)

Starting with a high enough prior probability to the point null (or, more correctly, to a small region around it), they show that an α significant difference can correspond to a posterior probability in H_0 that is not small. Where Bayesians take this as problematic for significance testers, the significance testers balk at the fact that use of the recommended priors can result in highly significant results being construed as no evidence against the null — or even “that the evidence favors H_0 .” If $n=1000$, a result statistically significant at the .05 level leads to a posterior to the null of .82! [Berger and Sellke, 1987]. Here, statistically significant results — results that we would regard as passing the non-null hypothesis severely — correspond to an *increase* in probability from the prior (.5) to the posterior.

What justifies this prior? The Bayesian prior probability assignment of .5 to H_0 , the remaining .5 probability being spread out over the alternative parameter space, (e.g., recommended by Jeffreys [1939]) is claimed to offer an “objective” assessment of priors: the priors are to be read off from a catalogue of favored “reference” priors, no subjective beliefs are to enter. It is not clear how this negative notion of objectivity secures the assurance we would want of being somewhere close to the truth. The Bayesians do not want too small a prior for the null since then evidence against the null is merely to announce that an improbable hypothesis has become more improbable. Yet the spiked concentration of belief (“weight”) in the null is at odds with the prevailing use of null hypotheses as simply a standard from which one seeks discrepancies. Finally, these examples where p-values differ from posteriors create a tension between the posterior probability in a testing context and the corresponding (highest probability) Bayesian confidence interval: the low posterior indicates scarce evidence against the null even though the null value is outside the corresponding Bayesian confidence interval [Mayo, 2005].

Some examples strive to keep within the frequentist camp: to construe a hypothesis as a random variable, it is imagined that we sample randomly from a population of hypotheses, some proportion of which are assumed to be true. The percentage “initially true” serves as the prior probability for H_0 . This gambit commits what for a frequentist would be a fallacious instantiation of probabilities:

50% of the null hypotheses in a given pool of nulls are true.

This particular null hypothesis H_0 was randomly selected from this pool.

Therefore $P(H_0 \text{ is true}) = .5$.

Even allowing that the probability of a randomly selected hypothesis taken from an “urn” of hypotheses, 50% of which are true, is .5, it does not follow that

this particular hypothesis, the one we happened to select, has a probability of .5, however probability is construed [Mayo, 1997; 2005; 2010b].¹⁰ Besides, it is far from clear which urn of null hypotheses we are to sample from. The answer will greatly alter whether or not there is evidence. Finally, it is unlikely that we would ever know the proportion of true nulls, rather than merely the proportion that have thus far not been rejected by other statistical tests! Whether the priors come from frequencies or from “objective” Bayesian priors, there are claims that we would want to say had passed severely that do not get a high posterior.

These brief remarks put the spotlight on the foundations of current-day reference Bayesians — arguably the predominant form of Bayesianism advocated for science. They are sometimes put forward as a kind of half-way house offering a “reconciliation” between Bayesian and frequentist accounts. Granted, there are cases where it is possible to identify priors that result in posteriors that “match” error probabilities, but they appear to mean different things. Impersonal or reference priors are not to be seen as measuring beliefs or even probabilities — they are often improper.¹¹ Subjective Bayesians often question whether the reference Bayesian is not here giving up on the central Bayesian tenets (e.g., [Dawid, 1997; Lindley, 1997]).

4 ERROR STATISTICS IS SELF-CORRECTING: TESTING STATISTICAL MODEL ASSUMPTIONS

The severity assessment of the primary statistical inference depends on the assumptions of the statistical model $\mathcal{M}$ being approximately true. Indeed, all model-based statistical methods depend, for their validity, on satisfying the model assumptions, at least approximately; a crucial part of the objectivity of error statistical methods is their ability to be used for this self-correcting goal.

Some critics would dismiss the whole endeavor of checking model assumptions on the grounds that:

#12 All models are false anyway.

This charge overlooks the key function in using statistical models, as argued by Cox [1995, p. 456]:

“... it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. ... The construction of idealized representations that capture important stable

¹⁰The parallel issue is raised by Bayesian epistemologists; see [Achinstein, 2010; Mayo, 2005; 2010c, 2010d].

¹¹Interestingly, some frequentist error statisticians are prepared to allow that reference Bayesian techniques might be regarded as technical devices for arriving at procedures that may be reinterpreted and used by error statisticians, but for different ends (see [Cox, 2006; Cox and Mayo, 2009; Kass and Wasserman, 1996]).

aspects of such systems is, however, a vital part of general scientific analysis.”

In order to obtain reliable knowledge of “important stable aspects” of phenomena, tests framed within approximately correct models will do — so long as their relevant error probabilities are close to those calculated. Statistical mis-specifications often create sizeable deviations between the calculated or *nominal* error probabilities and the *actual error probabilities* associated with an inference, thereby vitiating error statistical inferences. Since even Bayesian results depend on approximate validity of their statistical models, this might be an area for the Bayesian to employ non-Bayesian methods.

The error statistician pursues testing assumptions using three different types of tools: informal analyses of data plots, non-parametric and parametric tests, and simulation-based methods, including resampling.

Philosophers of science tend to speak of “the data” in a way that does not distinguish the different ways in which a given set of data are modeled, and yet such distinctions are crucial for understanding how reliable tests of assumptions are obtained. In using data to test model assumptions one looks, not at the reduced data in the test statistic (for primary testing), but rather the full data set $\mathbf{x}_0 := (x_1, x_2, \dots, x_n)$. For example, in test T_α , $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$ is a sufficient statistic for parameter μ . That means $\bar{X}$, together with its sampling distribution, contains all the information needed for those inferences. However, $\bar{X}$, by itself, does not provide sufficient information to assess the validity of the model assumptions underlying test T_α above. There are actually four distinct assumptions (table 1).

Table 1 - Simple Normal Model

		$X_k = \mu + u_k, \quad k \in \mathbb{N},$	
[1]	Normality:	$X_k \sim \mathbf{N}(\cdot, \cdot),$	} $k \in \mathbb{N}.$
[2]	constant mean:	$E(X_k) := \mu,$	
[3]	constant variance:	$Var(X_k) := \sigma^2,$	
[4]	Independence:	$\{X_k, \quad k \in \mathbb{N}\}$ is an independent process.	

The inferences about μ depend on the assumptions, but the tests of those assumptions should not depend on the unknowns. The idea underlying model validation is to construct Mis-Specification (M-S) tests using ‘distance’ functions whose distribution under the null (the model is valid) is known, and at the same time they have power against potential departures from the model assumptions. M-S tests can be regarded as posing ‘secondary’ questions to the data as opposed to the primary ones. Whereas primary statistical inferences take place within a specified (or assumed) model $\mathcal{M}$, the secondary inference has to put $\mathcal{M}$ ’s assumptions to the test; so to test $\mathcal{M}$ ’s assumptions, we stand outside $\mathcal{M}$, as it were. The generic form of the hypothesis of interest in M-S tests is:

H_0 : the assumptions of statistical model $\mathcal{M}$ hold for data $\mathbf{x}_0$,

as against the alternative not- H_0 , which, in general, consists of all of the ways one or more of $\mathcal{M}$'s assumptions can founder. However, this alternative $[\mathcal{P} - \mathcal{M}]$, where $\mathcal{P}$ denotes the set of all possible models that could have given rise to data $\mathbf{x}_0$, is too unwieldy. In practice one needs to consider a specific form of departure from H_0 , say H_r , in order to apply a statistical significance test to H_0 and test results must be interpreted accordingly. Since with this restricted alternative H_r , the null and alternative do not exhaust the possibilities (unlike in the N-P test), a statistically significant departure from the null would not warrant inferring the particular alternative in a M-S test H_r , at least not without further testing; see [Spanos, 2000].

In M-S testing, the logic of significance tests is this: We identify a test statistic $\tau(\mathbf{X})$ to measure the distance between what is observed $\mathbf{x}_0$ and what is expected assuming the null hypothesis H_0 holds, so as to derive the distribution of $\tau(\mathbf{X})$ under H_0 . Now the relevant p-value would be:

$$P(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); H_0 \text{ true}) = p,$$

and if it is very small, then there is evidence of violations of the assumption(s) in H_0 . We leave to one side here the particular levels counting as 'small'. A central asset of the error statistical approach to model validation, is its ability to compute the p-value, and other relevant error probabilities, now dealing with erroneous inferences regarding the assumptions.

Although the alternative may not be explicit in this simple (Fisherian) test, the interest in determining what violations have been ruled out with severity leads one to make them explicit. This may be done by considering the particular violations from H_0 that the given test is capable of probing. This goes beyond what, strictly speaking, is found in standard M-S tests; so once again the severity requirement is directing supplements. The upshot for interpreting M-S test results is this: If the p-value is not small, we are entitled only to rule out those departures that the test had enough capacity to detect. In practice, the alternatives may be left vague or made specific. We consider an example of each, the former with a non-parametric test, the latter with a parametric test.

4.1 Runs Test for IID

An example of a non-parametric M-S test for IID (assumptions [2]–[4]) is the well-known runs test. The basic idea is that if the sample $\mathbf{X} := (X_1, X_2, \dots, X_n)$ is random (IID), then one can compare the number of runs expected $E(R)$ in a typical realization of an IID sample with the number of runs observed $R = r$, giving rise to a test statistic:

$$\tau(\mathbf{X}) = \left| [R - E(R)] / \sqrt{\text{Var}(R)} \right|,$$

whose distribution under IID for $n \geq 20$ can be approximated by $N(0, 1)$. The number of runs R is evaluated in terms the residuals:

$$\hat{u}_k = (X_k - \bar{X}), \quad k = 1, 2, \dots, n,$$

where instead of the particular value of each observed residual one records its sign, a “+”, or negative, a “−”, giving rise to patterns of the form:

$$\begin{array}{cccccccccccccccc} ++ & - & ++ & - & +++ & - & + & - & + & -- & +++ & + & - & \cdots \\ \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} & \underbrace{\hspace{1.5cm}} \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & & \end{array}$$

The patterns we are interested in are called *runs*: a sub-sequence of one type (pluses only or minuses only) immediately preceded and succeeded by an element of the other type.

The appeal of such a non-parametric test is that its own validity does not depend on the assumptions of the (primary) model under scrutiny: we can calculate the probability of different numbers of runs just from the hypothesis that the assumption of randomness holds. As is plausible, then, the test based on R takes the form: Reject H_0 iff the observed R differs sufficiently (in either direction) from $E(R)$ — the expected R under the assumption of IID. The p-value is:

$$P(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \text{IID true}) = p.$$

However, since the test is sensitive to any form of departures from the IID assumptions, rejecting the null only warrants inferring a denial of IID. The test itself does not indicate whether the fault lies with one or the other or both assumptions. Combining this test with other misspecification analyses, however, can; [Mayo and Spanos, 2004].

4.2 A Parametric Test of Independence

Let us now compare this to a parametric M-S test. We begin by finding a way to formally express the denial of the assumption in question by means of a parameter value in an encompassing model. In particular, the dependence among the X_k ’s may be formally expressed as an assertion that the correlation between any X_i and X_j for $i \neq j$ is non-zero, which in turn may be parameterized by the following **AutoRegressive** (AR(1)) model:

$$X_k = \beta_0 + \beta_1 X_{k-1} + \varepsilon_t, \quad k = 1, 2, \dots, n.$$

In the context of this encompassing model the independence assumption in [4] can be tested using the parametric hypotheses:

$$H_0 : \beta_1 = 0, \text{ vs. } H_1 : \beta_1 \neq 0.$$

Notice that under H_0 the AR(1) model reduces to $X_k = \mu + u_t$; see table 1. Rejection of the null based on a small enough p-value provides evidence for a violation of independence. Failing to reject entitles us to claim that the departures against which the test was capable of detecting are not present; see [Spanos, 1999] for further discussion.

4.3 Testing Model Assumptions Severely

In practice one wants to perform a variety of M-S tests assessing different subsets of assumptions [1]–[4]; using tests which themselves rely on dissimilar assumptions. The secret lies in shrewd testing strategies: following a logical order of parametric and non-parametric tests, and combining tests that jointly probe several violations with deliberately varied assumptions. This enables one to argue, with severity, that when no departures from the model assumptions are detected, despite all of these distinct probes, the model is adequate for the primary inferences. The argument is entirely analogous to the argument from coincidence that let us rule out values of George's weight gain earlier on.

To render the probing more effective, error statisticians employ data analytic techniques and data plots to get hints of the presence of potential violations, indicating the most fruitful analytic tests to try. In relation to this some critics charge:

#13 Testing assumptions involves illicit data-mining.

The truth of the matter is that data plots provide the best source of information pertaining to potential violations that can be used to guide a more informative and effective probing. Far from being *illicit data mining*, it is a powerful way to get ideas concerning the type of M-S tests to apply to check assumptions most severely. It provides an effective way to probe what is responsible for the observed pattern, much as a forensic clue is used to pinpoint the culprit; see [Spanos, 2000]. The same logic is at the heart of non-parametric tests of assumptions, such as the runs test.

4.4 Residuals provide the Key to M-S testing

A key difference between testing the primary hypotheses of interest and M-S testing is that they pose very different questions to data in a way that renders the tests largely independent of each other. This can be justified on formal grounds using the properties of *sufficiency* and *ancillarity*; see [Cox and Hinkley, 1974].

It can be shown [Spanos, 2007] that, in many cases, including the above example of the simple Normal model, the information used for M-S testing purposes is independent of the information used in drawing primary inferences. In particular, the distribution of the sample for the statistical model in question simplifies as follows:

$$(3) \quad f(\mathbf{y}; \theta) \propto f(\mathbf{s}; \theta) \cdot f(\mathbf{r}), \quad \forall (\mathbf{s}, \mathbf{r}) \in \mathbb{R}_s^m \times \mathbb{R}_r^{n-m}.$$

where the statistics $\mathbf{R}$ and $\mathbf{S}$, are not only independent, but $\mathbf{S}$ is a *sufficient* statistic for $\theta := (\mu, \sigma^2)$ (the unknown parameters of the statistical model) and $\mathbf{R}$ is *ancillary* for θ , i.e. $f(\mathbf{r})$ does *not* depend on θ . Due to these properties, the primary inference about θ can be based solely on the distribution of the sufficient

statistic $f(\mathbf{s}; \theta)$, and $f(\mathbf{r})$ can be used to assess the validity of the statistical model in question.

In the case of the simple Normal model (table 1), the statistics $\mathbf{R}$ and $\mathbf{S}$ take the form:

$$\mathbf{S} := (\bar{X}, s), \text{ where } \bar{X} = \frac{1}{n} \sum_{k=1}^n X_k, \quad s^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2,$$

$$\mathbf{R} := (\hat{v}_3, \dots, \hat{v}_n), \quad \hat{v}_k = \frac{\sqrt{n}\hat{u}_k}{s} = \frac{\sqrt{n}(X_k - \bar{X})}{s} \sim \text{St}(n-1), \quad k = 3, 4, \dots, n,$$

where $(\hat{v}_1, \dots, \hat{v}_n)$ are known as *studentized residuals*; see [Spanos, 2007]. Note that the runs test, discussed above, relies on residuals because it is based on replacing their numerical values with their sign (+ or -). Likewise, the parametric M-S test for independence, placed in the context of the AR(1) model, can be shown to be equivalently based on the auxiliary autoregression in terms of the residuals:

$$\hat{u}_k = \beta_0 + \beta_1 \hat{u}_{k-1} + \varepsilon_t, \quad k=1, 2, \dots, n.$$

The above use of the residuals for model validation is in the spirit of the strategy in [Cox and Hinkley, 1974] to use the conditional distribution $f(\mathbf{x} | \mathbf{s})$ to assess the adequacy of the statistical model. What makes $f(\mathbf{x} | \mathbf{s})$ appropriate for assessing model adequacy is that when $\mathbf{s}$ is a sufficient statistic for θ , $f(\mathbf{x} | \mathbf{s})$ is free of the unknown parameter(s). The simple Poisson and Bernoulli models provide such examples [Cox, 2006, p. 33].

4.5 Further Topics, Same Logic

If one finds violations of the model assumptions then the model may need to be respecified to capture the information not accounted for, but the general discussion of respecification is beyond the scope of this entry; see [Spanos, 2006].

A distinct program of research for the error statistician is to explore the extent to which violations invalidate tests. Thanks to robustness, certain violations of the model assumptions will not ruin the validity of the test concerning the primary hypothesis.

Of particular interest in error statistics is a set of computer-intensive techniques known as resampling procedures, including permutation methods, the bootstrap and Monte Carlo simulations, which are based on empirical relative frequencies. Even without knowing the sampling distribution, one can, in effect generate it by means of these techniques. The logic underlying the generation of these simulated realizations is based on counterfactual reasoning: We ask, ‘what would it be like (in terms of sampling distributions of interest) were we to sample from one or another assumed generating mechanism?’ The results can then be used to empirically construct (by “brute force” some claim) the sampling distributions of any statistic of interest and their corresponding error probabilities.

This is particularly useful in cases where the sampling distribution of an estimator or a test statistic cannot be derived analytically, and these resampling methods

can be used to evaluate it empirically; see [Efron and Tibshirani, 1993]. The same pattern of *counterfactual reasoning* around which severity always turns is involved, thus unifying the methods under the error statistical umbrella. Here, however, the original data are compared with simulated replicas generated under a number of different data generating mechanisms, in order to discern the discrepancy between what was observed and “what it would be like” under various scenarios. It should be noted that model specification is distinct from model selection, which amounts to choosing a particular model within a prespecified family of models; see [Spanos, 2010].

BIBLIOGRAPHY

- [Achinstein, 2010] P. Achinstein. Mill’s Sins or Mayo’s Errors? pp. 170-188 in *Error and Inference*, (ed.) by D. G. Mayo and A. Spanos, Cambridge University Press, Cambridge, 2010.
- [Armitage, 1975] P. Armitage. *Sequential Medical Trials*, 2nd ed, Wiley, NY, 1975.
- [Barnard, 1972] G. A. Barnard. The Logic of Statistical Inference (review of Ian Hacking *The Logic of Statistical Inference*), *British Journal For the Philosophy of Science*, 23: 123-132, 1972.
- [Barnett, 1999] V. Barnett. *Comparative Statistical Inference*, 3rd ed., Wiley, NY, 1999.
- [Bernardo, 2005] J. M. Bernardo. Reference Analysis, pp. 17–90 in *Handbook of Statistics*, vol. 25: Bayesian Thinking, Modeling and Computation, D. K. Dey and C. R. Rao, (eds.), Elsevier, North-Holland, 2005.
- [Berger, 2004] J. Berger. The Case for Objective Bayesian Analysis, *Bayesian Analysis*, 1, 1–17, 2004.
- [Berger and Selke, 1987] J. Berger and T. Sellke. Testing a point-null hypothesis: the irreconcilability of significance levels and evidence, *Journal of the American Statistical Association*, 82: 112-122, 1987.
- [Berger and Wolpert, 1988] J. Berger and R. Wolpert. *The Likelihood Principle*. 2d ed., Institute of Mathematical Statistics, Hayward, CA, 1988.
- [Birnbaum, 1961] A. Birnbaum. Confidence Curves: An Omnibus Technique for Estimation and Testing, *Journal of the American Statistical Association*, 294: 246-249, 1961.
- [Birnbaum, 1962] A. Birnbaum. On the Foundations of Statistical Inference (with discussion), *Journal of the American Statistical Association*, 57: 269-326, 1962.
- [Birnbaum, 1969] A. Birnbaum. Concepts of Statistical Evidence, pp. 112-143 in S. Morgenbesser, P. Suppes, and M. White (eds.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, St. Martin’s Press, NY, 1969.
- [Casella and Berger, 1987] G. Casella and R. Berger. Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem, *Journal of the American Statistical Association*, 82: 106-111, 1987.
- [Cheyne and Worrall, 2006] C. Cheyne and J. Worrall, eds. *Rationality and Reality: Conversations with Alan Musgrave*, Studies in the History and Philosophy of Science, Springer, Dordrecht, 2006.
- [Cox, 1958] D. R. Cox. Some Problems Connected with Statistical Inference, *Annals of Mathematical Statistics*, 29: 357-372, 1958.
- [Cox, 1977] D. R. Cox. The Role of Significance Tests, (with discussion), *Scandinavian Journal of Statistics*, 4:49-70, 1977.
- [Cox, 1995] D. R. Cox. Comment on “Model Uncertainty, Data Mining and Statistical Inference,” by C. Chatfield, *Journal of the Royal Statistical Society*, A 158: 419-466, 1995.
- [Cox, 2006] D. R. Cox. *Principles of Statistical Inference*, Cambridge University Press, Cambridge, 2006.
- [Cox and Hinkley, 1974] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*, Chapman and Hall, London, 1974.

- [Cox and Mayo, 2010] D. R. Cox and D. G. Mayo. Objectivity and Conditionality in Frequentist Inference, pp. 276-304 in Mayo, D.G. and A. Spanos, *Error and Inference*, Cambridge University Press, Cambridge, 2010.
- [Dawid, 1997] A. P. Dawid. Comments on ‘Non-informative priors do not exist’, *Journal of Statistical Planning and Inference*, 65: 159-162, 1997.
- [Edwards et al., 1963] W. Edwards, H. Lindman, and L. Savage. Bayesian Statistical Inference for Psychological Research, *Psychological Review*, 70: 193-242, 1963.
- [Efron and Tibshirani, 1993] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*, Chapman and Hall, London, 1993.
- [Fisher, 1925] R. A. Fisher. *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, 1925.
- [Fisher, 1935] R. A. Fisher. *The Design of Experiments*, Oliver and Boyd, Edinburgh, 1935.
- [Fisher, 1955] R. A. Fisher. Statistical Methods and Scientific Induction, *Journal of the Royal Statistical Society*, B, 17: 69-78, 1955.
- [Fisher, 1956] R. A. Fisher. *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh, 1956.
- [Gigerenzer, 1993] G. Gigerenzer. The Superego, the Ego, and the Id in Statistical Reasoning, pp. 311-39 in Keren, G. and C. Lewis (eds.), *A Handbook of Data Analysis in the Behavioral Sciences: Methodological Issues*, Erlbaum, Hillsdale, NJ, 1993.
- [Godambe and Sprott, 1971] V. Godambe and D. Sprott, eds. *Foundations of Statistical Inference*, Holt, Rinehart and Winston of Canada, Toronto, 1971.
- [Hacking, 1965] I. Hacking. *Logic of Statistical Inference*, Cambridge University Press, Cambridge, 1965.
- [Hacking, 1980] I. Hacking. The Theory of Probable Inference: Neyman, Peirce and Braithwaite, pp. 141-160 in D. H. Mellor (ed.), *Science, Belief and Behavior: Essays in Honour of R.B. Braithwaite*, Cambridge University Press, Cambridge, 1980.
- [Hull, 1988] D. Hull. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*, University of Chicago Press, Chicago, 1988.
- [Jeffreys, 1939] H. Jeffreys. *Theory of Probability*, Oxford University Press, Oxford, 1939.
- [Kadane et al., 1999] J. Kadane, M. Schervish, and T. Seidenfeld. *Rethinking the Foundations of Statistics*, Cambridge University Press, Cambridge, 1999.
- [Kass and Wasserman, 1996] R. E. Kass and L. Wasserman. The Selection of Prior Distributions by Formal Rules, *Journal of the American Statistical Association*, 91: 1343-1370, 1996.
- [Kempthorne and Folks, 1971] O. Kempthorne and L. Folks. *Probability, Statistics, and Data Analysis*, The Iowa State University Press, Ames, IA, 1971.
- [Lehmann, 1986] E. L. Lehmann. *Testing Statistical Hypotheses*, 2nd edition, Wiley, New York, 1986.
- [Lehmann, 1993] E. L. Lehmann. The Fisher and Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? *Journal of the American Statistical Association*, 88: 1242-9, 1993.
- [Lindley, 1997] D. Lindley. Some comments on “Non-informative priors do not exist”, *Journal of Statistical Planning and Inference*, 65: 182-189, 1997.
- [Mayo, 1983] D. G. Mayo. An Objective Theory of Statistical Testing, *Synthese*, 57: 297-340, 1983.
- [Mayo, 1985] D. G. Mayo. Behavioristic, Evidentialist, and Learning Models of Statistical Testing, *Philosophy of Science*, 52: 493-516, 1985.
- [Mayo, 1996] D. G. Mayo. *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago, 1996.
- [Mayo, 1997] D. G. Mayo. Severe Tests, Arguing From Error, and Methodological Underdetermination, *Philosophical Studies*, 86: 243-266, 1997.
- [Mayo, 2003] D. G. Mayo. Could Fisher, Jeffreys and Neyman Have Agreed? Commentary on J. Berger’s Fisher Address, *Statistical Science*, 18, 2003: 19-24, 2003.
- [Mayo, 2005] D. G. Mayo. Evidence as Passing Severe Tests: Highly Probed vs. Highly Proved, pp. 95-127 in *Scientific Evidence*, P. Achinstein (ed.), Johns Hopkins University Press, 2005.
- [Mayo, 2006a] D. G. Mayo. Philosophy of Statistics, pp. 802-815 in S. Sarkar and J. Pfeifer (eds.), *Philosophy of Science: An Encyclopedia*, London: Routledge, 2006.
- [Mayo, 2006b] D. G. Mayo. Critical Rationalism and Its Failure to Withstand Critical Scrutiny, pp. 63-96 in C. Cheyne and J. Worrall (eds.) *Rationality and Reality: Conversations with Alan Musgrave*, Studies in the History and Philosophy of Science, Springer, Dordrecht, 2006.

- [Mayo, 2008] D. G. Mayo. How to Discount Double Counting When It Counts, *British Journal for the Philosophy of Science*, 59: 857–79, 2008.
- [Mayo, 2010a] D. G. Mayo. Learning from Error, Severe Testing, and the Growth of Theoretical Knowledge, pp. 28–57 in Mayo, D.G. and A. Spanos, *Error and Inference*, Cambridge University Press, Cambridge, 2010.
- [Mayo, 2010b] D. G. Mayo. An Error in the Argument from Conditionality and Sufficiency to the Likelihood Principle, pp. 305–314 in Mayo, D.G. and A. Spanos, *Error and Inference*, Cambridge University Press, Cambridge, 2010.
- [Mayo, 2010c] D. G. Mayo. Sins of the Epistemic Probabilist Exchanges with Peter Achinstein, pp. 189–201 in Mayo, D.G. and A. Spanos, *Error and Inference*, Cambridge University Press, Cambridge, 2010.
- [Mayo, 2010d] D. G. Mayo. The Objective Epistemic Epistemologist and the Severe Tester, in *Philosophy of Science Matters: The Philosophy of Peter Achinstein*, edited by Gregory Morgan, Oxford University Press, 2010.
- [Mayo and Kruse, 2001] D. G. Mayo and M. Kruse. Principles of Inference and their Consequences, pp. 381–403 in *Foundations of Bayesianism*, edited by D. Cornfield and J. Williamson, Kluwer Academic Publishers, Netherlands, 2001.
- [Mayo and Spanos, 2004] D. G. Mayo and A. Spanos. Methodology in Practice: Statistical Misspecification Testing, *Philosophy of Science*, 71: 1007–1025, 2004.
- [Mayo and Spanos, 2006] D. G. Mayo and A. Spanos. Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction, *The British Journal for the Philosophy of Science*, 57: 323–357, 2006.
- [Mayo and Cox, 2006] D. G. Mayo and D. R. Cox. Frequentist Statistics as a Theory of Inductive Inference, pp. 77–97 in *Optimality: The Second Erich L. Lehmann Symposium*, edited by J. Rojo, Lecture Notes-Monograph Series, vol. 49, Institute of Mathematical Statistics, Beachwood, OH, 2006.
- [Mayo and Spanos, 2010] D. G. Mayo and A. Spanos. *Error and Inference: Recent Exchanges on the Philosophy of Science, Inductive-Statistical Inference, and Reliable Evidence*, Cambridge University Press, Cambridge 2010.
- [Neyman, 1935] J. Neyman. On the Problem of Confidence Intervals, *The Annals of Mathematical Statistics*, 6: 111–116, 1935.
- [Neyman, 1952] J. Neyman. *Lectures and Conferences on Mathematical Statistics and Probability*, 2nd ed. U.S. Department of Agriculture, Washington. 1952.
- [Neyman, 1956] J. Neyman. Note on an Article by Sir Ronald Fisher, *Journal of the Royal Statistical Society, Series B (Methodological)*, 18: 288–294, 1956.
- [Neyman, 1957] J. Neyman. Inductive Behavior as a Basic Concept of Philosophy of Science, *Revue Inst. Int. De Stat.*, 25: 7–22, 1957.
- [Neyman, 1971] J. Neyman. Foundations of Behavioristic Statistics, pp. 1–19 in Godambe and Sprott, eds., 1971.
- [Neyman and Pearson, 1933] J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses, *Philosophical Transactions of the Royal Society, A*, 231: 289–337, 1933.
- [Neyman and Pearson, 1967] J. Neyman and E. S. Pearson. *Joint Statistical Papers of J. Neyman and E. S. Pearson*, University of California Press, Berkeley, 1967.
- [Pearson, 1955] E. S. Pearson. Statistical Concepts in Their Relation to Reality, *Journal of the Royal Statistical Society, B*, 17: 204–207, 1955.
- [Pearson, 1962] E. S. Pearson. Some Thoughts on Statistical Inference, *Annals of Mathematical Statistics*, 33: 394–403, 1962.
- [Peirce, 1931–5] C. S. Peirce. *Collected Papers*, vols. 1–6, edited by C. Hartshorne and P. Weiss, Harvard University Press, Cambridge, MA, 1931–5.
- [Poole, 1987] C. Poole. Beyond the Confidence Interval, *The American Journal of Public Health*, 77: 195–199, 1987.
- [Popper, 1959] K. Popper. *The Logic of Scientific Discovery*, Basic Books, NY, 1959.
- [Pratt, 1977] J. W. Pratt. Decisions as Statistical Evidence and Birnbaum’s Confidence Concept, *Synthese*, 36: 59–69, 1977.
- [Rosenkrantz, 1977] R. Rosenkrantz. *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*, Reidel, Dordrecht, 1977.
- [Rosenthal and Gaito, 1963] R. Rosenthal and J. Gaito. The Interpretation of Levels of Significance by Psychological Researchers, *Journal of Psychology*, 55: 33–38, 1963.

- [Royall, 1997] R. Royall. *Statistical Evidence: a Likelihood Paradigm*, Chapman and Hall, London, 1997.
- [Savage, 1962] L. Savage, ed. *The Foundations of Statistical Inference: A Discussion*, Methuen, London, 1962.
- [Sober, 2008] P. Sober. *Evidence and Evolution: The Logic Behind the Science*, Cambridge University Press, Cambridge, 2008.
- [Spanos, 1999] A. Spanos. *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press, Cambridge, 1999.
- [Spanos, 2000] A. Spanos. Revisiting Data Mining: ‘hunting’ with or without a license, *The Journal of Economic Methodology*, 7: 231-264, 2000.
- [Spanos, 2007] A. Spanos. Using the Residuals to Test the Validity of a Statistical Model: Revisiting Sufficiency and Ancillarity, Working Paper, Virginia Tech, 2007.
- [Spanos, 2006] A. Spanos. Where Do Statistical Models Come From? Revisiting the Problem of Specification, pp. 98-119 in *Optimality: The Second Erich L. Lehmann Symposium*, edited by J. Rojo, Lecture Notes-Monograph Series, vol. 49, Institute of Mathematical Statistics, Beachwood, OH, 2006.
- [Spanos, 2010] A. Spanos. Akaike-type Criteria and the Reliability of Inference: Model Selection vs. Statistical Model Specification, forthcoming, *Journal of Econometrics*, 2010.

SIGNIFICANCE TESTING

Michael Dickson and Davis Baird

Significance tests are supposed to provide inductive support to conclusions regarding processes that involve some element of chance, due to ignorance, sampling, or inherent probability. We sketch their history and logical structure. While significance tests are widely used in the social and physical sciences, they remain controversial. Indeed, there are serious objections to their use, especially in contexts that are not theoretically well-understood, or for which there are not well-established stochastic models.

1 INTRODUCTION

By definition, a probabilistic process allows for some chance variation in its output. For example, samples taken at random from a population of wage earners may vary with respect to some property (a ‘sample statistic’), such as the mean hourly wage. (The probabilistic process in that case is the sampling; the output is the sample; and the property that varies is the mean hourly wage.) Similarly, let the process be ‘flipping a coin ten times’; the output (a sequence of results, ‘H’ or ‘T’) may then vary with respect to the property ‘number of heads’.

Suppose that we have an hypothesis about the process, for example, about the distribution of hourly wages within the population, or the chance of getting ‘heads’ on a flip of the coin. While there are hypotheses that can be conclusively refuted by experiment (e.g., ‘the coin always comes up heads’), typical hypotheses about probabilistic processes are not of this sort. We may hypothesize, for example, that the mean hourly wage in the population is $\$X$, or that the coin has probability p ($0 < p < 1$) to come up ‘heads’. Sampling (or flipping) cannot usually refute these hypotheses conclusively. For example, letting $p = 3/4$, even a trial of the process that produces a string of 10 ‘tails’ in a row does not refute the hypothesis. The most we can hope for is to find strong inductive evidence against the hypothesis. (Of course, intuitively, a string of 10 ‘tails’ is such evidence.)

Significance tests are supposed to provide some way of quantifying the strength of such inductive evidence, thus quantifying the inductive ‘significance’ of the evidence. The central idea is to attach high ‘refutational’ significance to evidence (obtained as the result of a trial of the process in question) that the hypothesis makes extremely unlikely. Such is the case, for example, for our hypothesis $p = 3/4$ above, in light of the evidence of 10 ‘tails’ in a row. (The probability of this evidence on the hypothesis $p = 3/4$ is roughly 1 in a million.)

The pitfalls inherent in this form of argument are many, and do not stem merely from its inductive nature, but from the specific form that the argument takes. After

reviewing the history and logic of significance testing more carefully (Section 2), we will consider how one might interpret the results of a significance test (Section 3), and then discuss several objections to it (Section 4). We conclude (Section 5) that under some circumstances, the results of a significance test can be part of a strong inductive argument, but that very often the circumstances in which one *wishes* to appeal to significance tests are exactly those where these tests should not be compelling.

2 THE DEVELOPMENT AND LOGIC OF SIGNIFICANCE TESTS

2.1 *The Normal Distribution and the Z-test*

The normal distribution (of a variable, x), is given by the formula

$$\varphi(x) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right) e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where μ is the mean of the variable, and σ its standard deviation. Graphically, the distribution is the familiar ‘bell-shaped’ curve, centered at $x = \mu$, with width 2σ . By the late 18th century, this distribution was known to be the limit (as the number of trials grows without bound) of a binomial distribution (the distribution of results of trials of a process with just two possible outcomes, one having probability p , the other, therefore, $1 - p$). It was later used to develop a general ‘theory of errors’ of observation, where a ‘successful’ observation is associated with one value of the binomial variable, and an ‘erroneous’ observation is associated with the other value. The probability of an erroneous observation is thus given by the probability that characterizes the binomial distribution. (Bear in mind that the normal distribution approximates the binomial distribution only for large samples.)

The normal distribution became widely used to describe errors of observation, from astronomical observations to the measurements of the chests of Scottish soldiers [Hacking, 1990]. The ubiquitousness of this ‘exponential law of errors’ came under scrutiny, and rightly so, but under the appropriate circumstances, it can indeed be used to estimate the likelihood that a given observational result is due to random error.

In contemporary terms, the procedure is as follows. First, two conditions must be in place: (a) the mean and standard deviation of the variable of interest in the population must be known (or hypothesized); (b) the variable of interest must be known (or hypothesized) to be normally distributed within the population. We will take a sample from the population, and the question is: How likely is it that any deviation of the sample mean from the population mean is due to random chance? (If (b) fails, or is not known to hold, its (potential) failure can be mitigated to some extent by choosing samples of sufficiently large size, in cases where the sample means will then be approximately normally distributed for large samples — the case of a binomial distribution is an example, as we mentioned

above. Exactly how large the sample needs to be depends on the details of the case.)

Given a sample of observations of some variable, x , we then compute the sample mean, $\bar{x}$. The issue, then, is the extent to which we ought to believe that this sample mean should be ‘close’ to the population mean. The ‘null hypothesis’ is the hypothesis that the sample mean *is* close to the population mean, and that any deviations from the population mean are due to random fluctuation.

Does the evidence afforded by the sample speak against the null hypothesis? For a sample size of 1, we may answer this question by computing the ‘z-score’, the number of standard deviations that the sampled value is away from the population mean:

$$z = \frac{\bar{x} - \mu}{\sigma}$$

where μ and σ refer to the *population* mean and the *population* standard deviation, respectively.

The assumption that the variable x is normally distributed within the population, with mean μ and standard deviation σ , allows a straightforward calculation of the probabilities for various ranges of z-scores by integrating the normal distribution to determine the ‘area under the curve’ in the appropriate region. If the resulting probability is not too low, then one is often willing to chalk up any difference between $\bar{x}$ and μ to random error (rather than, say, some other, systematic, cause). If it is ‘too’ low (i.e., below some specified value), then one may wish, on those grounds, to reject the null hypothesis.

For larger samples, of size n ($n > 1$), we have to consider not the standard deviation of x , but the standard deviation of the means of samples of size n (still assuming that x has mean μ and standard deviation σ), and what matters then is how far our sample mean is from the population mean as measured in terms of *this* standard deviation. It turns out that this number is $\sigma/\sqrt{n}$ for samples of size n . Hence in this case the z-score is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.$$

We then proceed as above, for this z-score.

Although it would not have been described in these terms, the above analysis represents the state of the art by around the turn of the 20th century. It has some obvious, and potentially severe, limitations, most notably the fact that it does not apply to non-normally distributed variables, and it requires one to *know* the population mean and standard deviation, which are often the *objects* of investigation, and therefore not given from the start.¹ (On the other hand, the simple z-test can be appropriate in some cases. It is widely used, for example, in intelligence testing,

¹An alternative way to think about this test does not require one to know the population mean. One might instead wish to hypothesize the population mean, and then to calculate, based on that hypothesis, the probability that a given sample mean is ‘close to’ the true mean. If the sample is *known* to be random with respect to the variable of interest, then this test may help one to verify or reject hypotheses about the population mean.

where testers believe that they already know the population mean and standard deviation, and may wish to know whether the mean score of a given group of test takers is ‘significantly’ different from the mean, i.e., whether the group’s deviation from the mean can be ‘reasonably’ chalked up to chance variation.)

2.2 *Pearson and the Chi-Squared Test*

2.2.1 *Early History and Binomial Tests*

While there were precursors (notably Edgeworth’s (1885) attempt to develop a way to control for chance variation; cf. [Baird, 1987]), arguably the first modern example of a significance test was Karl Pearson’s [1900] chi-squared test. Pearson had a specific problem. In the early 1890s he had come to realize that the normal distribution does not always (or even very often) fit actual chance data. (Pearson makes this observation in numerous places, including several times about data collected by Galton in Pearson [1930], e.g., at p. 74. One modern real-world discussion of the ubiquity non-normality (in this case, in the context of quality control) is [Pyzdek, 2001].) He therefore developed a system of chance distributions, each obtainable via a variation in the parameters appearing in a ‘generalized probability curve’, a formula that generalizes the normal distribution. Pearson’s assumption — no longer held today — was that chance variation would have to be distributed according to one of the distributions in this system (Baird 1983a).

The very existence of alternatives to the normal distribution already raises the question of how well any given data fit a given distribution, and Pearson developed the chi-squared test to solve this problem. While we no longer use Pearson’s system of chance distributions, the chi-squared test remains an important test of the goodness of fit between gathered data and a proposed statistical distribution.

To see how it works, consider first the simple case of a chance process with just two possible outcomes (for example, flipping a coin). (What follows is not the chi-squared test itself, but a preliminary illustration of the logic of such tests.) Sampling from this process (flipping the coin) N times yields a probability of

$$\left(\frac{N!}{r!(N-r)!} \right) p^r (1-p)^{N-r}$$

for getting a sequence containing r many ‘heads’ and $N-r$ many ‘tails’, where p is the probability of getting a ‘heads’. An hypothesis about the (presumed binomial) distribution in this case amounts to a proposed value for p . A subsequent test of this hypothesis may consist in N samples (flips), the result being a sequence with r many ‘heads’. We may compute the probability of this result, under our proposed value for p , using the formula above. (Note: we are calculating the probability of getting r many heads, not the probability of a specific sequence.) If the probability of the actual result, thus determined, is exceptionally low, we may be inclined to reject the initial hypothesis. For example, if we propose $p = 1/2$ and subsequently get 20 heads in 100 flips, we may be inclined to reject our initial proposal, in favor

of one that asserts a much lower probability for heads on a single flip, and thus a much higher probability for getting 20 heads in 100 flips. (The probability of getting 20 heads in 100 flips, given $p = 1/2$, is of order 10^{-10}).

2.2.2 The Chi-Squared Test

This ‘significance test’ applies only in the simple case of a binomially distributed variable, and the calculations involved are feasible (or were in Pearson’s time) only for relatively small sample sizes.² Consider, now, the more general case of a variable with finitely many possible values, and an arbitrary hypothesized distribution, p_i , where i indexes the possible values of the variable, and p_i is the hypothesized probability that the variable will take the i^{th} value on any given trial. So the null hypothesis in this case is the hypothesis that the distribution of the variable of interest in the population is given by the p_i .³

Letting N be the sample size and n_i the number of times that the i^{th} value appears in the sample, Pearson defines

$$\chi^2 = \sum_i \frac{(n_i - Np_i)^2}{Np_i}.$$

(So Np_i is the expected frequency of the i^{th} value, given the hypothesis.) Note that this quantity gets larger as the actual data get further from the predicted result. Hence χ^2 is a measure, of sorts, of how much the actual data deviate from the prediction.

It turns out that, given N and the p_i , one can calculate the probability $p(\chi^2 \geq \chi_O^2)$, i.e., the probability that a chance process described by the probabilities p_i will produce the observed deviation χ_O^2 , or ‘worse’ (greater). Much as we might do in the binomial case discussed above, then, we may choose a probability, p , lower than which we are not willing to hold on to the original hypothesis. I.e., if our sample produces a χ_O^2 for which $p(\chi^2 \geq \chi_O^2)$ is lower than p , then we reject the hypothesis. (It is far from clear that Pearson intended the χ^2 -test to be understood in this way. He does speak in these probabilistic terms occasionally, but seems also to have considered the main point to be to provide a measure of goodness of fit, rather than a *probability* to be the basis for the acceptance or rejection of hypotheses. However, regardless of Pearson’s intentions, the χ^2 -test is now used in this way. See Baird 1983b.)

²There are good estimates for larger sample sizes — for example, a binomial distribution is well approximated by a normal distribution for large sample sizes — and these were known in Pearson’s time. However, the restriction to binomial distributions is still required, and remains a very severe limitation.

³Other sorts of hypothesis may be considered. For example, one may wish to hypothesize only the general form of the population distribution (e.g., that it is Poissonian). Here we consider the simple case of hypothesizing a specific distribution, as it is sufficient to examine the logical points that we wish to discuss.

2.2.3 Composite Hypotheses and Independence

Thus far, we have considered so-called ‘simple’ hypotheses, i.e., hypotheses that propose a specific distribution for the variable of interest. Often one is interested, instead, in ‘composite’ hypotheses, which do not propose a specific distribution, but a range of distributions. For example, one common hypothesis is that two variables are probabilistically independent of one another. Consider, for example, a case of sampling two binomial variables, $X \in \{0, 1\}$ and $Y \in \{0, 1\}$. Each sample produces one of four possible results: $\langle 0, 0 \rangle$, $\langle 0, 1 \rangle$, $\langle 1, 0 \rangle$, $\langle 1, 1 \rangle$. The problem may thus apparently be treated as sampling a single variable with four possible results. The hypothesis of independence, however, does not specify a single distribution for this 4-valued variable, but rather a family of distributions, in each of which the variables X and Y are probabilistically independent. (Of course, similar remarks hold for the multinomial case.)

It is far from clear what the relationship should be between significance tests for each of many simple hypotheses (to which the χ^2 -test most clearly applies) and inferences regarding the composite hypothesis comprised of these simple hypotheses. One may, on the one hand, wish to suppose that if each of the individual simple hypotheses should be rejected, then so should the composite hypothesis — after all, it is merely a disjunction of the individually rejected simple hypotheses. On the other hand, recall that rejection here is a probabilistic judgement: we reject an hypothesis because it makes the observed data very unlikely. Does the fact that each of several hypotheses makes the observed data unlikely entail that the group as a whole ought to be rejected? It is far from clear that the answer is ‘yes’ — that answer lies perilously close to the lottery paradox.

In some cases — including the important case of hypotheses of independence — this issue can be, at least for the moment, sidestepped. One may use observed data to make estimates of *specific* probabilities, in accord with the composite hypothesis (for example, the hypothesis of independence). The result is a simple hypothesis that can be tested in the usual way by a χ^2 -test.

A case of historical, theoretical, and practical importance is the ‘fourfold contingency table’ (nowadays called a ‘ 2×2 ’ table), where we wish to judge the independence of the two variables X and Y , as described above. We begin with the observed frequencies, described abstractly in Table 1.

	$X = 0$	$X = 1$	Total
$Y = 0$	a	b	$a + b$
$Y = 1$	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Table 1. Observed data for two binomial variables. The variables a, b, c, d are arbitrary natural numbers.

We may use this data to generate estimates of the four possible results, but in so

doing, we must also respect the hypothesis of independence. Hence, for example, the joint probability $p(\langle 0, 0 \rangle)$ must be the product of marginals:

$$p(\langle 0, 0 \rangle) = p(X = 0) \times p(Y = 0) = \left(\frac{a + c}{a + b + c + d} \right) \left(\frac{a + b}{a + b + c + d} \right).$$

The result is a simple hypothesis about the underlying (population) distribution, shown in Table 2. (Note that it is sufficient to calculate just one cell from Table 2, as the others can be determined from the result of that calculation together with the marginal totals.)

	$X = 0$	$X = 1$	Total
$Y = 0$	$\frac{(a+c)(a+b)}{a+b+c+d}$	$\frac{(b+d)(a+b)}{a+b+c+d}$	$a + b$
$Y = 1$	$\frac{(a+c)(c+d)}{a+b+c+d}$	$\frac{(b+d)(c+d)}{a+b+c+d}$	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Table 2. Expected (predicted) data for two independent binomial variables.

A standard χ^2 -test may then be applied to the observed and hypothesized data in Tables 1 and 2. We discuss this case further below (Section 2.4).

2.3 Gossett and the T -test

In 1899, William Gossett went to work for the Guinness brewery. One of the problems that he took on there was the selection of the best yielding varieties of barley. The problem is statistical in nature because one has, to begin, a chance process — the raising of barley (the variable of interest is the yield). In order to learn about the latest methods for estimating population statistics (in this case, yield) from sample statistics (measured yield of a given variety), Gossett spent two terms (1906-07) in Pearson's statistics laboratory.

Pearson's methods were what Gossett needed, almost. As we noted above, the χ^2 -test works well only for large sample sizes. However, Gossett was working with relatively small sample sizes, and thus needed to develop a new method. (Pearson had not faced this problem — he was largely concerned with biometrics, where one typically has very large sample sizes.) The result was a new significance test, suitable for small sample sizes, the t -test. Because Guinness was concerned

with its employees revealing trade secrets, Gossett published this work under the pseudonym ‘Student’ [Gossett, 1908], and hence the test is often called “Student’s t -test”.

As was common at the time (and remains so), Gossett assumed that the variable of interest is distributed normally in the population. For large sample sizes, this assumption more or less assures one that the sample means taken from the population will be distributed normally, which fact makes the reliable determination of the probability (given an hypothesis) of the actual data straightforward, *if* one has a good estimate of the standard deviation in the population. However, for small sample sizes, the sample in general affords a relatively poor estimate of the population standard deviation. Gossett’s main accomplishment was to define a statistic whose distribution could be determined exactly, even when the population standard deviation is being estimated from a relatively small sample.

This statistic is denoted ‘ t ’, and is defined by

$$t = \frac{\bar{x} - \mu}{s}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and μ is the hypothesized population mean. The statistic t measures ‘how many sample standard deviations the observed (sample) mean is away from the hypothesized mean’. Given the hypothesis, larger values of t are increasingly less probable. As we mentioned, Gossett was able to determine (by “guessing” — his own word) a probability distribution for this statistic. Hence, given a hypothesized mean (and the assumption that the population is normally distributed), a sample mean, and a sample standard deviation, one can determine the probability that t is equal to or greater than some observed value for t . The novelty here is that this probability tracks the probability that the hypothesis confers on observed sequences even for small sample sizes.

Gossett did not have a mathematically rigorous proof for the distribution of t . In 1912, as a Cambridge undergraduate, R. A. Fisher, provided the proof (published as [Fisher, 1915]). However, it was recognized by both Gossett and Fisher that similar results for other statistics were much in need. Fisher had tackled the problem of correlation coefficients, but in his [1922], partly prompted by Gossett (in private communication), he acknowledged that several important cases remained to be resolved for small samples, including multivariate correlation coefficients, regression coefficients, and others. Many of these cases were later solved by Fisher, using (geometric) methods similar to the one he used to derive the t -distribution.

Another direction of extension is raised by the question to what extent these various distributions depend on the assumption that the population is normally distributed. Gossett had, in 1923, tried to interest Fisher in the problem of ‘robustness’ (as it would later be called), to no avail. He did the same with E.S. Pearson (son of Karl) 3 years later. E.S. Pearson and his coworkers began testing the robustness of Gossett’s and Fisher’s results for non-normal distributions

experimentally (using tables of random numbers to perform simulated random sampling from non-normal distributions), ultimately showing which of the results were robust, and which not. The t -distribution is robust, but some of the other distributions derived by Fisher are not. Theoretical work confirming the results of the simulations occurred much later (see, notably, [Box, 1953]). (On the other hand, Fisher was apparently offended by the very idea that the point even needed consideration. There was some friction, as a result, between him and E.S. Pearson. Gossett somehow managed to stay friendly with both of them. See [Lehmann, 1999].)

2.4 Fisher's Degrees of Freedom

Pearson's χ^2 -test is used in a variety of circumstances. In some cases, theory provides one with a well-defined hypothesized distribution, and in such cases, all agree that the test should proceed as we describe above. However, frequently one uses the data itself to provide estimates for the parameters of an hypothesized type of distribution (in particular, the probabilities or frequencies for the various values of the variable of interest). Indeed, Pearson himself pioneered methods for estimating the parameters of an hypothesized distribution from the observed data. Above (Section 2.2.3), we saw a case where a composite hypothesis was convertible to a simple hypothesis susceptible to a χ^2 -test by estimating frequencies. In such cases, it is unclear exactly how many parameters of the distribution are being tested for 'goodness of fit'. For example, one of the probabilities that is being estimated is *fixed* by the estimates for the others, because the probabilities of all possible values must sum to 1.

A specific example is both historically important and illustrative. Yule and Greenwood [1915] came across this issue in an analysis of the effectiveness of cholera vaccination. The (null) hypothesis of interest here is the composite hypothesis that the 'vaccinated' variable is independent of the 'contracts disease' variable. As discussed in Section 2.2.3, in order to apply the χ^2 -test in its usual form, they had to estimate the relevant probabilities, using the data that they had collected (rather than consulting some theory that is independent of that data). Following essentially the procedure illustrated in Tables 1 and 2, they then employed the χ^2 -test, using the distribution of χ^2 for a 4-valued variable. They also performed a more traditional statistical analysis (a 'standard error' analysis) under the assumption that the populations are distributed normally (an assumption not, of course, required by the χ^2 -test). The results differ.

While there was already in the literature (largely due to Pearson) ample reason to be suspicious of the assumption of normality, this discrepancy remained a thorn in the side of the χ^2 -test as described by Pearson. Yule and Greenwood themselves observed that the standard error analysis agreed with the χ^2 -test under the assumption that the variable of interest has only 2 (rather than 4) possible values, and they proposed an explanation for the discrepancy, which was later given a stronger mathematical foundation by Fisher, who was able to derive an exact

distribution for χ^2 even under the circumstance that the parameters describing the distribution of the variable of interest in the population are *estimated* from the data, rather than known, or given by theory. Fisher's result is thus similar to Gossett's (which demonstrates an exact distribution for t in the population, even though t contains an *estimate* of the population standard deviation). It turns out that in the case considered by Yule and Greenwood, the distribution of χ^2 , when the population (specifically, frequency) parameters are estimated from the data, is given by Pearson's distribution for χ^2 under the assumption of a two-valued, rather than a four-valued, variable of interest.

Fisher's result relies on the notion of 'degrees of freedom'. Recall that in the case of the fourfold contingency table, fixing one cell of the table (by assuming independence) fixes the other three cells. In other words, the hypothesis of independence (plus the use of observed data to estimate frequencies) determines the probability for *one* value of the (composite) variable under investigation, and simple mathematics fixes the others. This fact strongly suggests that we should take ourselves to be estimating a *single* binomial parameter describing the compound distribution of the pair of variables, rather than the four parameters illustrated in the table, or the two binomial parameters that gave rise to the inquiry in the first place. Fisher introduced the term 'degrees of freedom' to describe the number of parameters required to describe all possible variations from the hypothesis under examination. He further argued that when calculating the probability that a given deviation from the hypothesis is due to chance variation, the proper distribution to use for χ^2 is the one obtained by assuming that the number of parameters is equal to the degrees of freedom.

Debate and controversy ensued. Pearson was apparently concerned more with measures of goodness of fit, while Fisher was concerned with the probability that an hypothesis confers on an observed statistics. (See [Baird, 1983b].) Pearson thus never acknowledged that Fisher's 'degrees of freedom' is a correct way to choose a distribution for χ^2 . Given his understanding of the meaning of χ^2 and its 'distributions', he may have been correct, but the statistical tradition has clearly followed Fisher on this point.

2.5 *Complications in the Application of Significance Tests*

2.5.1 *Data 'Too Good to be True'*

In the early 20th Century, Gregor Mendel's [1865] work on plant hybridization was 'rediscovered', and indeed Mendelian genetics was at the center of ongoing debates between self-proclaimed 'Darwinists' and 'Mendelians'. (Franklin *et al.* [2008] provide an exhaustive review and analysis of the story to come.) It was therefore natural for Fisher — who hoped and argued for a reconciliation between 'Mendelians' and the 'Darwinists' — to examine Mendel's work closely, from a statistical point of view.

While taking quite a positive view of Mendel's work overall, Fisher was critical of a few aspects of it, the most important, for our purposes, being the 'goodness

of fit' between Mendel's reported observations and his theoretical prediction. The experiments in question were aimed at determining the ratio of heterozygous to homozygous individuals (in this case, pea plants) exhibiting a given dominant trait. (We simplify matters here to get to the basic point — see [Franklin *et al.*, 2008] for the gory details, which are in fact quite important to the ultimate judgment one might make about Mendel's work, but not to the point we wish to make here.) For each individual, Mendel and his associates examined 10 of its progeny for the recessive trait, and when at least one exhibited the recessive trait, they judged the parent to be heterozygous. Mendel's theory predicts that the ratio of heterozygous to homozygous individuals is 2:1.

In one of his experiments, Mendel reports, out of 600 plants, 399 heterozygotes and 201 homozygotes. This data is, of course, extremely close to the theoretically predicted frequencies of 400 and 200. Indeed, Fisher concludes, the data are too good to be true. The value of χ^2 in this case is a suspiciously low 7.5×10^{-3} . The probability of getting data this good, if Mendel's underlying probabilistic (binomial) model is correct, is very low, in the sense that *if* Mendel's model is correct, then the probability of finding such a low value for χ^2 is almost zero.

The issue is further complicated by the fact, observed by Fisher, that if the true ratio in the population is 2:1, then Mendel's experimental methodology would yield an observed ratio of approximately 1.7:1, as some heterozygous parents will go undetected, because their offspring will all, by chance, have the dominant trait. By Fisher's lights, this observation makes matters worse for Mendel. He calculated χ^2 under the assumption that the expected observed ratio is 1.7:1 (which, again, is what it *would* be if the 2:1 law is correct), and finds the value to be unacceptably high. In other words, Mendel's data does not, apparently, fit this expected result very well.

Prima facie, it looks very much like Mendel's data were cooked. Assuming Mendel's 2:1 law (which is not in doubt, here), we would expect the observed results to be close (but not *too* close!) to the 1.7:1 ratio. Mendel, however, expected to observe a 2:1 ratio. His data are, in fact, extremely close (*too* close?) to this 2:1 expectation. It looks for all the world as if the 'uncooked' data were somewhat close to the 1.7:1 ratio, then 'cooked' to make them very close to the 2:1 ratio. This conclusion is, in fact, the one that Fisher drew (though not being able to malign Mendel, whose work Fisher admired, he speculated that the blame lie with an over-zealous assistant). It is not uncommon to find this conclusion in the literature today as well, though the matter remains controversial.

2.5.2 Some Illustrative Issues With Fisher's Analysis

The story does not end here. There are several issues remaining to consider. Here we mention three, and refer the reader to [Seidenfeld, 2008] for an extensive discussion.

First, much depends on how one organizes the data into 'samples'. Consider: suppose a fair coin comes up 'heads' exactly 500 times in 1000 independent flips.

This outcome is rare. However, *any* specific outcome (i.e., any specific number of ‘heads’) is rare in this case, and it is far from clear that we should throw out the hypothesis of fairness based on a single sample, unless the sample suggests a strong *bias* (either for or against ‘heads’). But suppose we did 10 samples of 100 flips each, and got exactly 50 ‘heads’ each time. Now matters are beginning to look fishy. While exactly 50% ‘heads’ may not be cause for suspicion in a single sample, its recurrence 10 times should raise eyebrows. The problem here is that there are no compelling principles guiding the individuation of samples (of a process whose trials are independent); but if arbitrary divisions of (independent) trials into multiple samples is permitted, then we are in trouble, because we could easily divide the 1000 flips into samples in such a way that the individual samples have all sorts of unlikely statistical properties, such as nearly always alternating ‘heads’ with ‘tails’.

Second, in order for a significance test to provide decent evidence against a given null hypothesis, it is at least helpful, if not compulsory, to know which other hypotheses are in play, and whether they have less troublesome values for χ^2 , and whether they are plausible (or implausible) for other reasons. For suppose that a given hypothesis, H , has a particularly improbable value for χ^2 , but that all *other* hypotheses that we are willing to consider *also* have improbable values for χ^2 . Then there is little reason to reject any of them on the grounds of its value for χ^2 .

However, the real problem here is not whether such hypotheses (predicting a ‘reasonable’ value for χ^2) exist (generally, mathematically, they do), but whether there is any *one* hypothesis that stands out as *the* alternative to the null hypothesis. In general, there is not. In the case of Mendel’s peas, one alternative is that the data were ‘cooked’, but another is that the trials were not, after all, independent and identically distributed (i.i.d.). (Seidenfeld [2008], for example, argues that in fact the trials may not have been i.i.d.) Of course, one may take as the ‘alternative’ the bare denial of the null hypothesis (as perhaps Fisher would have one do), but in that case, one must be extremely modest about what is being asserted at the end — one cannot, for example, claim to have shown that the data were ‘cooked’ (as opposed to, for example, coming from trials that were not i.i.d.).

Third, it is not clear why one must choose the χ^2 statistic to determine the likelihood that the hypothesis confers on the data. χ^2 , like any other statistic, is, mathematically, just a number computed from the observed data. But *any* outcome of an experiment can be made to appear ‘rare’ when it is redescribed in terms of a (‘suitably’ chosen) statistic. An extreme case here would (in many cases) be to use as one’s p -value the probability that the hypothesis confers on the data itself. (Nevermind for the moment that this number is often very difficult to calculate, which is partly why we opted for statistics in the first place.) In many cases this choice will yield extremely low p -values for *any* observed data (which, again, is partly why we opted to calculate the probability of a statistic). For example, the hypothesis that a coin is fair confers exactly the same (extraordinarily low) probability on *every* $N = 100$ sequence of ‘heads’ and ‘tails’. Hence we are

driven to consider the probability of certain ‘well-chosen’ statistics computed from the data. However, any choice of a statistic introduces additional assumptions into the discussion, and in the end a discrepancy ‘too unlikely to occur by chance’ between the data and an hypothesis may be chalked up to these assumptions.

For example, the appropriateness of measuring the ‘likelihood’ of a given sequence of heads and tails by considering the number of heads it contains depends on whether the trials are i.i.d.. If the trials are i.i.d., then we would assign very low probability to a sequence of all heads of any appreciable length (assuming that the probability of ‘heads’ on any given trial is $1/2$). If they are not — for example (to take the extreme case), if trials 2 and beyond are guaranteed to match the result of trial 1, then the sequence of all heads would have quite a high probability ($1/2$, if the probability of ‘heads’ on any given trial is $1/2$).

Our purpose here is not to resolve these issues, but to raise them. The choices that one makes in designing a significance test can have a significant outcome on the result.

2.5.3 Two Types of Test

Recall that the χ^2 -test measures, roughly, ‘how well the observed data fit the hypothesized distribution’. It is insensitive to whether too many of the observed values are ‘too high’ or ‘too low’, counting both cases as ‘misfits’. In general, a test is ‘two-tailed’ when the null hypothesis will be rejected if the value of the test statistic (such as a sample mean) is either too small or too large. Otherwise, the test is ‘one-tailed’. (The terms ‘one-sided’ and ‘two-sided’ are sometimes used, especially when the underlying population distribution is known or suspected to be poorly described as having ‘tails’.)

Whether a significance test should be one-tailed or two-tailed is not always clear — it often depends on the ultimate aim of the testing. For example, one-tailed tests are often appropriate in cases where one is interested in whether a given intervention has a positive effect on some condition, for example, in medical testing. The point is that one rejects the null hypothesis of chance variation only when the evidence points in favor of the intervention’s improving the condition in question. One will, in that case, also reject the null hypothesis when the evidence happens to point in favor of the intervention’s having a negative effect, but if the plan is to adopt the intervention only if the null hypothesis is rejected, then arguably this failure to detect negative effects is acceptable.

However, notice that the actual (numerical) significance levels mean different things in the two cases. In a one-tailed test, failing the test at the 5% level means that the sample statistic was in the top (alternatively, bottom) 5% of values within those predicted by the model. In a two-tailed test, failure means that the sample statistic was either in the top 2.5% or the bottom 2.5% of values within those predicted by the model.

This fact can lead to misinterpretation. Consider again the medical testing example, and suppose that one researcher uses a one-tailed test, while the other

uses a two-tailed test. They both find that the test statistic has a p -value of 4%. The one-tailed test will *reject* the null hypothesis in this case, and conclude that the intervention has a positive effect, while the two-tailed test will not reject the null hypothesis. (See [Bland and Bland, 1994] for a medical example along these lines.)

2.5.4 A Common Logic

We pause, here, to notice that significance tests, as we have described them and as they are commonly employed today, are united by a common logic (see [Baird, 1987]).

Once a probabilistic process has been identified, significance testing involves four steps. The first is the stipulation of a ‘null hypothesis’, which, generically, states that the probabilistic nature of the process itself is sufficient to account for the results of any sample of the process. In other words, nothing other than the variation that occurs by chance need be invoked to explain the results of any given trial, or the variation in results from one trial to another.

Of course, ‘by chance’ must be carefully defined, by means of some probabilistic model of the process, one that specifies the probabilities that underlie the results. ‘Significance’ will then be assessed relative to this model, and it is this model, and only this model, that is the target of inductive conclusions based on a significance test. (To illustrate: Seidenfeld’s [2008] concern about whether Mendel’s trials are i.i.d. is a concern about Fisher’s model, not about his analysis of significance relative to the model that he (and, apparently, Mendel) presumed).

For example, the coin-flipping process mentioned above may be modeled as a binomial distribution defined by the parameters p (the probability of getting a heads), N (the number of flips), and the probabilistic independence of the flips. Sampling N hourly wages from a population of workers may be modeled, semi-formally, as the random drawing of N balls from an urn containing a given proportion, p_n , of balls for each possible hourly wage, w_n . The null hypothesis specifies values for the parameters of these models (for example, $p = 3/4$ in the coin flipping process), in virtue of which it specifies a distribution for some variable or variables of interest (for example, number of ‘heads’, or hourly wage).

The second step in significance testing is to find a statistic that appropriately summarizes the results of trials from the process in question. Appropriateness comes down to two things: (1) It must rank all the possible outcomes of the process with regard to how adequately the chance model accounts for their occurrence; (2) the model provided by the null hypothesis must provide well-defined probabilities that order the possible outcomes in the same way. The first demand ensures that evidence collected from trials of the process will bear on the adequacy of the model provided by the null hypothesis. The second demand permits one to compare results from different significance tests meaningfully. Clearly, if these two demands are met by the test statistic, then the outcomes of the process for which the model provided by the null hypothesis cannot readily account will tend to have

low probability.

For example, in the coin-flipping case, we can choose ‘frequency of heads’ as our test statistic. The binomial model for which $p = 3/4$, $N = 10$, and the flips are independent then supplies probabilities for all possible results. These probabilities rank a result lower the further it is from 7.5 heads. In the case of sampling hourly wages, a natural choice of sample statistic would be the mean hourly wage. A model that specifies a distribution (in the total population) of wages will, by virtue of this distribution, generate probabilities for all possible sample means, and these probabilities will tend to be lower the further the sample mean is away from the hypothesized population mean, though it is crucial to bear in mind that the details of the relationship between the probability of a given sample mean and its distance from the population mean may be highly dependent on the shape of the distribution. We return to this point below.

The third step in significance testing is establishing a level of significance. A trial of the process is performed, and the test statistic is calculated. The level of significance of the outcome is the probability (as determined by the model given by the null hypothesis) of any result as bad *or worse* than the observed result (understood in terms of the ranking provided by the test statistic). The smaller this probability, the stronger the evidence against the null hypothesis. This probability is often referred to as the ‘ p -value’ of a test.

Finally, the fourth step is to reject, or fail to reject, the null hypothesis. When the observed result of the trial tells *significantly* against the null hypothesis, then the null hypothesis is rejected.

Logically, then, significance testing, in its most conservative form, appears as a kind of ‘inductive *modus tollens*’. The null hypothesis (the ‘antecedent’) implies a probability distribution, $p : \mathbb{R} \rightarrow [0, 1]$, for some test statistic. If the actual value, r , of the test statistic, together with all values, Δ_r , ‘worse than’ it, has ‘sufficiently low’ probability (i.e., if it is lower than some stipulated threshold, τ) according to this distribution, then the null hypothesis is rejected:

$$\text{Null Hypothesis} \Rightarrow p : \mathbb{R} \rightarrow [0, 1]$$

$$p(\Delta_r) < \tau$$

$$\neg \text{Null Hypothesis.}$$

We forego comments on the strength of this argument until Section 3. For now, two further observations are crucial to bear in mind.

First, as we noted above, it is common, and probably almost always correct, to consider ‘all values worse than r ’, rather than just r itself. Otherwise, we could easily falsely reject a null hypothesis based on very good evidence *in its favor*. Recall the example above: getting exactly 500 ‘heads’ in 1000 flips of a fair coin presumably is not evidence *against* the coin’s being fair. (On the other hand, getting exactly 500 flips in 1000 trials *repeatedly* may raise questions not about the parameters of the hypothesized model, but about its basic form, for example,

the assumption of i.i.d. trials — see Section 2.5.2 above.) The point is that the exact probability of a given single value of the test statistic is a function of how finely the outcome space has been divided. To avoid this misunderstanding of inductive evidence, significance tests look to the class of results as bad or worse than that obtained.

Second, if we fail to reject the null hypothesis, we ought not, logically, conclude anything whatsoever about the null hypothesis itself, except that it was not rejected by the given trial. The analogy with *modus tollens* is apt in this context: just as one would not conclude P from $P \rightarrow Q$ and Q , so also one ought not conclude that the null hypothesis is true in cases where the disagreement between a trial and the null hypothesis is determined to be ‘statistically insignificant’. We return to this point below.

2.6 Subsequent Developments

2.6.1 Neyman and Pearson’s Approach

In Fisher’s approach to significance testing, statistics were chosen largely on the grounds that an exact population distribution could be calculated (from the available data) for them, at least under given assumptions (about the population). But apart from practical considerations, why choose one statistic rather than another?

This question was addressed by Jerzy Neyman and E.S. Pearson, initially jointly [Neyman and Pearson, 1933], and later separately (e.g., [Neyman, 1938; 1955; Pearson, 1955; 1962]). To do so, they consider not merely a null hypothesis, but a class of alternatives to it, and consider, amongst all of these hypotheses, two probabilities: the probability of falsely rejecting the null hypothesis (‘Error I’) and the probability of falsely accepting the null hypothesis (‘Error II’). Given a fixed probability for Error I, they defined the ‘best’ test as the one employing whichever hypothesis minimizes the probability of Error II. Neyman and Pearson were able to solve this problem (i.e., to determine which test is best by this definition) for simple cases, and later work extended their results in various ways to more complex cases.

As Fisher would point out in his relentless opposition to these ideas, the calculation of the probability of Error I is based on the null hypothesis, without consideration of any alternative hypotheses. However, the calculation of the probability of Error II requires the specification of a class of alternative hypotheses. On Fisher’s view, this fact rendered these probabilities unable to support the acceptance of an hypothesis in light of a test.

Indeed, the issue of just what statistical tests are supposed to achieve seems to have been at the center of the controversy between Fisher and Neyman and Pearson. Neyman and Pearson often spoke in terms of ‘acceptance’ and ‘rejection’ of hypotheses, while Pearson argued that statistical tests are meant only to support rejection: “[the calculation of probabilities of Error II] would never have come into consideration in the theory only of tests of significance, had the logic of such tests not been confused with that of acceptance procedures” [Fisher, 1947, 17]. This

difference was formulated by the protagonists in terms of ‘inductive inference’ (Fisher) versus ‘inductive behavior’ (Neyman and Pearson). Fisher’s view was that statistical tests support inductive inferences; and because the probability of Error II is always relative to a chosen class of hypotheses, statistical tests can never support acceptance *tout court*, but only a kind of ‘conditional’ acceptance that Fisher thought pointless. Neyman’s view was that statistical tests support inductive behavior, and as behavior *requires* one to accept some hypothesis or other (one must behave as if *something* were the case), we do the best we can based on our statistical tests, and argue that in the long run we’ll come out alright, probably. (Pearson shied away from this issue, though the phrase does appear even in their early joint work. For the view that ‘inference versus behavior’ was not, however, at the heart of the dispute, and that Pearson therefore remained ‘true’ to the original motivation for the approach, see [Mayo, 1992].)

The story is complicated further by what appear to have been fundamental differences in their understandings of probability, modeling, and the relationship between statistical judgements and action, among other things. (See [Hacking, 1965; Kyburg, 1974; Seidenfeld, 1979; Gigerenzer *et al.*, 1989; Lenhard, 2006]. For the view that the differences between them are largely rhetorical rather than substantial, see [Lehman, 1993].)

2.7 The Institutionalization of Significance Tests

Statistical significance has become the gold standard in many academic disciplines. (Gigerenzer [1993] tells the story in the case of psychology.) Many major journals in social science, for example, require — either officially or in practice — that publishable studies demonstrate a statistically significant effect (i.e., the data must depart ‘significantly’ from some specified null hypothesis, for some statistic). Often, the level of significance (commonly, 5% or 1%) is dictated as well. This transformation of social science (and it has spread to other data-intensive disciplines as well) occurred rapidly. By one account, it was complete by 1940 or so: “[statisticians] have already overrun every branch of science with a rapidity of conquest rivaled only by Attila, Mohammed, and the Colorado beetle” [Kendel, 1942, 69].

In the light of vocal critics (see Section 4 below), the issue of whether and to what extent to continue these requirements has been taken up by some major academic societies, such as the American Psychological Association (see [Azar, 1996]), although significance tests continue to play much the same role as they had throughout at least the latter half of the 20th century.

3 (MIS)INTERPRETING SIGNIFICANCE TESTS

As has been frequently noted in the literature, there are numerous ways to misinterpret the results of a significance test. (Our discussion will only scratch the surface. See, for example, [Hacking, 1965; Kyburg, 1974; Seidenfeld 1979] for

some classic and extensive discussions.) Carver [1978], for example, lists three common ‘fantasies’ about the p -value delivered by significance tests:

1. p is the probability that the result of the trial (sample) is due to chance. A small value for p indicates that the result is not due to chance, while a relatively large value for p indicates that the result is due to ‘chance alone’.
2. $1 - p$ is the reliability of the result of the trial (sample). More precisely, it conveys information about replicability: it is the probability of getting the same result if the experiment were repeated. Researchers may claim that a significant difference (between the model provided by the null hypothesis and the observed result) is replicable if $1 - p$ is large.
3. p is the probability that the null hypothesis is true.

None of these interpretations is correct, though research would be far easier if they were (hence the term ‘fantasy’). Indeed, researchers are often seduced into adopting one of these interpretations. Numerous examples have been noted by various critics of significance testing. (See, e.g., [Gigerenzer, 1993; Cohen, 1994; Falk and Greenbaum, 1995] for several citations and discussions of such misunderstandings. Cf. [Nickerson, 2000] for an extensive discussion.)

It is instructive to consider why these interpretations are incorrect. We take them each in turn.

1. Strictly, p is the probability of the result as given by the model provided by the null hypothesis. Even in cases where the null hypothesis is the claim that the process in question operates randomly according to some distribution, is this p the probability that the result is due to chance? No! Consider the case of flipping a coin 10 times. Suppose we get 6 heads in 10 flips, and the null hypothesis is that the process is binomially distributed with probability 0.5. The null hypothesis thus assigns a probability of roughly 0.75 to getting either 6 or more, or 4 or fewer, heads. (Hence by any reasonable measure of significance, this result does not tell against the null hypothesis.) But is this number (75%) the probability that our result is due to chance? Certainly not. Indeed, it is hard to know how one could ever go about calculating such a probability. What, for example, is the probability that a covert agent, knowing about our coin flipping experiment, controls the flips and has arranged in advance to make the numbers come out ‘right’? What is the probability that the coin is weighted slightly in favor of heads? Or significantly in favor of heads? The fact that the null hypothesis assigned probability 0.75 to the actual result simply does not bear on these other questions.
2. Neither is $1 - p$ the reliability of the result. *If* the null hypothesis is true, then p does in fact represent the probability that we will get the same result (with respect to the test statistic), should we perform another trial. But of

course we do not know that the null hypothesis is true, even if p is small. We know only that it has not (yet?) been rejected. We could obtain the result we did (or any string of such results) for reasons that have nothing to do with the truth of the null hypothesis.

3. It is very seductive to think that p is the probability that the null hypothesis is true. Doing so appeals to our desire to make a decision, to come to a conclusion, even if probabilistic, about the null hypothesis (see, e.g., Gigerenzer 1993). But it is also wrong. There is a difference between Bayesian updating (which should after all give the posterior probability of the hypothesis regardless of one's commitment to Bayesianism writ large) and significance testing — see [Berger and Sellke, 1987] for an example in which after updating on the new evidence (from a random sample), the hypothesis still has probability > 0.5 , and yet a standard significance test would reject the hypothesis 'at the 5% level'. The failure of this interpretation also illustrates why the others are wrong — for example, if high p indicated that the null hypothesis is true, then it would also be the probability of getting the same result on a second trial.

In addition to avoiding these interpretive pitfalls, there are other issues to consider. Here we mention two.

1. Subjectivists and relative frequentists interpret the probabilities generated by significance tests differently. Indeed, whatever the pitfalls of subjectivism may be, one advantage (ironically, as significance testing is often associated with frequentism), is that the subjectivist is less likely to make the errors mentioned above. There already is a subjectivist account of the posterior probability of any hypothesis, subsequent to the collection of new evidence, namely, Bayesian updating. (Of course, this observation raises the ugly issue of the status of subjectivist Bayesian priors, but we ignore that issue here.)

On the other hand, it is not entirely clear what the relative frequentist ought to say about the probabilities produced by a significance test in at least some cases. The relative frequentist often must appeal to some idealized, infinite, ensemble, or an idealized infinite sequence of trials, in order to make sense of probabilities as long-run limiting relative frequencies. There are (at least) two well-known and much discussed problems with this view: the actual relative frequency in a finite ensemble or sequence can differ arbitrarily from the limiting relative frequency; and even given an infinite ensemble or sequence, there is no guarantee that the required limits exist. Both of these issues raise the question of what grounds the probabilities in this approach — why believe that observed frequencies have anything to do with 'probabilities', or indeed that probabilities, understood in this way, exist at all? (One answer to these questions requires the appeal to an applicable theory or model that supplies probabilities, and is supposed to govern the physical processes under examination.)

These issues are important in the context of significance testing. How ought a relative frequentist interpret the probability that the model provided by a null hypothesis assigns to a given sample? One may be tempted to say that these probabilities represent the limiting relative frequency of infinitely many samples, or trials of the process, but very often this interpretation is meaningless (not to mention its other problems). For example, suppose that the process of interest is sampling from some finite population (say, all wage earners in the United States). There are only finitely many *possible* samples, in that case. It is not clear what sense there is in talking about infinitely many samples in this case, since long ‘before’ we ever had infinitely many samples in hand, we would have sampled the entire population in every possible way.

Again, one answer to these questions is to say that we are in fact sampling from some theoretical population that is infinite, and whose statistical properties are ultimately given by some theory. For example, suppose we are interested in the mean velocity of randomly selected sets of fundamental particles. The actual number of particles in the universe is (let us suppose) finite, and thus (at some appropriate level of coarse-graining) there are only finitely many possible samples of these particles. But we may imagine an infinite ensemble of particles in which the velocity is distributed according to some physical law. If the actual particles in the universe are in fact governed by this law, then one might interpret long-run relative frequencies in terms of an *imagined*, or purely theoretical, infinite sequence of samples from the theoretical infinite ensemble. In other words, with a well-established (or at any rate, accepted) theory in place, it may be possible for the relative frequentist to make sense of the probabilities of a significance test in these theoretical terms. This does not settle the issue of what conclusions, if any, to draw from the test, but it helps with the issue of interpretation.

2. The issue of causation poses a different collection of issues. Does a small value for p indicate a causal relationship of some sort? For example, in the case of intelligence testing (Section 2.2), if z is large enough (hence p small enough), ought we conclude that something about the manner in which the sample was selected is *causally* responsible for the significant difference between the score in the sample and the score in the population?

Suffice it to say that the issue of how and when to draw causal inferences from statistical data is thorny, and hotly debated.⁴ At the very least, we must acknowledge that it will be extremely difficult — perhaps impossible without the appeal to some theory that involves the variable of interest — to identify a specific causal relationship on the basis of a significant difference between a sample and a population, or between two populations. Even the assumption

⁴There is a tradition amongst at least some statisticians of setting it aside completely. Box is often quoted in this context — for example: “Essentially, all models are wrong, but some are useful” [Box and Draper, 1987, 424].

that a significant difference must be due to some (perhaps unknown) causal difference is more an article of faith than a well-established principle. (The literature on causal inference in a probabilistic context is vast. Some recent monographs include [Humphreys, 1989; Eells, 1991; Spirtes *et al.*, 2000; Pearl, 2000].)

A significance test *can* provide inductive evidence that in a certain model — typically, a model specifying “random chance”, as spelled out by the model — is not a good explanation for the data at hand. However, this result does not allow one to distinguish between the various alternatives, such as that the data are the result of some causal process, or some different chance model.

4 OBJECTIONS TO SIGNIFICANCE TESTS

There are numerous objections to significance testing in the literature, though surprisingly, the controversial nature of significance testing rarely shows up in textbooks. (For example, Gliner, Leech and Morgan [2002] examined how significance testing was treated in a variety of textbooks in the social sciences and found the treatments to be uniformly uncritical.) Objections have ranged from cautionary notes (“caution is necessary”, [Mohr, 1990, 74]) to warnings of dangerous irrelevance (“Statistical significance is . . . a diversion from the proper objects of scientific study”, [McCloskey and Ziliak, 2008a, 2]) to outright ridicule (“surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students”, [Rozeboom, 1997, 335]). (Some further notable critiques include [Berkson, 1938; Bakan, 1966; Meehl, 1967; 1990; Carver, 1978; Folger, 1989; Shaver, 1993; Thompson, 1993; Cohen, 1994; Krantz, 1999; Altman, 2004; Ziliak and McCloskey, 2004]. There are many others, including several papers found in the edited collection Morrison and Henkel 1970.) We cannot possibly discuss the full range of critiques, as they have been the subject of full-length books. We focus here on some common or important problems and refer the reader to the extensive literature only hinted at in the citations above for full discussion.

4.1 *The Logic Fails*

While it may be seductive, the argument form suggested in Section 2.5.4 is just not a strong inductive argument (see, e.g., [Baird, 1984]). As we noted there, the general form is: (P1) if p then q has low probability; (P2) in fact, q ; (C) therefore, p has low probability (or, more conservatively, q is evidence against p). But despite its similarity with *modus tollens* (which may be the source of its seductiveness [Seidenfeld, 1979]), this inductive version of *modus tollens* is a bad argument — not merely deductively invalid, but inductively *weak*, i.e., not to be endorsed at this level of generality. For consider the following example: (P1) If John is a soccer player, there is a low probability that he is a goal keeper; (P2)

in fact, John is a goal keeper; (C) therefore, John is likely not a soccer player (or, his being so is evidence against his being a soccer player). To the contrary, (P2) is quite conclusive evidence in *favor* of John's being a soccer player. (Other examples along similar lines are easily constructed.)

This point illustrates the general trouble one can get into when drawing analogies between deductive and inductive reasoning. Similar cases are much discussed. For example, Salmon [1984] and others have suggested that explanations can *lower* the probability of the event or facts to be explained. That is, after having come to know the explanation for fact F , one might find that one's assessment of the probability of F is lower than it was before. (Of course, there are delicate issues here regarding what one *means* by the 'probability of F ', given that F is presumed to be a fact.) Presumably the analogous thing cannot happen for non-probabilistic explanations — if a proposed explanation *entails* the *denial* of F , then presumably the explanation fails. What goes for explanation goes for evidence too: the fact that a null hypothesis confers low probability on some piece of evidence ought not necessarily be taken as evidence against the hypothesis. The example above illustrates why.

4.2 Disagreement with Bayesian Updating

The deviation between the results of a significance test and Bayesian updating (see Section 3) also constitutes an objection against significance testing, because a significance test may reject an hypothesis (in light of some sampling evidence, for example) even if that hypothesis continues, after Bayesian updating, to have high probability. Such will often be the case when the prior probability of the hypothesis is quite high.

Consider, for example, a medical test for some condition, C , whose prevalence in the relevant population is very low, say 1 in 100,000. Our hypothesis is that a given patient does *not* have C . Suppose that the test has sensitivity 90% and specificity 97%. (That is, $\Pr(+|C) = 0.90$ and $\Pr(-|\neg C) = 0.97$, where $\pm$ indicates a positive/negative test result.) A positive test for C is therefore statistically significant at the 5% level — the 'null hypothesis' (that the patient does not have C), together with this information about the test, predicts a probability of less than 0.05 for the positive test result. The logic of significance testing would therefore suggest that we reject the null hypothesis (i.e., assert that the patient has the condition) in the case of a positive test.⁵ However, Bayesian updating (which in this case clearly provides the correct analysis) yields a posterior probability for the null hypothesis of nearly 1. The cause of the discrepancy is clear: significance testing does not take into account the prior probability of the null hypothesis. It treats the most fantastical, unlikely, null hypothesis on a par with well-established theories.

The example makes it clear that this issue is not just an esoteric roadblock

⁵To be clear: the point here is about the logic of significance testing. We are not claiming that anyone would in fact make this mistake.

to the full exposition of a logic of discovery. Real decisions must be made on the basis of these such inductive inferences (see Section 4.3), and adopting the ‘wrong’ methodology can be literally fatal. (See [McCloskey and Ziliak, 2008a] — the subtitle of that book, “How the Standard Error Costs Us Jobs, Justice, and Lives”, raises this issue pointedly.) Indeed, one of the sharpest critiques of the recent movement towards ‘evidence-based’ medicine (which relies heavily on statistical analysis, including significance testing) is precisely that the statistical methods available to us in this context (notably, including significance testing) are simply not up to the task of rationally guiding our actions. This fact has even been expressed in Popperian terms: The method of significance testing when properly pursued is, as we note below, a form of Popperian falsificationism, which is not up to the task of grounding medical practice. (See [Daly, 2005] for an account of evidence-based medicine, and [Shahar, 2003] for the point about Popperian methodology.)

4.3 *Grounding Practical Reason*

As we just suggested, significance testing properly pursued is just a particularly simple form of Popperian falsificationism, in which we reject a null hypothesis in light of ‘evidence’ against it (if even that much inference is licensed — see Sections 4.2 and 4.1), but say nothing about the null hypothesis otherwise. (Popper’s considered position may be somewhat more complex. The literature here is vast. See Popper 1959, 1963, and the exposition in Miller 1994.) Many researchers have commented on the emptiness of a science that is restricted to a simple Popperian methodology: plain falsificationism cannot account for the fact that a primary aim of science, and necessarily so for applied science, is to make *positive* assertions about the world, assertions on which one can — rationally! — act. (Again, the literature is vast, and the debate perhaps intractable. See, for example, [Jeffrey, 1975; Salmon, 1981; Rothbart, 2007] for the criticism mentioned here, and on the intractability of the debate, [Swann, 1988].) Despite Popper’s understandable epistemological scruples about inductive inference, in the end, the practical value of science must rest on its ability to ground action in a rational way, and this requires some reasonable logic of induction that can provide such grounding. Significance testing cannot do so, when it is restricted to its logically permitted role (if even this much is permitted) of licensing the rejection but not the acceptance of scientific hypotheses. (Recall from Section 2.6.1 that exactly this issue may be at the heart of the dispute between Fisher and Neyman-Pearson.)

A further potential problem (with the *application* of significance testing) is that even if one were to ignore the logical issues mentioned above — indeed, even if one were to set aside Popperian scruples about induction — it is far from clear that the result would be a science of *practical* value. For example, it is widely recognized that statistical significance and economic significance are independent of one another. (How widely recognized this point is in the current actual *practice* of economics is a hotly debated issue. See, e.g., [McCloskey, 1998], the highly

critical article by Hoover and Siegler 2008a, the reply by [McCloskey and Ziliak, 2008b], and the reply to the reply by [Hoover and Siegler, 2008b].) For example, a statistically significant difference in the effectiveness of two different investment strategies may be economically insignificant.

The point here is that statistical significance may by itself be completely insensitive to the *relevant* units of analysis (e.g., [Snyder and Lawson, 1993; Ziliak and McCloskey, 2004; Kirk, 2007]). It is constrained not by practically relevant numerical measures (costs, levels of effectiveness, and so on), but by such factors as sample sizes and standard deviations. For example, a very large (in practical terms) difference between two populations may be ‘undetectable’ by significance testing if our samples are small. Similarly, a very small (in practical terms) difference may show up as a strongly significant difference (statistically) if the samples are very large.

These points have practical import. For example, in *Castaneda v. Partida*, 430 U.S. 482 (1977), the Supreme Court established a guideline of “2 or 3 standard deviations” (p -values of 0.05 to 0.01) for determining whether an employer has engaged in employment discrimination. (For a discussion of statistical issues in employment litigation, see, e.g., [Zink and Gutman, 2005; Siskin and Trippi, 2005, 136–143].) But this standard clearly has its limitations (as the Court has acknowledged in more recent cases). For example, an employer with only 10 employees may be discriminating, but the size of the population is sufficiently small that only the most egregious cases would be detectable by a significance test. On the other hand, a careful statistical significance test could detect extremely small salary differences (say, differences of only a few dollars per year), or differences (say, 1 in 10,000) in rates of layoffs, among subgroups of employees of a very large employer. Such differences may quite reasonably be judged statistically, but not economically, or even legally, significant. (Would an employer *really* ‘discriminate’ by paying certain groups only a few dollars less per year, or laying off just 1 in 10,000 ‘extra’ member of a given group, or is some other explanation more likely?) Examples from medicine abound as well. Suppose that in a clinical trial, a cholesterol-lowering drug was found to lower cholesterol by .05mgdL. In a well-designed trial, this result may be strongly statistically significant, but it is doubtful that the drug’s effect should be taken to be medically significant.

Those who oppose significance testing on the grounds that statistical and practical significance differ are apt to advocate for a scientific focus on the *magnitude*, rather than the bare *existence*, of differences amongst populations, or differences between samples and the population from which they are taken. This sort of scientific activity requires formal methods other than significance testing. (See, e.g., [Thompson, 1998; Miller and van der Meulen Rodgers, 2008].)

4.4 *The Problem of Auxiliary Hypotheses*

As has been well established within philosophy of science, the logical gap between simple hypotheses and empirical fact is vast — various auxiliary hypotheses are

always in play. This point goes back at least to Duhem [1914/1954], and was made famous in Anglo-American philosophical circles in a somewhat stronger version by Quine [1951]. In a different guise, the point has been argued at length by Cartwright [1983]. The general idea is easy enough to establish: if one's theory predicts, for example, that the temperature of a given sample of coffee should be 135 °F by 4:00pm then any test of that theory is necessarily going to involve several auxiliary assumptions, such as that the temperature was indeed measured at 4:00pm (the clock was working), the thermometer is accurate, and so on, including many other, more complex and subtle, theoretical, experimental, technological, logical, and epistemological hypotheses, many of which may go unrecognized. Whenever a piece of empirical evidence threatens to overturn some hypothesis, one must keep in mind that these sorts of auxiliary hypothesis may be to blame instead.

Again, the point is not of merely esoteric theoretical interest. To take the most obvious example: the application of significance testing to survey sampling typically requires an assumption that the sample was 'random' with respect to the variable of interest. Suppose, for example, that we wished to determine whether a given intervention affects health in statistically significant way. The usual procedure is via a double-blind test, involving a random sample of subjects some of whom are given the intervention, others not. (See [Cartwright, 2007] and references therein for an extended critique of this sort of testing, however.) If the samples are not random, then the results require further scrutiny.

There are, of course, many famous examples of 'random' samples that were not so [Wheeler, 1976]. The problem of obtaining a random sample is especially intractable in situations where a decent theory of the matter under investigation is unavailable. In that case, one simply does not know which variables are relevant to the variable of interest — which other properties of the units being selected may influence the variable of interest — and thus one does not know how to control for them to produce a truly random sample. It is also worth bearing in mind that in many testing situations, samples (for example, those in need of treatment for some condition) are convenience samples (those willing and available to take part in the treatment), whose relationship to a larger 'population' (whatever it may be) with respect to the variables of interest may be quite unclear.

Of course, there are plenty of other auxiliary hypotheses that are generally required in an application of significance testing. Sometimes these hypotheses can themselves be verified to reasonable precision, but they may also be simply articles of faith.

4.5 Which Level of Statistical Significance Matters?

Suppose that, at least for some specified purpose and experimental design, we have agreed that statistical significance testing is the appropriate way to determine whether some evidence supports or disconfirms a given null hypothesis. We must still decide what *level* of significance we require. It is common to choose either the

1% or the 5% level. It is even more common to suppose that this choice matters, but in fact often the difference is itself statistically *insignificant*.

The point is easiest to see in the context of comparing studies. (We have modified the following example from Gelman and Stern [2006], who discuss this point, and the appearance of this mistake in the literature, in further detail.) Consider two studies of the effectiveness of a pest control on crop harvest. In the first, a difference is found between the control group (no pest control applied) and the treatment group (pest control applied) of 25 pounds per acre of harvest. (So the statistic in this case is a difference of sample means.) The standard error is 10 pounds. (The standard error associated with a random sample is the standard deviation of the distribution of all sample means generated by random samples of the same size, from the same population.) This result is thus statistically significant at around the 1% level. In the second study, a difference of 10 pounds per acre is found, with the same standard error of 10 pounds. This result is not statistically significant.

What should we conclude from these results? Is one of these studies better than the other? And what if the two tests were for two different pest controls? Should we conclude that the first is effective, but the second not (or not shown to be)? Any difference in one's judgment of the two tests would presumably rest on the difference in statistical significance achieved in them. However, that difference itself is not statistically significant. The expected difference between the two studies (i.e., the expected value of the difference between the means generated by the two different samples) is 15, with a standard error of roughly 14. In other words, the difference between these two studies is itself not statistically significant. Hence it is unclear what statistical evidence would ground the preference for one over the other.

Furthermore, consider a third study, with a larger sample size. Suppose it reveals an effect of 2.5 pounds, with a standard error of 1 pound. The statistical significance of this study is the same as that of the first study, but the difference between them is also statistically significant. Moreover, if we were to consider just the null hypothesis that there is an effect, we might be inclined to say that the third study replicates the first, when in fact the magnitude of the effect found in the third study is quite different from that found in the first study — and this difference is statistically significant! The lesson here is that comparing studies of statistical significance is at best a dicey business. (Note also that the issue of the size of the effect has resurfaced — see Section 4.3.)

5 CONCLUSION

Certainly there are defenders of significance testing (e.g., [Abelson, 1995; 1997; Tukey, 1991]), and although we have focused here on the myriad ways in which significance tests can be misleading, or can be misused or even abused, it would be wrong to say that a well-designed significance test tells one nothing. Exactly how much it does tell one is a matter for further discussion.

We will not engage in that discussion here in detail, except to make two observations. The first is that there is an important similarity between this issue and the issue of the strength of hypothetico-deductive reasoning in general. Of course, views on hypothetico-deduction in science span the range from utter rejection (e.g., Popper) to claims that it is *the* central form of reasoning in science (e.g., [Lawson, 2000]). Perhaps more interesting for our purposes are those who take some middle path, cautiously endorsing hypothetico-deduction under appropriate circumstances (e.g., [McMullin, 1984]). Such authors must address the question under what conditions an inference to the (likely) truth of an hypothesis is licensed by the observation of consequences predicted by that hypothesis. Such accounts are too subtle and lengthy to be pursued here. (We refer the reader to [McMullin, 1984] for an example.)

The similarity to significance testing should be clear — the interpretive and logical problems that we have discussed here with significance testing do not constitute a rejection of the entire enterprise, but present a challenge to specify (either generally, or in the context of a specific case) the conditions under which inferences based on significance testing are inductively strong. This is an open project.

This similarity between hypothetico-deductive reasoning and significance testing leads us to our second observation, that the plausibility of such inferences (in both cases) may be greatly enhanced by the theoretical or experimental context in which the inference is made. Significance testing runs into trouble when researchers collect data without much of a clue about the underlying mechanisms that give rise to the data. On the other hand, in situations where a great deal of knowledge about the possible causes of an effect are in place, it is far easier to avoid a mistaken inference or interpretation.

However, it must be emphasized that in such cases, a great deal *more* than statistical significance is at work, and it is fair to wonder just how much inferential work is being done by significance (or lack of it) alone. The discussion above makes it clear, in any case, that bare statistical significance (or lack of it) alone is a very shaky basis on which to draw inferences, despite the siren call of those inferences.

ACKNOWLEDGEMENTS

Thanks to Prasanta Bandyopadhyay and an anonymous referee for very helpful comments on an earlier draft, which has been much improved because of their remarks.

BIBLIOGRAPHY

- [Abelson, 1995] R. P. Abelson. *Statistics as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum, 1995.
- [Abelson, 1997] R. P. Abelson. On the Surprising Longevity of Flogged Horses: Why There Is a Case for the Significance Test, *Psychological Science* 8:12–15, 1997.
- [Altman, 2004] M. Altman. Statistical Significance, Path Dependency, and the Culture of Journal Publication, *Journal of Socio-Economics* 33:651–663, 2004.

- [Azar, 1997] B. Azar. APA Task Force Urges a Harder Look at Data, *APA Monitor*, 28:26, 1997.
- [Bakan, 1966] D. Bakan. The Test of Significance in Psychological Research, *Psychological Bulletin* 66:423–437, 1966.
- [Bakan, 1984] D. Bakan. Tests of Significance Violate the Rule of Implication, in P.D. Asquith and P. Kitcher (eds.), *PSA 1984* 1:81–92, 1984.
- [Baird, 1983a] D. Baird. Conceptions of Scientific Law and Progress in Science, in N. Rescher (ed.), *The Limits of Lawfulness*. Lanham, MD: University Press of America, 33–41, 1983.
- [Baird, 1983b] D. Baird. The Fisher/Pearson Chi Squared Controversy: A Turning Point for Inductive Inference, *The British Journal for the Philosophy of Science* 34:105–118.
- [Baird, 1987] D. Baird. Significance Tests: History and Logic, in N. L. Johnson and S. Kotz (eds.), *Encyclopedia of Statistical Sciences*, vol. 8. New York: John Wiley & Sons, 466–471, 1987.
- [Berkson, 1938] J. Berkson. Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test, *Journal of the American Statistical Association* 33:526–542, 1938.
- [Bland and Bland, 1994] J. M. Bland and D. G. Bland. Statistics Notes: One and Two Sided Tests of Significance. *British Medical Journal* 309:248, 1994.
- [Box, 1953] G. E. F. Box. Non-normality and Test on Variances, *Biometrika* 40:318–335, 1953.
- [Box and Draper, 1987] G. E. P. Box and N. R. Draper. *Empirical Model-Building and Response Surfaces*. Oxford: John Wiley & Sons, 1987.
- [Cartwright, 1983] N. Cartwright. *How the Laws of Physics Lie*. Oxford: The Clarendon Press, 1983.
- [Cartwright, 2007] N. Cartwright. Are RCTs the Gold Standard?, *BioSocieties* 2:11–20, 2007.
- [Carver, 1978] R. P. Carver. The Case Against Statistical Significance Testing, *Harvard Educational Review* 48:378–399, 1978.
- [Cohen, 1994] J. Cohen. The Earth is Round ($p < .05$), *American Psychologist* 49:997–1003, 1994.
- [Daly, 2005] J. Daly. *Evidence-Based Medicine and the Search for a Science of Clinical Care*. California/Milbank Books on Health and the Public, vol. 12. Berkeley, CA: University of California Press, 2005.
- [Duhem, 1914/1954] P. Duhem. *La Théorie Physique Son Objet et Sa Structure*. Paris, Chevalier et Rivière. 1914. Translated by Phillip Wiener as *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press, 1954.
- [Eells, 1991] E. Eells. *Probabilistic Causality*. Cambridge: Cambridge University Press, 1991.
- [Fisher, 1915] R. A. Fisher. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population, *Biometrika* 10:507–521, 1915.
- [Fisher, 1922] R. A. Fisher. On the Mathematical Foundations of Theoretical Statistics, *Philosophical Transactions of the Royal Society, London, Series A* 222:309–368, 1922.
- [Fisher, 1936] R. A. Fisher. Has Mendels Work Been Rediscovered?, *Annals of Science* 1:115–137, 1936.
- [Fisher, 1947] R. A. Fisher. *The Design of Experiments*, Fourth Edition. New York: Hafner Press, 1947.
- [Folger, 1989] R. Folger. Significance Tests and the Duplicity of Binary Decisions, *Psychological Bulletin* 106:155–160, 1989.
- [Franklin et al., 2008] A. Franklin, A. W. F. Edwards, and D. J. Fairbanks. *Ending the Mendel-Fisher Controversy*. Pittsburgh: University of Pittsburgh Press, 2008.
- [Gelman and Stern, 2006] A. Gelman and H. Stern. The Difference Between ‘Significant’ and ‘Not Significant’ Is Not Itself Statistically Significant, *American Statistician* 60: 328–331, 2006.
- [Gigerenzer, 1993] G. Gigerenzer. The Superego, the Ego, and the Id in Statistical Reasoning, in G. Keren and C. Lewis (eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale, NJ: Erlbaum, 311–339, 1993.
- [Gigerenzer et al., 1989] G. Gigerenzer, Z. Swijtink, and T. Porter. *The Empire of Chance*. Cambridge: Cambridge University Press, 1989.
- [Gliner et al., 2002] J. A. Gliner, N. L. Leech, and G. A. Morgan. Problems with Null Hypothesis Significance Testing (NHST): What Do the Textbooks Say?, *The Journal of Experimental Education* 71:83–89, 2002.
- [Gossett, 1908] W. Gossett. The Probable Error of a Mean, *Biometrika* 6:1–25, 1908.
- [Hacking, 1965] I. Hacking. *Logic of Statistical Inference*. New York: Cambridge University Press, 1965.

- [Hacking, 1990] I. Hacking. *The Taming of Chance*. New York: Cambridge University Press 1990.
- [Hoover and Siegler, 2008a] K. D. Hoover and M. V. Siegler. Sound and Fury: McCloskey and Significance Testing in Economics, *Journal of Economic Methodology* 15:1–37, 2008.
- [Hoover and Siegler, 2008b] K. D. Hoover and M. V. Siegler. The Rhetoric of ‘Signifying Nothing’: A Rejoinder to Ziliak and McCloskey, *Journal of Economic Methodology* 15:57–68, 2008.
- [Humphreys, 1989] P. Humphreys. *The Chances of Explanation: Causal Explanations in the Social, Medical, and Physical Sciences*. Princeton: Princeton University Press. 1989.
- [Jeffrey, 1975] R. C. Jeffrey. Probability and Falsification: Critique of the Popper Program, *Synthese* 30:95–117, 1975.
- [Kendall, 1942] M. G. Kendall. On the Future of Statistics *Journal of the Royal Statistical Society* 105:69–80, 1942.
- [Kirk, 2007] R. E. Kirk. Effect Magnitude: A Different Focus, *Journal of Statistical Planning and Inference* 137:1634–1646, 2007.
- [Krantz, 1999] D. H. Krantz. The Null Hypothesis Testing Controversy in Psychology, *Journal of the American Statistical Association* 94:1372–1381, 1999.
- [Kyburg, 1974] H. E. Kyburg. *The Logical Foundations of Statistical Inference*. Boston: D. Reidel, 1974.
- [Lawson, 2000] A. Lawson. The Generality of the Hypothetico-Deductive Method: Making Scientific Thinking Explicit, *American Biology Teacher* 62:482–495, 2000.
- [Lehman, 1993] E. L. Lehman. The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?, *Journal of the American Statistical Association* 88:1242–1249, 1993.
- [Lehman, 1999] E. L. Lehman. ‘Student’ and Small-Sample Theory, *Statistical Science* 14:418–426, 1999.
- [Lenhard, 2006] J. Lenhard. Models and Statistical Inference: The Controversy between Fisher and Neyman-Pearson, *The British Journal for the Philosophy of Science* 57:69–91, 2006.
- [Mayo, 1992] D. G. Mayo. Did Pearson Reject the Neyman-Pearson Philosophy of Statistics?, *Synthese* 90:233–262, 1992.
- [McCloskey, 1998] D. N. McCloskey. *The Rhetoric of Economics*. Madison, WI: University of Wisconsin Press, 1998.
- [McCloskey and Ziliak, 2008a] D. N. McCloskey and S. Ziliak. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, 2008.
- [McCloskey and Ziliak, 2008b] D. N. McCloskey and S. Ziliak. Signifying Nothing: Reply to Hoover and Siegler, *Journal of Economic Methodology* 15:39–55, 2008.
- [McMullin, 1984] E. McMullin. A Case for Scientific Realism, in Jarrett Leplin (ed.), *Scientific Realism*. Berkeley, CA: University of California Press, 8–40, 1984.
- [Meehl, 1967] P. E. Meehl. Theory Testing in Psychology and in Physics: A Methodological Paradox, *Philosophy of Science* 34:103–115, 1967.
- [Meehl, 1990] P. E. Meehl. Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles That Warrant It, *Psychological Inquiry* 1:108–141 1990.
- [Mendel, 1909/1865] G. Mendel. Versuche über Pflanzen-Hybriden, *Verhandlungen des naturforschenden Vereines in Brünn* 4:3–47, 1865. Reprinted and translated as Experiments in Plant-Hybridization in Bateson (1909), *Mendel’s Principles of Heredity*, 318–361, 1909.
- [Miller, 1994] D. Miller. *Critical Rationalism: A Restatement and Defence*. Chicago: Open Court, 1994.
- [Miller and van der Meulen Rodgers, 2008] J. E. Miller and Y. van der Meulen Rodgers. Economic Importance and Statistical Significance: Guidelines for Communicating Empirical Research, *Feminist Economics* 14:117–149, 2008.
- [Mohr, 1990] L. B. Mohr. *Understanding Significance Testing*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-073. Newbury Park, CA: Sage, 1990.
- [Neyman, 1938] J. Neyman. L’Estimation Statistique Traitée Comme in Problème Classique de Probabilité, *Actualités Scientifiques et Industrielles* 739:25–57, 1938.
- [Neyman, 1955] J. Neyman. The Problem of Inductive Inference, *Communications in Pure and Applied Mathematics* 8:13–46, 1955.
- [Neyman and Pearson, 1933] J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses, *Philosophical Transactions of the Royal Society of London, Series A* 231:289–337, 1933.
- [Nickerson, 2000] R. S. Nickerson. Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy, *Psychological Methods* 5:241–301, 2000.

- [Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press, 2000.
- [Pearson, 1900] K. Pearson. On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it Can be Reasonably Supposed to Have Arisen from Random Sampling, reprinted in Pearson (1948), 1900.
- [Pearson, 1930] K. Pearson. *The Life, Letters, and Labours of Francis Galton*, vol. 3a. Cambridge: Cambridge University Press, 1930.
- [Pearson, 1948] K. Pearson. *Karl Pearson's Early Statistical Papers*. E. S. Pearson (ed.). Cambridge: Cambridge University Press, 1948.
- [Pearson, 1955] E. S. Pearson. Statistical Concepts in Their Relation to Reality, *Journal of the Royal Statistical Society, Series B* 17:204–207, 1955.
- [Pearson, 1962] E. S. Pearson. Some Thoughts on Statistical Inference, *Annals of Mathematical Statistics* 33:394–403, 1962.
- [Popper, 1959] K. Popper. *The Logic of Scientific Discovery*. London: Hutchinson, 1959.
- [Popper, 1963] K. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge, 1963.
- [Pyzdek, 2001] T. Pyzdek. Non-Normal Distributions in the Real World, *Journal of the RAC* Second Quarter, 2001:10–17, 2001.
- [Quine, 1951] W. V. O. Quine. Two Dogmas of Empiricism, *The Philosophical Review* 60:20–53, 1951.
- [Rothbart, 2007] D. Rothbart. Popper Against Inductivism, *Dialectica* 34: 121–128, 2007.
- [Rozeboom, 1997] W. W. Rozeboom. Good Science is Abductive, Not Hypothetico-Deductive, in L. L. Harlow, S. A. Mulaik, and J. H. Steiger (eds.), *What If There Were No Significance Tests?*. Hillsdale, NJ: Erlbaum, 335–391, 1997.
- [Salmon, 1981] W. Salmon. Rational Prediction, *British Journal for the Philosophy of Science* 32: 115–125, 1981.
- [Salmon, 1984] W. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press, 1984.
- [Seidenfeld, 1979] T. Seidenfeld. *Philosophical Problems of Statistical Inference*. Boston: D. Reidel, 1979.
- [Seidenfeld, 2008] T. Seidenfeld. , P's in a Pod: Some Recipes for Cooking Mendel's Data, in Franklin et al. (2008), 215–257, 2008.
- [Shahar, 2003] E. Shahar. A Popperian Perspective of the Term 'Evidence-Based Medicine', *Journal of Evaluation in Clinical Practice* 3:109–116, 2003.
- [Shaver, 1993] J. P. Shaver. What Statistical Significance Testing Is, and What It Is Not, *Journal of Experimental Education* 61:293–316, 1993.
- [Siskin and Trippi, 2005] B. R. Siskin and J. Trippi. Employment Discrimination Litigation: Behavioral, Quantitative, and Legal Perspectives, in Frank J. Landy and Eduardo Salas (eds.), *Employment Discrimination Litigation: Behavioral, Quantitative, and Legal Perspectives*. San Francisco, CA: Wiley, 133–166, 2005.
- [Snyder and Lawson, 1993] P. Snyder and S. Lawson. Evaluating Results Using Corrected and Uncorrected Effect Size Estimates, *Journal of Experimental Education* 61:293–316, 1993.
- [Spirtes et al., 2000] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. Cambridge, MA: M.I.T. Press, 2000.
- [Student, 1908] Student [William S. Gosset]. The Probable Error of a Mean, *Biometrika* 6:1–25, 1908.
- [Swann, 1988] A. J. Swann. Popper on Induction, *British Journal for the Philosophy of Science*, 367–373, 1988.
- [Thompson, 1993] B. Thompson. The Use of Statistical Significance Tests in Research: Bootstrap and Other Alternatives, *Journal of Experimental Education* 61:361–377, 1993.
- [Thompson, 1998] B. Thompson. Statistical Significance and Effect Size Reporting: Portrait of a Possible Future, *Research in the Schools* 5:33–38, 1998.
- [Tukey, 1991] J. W. Tukey. The Philosophy of Multiple Comparisons, *Statistical Science* 6:100–116, 1991.
- [Wheeler, 1976] M. Wheeler. *Lies, Damn Lies and Statistics: The Manipulation of Public Opinion in America*. New York: Dell Publishing, 1976.
- [Yule and Greenwood, 1915] G. U. Yule and M. Greenwood. The Statistics of Anti-Typhoid and Anti-Cholera Inoculations and the Interpretation of Such Statistics in General, *Royal Society of Medicine Proceedings, Section of Epidemiology and State Medicine* 8(II):113–194, 1915.

- [Ziliak and McCloskey, 2004] S. Ziliak and D. N. McCloskey. Size Matters: The Standard Error of Regressions in the *American Economic Review*, *Econ Journal Watch* 1:331–358, 2004.
- [Ziliak and McCloskey, 2004] S. Ziliak and D. N. McCloskey. Significance Redux, *Journal of Socio-Economics* 33:665–675, 2004.
- [Zink and Gutman, 2005] D. L. Zink and A. Gutman. Statistical Trends in Private Sector Employment Discrimination Suits, in Frank J. Landy and Eduardo Salas (eds.), *Employment Discrimination Litigation: Behavioral, Quantitative, and Legal Perspectives*. San Francisco, CA: Wiley, 101–131, 2005.

This page intentionally left blank

Bayesian Paradigm

Subjective Bayesianism

Objective Bayesianism

Confirmation Theory and Challenges to it
Bayesianism as a Form of “Logic”

This page intentionally left blank

THE BAYESIAN DECISION-THEORETIC APPROACH TO STATISTICS

Paul Weirich

Statistical inferences rely on probabilities. Probabilities also guide decisions. What implications does probabilities' role in decisions have for their role in statistical inferences? Bayesian statistics employs probabilities that are relative to evidence. These same probabilities ground decisions. Consequently, Bayesian statistics studies decisions to illuminate the probabilities it uses to analyze statistical inferences.

Section 1 explains the Bayesian decision-theoretic approach to statistics. Section 2 describes Bayesian probabilities' function in representations of decision problems. Section 3 considers how to use the relationship between Bayesian probabilities and decisions to illuminate these probabilities. It takes them to be rational degrees of belief and does not define them in terms of preferences. Section 4 explores the properties of Bayesian probabilities. Its main conclusion is that these probabilities respond only to evidence, and not to any nonevidential goal. Section 5 treats Bayesian conditionalization and presents strategies for meeting objections. Section 6 compares Bayesian and classical statistics. It defends Bayesian statistics against the charge of inordinate subjectivism.

1 BAYESIANISM IN INFERENCE AND DECISION

The term Bayesianism has multiple meanings. This should not be surprising. Ambiguity is a common feature of widely used technical terms. In one familiar sense Bayesianism analyzes statistical inferences and decisions made using probabilities. This section describes the key components of Bayesianism understood this way.

1.1 Bayesian Methods

Bayesianism advances conditionalization as a rule of probabilistic inference and maximization of expected utility as a decision rule. Conditionalization uses the probability of a hypothesis conditional on a possible bit of evidence to update the hypothesis's probability if that evidence arrives. A decision's expected utility is a probability-weighted average of the utilities of the decision's possible outcomes. A decision maximizes expected utility if and only if no alternative decision has greater expected utility. Later sections explain these tenets of Bayesianism.

Bayesianism derives its name from Bayes's theorem, which shows how to compute the probability of a hypothesis given some evidence. A simple version of the theorem states that if H is a hypothesis, E is evidence, P is a probability function, and $P(X|Y)$ stands for the probability of X given Y , then $P(H|E) = P(E|H)P(H)/P(E)$. Everyone accepts Bayes's theorem as a component of probability theory, but classical statisticians do not accept its use in statistical reasoning. In the course of statistical reasoning classical statistics, unlike Bayesian statistics, does not compute the posterior probability of a hypothesis H after acquiring evidence E , a quantity equal to $P(H|E)$ according to Bayesian conditionalization. The posterior probability's computation using Bayes's Theorem requires that the prior probability of the hypothesis, $P(H)$, and the prior probability of the evidence, $P(E)$, are both meaningful and have well-defined values. Classical statistics holds that these prior probabilities are generally not available to guide statistical inference. In contrast, Bayesians hold that these probabilities, properly construed, are available. Hence, the controversy concerns applications of Bayes's theorem and the interpretation of the probabilities it regulates rather than the theorem itself.

A statistical test of a hypothesis collects statistical data to evaluate the hypothesis. The Bayesian decision-theoretic approach to statistics uses a statistical test along with prior information to evaluate a hypothesis. Bayesians methods generate the hypothesis's probability at the conclusion of the test. First, they use judgment to assign prior probabilities to the hypothesis and to the test's possible evidential outcomes. Then, in light of the test's actual outcome E , they employ Bayes's theorem to update the probability of the hypothesis from its prior value, $P(H)$, to a new value equal to $P(H|E)$. A last step may be to ground a decision on the results of the statistical test. For example, a pharmaceutical company may use a statistical test to decide whether to market a new drug. The success of marketing the drug depends on the drug's effectiveness. So statistical tests providing information about the drug's effectiveness may direct the company's decision about marketing the drug.

Some statisticians have employed decision methods directly to statistical inference. Lindley [1971] presents an account of decision making that analyzes gathering information and applies Bayes's Theorem to update probabilities. Lindley's approach includes assessment of the value of gathering new information for a decision.

DeGroot [1970] treats as a decision problem selection of an experiment to perform. According to DeGroot, "A statistician attempts to acquire information about the value of some parameter. His reward is the amount of information, appropriately defined, that he obtains about this value through experimentation. He must select an experiment from some available class of experiments, but the information that he will obtain from any particular experiment is random" (p. 88). DeGroot maintains that a statistician should choose experiments that minimize the risk of being mistaken about the truth of a hypothesis. He derives this policy from Wald [1950].

Berger [1993] advocates Bayesian statistics, which employs prior probabilities,

because its methods incorporate information available prior to testing. Prior probabilities often settle whether one should believe a hypothesis after testing. Suppose, for example, that one tests the hypothesis that a musician can distinguish scores written by Haydn from scores written by Mozart. The musician is shown ten scores written by either Haydn or Mozart. He correctly identifies the composer of each. Next, suppose that one tests the hypothesis that a gambler can predict whether a coin toss yields Head or Tails. A fair coin is tossed ten times, and the gambler correctly predicts the result of each toss. The two tests are structurally similar. However, using prior information, one may conclude that the musician is indeed an expert but that, despite his lucky streak, the gambler is a charlatan. Reasonable statistical inference should not ignore such prior information, and Bayesian statistics folds it smoothly into probabilistic reasoning.

Gregory [2006] promotes Bayesian statistics for physics and astronomy because, he claims, it combines deductive and inductive argumentation, and so is more powerful than classical statistics. He treats probability theory as a logic of science, following Jaynes [2003].¹

1.2 Subjective Probability

The type of probability Bayesian statistics employs distinguishes it from classical statistics. Bayesians, in contrast with classical statisticians, employ probabilities that are relative to evidence. These probabilities are subjective because people have different bodies of evidence. Such informationally sensitive probabilities are also often called personal, evidential, or epistemic probabilities. Bayesians differ about the constraints evidence imposes on a person's subjective probabilities, but most hold that it allows some latitude so that two rational people with the same evidence may assign subjective probabilities differently. For example, two rational physicians may reach different diagnoses of a patient's ailment despite having the same data and medical training. However, Bayesians show that the mounting force of incoming evidence will often eventually trump personal epistemic proclivities. As two Bayesians who assign different initial probabilities to a hypothesis update their assignments in response to new relevant evidence, their updated probability assignments will usually converge to the same value, provided that the incoming evidence is ample.²

¹A hypothesis's likelihood on the data is the data's probability given the hypothesis. Classical statistics rejects the likelihood principle as well as Bayesian methods. Edwards [1992: 30] states the likelihood principle as follows: "Within the framework of a statistical model, *all* the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data" (emphasis in the original).

²Savage [1954/1972: Sec. 3.6] presents a theorem about convergence. It shows that evidence reduces the importance of initial probability assignments. He holds that his theorem supports convergence not only to agreement but also to truth. About a typical person, he claims (p. 50), "With the observation of an abundance of relevant data, the person is almost certain to become highly convinced of the truth, and ... knows this to be the case." James Hawthorne's contribution to this volume, "Confirmation Theory," extends Savage's results.

Subjective probabilities, being relative to evidence, differ from objective probabilities that are independent of evidence. Objective probabilities do not change as evidence changes. They depend on natural phenomena and not evidence about them. To illustrate the difference between subjective and objective probabilities, consider the probability of obtaining a black ball during a random draw from an urn with 40 black balls and 60 red balls. The objective probability is 40%. If a person does not know the urn's contents, the person's subjective probability may have a different value. If he thinks that the urn has 50 black balls and 50 red balls, his subjective probability of drawing black may be 50%.

For another example, consider a juror hearing testimony in court during a criminal case. She may favor the defendant's innocence after hearing the defense attorney's presentation. Then after hearing the prosecuting attorney's presentation, she may favor the defendant's guilt. The subjective probability of guilt varies as evidence is presented. Nonetheless, during the presentation of evidence, the objective probability of guilt is constant. It is either 0 or 1 depending on whether the defendant in fact committed the crime. A person's subjective probability assignment to a proposition represents the person's strength of belief that the proposition is true. It ranges from 0, which represents certainty that the proposition is false, to 1, which represents certainty that the proposition is true, and responds to evidence bearing on the proposition's truth.

Objective probabilities may change over time. The objective probability that a random draw from an urn yields a black ball changes over time if the urn has a hole through which its mixture of black and red balls spills. Also, an analogue of the principle of conditionalization holds for objective probabilities. It asserts that if an event E influences the objective probability of a hypothesis H , then, if E occurs and nothing else of relevance occurs, the new objective probability of H equals the previous conditional objective probability of H on E . Despite these similarities, subjective and objective probabilities have different grounds and consequently different basic features.

2 DECISION PROBLEMS

Granting the distinction between objective and subjective probabilities, Bayesianism requires a precise account of the subjective probabilities that it uses. Various accounts are on the market. This section presents the foundations of the most common account, which explicates subjective probabilities by describing their relation to preferences among options in decision problems. The section characterizes decision problems, their representations, and relations among components of their representations such as subjective probabilities.

2.1 *Preferences*

In a decision problem an agent must select an act from all the acts that she may perform. In standard cases, she has a preference ranking of her options. These

are possible acts or, more precisely, possible decisions to perform those acts. The agent's preference ranking of options guides her choice. The usual representation of a decision problem lists an agent's options, possible outcomes of adopting those options, and states of the world that influence options' outcomes. The probabilities of the states yield the probabilities of an option's possible outcomes. The decision problem's representation exhibits the relation between preferences among options and the probabilities and utilities of options' possible outcomes. An option's position in an agent's preference ranking varies with the probabilities and utilities of the act's possible outcomes.

In typical decision problems the state of the world affects the outcome that an act produces. However, the agent often does not know the state of the world. Suppose that a set of mutually exclusive and jointly exhaustive possible states represents all features of the world that affect the various options' outcomes. For each possible state, an option has a specific outcome. Given a decision problem's representation in terms of options, states, and outcomes, standard decision principles apply. For example, in special cases a simple comparison of possible outcomes indicates a preference between a pair of options. If, for example, the outcome of one option is superior to the outcome of another option in each possible state, then the first option strictly dominates the second option. If an option's adoption does not influence the state that obtains, then a rational ideal agent prefers the first, dominating option to the second, dominated option.

There are many ways of elaborating a decision problem's representation in terms of options, states, and outcomes. A major issue concerns the nature of possible outcomes of an agent's options. These possible outcomes are objects of an agent's preferences and also objects of an agent's probability assignments. Among the candidates for outcomes are dated commodity-bundles, events, and propositions. A preference between a pair of dated commodity-bundles is, more precisely, a preference to acquire one dated commodity-bundle rather than the other. Because acquisition of a dated commodity-bundle is an event, taking events rather than dated commodity-bundles as objects of preference better represents the psychology of preference.

The laws of probability govern a Boolean algebra of events formed by applying to basic events operations that yield compound events. Events of many varieties may yield such an algebra. Selecting events with special features to be the objects of probability advances nonstructural, epistemic goals of probabilities. Because the subjective probability an agent assigns to an event depends on the agent's grasp of the event, subjective probabilities fit abstract events-types better than concrete events.

Taking outcomes to be propositions, rather than events-types, is a further improvement, because it makes the objects of preference more fine-grained and responsive to an agent's information. Oedipus did not know that Jocasta was his mother. This explains why he preferred marrying Jocasta to not marrying her but did not prefer marrying his mother to not marrying her. The proposition that he marries Jocasta differs from the proposition that he marries his mother,

even if the corresponding events-types are the same. Taking preferences to compare propositions thus improves fidelity to psychology. Many Bayesians, such as Jeffrey [1965/1983], take propositions rather than events-types as the objects of preference, or they take events-types to be finely individuated by the propositions that describe them. Psychological considerations concerning an agent's grasp of propositions motivate making objects of preference even more fine-grained than are propositions, but this essay puts aside that complication and treats outcomes and other objects of preferences and probability assignments, such as options and states, as propositions.

Using this account of options, states, and outcomes in a decision problem, let us explore the relation between preferences among options and probabilities of states. Suppose that a person at a particular time has preferences among bets. Those preferences may reveal the person's probability assignment. Suppose, for example, that he considers a bet that S and a bet that T , and the stakes for each bet are the same. If he prefers the bet that S to the bet that T , then that preference indicates that he assigns a higher probability to S than to T .

In this way, by considering bets for many possible gambles, one may infer a person's probabilities from his preferences among gambles. For example, suppose that a person is just as ready to bet that it will snow as that it will not snow. Because $P(S \text{ or } \sim S) = 1$ and $P(S) + P(\sim S) = P(S \text{ or } \sim S)$, if $P(S) = P(\sim S)$ then both probabilities equal 0.5. Given four mutually exclusive and jointly exhaustive equiprobable events, one may similarly infer that the probability of each is 0.25. Also, given eight mutually exclusive and jointly exhaustive equiprobable events, one may infer that the probability of each is 0.125. One may measure an arbitrary event's probability as finely as one pleases by comparing it to events on scales formed with disjunctions of equiprobable events.

The utility an agent assigns to a possible outcome of an option is a number that represents the possible outcome's subjective value to the agent. A gamble's expected utility depends on the probabilities and utilities of its possible outcomes. Winning the gamble has a probability and utility. Losing the gamble also has a probability and utility. The product of an outcome's probability and utility indicates the subjective value of the prospect of that outcome. The sum of these products for a gamble's possible outcomes is the gamble's expected utility.

In general, an option's expected utility depends on a set of mutually exclusive and jointly exhaustive possible states. For each state, one considers the outcome of the option in the state and forms the product of the outcome's probability and its utility. The option's expected utility is the sum of the products. Any set of states that are mutually exclusive and jointly exhaustive yields a corresponding set of possible outcomes. That set of possible outcomes in turn yields the option's expected utility. The option's expected utility is the same regardless of the set of mutually exclusive and jointly exhaustive states that one employs to assess it.

An agent's utility assignment depends on the agent's personal goals, and so is subjective. Because utility is sensitive to an agent's information as well as his personal goals, an option's utility equals its expected utility. This basic principle

of utility analysis connects the probabilities and utilities of an option's possible outcomes to the option's position in a rational agent's preference ranking of her options. A rational agent's preferences among options agrees with the options' utilities and therefore with their expected utilities.

2.2 Representation Theorems

A decision problem's presentation lists options, states, and outcomes. An agent has preferences among outcomes and options. She also assigns probabilities to states and utilities to outcomes and options. How are these preferences, probabilities, and utilities related?

Representation theorems show that if a person's preferences comply with certain plausible axioms for relations among preferences, then the person's preferences entail a probability assignment to possible states of the world. Some preference axioms express requirements of rationality. For example, one axiom requires that preferences be transitive. Other axioms are structural. Compliance with them ensures a rich structure of preferences. For example, one structural axiom requires that, for any two options such that a person prefers one to the other, a third option has an intermediate preference rank. To illustrate, suppose that a person prefers betting that Heads turns up on a coin toss to betting that a six turns up on a roll of a die. Then for another proposition such as that drawing a card from a standard deck yields a diamond, she prefers betting that Heads turns up to betting that a diamond is drawn, and prefers betting that a diamond is drawn to betting that a six turns up.

The normative axioms are principles of rationality for the preferences of an ideal agent. Meeting them is necessary for having preferences that follow expected utilities. The nonnormative, structural axioms are not principles of rationality. Meeting them is not necessary for having preferences that follow expected utilities. The structural axioms ensure that preferences among options are extensive enough to derive probabilities of states from those preferences assuming that the preferences follow expected utilities.

A typical representation theorem shows that if a person's preferences concerning gambles satisfy the preference axioms, then there is a unique assignment of probabilities to possible states such that a person prefers betting that one possible state obtains to betting that another possible state obtains just in case the expected utility of the first bet exceeds the expected utility of the second bet.

Some very general representation theorems show that one may infer an agent's probabilities and utilities simultaneously. They show that if an agent's preferences concerning gambles satisfy certain axioms, then there is a unique assignment of probabilities and also a unique assignment of utilities (given a choice of scale) such that the agent prefers one gamble to another just in case the first's expected utility exceeds the second's expected utility. Savage [1954/1972] presents a famous representation theorem of this sort. His proof uses techniques that Ramsey [1931/1990] pioneered. Section 7, an appendix, presents the axioms of preference that Savage's

representation theorem assumes.

A general representation theorem does not justify the laws of probability and the expected utility principle. As explained, it shows only that given any preference ordering of options that satisfies the preference axioms, there is a unique assignment of numbers to possible states of the world complying with the laws of probability, and (given a choice of scale) a unique assignment of numbers to outcomes complying with the laws of utility, such that the expected utilities of options (as computed from these probability and utility assignments) accurately reflect the preference ordering of options. The expected utility principle, for example, requires that an option's utility equal its expected utility. It is a normative principle. Its justification requires normative principles involving probabilities and utilities and not just a representation theorem's normative axioms of preference.

2.3 *Refinements*

The standard formula for an option's expected utility assumes that the option's realization does not affect a state's probability. If an option's realization affects a state's probability, then the probability of the state given the option replaces the probability of the state. The probability of a state S given an option O by definition is a ratio of nonconditional probabilities: $P(S|O) = P(S \& O)/P(O)$. The ratio's being high does not establish that O 's occurrence will probably bring about S . The option O may be a sign of the state S without having any causal influence on S . For example, getting ten Heads on ten tosses in a row may be a sign that a coin is biased for Heads. However, the string of Heads does not causally influence the result of the eleventh toss. Also, although the probability of measles given the characteristic rash is high, the rash does not cause measles. It is just a symptom of the disease. Causal decision theory, as expounded by Joyce [1999], for example, teaches that conditional probabilities used to compute expected utilities should not be standard conditional probabilities because those conditional probabilities do not distinguish causation from mere correlation. Expected utilities that direct action require a type of conditional probability that attends only to causation.³

The utilities of options, given by options' expected utilities, are the grounds of the decision principle of utility maximization. It requires adopting an option with a utility at least as great as the utility of any other option. However, if an option's realization furnishes evidence about which possible state of the world obtains, a decision principle should take account of this evidence, and standard decision theory needs revision to accommodate it. One such revision advances the principle of self-ratification — a principle that generalizes the principle of utility maximization. It requires adopting an option that maximizes utility assuming the option's realization.

To illustrate the difference between utility maximization and self-ratification, consider a person playing a game of matching pennies and trying for a match.

³Freedman [1997: 120–121] also distinguishes two types of conditional probability. One type responds to correlation, and the other responds to causation.

He thinks that his opponent is equally likely to show Heads or to show Tails. His showing Heads thus maximizes utility. However, his opponent is a clever psychologist. If he shows Heads, then that act is evidence that his opponent has predicted it. Given the evidence the act provides, his opponent is likely to show Tails. Thus, his showing Heads does not maximize utility given its realization. It is not self-ratifying.

In some decision problems, no option is self-ratifying. Hence, a more general decision principle is required – one that recommends adopting an option that is self-supporting in a broader sense than being self-ratifying. An option is self-supporting in that broader sense if and only if given its adoption one lacks a sufficient reason for adopting an alternative. Suppose that no option is self-ratifying. Some option may nonetheless be self-supporting. Although given its adoption another option has higher utility, that reason for the other option is not sufficient if all options generate reasons for alternatives.

Decision theorists distinguish standards of evaluation for acts from rational procedures for selecting acts. Whether probability should always have a place in rational procedures for selecting acts is a matter of debate. In some cases the procedure of utility maximization is computationally complex, and its computational cost is not justified by its benefits. Probabilities should not guide action when the guidance they supply has an excessive cost. However, probabilities clearly have a role to play in idealized normative standards of evaluation, where the cost of evaluation is not relevant. For example, the standard of utility maximization may appropriately evaluate a decision even in cases where the standard's application requires high-cost, complex calculations of expected utilities.

3 DEGREES OF BELIEF AND DESIRE

The previous section reviews a decision problem's standard representation and also relationships between the representation's components, in particular, the relationship between probabilities and utilities of options' possible outcomes and preferences among options. How should Bayesian statistics use those relationships to illuminate subjective probabilities?

3.1 *Definition*

A representation theorem assumes that preferences follow expected utilities and then derives probabilities and utilities from the preferences. Some theorists use a representation theorem to ground explicit definitions of subjective probabilities and utilities. They define a person's probabilities and utilities as the values of the functions that the person's preferences imply when the preferences satisfy the theorem's axioms. Accordingly, preferences among options agree with the options' expected utilities by definition. So the injunction to form preferences in accord with expected utilities has no normative force.

The previous section distinguishes a representation theorem's normative and merely structural axioms of preference. The normative axioms of preference are principles of rationality, and complying with them, given the nonnormative, structural axioms, is equivalent to having preferences that imitate agreement with expected utilities. A stronger normative principle requires not only that preferences behave as if they agree with expected utilities, but also requires that preferences agree with expected utilities. This stronger principle assumes that probabilities and utilities are defined independently of preferences and the expected utility principle.

If probabilities and utilities are defined in terms of preferences among options, then methods of using them to form preferences among options seem otiose because those preferences generate the probabilities and utilities. The probabilities and utilities do not generate the preferences. However, in special cases a two-step method, proceeding as follows, generates preferences among options. First, obtain probabilities of states and utilities of outcomes from preferences among a set of options. Suppose that additional options exist involving states and outcomes relevant to the original options. Then one may perform a second step generating preferences among the additional options. One may obtain the additional options' expected utilities using probabilities and utilities already measured and use those expected utilities to form preferences among the additional options. Forming the new preferences this way ensures that the expanded set of preferences complies with the preference axioms and is consistent with the preference ordering of the original options. This method of preference formation assumes that different options may have the same outcomes and that preferences among the original options remain constant during formation of preferences among the additional options.

A person's basic preferences among features of possible worlds, such as personal health, explain the person's preferences among possible worlds and preferences among outcomes. Defining probabilities and utilities in terms of preferences among options does not yield an account of a person's basic preferences among features of possible worlds. The representation theorems work with preferences no more basic than preferences among possible worlds. The utility functions they yield do not analyze the utility of a possible world into the utilities of its features. Possible worlds are the atoms of the preference structures a representation theorem analyzes.

Moreover, satisfying a representation theorem's structural axioms generally requires more preferences than a typical person has. When confronted with gambles concerning unstudied states, many people reasonably do not form preferences among those gambles. If forced to select a gamble, a person's selection does not reveal a genuine preference. For example, a person may not have preferences concerning bets about the number of votes cast in the last presidential election. If that person is forced to select a bet about that number, she may select a bet at random so that her selection does not reflect a preference.

Probabilities and utilities inferred from one set of a person's preferences may differ from probabilities and utilities inferred from another set of the same person's

preferences. To handle this problem, a definition of probabilities and utilities in terms of preferences may restrict itself to cases in which a person's preferences are coherent in the sense that all sets of the person's preferences implying probabilities and utilities imply compatible probabilities and utilities. However, definitions of probabilities and utilities in terms of a person's preferences typically take a different approach. They use a comprehensive set of preferences. That is, they use all preferences that the person has. They assume that those preferences are coherent in the sense of complying with standard normative axioms of preference.

If preferences define probabilities, then updating probabilities is just updating preferences. Representation theorems need additional axioms of preference governing change of preferences over time to permit inference of probabilities from preferences holding at different times. The handiest axiom requires that preferences be updated so that it is as if they agree with expected utilities after updating probabilities by conditionalization.

As this section explains, defining probabilities in terms of preferences creates problems. An alternative to using a representation theorem to ground an explicit definition of probabilities in terms of preferences is to take probabilities as rational degrees of belief attaching to propositions. Rational degrees of belief conform to the laws of probability, as Section 4 argues. They yield a suitable interpretation of the probabilities that those laws govern.

Assigning a degree of belief to a proposition is having a psychological attitude toward that proposition. The attitude is a mental state not defined in terms of preferences but inferable from them if the agent is rational, cognitively ideal, and in ideal circumstances for forming degrees of belief.⁴ Giving degrees of belief an interpretation conceptually independent of preferences is congenial with nonbehavioristic accounts of mental states. Of course, such an interpretation must explain the propositional attitude a degree of belief represents. The next subsection sketches an explanation.

Nothing in the Bayesian decision-theoretic approach to statistics requires that probabilities and utilities be defined in terms of preference. It suffices that probabilities and utilities be inferable from preferences. The representation theorems establish their inferability from preferences assuming that preferences satisfy standard axioms. Because of problems that definitions using preferences generate, the Bayesian decision-theoretic approach to statistics is on firmer ground if it takes probabilities and utilities to be expressions of attitudes toward propositions and takes probabilities and utilities as inferable from preferences, but not defined in terms of preferences.

3.2 *Inference*

A degree of belief attached to a proposition is a degree of belief that the proposition is true. Degrees of beliefs may be implicitly defined by the theories to which they

⁴These idealizations skirt objections that Maher [forthcoming] raises. Weirich [2004] shows how to weaken the idealizations without losing precision.

belong, as in Weirich [2001]. Degrees of belief are quantitative representations of belief states, but do not presume that belief states are themselves quantitative. The belief states and their representations have many independent features. For example, two belief states, one resting on more extensive evidence than the second, may receive the same quantitative representation but may behave differently in response to new information. The more extensively supported belief state may change less rapidly than the less extensively supported belief state changes. A belief state is more complex than representation by a single number indicates. Furthermore, although traditionally degrees of belief use the real number system, belief states may have features that warrant alternative representations. Nonstandard numerical analysis inspires representations of belief states that accommodate infinitesimal degrees of belief.

Taking probabilities as rational degrees of belief strengthens the norms that decision theory imposes. As mentioned, it makes preferences' agreement with expected utilities a normative requirement, not a definitional truth. For example, if betting has higher expected utility than not betting, then the normative principle says that one should prefer betting. In contrast, the corresponding definitional truth holds that by the meaning of expected utility, one prefers betting to not betting. Otherwise, it is false that betting's expected utility exceeds not betting's expected utility.

Assuming that probabilities are rational degrees of belief, that one option has higher expected utility than another explains why a rational person prefers the first option to the second. Expected utilities justify preferences. Another advantage of this interpretation of probabilities is that one may calculate expected utilities to form preferences without extracting probabilities and utilities from preferences already formed. The normative principle to follow expected utility applies to a single preference and does not require constant preferences among some options to generate probabilities of states. Preferences may change from one moment to the next, and need not be the same throughout a period of time during which they entail probabilities and utilities that generate new preferences. Normative principles of preference formation require fewer resources if probabilities are not defined in terms of preferences.

Furthermore, taking probabilities as rational degrees of belief yields a richer account of the factors that affect preferences among options. Because of reliance on representation theorems, some Bayesians constrain an option's outcome so that options may more easily share an outcome. This facilitates extracting outcomes' probabilities and utilities from preferences among options. However, constraining outcomes may exclude factors that affect rational preferences among options. For instance, a rational preference concerning an option may take account of the risk the option runs and the agent's attitude toward that risk. The risk may depend on features of the option such as the agent's distribution of degrees of belief over the option's possible outcomes. Options with a different distribution do not share the same risk. Constraining outcomes to promote shared outcomes conceals the grounds of an agent's preferences. A representation theorem using

such constrained outcomes shows only that an agent's preferences are *as if* they agree with expected utilities and *as if* the agent cares only about the factors in the constrained outcomes. In contrast, taking probabilities as rational degrees of belief permits a comprehensive account of an option's possible outcomes that includes factors such as risk. This comprehensiveness is vital for decision theory because the normative principle of expected-utility maximization is sound only if possible outcomes are comprehensive.

Taking probability and utility as implicitly defined theoretical terms retains the value of representation theorems. These theorems still show that given their assumptions one may infer probabilities and utilities from the rational preferences of an ideally situated ideal agent. Probabilities and utilities still have that grounding in preferences. Other means of inferring probabilities are also possible, however. A person may use the laws of probability to infer some probabilities from others. Also, a person may use introspection to identify some probabilities. For example, he may know that he is certain of some state of the world and so assigns 1 as its probability. He may infer the probability's value without extracting a complete probability assignment from his preferences. Moreover, one may infer probabilities from their causes as well as from their effects. For example, one may infer some probabilities from an agent's evidence. Suppose that an agent knows the objective probability of an event. If he is rational, one may infer that the probability he assigns to the event has the same value as its objective probability. This inference invokes a version of Lewis's [1986: 87] Principal Principle, which moves from knowledge of a proposition's objective probability to a corresponding subjective probability assignment to the proposition.

One may infer a person's degrees of belief from a small set of her preferences. Suppose that a person is willing to buy or sell for \$0.40 a bet that pays \$1 if the state S holds and \$0 if it does not. These betting odds yield a betting quotient of 40%. Given ideal conditions, one may infer that the person's degree of belief that S holds equals 40%. Only that value is compatible with expected-utility maximization using degrees of belief. In general, betting quotients equal degrees of belief. Some theorists take the equality of degrees of belief and betting quotients as a definition of degrees of beliefs. The definition grounds degrees of belief as solidly as does any representation theorem. Also, the definition yields degrees of belief in cases where an expected-utility representation of all preferences does not exist. This extension of degrees of belief is attractive because it extends normative decision principles to more cases. It applies those principles not only when an expected-utility representation of preferences exists but also in other circumstances.

Despite the advantages of defining degrees of belief in terms of betting quotients, a theory of rationality does better, all things considered, taking degrees of belief as implicitly defined theoretical entities. Then it may advance the normative principle that betting quotients should equal degrees of belief. A broad psychological account of degrees of belief enhances a theory of rationality's normative power.

4 PROBABILITY AXIOMS

For degrees of belief to be probabilities, they have to conform to the probability axioms and hence to all the laws of probability. Taking probabilities as rational degrees of belief requires arguing that if degrees of belief are rational, they conform to the axioms, at least in the case of an ideal agent without cognitive limits and in ideal conditions for forming degrees of belief.

4.1 *The Dutch Book Argument*

A common way of arguing for conformity to the probability axioms examines the consequences of using degrees of belief to guide action if they fail to conform to the axioms. De Finetti [1937/1964] pursues this line of argument. He takes an agent's degrees of belief to govern the odds that the agent posts for buying and selling gambles. Consequently, if degrees of belief do not comply with the probability axioms, then the agent's betting quotients license a Dutch book. This is a system of bets that guarantees a net loss. The case for compliance with the probability axioms is called the Dutch book argument. Kemeny [1955] shows that compliance with the probability axioms is necessary and sufficient for preventing Dutch books.

In the Dutch book argument, one may assume that the bookie has the same information that the bettor has and that no gain elsewhere compensates for a system of bets that guarantees a loss. Bets may be on ethically neutral states that do not affect enjoyment of gains and on states eventually within the better's and the bookie's ken so that bets are settled. Making the argument airtight requires many background assumptions, as Kyburg [1983: 81–85] notes. Skyrms [1990: Chap. 5] views the argument as only a loose dramatization of the incoherence of having degrees of belief not conforming to the probability axioms. Christensen [1996] presents an account of that incoherence.

A number of philosophers criticize Dutch book arguments. Kennedy and Chihara [1979], for example, argue that Dutch book arguments fail to establish the intended conclusion that conformance of belief strengths to the probability axioms is necessary and sufficient for rational betting. However, this section presents a different objection, one that grants that a careful version of the Dutch book argument establishes its conclusion about conformity to the laws of probability. The objection maintains that nonetheless the argument does not justify the probability laws for degrees of belief. Because the laws are epistemic principles for degrees of belief, a justification of the laws requires epistemic reasons for complying with the laws. Dutch book arguments offer pragmatic reasons for compliance with the laws of probability. A proper epistemic justification gives a person reasons to comply with the laws even if she does not gamble and does not post odds for bets. Rosenkrantz [1981], for instance, distinguishes cognitive and pragmatic goals. As he argues, “We need . . . an argument to show why incoherent beliefs are irrational from the perspective of the agent's purely cognitive goals” (p. 2.1-4). Dutch book arguments supply no such epistemic justification for believing probabilistically.

Suppose that a representation theorem grounds a method of inferring probabilities and utilities. Then it also furnishes an argument for compliance with the probability laws. It shows that, given a rich preference structure, having degrees of belief that violate a probability law entails violating a normative preference axiom if degrees of belief yield options' expected utilities. However, because this argument depends on axioms of preference, it is not purely epistemic.

Probability has a role in computing expected utilities that guide action. So it has a pragmatic role as well as an epistemic role. Why should one want a purely epistemic argument that rational degrees of belief comply with the laws of probability? One may infer that rational degrees of belief comply with probabilistic laws given their role in action-guiding principles. However, degrees of belief epistemically assess propositions for truth. A pragmatic argument does not explain why it is epistemically appropriate that the laws of probability govern degrees of belief. So a successful epistemic argument offers more than a pragmatic argument does. It offers an explanation of the epistemic force of the probability laws for degrees of belief. It does this by revealing epistemic reasons for having degrees of belief that follow the probability laws. Besides establishing that degrees of belief conform to the laws of probability, it epistemically explains their conformity to those laws.

4.2 Calibration

An epistemic argument for the probability laws may appeal to epistemic goals. One such epistemic goal for degrees of belief is called calibration. An agent's degrees of belief are calibrated if, for each proposition, the agent's degree of belief that the proposition is true equals the proposition's objective probability. Because objective probabilities obey the laws of probability, an agent's degrees of belief equal the corresponding objective probabilities only if those degrees of belief also obey the probability laws. Thus, the epistemic goal of calibration furnishes an appropriate epistemic reason for having degrees of belief that conform to the probability laws.

Calibration is an epistemic goal for degrees of belief in much the same way that the epistemic goal of believing all and only true propositions is an epistemic goal. A person falls short of the goal for belief when she believes a false proposition or fails to believe a true proposition. Calibration is a suitable epistemic goal given that objective probabilities, not truth-values, are physically accessible. A person falls short of this goal for degrees of belief if her degree of belief that a proposition is true is more or less than the proposition's objective probability. A person also fails to reach the goal of calibration with respect to a proposition if she fails to form a degree of belief that the proposition is true. Reaching the goal of calibration has epistemic value, and the further a person departs from the goal, the less epistemic value her degrees of belief have.

Shimony [1988: 80] advances a similar epistemic argument for compliance with the probability laws. He says, "The crux of [his proposal] is the rough idea that epistemic probability is somehow an estimate of relative frequency." Epistemic probability is another term for rational degree of belief, and relative frequency is

a kind of objective probability. Because people are not in a position to achieve perfect calibration, a rational degree of belief is an estimate of an objective probability. A revised calibration argument shows that the probability laws govern rational estimates of objective probabilities, too.

Rosenkrantz [1981] favors a calibration argument that uses Brier scores as a means of calculating how well calibrated degrees of belief are. Van Fraassen (1983) and Joyce (1998) elaborate calibration arguments, too. For present purposes, let us grant the epistemic goal of having degrees of belief that match objective probabilities. Then an important elaboration of this idea is an account of the best means of pursuing the goal of calibration given our ignorance of objective probabilities. Let us also suppose for present purposes that the pursuit of calibration is best accomplished by making estimates that obey the probability laws. Thus, successful degrees of belief accord with the laws of probability.

Granting all these background assumptions, a problem nonetheless looms for this sort of justification of conforming belief strengths to the laws of probability. Forming degrees of belief with an eye on the epistemic goal of calibration runs a risk of forming degrees of belief that do not guide action well. To introduce the problem, I shall first show how belief formation guided by epistemic goals creates a similar risk.

Epistemologists put aside pragmatic reasons for belief, although they acknowledge that pragmatic considerations may justifiably influence belief because of belief's role in communication and other areas besides epistemic inquiry. Their discipline focuses on epistemic reasons for belief and epistemic justification of belief. Some accounts of epistemic justification of belief draw on accounts of rational pursuit of epistemic goals such as holding true beliefs. Levi [1967], for instance, proposes forming beliefs that maximize epistemic utility. Maher [1993: Chaps. 6–8] similarly advocates accepting a hypothesis according to its epistemic utility. A belief's epistemic utility depends on the epistemic gain if it is true and the epistemic loss if it is false. Factors such as informational value affect epistemic gain.

However, forming beliefs because of their informational value makes beliefs poor guides to action. Suppose that a hypothesis H has high informational value so that believing it maximizes epistemic utility despite the risk of error. An agent forms the belief that H and then uses the belief to direct action. Suppose that a possible outcome of a particular act ranks high all things considered, including attainment of epistemic goals such as informational value. That outcome obtains if H is true, and, because the agent believes that H , she performs that act. Informational value's desirability then influences the attractiveness of the act twice, once in making belief that H attractive and a second time in making the act's expected outcome attractive. The result is double counting informational value and overvaluing the act.

Levi's [2006] method of belief and degree of belief formation is susceptible to this problem. He treats beliefs as certainties and uses beliefs as evidence grounding degrees of belief that direct action. Epistemic goals influence belief formation,

formation of degrees of belief, and then action.

According to Levi, whether an agent's inquiry should culminate in accepting a hypothesis H depends on the expected epistemic value for the agent of believing that H . His account of the expected epistemic value of a belief uses credal probabilities. A credal probability is a type of subjective probability that is relative to evidence. Suppose that Q_K is a probability function that for an agent assigns credal probabilities to propositions according to her current corpus of knowledge K . Let α be the agent's degree of caution in belief formation, and suppose that a value assignment V_α assigns for the agent an epistemic value (with respect to α) of believing a hypothesis given a possible truth-value for it. Then the following formula expresses the agent's expected epistemic value with respect to α of believing that H .

$$EV_\alpha(H) = Q_K(H)V_\alpha(H, \text{ true}) + Q_K(\sim H)V_\alpha(H, \text{ false})$$

Rival hypotheses have expected epistemic values also. An agent should accept H only if the expected epistemic value of believing H is at least as great as the expected epistemic value of believing any rival hypothesis.

If probabilities guide action but rest on beliefs adjusted to promote informational value, then epistemic goals are double counted. They count once in adjustments of beliefs and a second time in decisions about action. Action thus accords excessive weight to an agent's epistemic goals. An agent who follows Levi's principle of belief formation may be exploited as a result. Suppose, for example, that she boosts from 0.8 to 1.0 the probability of a hypothesis because accepting the hypothesis maximizes expected epistemic value in light of the hypothesis's high informational value. Before accepting the hypothesis, she may sell for \$0.80 a ticket that pays a dollar if the hypothesis is true and otherwise nothing. After accepting the hypothesis (because acceptance amounts to certainty), she may buy back the ticket for \$1.00. She loses \$0.20 from these transactions. The selling and buying back indicate a preference reversal that occurs without any change in basic goals or evidence. The combination of transactions is irrational.

The only epistemic justification of belief is fit with evidence. Even the epistemic goal of true belief does not justify belief because the goal may lead to a belief that does not fit with evidence. It may prompt excessive epistemic risk-taking. A belief that aims for informational value and not just fit with evidence lacks epistemic justification despite having purely epistemic goals. Some epistemic goals for beliefs may make beliefs unsuitable guides to action. Action-guiding beliefs should maintain fit with evidence.

The kinds of problems that affect formation of beliefs according to epistemic goals also affect formation of degrees of belief to achieve the goal of calibration. Suppose that an agent initially assigns a degree of belief 0.5 to the state S and then discovers, without acquiring additional empirical evidence, that the best prospect of calibration obtains if she boosts to 0.6 her degree of belief that S holds. Then she may sell a dollar gamble that S for \$0.50 and afterwards buy it back for \$0.60, losing \$0.10 without any change in empirical evidence concerning S . The influence

of the goal of calibration leaves the agent open to exploitation.

Letting calibration guide degrees of belief tends towards a form of wishful thinking. Aiming for calibration, an agent may adjust a degree of belief to enhance the prospect of calibration and not just to respond to the strength of evidence. Degrees of belief formed to match objective probabilities may fail to match strengths of evidence. An ideal agent in ideal circumstances should not let anything other than evidence direct probability assignments that control action. Only the goal of representing strength of evidence suits probability's role in guiding action.

Letting epistemic goals influence formation of beliefs and degrees of belief makes belief states the result of pursuit of those goals. An agent risks failure to respond well to strength of evidence if she aims for calibration. Beliefs and degrees of belief guide action best if they just represent evidence's strength (as the agent evaluates it if there is room for judgment). Belief states are not in an agent's direct control, and that makes them suited for guiding decisions that aim for pragmatic goals. Decision principles rely on assessments of strength of evidence, and degrees of belief provide those assessments. If degrees of belief are to guide action well, they should respond only to strength of evidence and not also to calibration. Degrees of belief pursuing calibration may misrepresent strength of evidence, and so fail to guide action well.

Suspension of judgment is often reasonable when evidence is scarce. Similarly, not forming a degree of belief is often reasonable when evidence is scarce. The goal of calibration encourages forming a degree of belief when one ought to suspend judgment because evidence does not warrant precision. It encourages spurious precision of the sort Kaplan [1996, Chap. 1] deplores. Just as an agent may rashly form a belief for the sake of informational value, an agent may rashly form a degree of belief for the sake of calibration. It is better to suspend judgment and forgo prospects of calibration if evidence does not support the degree of belief, that is, if the degree of belief does not accord with strength of evidence. Calibration is a purely epistemic goal and offers a purely epistemic argument for the probability laws. Still, it does not justify the laws. It does not explain their truth. Calibration is not the proper epistemic goal for a justification of the probability laws. Rather, the proper goal is having degrees of belief that accord with evidence.

I assume without argument that the epistemic goal for degrees of belief is matching strength of evidence and only say a few words to clarify this goal. Its consequences are straightforward. For example, if one's evidence for a hypothesis is strong, then ideally one has a high degree of belief that the hypothesis is true. Rational degree of belief may diverge from strength of evidence in some technical senses. However, in the sense of strength of evidence that measures confirmation with respect to total evidence, for an ideal agent in ideal circumstances rational degree of belief matches strength of evidence. The goal of matching strength of evidence is a subsidiary epistemic goal. It serves the primary goal having degree of belief 0 in falsehoods and degree of belief 1 in truths. Given our normal condition of incomplete information, the best means of pursuing the primary goal is pursuing the secondary goal.

Perhaps the correct way to pursue the goal of calibration, given ignorance of objective probabilities, is to form degrees of belief so that they represent strength of evidence. An agent achieves this subsidiary goal when, for each proposition for which the agent has a degree of belief, the agent's degree of belief that the proposition is true matches the agent's strength of evidence for the proposition. If the goal of calibration leads to degrees of belief representing strength of evidence, then pursuing it makes degrees of belief suited to their role in the evaluation of action. Degrees of belief responding to and only to that epistemic goal will not go wrong. However, even if the goal of calibration requires that degrees of belief match strength of evidence, it does not justify the laws of probability. The complaint against Dutch book arguments applies here as well. One may only infer that degrees of belief conform to the probability laws because they pursue well the goal of calibration. That inference is not an explanation of the rationality of their compliance with the laws. The goal of calibration is epistemic, but the argument from calibration fails for the same reason that a pragmatic argument fails. An explanation of the probability laws must show why degrees of belief representing strength of evidence satisfy those laws. An epistemic justification of the laws must show that the goal of representing strength of evidence leads to degrees of belief that comply with those laws. Not every epistemic argument serves this purpose. Because the epistemic job of degrees of belief is representation of strength of evidence, an explanatory argument must show that doing this job requires conformity with the probability laws.

To highlight the problem, suppose that rational pursuit of calibration requires degrees of belief that minimize expected deviation from objective probabilities. The measure of deviation may be designed so that aiming to minimize it requires using degrees of belief that fit evidence. Then perhaps it follows that if degrees of belief are to minimize expected deviation, they must satisfy the probability laws. However, such an argument still does not justify their compliance with the probability laws. Their compliance is only an artifact of the measure of deviation, so the measure of deviation needs justification for the argument to work.

Belief and degree of belief are passive states. They should respond to evidence and not to informational value, calibration, or any goal whether pragmatic or epistemic. An agent should only indirectly control formation of degrees of belief by taking reasonable steps to make the formation process produce degrees of belief that represent strength of evidence. A justification of the probability laws for degrees of belief, an explanatory argument, must show that degrees of belief do not represent strength of evidence unless they conform to the laws of probability.

Proper control may produce justified degrees of belief. One may control degrees of belief so as to make them reflect strength of evidence. Then they are perfect for their role in guiding action. However, controlling degrees of belief to make them represent strength of evidence may have greater costs than benefits, just as removing inconsistencies in beliefs may have greater costs than benefits. Whether an agent should take steps to ensure that her system of degrees of belief has some desirable property depends on the expected costs and benefits of taking such steps.

Similarly, pursuit of calibration may not be a worthwhile exercise of belief-control all things considered.

Perhaps it will shed light on the issue to look at belief-control from the perspective of artificial intelligence. How should one design the formation of degrees of belief for a utility-maximizing robot with multiple goals? A good design incorporates a mechanism for forming degrees of belief according to evidence. It does not let the robot control the formation process. If the robot controls that process, it may well use its control to promote goals besides having degrees of belief that represent strength of evidence. A utility maximizer exercises any control it has to improve its prospects of meeting all its goals. It never acts for the sake of one goal without regard for its other goals. Hence if the robot's mechanism for forming degrees of belief is to have only the goal of representing evidence, then the robot should not control the belief-forming mechanism. Similarly, people have better prospects for degrees of belief that match strength of evidence because they do not directly control formation of degrees of belief.

5 CONDITIONALIZATION

Bayesianism advances the principle of conditionalization for updating probabilities as evidence changes. The principle uses standard conditional probabilities to update nonconditional probabilities when new evidence arrives. This section presents conditionalization and then examines some common misunderstandings of, and objections to, conditionalization.

5.1 *Conditionalization's Features*

According to the principle of conditionalization, when one acquires new evidence E (and no other evidence), the probability of a hypothesis H given the new evidence E , $P_E(H)$, equals the former probability of the hypothesis conditional on the evidence. That is, $P_E(H) = P(H|E) = P(H \& E)/P(E)$. The equality $P_E(H) = P(H|E)$ is a principle of inductive inference. It is not a definition because the first probability exists only after acquisition of the evidence E , whereas the second probability exists prior to acquisition of that evidence.

Applications of conditionalization distinguish prior and posterior probabilities. $P(H)$ and $P(E)$ are called prior probabilities. They are probabilities assigned prior to acquisition of new data and are updated with the acquisition of new data. A prior probability such as $P(E)$ may but need not be calculated using a formula such as $P(E) = P(E|H)P(H) + P(E|\sim H)P(\sim H)$. It represents a rational agent's degree of belief and may be inferred from the agent's betting behavior. $P_E(H)$ is the probability of H after acquiring exactly evidence E . It is a posterior probability as opposed to a prior probability.⁵

⁵The probability of two events together is called a *joint* probability, and the probability of one event alone is called a *marginal* probability. Thus, sometimes $P(H \& E)$ is called a joint probability, and $P(H)$ and $P(E)$ are called marginal probabilities. However, in the context of

To illustrate, suppose that an agent becomes certain that a piece of blue litmus paper turned red after immersion in a liquid. Accordingly, the agent's probability assignment to that proposition equals 1. The proposition's truth is a newly acquired bit of evidence. The principle of conditionalization claims that, after acquiring just that evidence, the probability of the hypothesis that the liquid is acid equals the probability of the hypothesis given that evidence.

Standard Bayesian conditionalization assigns newly learned evidence E a probability of 1 because it understands acquisition of evidence that E as coming to be certain that E . However, Bayesian conditionalization may be generalized to cases where a person's observational experience leaves her less than certain that an evidential proposition E is true. Jeffrey [1965/1983] shows how to generalize conditionalization for cases in which an experience changes a person's probability assignment but does not increase any evidential proposition's probability all the way to 1. For example, an observation of color in bad light may influence a person's probability assignment that way. According to Jeffrey's generalization of conditionalization, when an experience changes a person's probability for an evidential proposition E from a previous value $P_{old}(E)$ to a new probability value $P_{new}(E)$, the new conditional probability of a hypothesis H given the truth of E will generally not change – that is, $P_{new}(H|E) = P_{old}(H|E)$; and similarly for the negation of E , $P_{new}(H|\sim E) = P_{old}(H|\sim E)$. Then, by a theorem of probability theory, the new probability of H based on the experience itself $P_{new}(H)$ equals $P_{old}(H|E)P_{new}(E) + P_{old}(H|\sim E)P_{new}(\sim E)$. Thus, the Bayesian approach to updating on evidence need not be limited to cases in which a person becomes certain of an evidential proposition.

The principle of conditionalization is handicapped in case the new evidence has a prior probability equal to zero. Then the probability of a hypothesis conditional on it is undefined. Supplementary principles may handle that special case. They may authorize a fresh start for probability assignments so that the prior probability of the evidence is discarded and a new, positive assignment takes its place. Also, conditional probabilities with respect to conditions of probability zero may be defined in terms of hypothetical conditionals.

The second approach, the appeal to hypothetical probability assignments, may be warranted, in addition, as a method of addressing the problem of old evidence. That is the problem of explaining how, given the principle of conditionalization, evidence may confirm a hypothesis when the prior probability of the evidence equals 1 because the evidence was known in advance of the hypothesis's formulation. The evidence does not boost the probability of the hypothesis according to the formula $P_E(H) = P(H|E) = P(H \& E)/P(E)$ because if $P(E)$ equals 1, then $P_E(H) = P(H)$. Nonetheless, evidence on hand may support a hypothesis. For example, facts about the planet Mercury's orbit support general relative theory although those facts were known before Einstein formulated his theory. Conditionalization may handle this problem by comparing the hypothetical probability that a hypothesis has independently of a bit of evidence to the probability that

conditionalization, $P(H)$ and $P(E)$ are called prior probabilities.

the hypothesis has given that evidence.

A representation of a dynamic system of preferences that uses probabilities may make conditionalization a definitional truth. However, taking degrees of belief as implicitly defined theoretical entities makes conditionalization a norm for degrees of belief. Dynamic versions of the Dutch book argument, presented for example by Teller [1973] and Skyrms [1990], offer pragmatic reasons for complying with the principle of conditionalization. An epistemic justification of the principle, that is, a justification that grounds the principle in degree of belief's role of representing strength of evidence, may be more desirable for the reasons Section 4 reviews. One attempt to justify conditionalization shows that conditionalization is very likely to lead eventually to high belief strengths for true hypotheses and low belief strengths for false hypotheses. However, this approach does not treat cases in which epistemic concerns focus on the present. In that case conclusions about the future carry no weight. A more promising general approach argues that strength of evidence obeys the principle of conditionalization and that rational degrees of belief match strength of evidence given standard idealizations.

5.2 *Objections to Conditionalization*

The literature presents many problem cases for the principle of conditionalization. Bayesians claim that ideal agents in ideal circumstances comply with the principle. Ideal agents are aware of all logical truths and do not suffer memory losses or other cognitive deficiencies. Some problem cases, such as those Arntzenius [2003] presents, consider whether conditionalization applies when idealizations are relaxed.

Even when all idealizations are satisfied, conditionalization still faces problem cases. In such cases, often the best defense notes that the principle requires conditionalization with respect to exactly the new evidence and not some approximation to it. I elaborate this point with a brief look at Nicod's principle of confirmation.

Some probability theorists, such as Howson and Urbach [1989: 90], advance counterexamples to Jean Nicod's principle of confirmation. That principle holds that discovering a true instance of a generalization confirms the generalization. A typical counterexample goes as follows. Suppose that a person believes that unicorns do not exist. Keeping in mind that a truth-functional conditional with a false antecedent is true, he believes the generalization that all unicorns are white, taken as $(\forall x)(Ux \rightarrow Wx)$, where the arrow expresses the familiar truth-functional connective for formation of a conditional. Then the person discovers a white unicorn, Alexander. His degree of belief that the generalization is true declines. He rejects the generalization because he believes that unicorns, given that they exist, come in various colors, as horses do, and are not exclusively white. In this case a positive instance of a generalization lowers the generalization's probability.

The example not only challenges Nicod's principle but also challenges the principle of conditionalization. According to conditionalization, the probability of a hypothesis H after acquiring evidence that E equals the probability of the hypoth-

esis conditional on the evidence. Let H be the generalization $(\forall x)(Ux \rightarrow Wx)$, and let E be the positive instance $(Ua \rightarrow Wa)$ formed by letting a , for Alexander, replace x . By Bayes's Theorem, $P(H|E) = P(E|H)P(H)/P(E)$. Suppose that $P(H)$ and $P(E)$ are nonzero so that probabilities conditional on H and on E are defined. Because H entails E , $P(E|H) = 1$. Hence $P(H|E) = P(H)/P(E)$. $P(H)/P(E) > P(H)$ because $P(E) < 1$ prior to the discovery that E is true. Therefore $P(H|E) > P(H)$. However, in the example, the probability of the generalization after acquiring the evidence is lower than the prior probability of the hypothesis, namely, $P(H)$. So, apparently, updating does not replace $P(H)$ with $P(H|E)$. It does not proceed by conditionalization. Nonetheless, it is rational.

Interpreting confirmation as an increase in probability, as explained and refined in Horwich [1982: Chap. 3], conditionalization entails Nicod's principle. A generalization entails each instance. Hence if H is a generalization and E is any instance of it, $P(H|E) > P(H)$, provided that $P(H) \neq 0$ and $0 < P(E) < 1$. Recall that $P_E(H)$ is the probability of H after learning that E , which according to conditionalization equals the prior probability of H under the assumption that E , that is, $P(H|E)$. Given that the inequality $P_E(H) > P(H)$ entails E 's confirmation of H , conditionalization entails that a positive instance confirms a generalization. Any counterexample to Nicod's principle is also a counterexample to conditionalization.

Paying attention to the details of conditionalization and Nicod's principle sorts out the trouble. Conditionalization requires updating according to a probability conditional on the total evidence newly acquired. In the example, the instance $(Ua \rightarrow Wa)$ is not the total new evidence. Observation of the white unicorn Alexander yields evidence that $(Ua \& Wa)$. This conjunction is stronger than the instance. Given that E is $(Ua \& Wa)$, $P(E|H)$ does not equal 1 because H does not entail E . Hence $P(H|E) > P(H)$ does not follow. The counterexample to conditionalization dissolves. The counterexample to Nicod's principle also dissolves given that the principle treats only cases in which a positive instance of a generalization is the total evidence newly acquired. Under a charitable interpretation, the principle has this implicit restriction because, obviously, if one were to acquire knowledge of a generalization's positive instance along with other evidence, such as a raft of negative instances, the generalization need not be confirmed.

Some objections to conditionalization trade on a common heuristic for calculating $P(H|E)$. To obtain its value, one imagines the probability of H if one were to learn that E . By definition, $P(H|E)$ is the ratio $P(H \& E)/P(E)$. This ratio equals the probability that H under the assumption that E . It need not equal the probability of H if E were learned. Learning that E often carries the extra information that one learned that E . In most cases that extra information is irrelevant, and $P(H|E)$ equals the probability that H given that one learns that E . However, in some special cases the equality fails. In those cases a defense of conditionalization notes the distinction between conditionalizing on E and conditionalizing on the information one would have if one were to learn that E .

Bertrand's box paradox presents another challenge to conditionalization. Imag-

ine three boxes with two drawers apiece. Each drawer of the first box contains a gold medal. Each drawer of the second box contains a silver medal. One drawer of the third contains a gold medal, and the other drawer a silver medal. At random, a box is selected and one of its drawers is opened. If a gold medal appears, what is the probability that the third box was selected? The probability seems to be $1/2$. By Bayes's theorem and conditionalization its probability is $1/3$.

The inclination to assign probability $1/2$ to selection of the third box may generate a temptation to reject conditionalization. However, one may explain away the inclination by attending to the process that yields a gold medal. Two selection processes come to mind. One may see obtaining a gold medal as selection of a box or as selection of a drawer. The latter correctly describes the effect of the random process that selects a box and then a drawer. Given that process, a gold medal is less probable from the third box than from the first, so the third box is less probable than the first. In many cases that seem to throw doubt on conditionalization, multiple ways of understanding an underlying selection mechanism fit the case. The various mechanisms yield different conditional probabilities to use when updating. Fully specifying the underlying selection mechanism resolves the challenge to conditionalization.

6 BAYESIAN AND CLASSICAL STATISTICS COMPARED

The previous sections present Bayesian statistics and its components. This section briefly compares Bayesian and classical statistics.

6.1 *Subjectivity*

Proponents of classical statistics often fault Bayesian methods for their reliance on subjective probabilities, in particular, subjective probabilities of a hypothesis and evidence prior to statistical tests of the hypothesis. Critics generally do not object to the kind of subjectivity that arises from investigators' having different evidence. Although that relativity to evidence makes probability subjective, it is compatible with personal evidence's objectively settling the hypothesis's probability for a person. Critics object to another kind of relativity. They decline to leave latitude for the exercise of epistemic taste in assigning subjective prior probabilities. According taste a role in that assignment makes subjective probabilities dependent on personal psychology, not just on evidence. The usual defense of Bayesianism points out that some reliance on personal judgment is unavoidable in statistical inference. Classical statistics, although it uses only objective probabilities, also employs some procedures that depend on personal judgment, as the next subsection explains.

6.2 Statistical Tests

A classical statistical test of a hypothesis assigns a probability neither to the hypothesis tested nor to the test's actual outcome. It assigns probabilities only to possible outcomes of the test given the hypothesis. A hypothesis is testable only if its supposition yields well-defined objective probabilities for a test's possible outcomes. A test identifies a critical region formed by a set of possible outcomes. The test selects a critical region so that obtaining an outcome in the critical region has low objective probability given the hypothesis. According to the hypothesis, it is unlikely that the test's outcome lies in the critical region. If the actual outcome falls into the critical region, then the test rejects the hypothesis, so the critical region may also be called the rejection region. For example, take the hypothesis that a coin has two Heads. When tossing the coin, the outcome Tails forms a critical region. A toss that yields Tails produces an outcome in that critical region. Obtaining that outcome thus leads to rejection of the hypothesis.

A typical *significance test* of the sort R. A. Fisher pioneered may investigate, using a critical region for possible outcomes, whether a coin is fair. Fisher called the hypothesis being tested the null hypothesis. Suppose that an experimenter tosses the coin 20 times. One test statistic, a number derived from the test's outcome, is the number of Heads. For a proposed test, one may compute the probability distribution of the number of Heads given the null hypothesis. Suppose that an experimenter sets the significance level at 5%, which means that she decides, in advance of performing the experimental test, to reject the hypothesis given that the test statistic falls into a critical region, containing outcomes of lowest probability given the hypothesis, that has an objective probability of only 5%. Then, on a test consisting of 20 tosses, the test will lead to rejection of the null hypothesis if the number of Heads is more than 14 or less than 6. Otherwise, it does not lead to rejection of the null hypothesis that the coin is fair. Such a significance test is a cornerstone of classical statistics.

Classical statisticians are quick to point out that rejection does not justify belief that the hypothesis is false; and nonrejection does not justify belief that the hypothesis is true. For one thing, besides the test result, there may be other relevant evidence, for example, results from other tests. Similarly, an outcome's falling inside the critical region does not justify concluding that the hypothesis is improbable. Nor does its falling outside the critical region justify concluding that the hypothesis is probable. Whether the hypothesis is improbable or probable in light of the test depends on prior probabilities that classical statistics does not entertain because it takes them to be non-objective and illegitimate.

Does nonrejection justify willingness to act as if the hypothesis is true? That depends on circumstances. One should not unreservedly act as if the hypothesis is true because it may be false despite test results supporting it. Even a very thorough statistical test with excellent credentials may on occasion fail to reject a false hypothesis. To reduce the risk of acting wrongly, an experimenter may adjust the significance level that triggers rejection in a way that reflects the hypothesis's

practical importance. She may set the significance level at 5% instead of 1% to make rejection more likely if nonrejection will lead to a weighty decision. However, that kind of pragmatism is only a rough guide to action. It is better to use the probability of the hypothesis to guide action, if that probability exists. Having that probability, the experimenter is in a position to apply the traditional decision principle of maximizing expected utility.

Bayesians often point out that classical statistics has subjective components, and in that regard is in just as much “trouble” as Bayesian statistics. Suppose that in the example of tossing a coin 20 times, the number of Heads is 5. Given that the coin is fair, that is an improbable result, but so is any other number of Heads. Each possible number of Heads in 20 tosses is improbable given that the coin is fair, although some numbers of Heads are less improbable than are other numbers. For example, although both 10 Heads and 20 Heads are improbable numbers of Heads, 10 Heads is less improbable than 20 Heads because many sequences of tosses produce 10 Heads, whereas only one sequence produces 20 Heads.

Rejection of the null hypothesis depends on the probability of a class of events to which a test’s outcome belongs, not just on the outcome’s probability given the hypothesis. A significance test rejects the hypothesis because the test’s outcome and all those possible outcomes with equal or less probability have probabilities that sum to less than some critical value such as 5%. The test rejects the coin’s fairness if and only if the number of Heads falls into the critical region of low probability. However, there are many other regions of equally low probability. Using a particular critical region is a matter of subjective choice. Howson and Urbach [1989: 160] report, “It seemed to Neyman and Pearson that if Fisher were merely interested in seeing whether the outcome of a trial fell in a region of low probability, he could, for example, have chosen for that region a narrow band in the centre of a bell-shaped distribution just as well as a wide band in its tails.”

A test statistic for a hypothesis’s statistical test summarizes the possible outcomes that the test may yield. Different test statistics use different ways of describing possible outcomes. A test statistic usually discards much information about a test’s actual outcome. Suppose that the example’s test of a coin yields 6 Heads. That the test produced 6 Heads in 20 tosses is just one description of its outcome. Including the order of Heads and Tails in the 20 tosses yields another description. Statistical tests may use a wide variety of test statistics – many of which may seem quite “unnatural.” A test statistic may, for example, count as a single possible outcome the disjunctive result of obtaining 5 Heads or 10 Heads. It may also count as a single possible outcome the disjunctive result of obtaining 14 or 15 Heads. Using these two disjunctive outcomes to reduce the number of possible outcomes may affect whether the null hypothesis is rejected at the 5% level, as Howson and Urbach [1989: 130–132] explain. So test results (rejection or nonrejection) depend on which test statistic is selected.

A test statistic is sufficient just in case it includes all information about possible test outcomes that is relevant to the hypothesis being tested. A minimally sufficient statistic has all the relevant information and no more than is relevant.

Some statisticians respond to the dependency of a test's result on the test statistic selected by proposing use of only a minimally sufficient test statistic. However, whether this proposal adequately deals with the problem is controversial. In any case, Bayesians argue that classical statistics does not objectively ground selection of a test statistic and that selection of a test statistic is a subjective choice.

The case that classical statistics has subjective components is a strong one. A statistical test's critical region, test statistic, and significance level appear to be settled subjectively. So, arguably, classical statistics has no clear advantage over Bayesian statistics with respect to objectivity.

Bayesianism is a robust theory of statistical inference. It has resources for meeting common objections to it. Its elaboration and refinement contribute enormously to the discipline of statistics and to the multitude of disciplines that employ statistical methods.⁶

7 APPENDIX: SAVAGE'S REPRESENTATION THEOREM

This appendix presents the axioms of preference on which Savage's [1954/1972] famous representation theorem relies.

Savage's representation theorem assumes a set of states S with elements $s, s', \dots$ and subsets $A, B, C, \dots$, and also a set of consequences F with elements $f, g, h, \dots$. For an agent, acts are arbitrary functions $\mathbf{f}, \mathbf{g}, \mathbf{h}, \dots$ from S to F . For acts $\mathbf{f}$ and $\mathbf{g}$, the expression $\mathbf{f} \leq \mathbf{g}$ means that the agent does not prefer $\mathbf{f}$ to $\mathbf{g}$. Savage's theorem derives from postulates P1–P7, relying on definitions D1–D5.

- P1 The relation $\leq$ is a simple ordering.
- D1 $\mathbf{f} \leq \mathbf{g}$ given B , if and only if $\mathbf{f}' \leq \mathbf{g}'$ for every $\mathbf{f}'$ and $\mathbf{g}'$ that agree with $\mathbf{f}$ and $\mathbf{g}$, respectively, on B and with each other on $\sim B$ and $\mathbf{g}' \leq \mathbf{f}'$ either for all such pairs or for none.
- P2 For every $\mathbf{f}, \mathbf{g}$, and B , $\mathbf{f} \leq \mathbf{g}$ given B or $\mathbf{g} \leq \mathbf{f}$ given B .
- D2 $g \leq g'$; if and only if $\mathbf{f} \leq \mathbf{f}'$, when $f(s) = g, f'(s) = g'$ for every $s \in S$.
- D3 B is null, if and only if $\mathbf{f} \leq \mathbf{g}$ given B for every $\mathbf{f}, \mathbf{g}$.
- P3 If $f(s) = g, f'(s) = g'$ for every $s \in B$, and B is not null; then $\mathbf{f} \leq \mathbf{f}'$ given B , if and only if $g \leq g'$.
- D4 $A \leq B$; if and only if $\mathbf{f}_A \leq \mathbf{f}_B$ or $g \leq g'$ for every $\mathbf{f}_A, \mathbf{f}_B, g, g'$ such that: $f_A(s) = g$ for $s \in A, f_A(s) = g'$ for $s \in \sim A, f_B(s) = g$, for $s \in B, f_B(s) = g'$ for $s \in \sim B$.
- P4 For every $A, B, A \leq B$ or $B \leq A$.

⁶I thank Kenneth Boyce for bibliographical research for this essay and also thank Prasanta Bandyopadhyay and James Hawthorne for helpful comments on a preliminary draft.

- P5 It is false that, for every $f, f', f \leq f'$.
- P6 Suppose it false that $\mathbf{g} \leq \mathbf{h}$; then, for every f , there is a (finite) partition of S such that, if $\mathbf{g}'$ agrees with $\mathbf{g}$ and $\mathbf{h}'$ agrees with $\mathbf{h}$ except on an arbitrary element of the partition, $\mathbf{g}'$ and $\mathbf{h}'$ being equal to f there, then it will be false that $\mathbf{g}' \leq \mathbf{h}$ or $\mathbf{g} \leq \mathbf{h}'$.
- D5 $\mathbf{f} \leq g$ given $B(g \leq \mathbf{f}$ given B); if and only if $\mathbf{f} \leq \mathbf{h}$ given $B(\mathbf{h} \leq \mathbf{f}$ given B), when $h(s) = g$ for every s .
- P7 If $\mathbf{f} \leq g(s)$ given $B(g(s) \leq \mathbf{f}$ given B) for every $s \in B$, then $\mathbf{f} \leq \mathbf{g}$ given $B(\mathbf{g} \leq \mathbf{f}$ given B).

Using the postulates or axioms of preference, Savage shows that there is a probability function over states and a utility function over consequences (given a choice of scale) such that preferences among acts follow acts' expected utilities. Savage (Sec. 3.3) first shows that a quantitative probability function represents preferences concerning gambles on events, or sets of states. Then he (Sec. 5.4) shows that a utility function (given a choice of scale) represents preferences concerning acts in general, assuming the probability function earlier established and the equality of the act's utility with its expected utility.

BIBLIOGRAPHY

- [Arntzenius, 2003] F. Arntzenius. Some Problems for Conditionalization and Reflection. *Journal of Philosophy* 100: 356–370, 2003.
- [Berger, 1993] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Second edition. New York: Springer, 1993.
- [Christensen, 1996] D. Christensen. Dutch Book Arguments Depigmatized: Epistemic Consistency for Partial Believers. *Journal of Philosophy* 93: 450–479, 1996.
- [De Finetti, 1964] B. De Finetti. Foresight: Its Logical Laws, Its Subjective Sources. In H. E. Kyburg, Jr. and H. E. Smokler, eds., *Studies in Subjective Probability*, pp. 93–158. New York: Wiley, 1964. Originally published in 1937 as *La Prévision: Ses Lois Logiques, Ses Sources Subjectives*. *Annales de l'Institut Henri Poincaré*, 7: 1–68.
- [DeGroot, 1970] M. DeGroot. *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.
- [Edwards, 1992] A. W. F. Edwards. *Likelihood*. Expanded edition. Baltimore: Johns Hopkins University Press, 1992.
- [Gregory, 2005] P. C. Gregory. *Bayesian Logical Data Analysis for the Physical Sciences*. Cambridge: Cambridge University Press, 2005.
- [Horwich, 1982] P. Horwich. *Probability and Evidence*. Cambridge: Cambridge University Press, 1982.
- [Howson and Urbach, 1989] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*. LaSalle, IL: Open Court, 1989.
- [Freedman, 1997] D. A. Freedman. From Association to Causation via Regression. In V. R. Vaughn and S. P. Turner, eds., *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, pp. 113–161. Notre Dame, IN: University of Notre Dame Press, 1997.
- [Jaynes, 2003] E. T. Jaynes. *Probability Theory—The Logic of Science*. G. L. Bretthorst, ed. Cambridge: Cambridge University Press, 2003.
- [Jeffrey, 1983] R. Jeffrey. *The Logic of Decision*. Second edition. Chicago: University of Chicago Press. First edition, 1965. New York: McGraw-Hill, 1983.

- [Joyce, 1998] J. M. Joyce. A Nonpragmatic Vindication of Probabilism. *Philosophy of Science* 65: 575–603, 1998.
- [Joyce, 1999] J. M. Joyce. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press, 1999.
- [Kaplan, 1996] M. Kaplan. *Decision Theory as Philosophy*. Cambridge: Cambridge University Press, 1996.
- [Kemeny, 1955] J. Kemeny. Fair Bets and Inductive Probabilities. *Journal of Symbolic Logic* 20: 263–273, 1955.
- [Kennedy and Chichara, 1979] R. Kennedy and C. Chihara. The Dutch Book Argument: Its Logical Flaws, its Subjective Sources. *Philosophical Studies* 36: 19–33, 1979.
- [Kyburg, 1983] H. Kyburg. *Epistemology and Inference*. Minneapolis, MN: University of Minnesota Press, 1983.
- [Levi, 1967] I. Levi. *Gambling with Truth: An Essay on Induction and the Aims of Science*. New York: Knopf, 1967.
- [Levi, 2006] I. Levi. 2006. Replies. In E. Olsson, ed., *Knowledge and Inquiry: Essays on the Pragmatism of Isaac Levi*, pp. 227–380. Cambridge: Cambridge University Press, 2006.
- [Lewis, 1986] D. Lewis. *Philosophical Papers*. Vol. 2. New York: Oxford University Press, 1986.
- [Lindley, 1971] D. V. Lindley. *Making Decisions*. London: Wiley, 1971.
- [Maher, 1993] P. Maher. *Betting on Theories*. Cambridge: Cambridge University Press, 1993.
- [Maher, forthcoming] P. Maher. Bayesian Probability. *Synthese*, forthcoming.
- [Ramsey, 1990] F. P. Ramsey. Truth and Probability. In D. H. Mellor, ed., *Philosophical Papers*, pp. 52–109. Cambridge: Cambridge University Press. Originally published in 1931. In R. B. Braithwaite, ed., *Foundations of Mathematics and other Essays*, pp. 156–198. London: Routledge & Kegan Paul, 1990.
- [Rosenkrantz, 1981] R. D. Rosenkrantz. *Foundations and Applications of Inductive Probability*. Atascadero, CA: Ridgeview Publishing Company, 1981.
- [Savage, 1972] L. J. Savage. *The Foundations of Statistics*. Second edition. New York: Dover. First edition, 1954. New York: Wiley, 1972.
- [Skyrms, 1990] B. Skyrms. *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press, 1990.
- [Shimony, 1988] A. Shimony. An Adamite Derivation of the Calculus of Probability. In J. H. Fetzer, ed., *Probability and Causality*, pp. 79–89. Dordrecht: Reidel, 1988.
- [Teller, 1973] P. Teller. Conditionalization and Observation. *Synthese* 26: 218–258, 1973.
- [van Fraassen, 1983] B. Van Fraassen. Calibration: A Frequency Justification for Personal Probability. In R. S. Cohen and L. Laudan, eds., *Physics, Philosophy, and Psychoanalysis: Essays in Honor of Adolf Grünbaum*, pp. 295–319. Dordrecht: Reidel, 1983.
- [Wald, 1950] A. Wald. *Statistical Decision Functions*. New York: Wiley, 1950.
- [Weirich, 2001] P. Weirich. *Decision Space: Multidimensional Utility Analysis*. Cambridge: Cambridge University Press, 2001.
- [Weirich, 2004] P. Weirich. *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances*. New York: Oxford University Press, 2004.

This page intentionally left blank

MODERN BAYESIAN INFERENCE: FOUNDATIONS AND OBJECTIVE METHODS

José M. Bernardo

The field of statistics includes two major paradigms: frequentist and Bayesian. Bayesian methods provide a *complete* paradigm for both statistical inference and decision making under uncertainty. Bayesian methods may be derived from an axiomatic system and provide a *coherent* methodology which makes it possible to incorporate relevant initial information, and which solves many of the difficulties which frequentist methods are known to face. If no prior information is to be assumed, a situation often met in scientific reporting and public decision making, a formal initial prior function must be mathematically derived from the assumed model. This leads to *objective* Bayesian methods, objective in the precise sense that their results, like frequentist results, only depend on the assumed model and the data obtained. The Bayesian paradigm is based on an interpretation of probability as a *rational conditional measure of uncertainty*, which closely matches the sense of the word ‘probability’ in ordinary language. Statistical inference about a quantity of interest is described as the modification of the uncertainty about its value in the light of evidence, and Bayes’ theorem specifies how this modification should precisely be made.

1 INTRODUCTION

Scientific experimental or observational results generally consist of (possibly many) sets of data of the general form $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, where the $\mathbf{x}_i$ ’s are somewhat “homogeneous” (possibly multidimensional) observations $\mathbf{x}_i$. Statistical methods are then typically used to derive conclusions on both the nature of the process which has produced those observations, and on the expected behaviour at future instances of the same process. A central element of *any* statistical analysis is the specification of a *probability model* which is assumed to describe the mechanism which has generated the observed data D as a function of a (possibly multidimensional) parameter (vector) $\omega \in \Omega$, sometimes referred to as the *state of nature*, about whose value only limited information (if any) is available. All derived statistical conclusions are obviously conditional on the assumed probability model.

Unlike most other branches of mathematics, frequentist methods of statistical inference suffer from the lack of an axiomatic basis; as a consequence, their proposed desiderata are often mutually incompatible, and the analysis of the same data may well lead to incompatible results when different, apparently intuitive

procedures are tried; see Lindley [1972] and Jaynes [1976] for many instructive examples. In marked contrast, the Bayesian approach to statistical inference is firmly based on axiomatic foundations which provide a unifying logical structure, and guarantee the mutual consistency of the methods proposed. Bayesian methods constitute a *complete* paradigm to statistical inference, a scientific revolution in Kuhn's sense.

Bayesian statistics only require the *mathematics* of probability theory and the *interpretation* of probability which most closely corresponds to the standard use of this word in everyday language: it is no accident that some of the more important seminal books on Bayesian statistics, such as the works of de Laplace [1812], Jeffreys [1939] or de Finetti [1970] are actually entitled "Probability Theory". The practical consequences of adopting the Bayesian paradigm are far reaching. Indeed, Bayesian methods (i) reduce statistical inference to problems in probability theory, thereby minimizing the need for completely new concepts, and (ii) serve to discriminate among conventional, typically frequentist statistical techniques, by either providing a logical justification to some (and making explicit the conditions under which they are valid), or proving the logical inconsistency of others.

The main result from these foundations is the mathematical *need* to describe by means of probability distributions all uncertainties present in the problem. In particular, unknown parameters in probability models *must* have a joint probability distribution which describes the available information about their values; this is often regarded as *the* characteristic element of a Bayesian approach. Notice that (in sharp contrast to conventional statistics) *parameters are treated as random variables* within the Bayesian paradigm. This is not a description of their variability (parameters are typically *fixed unknown* quantities) but a description of the *uncertainty* about their true values.

A most important particular case arises when either no relevant prior information is readily available, or that information is subjective and an "objective" analysis is desired, one that is exclusively based on accepted model assumptions and well-documented public prior information. This is addressed by *reference analysis* which uses information-theoretic concepts to derive formal *reference* prior functions which, when used in Bayes' theorem, lead to posterior distributions encapsulating inferential conclusions on the quantities of interest solely based on the assumed model and the observed data.

In this article it is assumed that probability distributions may be described through their probability density functions, and no distinction is made between a random quantity and the particular values that it may take. Bold italic roman fonts are used for *observable* random vectors (typically data) and bold italic greek fonts are used for unobservable random vectors (typically parameters); lower case is used for variables and calligraphic upper case for their dominion sets. Moreover, the standard mathematical convention of referring to *functions*, say f and g of $\mathbf{x} \in \mathcal{X}$, respectively by $f(\mathbf{x})$ and $g(\mathbf{x})$, will be used throughout. Thus, $\pi(\theta|D, C)$ and $p(\mathbf{x}|\theta, C)$ respectively represent general *probability densities* of the unknown parameter $\theta \in \Theta$ given data D and conditions C , and of the observable random

vector $\mathbf{x} \in \mathcal{X}$ conditional on θ and C . Hence, $\pi(\theta|D, C) \geq 0$, $\int_{\Theta} \pi(\theta|D, C) d\theta = 1$, and $p(\mathbf{x}|\theta, C) \geq 0$, $\int_{\mathcal{X}} p(\mathbf{x}|\theta, C) d\mathbf{x} = 1$. This admittedly imprecise notation will greatly simplify the exposition. If the random vectors are discrete, these functions naturally become probability mass functions, and integrals over their values become sums. Density functions of specific distributions are denoted by appropriate names. Thus, if x is a random quantity with a normal distribution of mean μ and standard deviation σ , its probability density function will be denoted $N(x|\mu, \sigma)$.

Bayesian methods make frequent use of the the concept of logarithmic divergence, a very general measure of the goodness of the approximation of a probability density $p(\mathbf{x})$ by another density $\hat{p}(\mathbf{x})$. The Kullback-Leibler, or *logarithmic divergence* of a probability density $\hat{p}(\mathbf{x})$ of the random vector $\mathbf{x} \in \mathcal{X}$ from its true probability density $p(\mathbf{x})$, is defined as $\kappa\{\hat{p}(\mathbf{x})|p(\mathbf{x})\} = \int_{\mathcal{X}} p(\mathbf{x}) \log\{p(\mathbf{x})/\hat{p}(\mathbf{x})\} d\mathbf{x}$. It may be shown that (i) the logarithmic divergence is non-negative (and it is zero if, and only if, $\hat{p}(\mathbf{x}) = p(\mathbf{x})$ almost everywhere), and (ii) that $\kappa\{\hat{p}(\mathbf{x})|p(\mathbf{x})\}$ is invariant under one-to-one transformations of $\mathbf{x}$.

This article contains a brief summary of the mathematical foundations of Bayesian statistical methods (Section 2), an overview of the paradigm (Section 3), a detailed discussion of objective Bayesian methods (Section 4), and a description of useful objective inference summaries, including estimation and hypothesis testing (Section 5).

Good introductions to objective Bayesian statistics include Lindley [1965], Zellner [1971], and Box and Tiao [1973]. For more advanced monographs, see [Berger, 1985; Bernardo and Smith, 1994].

2 FOUNDATIONS

A central element of the Bayesian paradigm is the use of probability distributions to describe all relevant unknown quantities, interpreting the probability of an event as a conditional measure of uncertainty, on a $[0, 1]$ scale, about the occurrence of the event in some specific conditions. The limiting extreme values 0 and 1, which are typically inaccessible in applications, respectively describe impossibility and certainty of the occurrence of the event. This interpretation of probability includes and extends all other probability interpretations. There are two independent arguments which prove the mathematical inevitability of the use of probability distributions to describe uncertainties; these are summarized later in this section.

2.1 Probability as a Rational Measure of Conditional Uncertainty

Bayesian statistics uses the word *probability* in precisely the same sense in which this word is used in everyday language, as a *conditional measure of uncertainty* associated with the occurrence of a particular event, given the available information

and the accepted assumptions. Thus, $\Pr(E|C)$ is a measure of (presumably rational) belief in the occurrence of the *event* E under *conditions* C . It is important to stress that probability is *always* a function of two arguments, the event E whose uncertainty is being measured, and the conditions C under which the measurement takes place; “absolute” probabilities do not exist. In typical applications, one is interested in the probability of some event E given the available *data* D , the set of *assumptions* A which one is prepared to make about the mechanism which has generated the data, and the relevant contextual *knowledge* K which might be available. Thus, $\Pr(E|D, A, K)$ is to be interpreted as a measure of (presumably rational) belief in the occurrence of the *event* E , given data D , assumptions A and any other available knowledge K , as a measure of how “likely” is the occurrence of E in these conditions. Sometimes, but certainly not always, the probability of an event under given conditions may be associated with the relative frequency of “similar” events in “similar” conditions. The following examples are intended to illustrate the use of probability as a conditional measure of uncertainty.

Probabilistic diagnosis. A human population is known to contain 0.2% of people infected by a particular virus. A person, *randomly selected* from that population, is subject to a test which is from laboratory data known to yield positive results in 98% of infected people and in 1% of non-infected, so that, if V denotes the event that a person carries the virus and $+$ denotes a positive result, $\Pr(+|V) = 0.98$ and $\Pr(+|\bar{V}) = 0.01$. Suppose that the result of the test turns out to be positive. Clearly, one is then interested in $\Pr(V|+, A, K)$, the *probability* that the person carries the virus, given the positive result, the assumptions A about the probability mechanism generating the test results, and the available knowledge K of the prevalence of the infection in the population under study (described here by $\Pr(V|K) = 0.002$). An elementary exercise in probability algebra, which involves Bayes’ theorem in its simplest form (see Section 3), yields $\Pr(V|+, A, K) = 0.164$. Notice that the four probabilities involved in the problem have *the same interpretation*: they are all conditional measures of uncertainty. Besides, $\Pr(V|+, A, K)$ is *both* a measure of the uncertainty associated with the event that the particular person who tested positive is actually infected, *and* an *estimate* of the proportion of people in that population (about 16.4%) that would eventually prove to be infected among those which yielded a positive test. $\triangleleft$

Estimation of a proportion. A survey is conducted to estimate the proportion θ of individuals in a population who share a given property. A random sample of n elements is analyzed, r of which are found to possess that property. One is then typically interested in using the results from the sample to establish regions of $[0, 1]$ where the unknown value of θ may plausibly be expected to lie; this information is provided by *probabilities* of the form $\Pr(a < \theta < b|r, n, A, K)$, a conditional measure of the uncertainty about the event that θ belongs to (a, b) *given* the information provided by the data (r, n) , the assumptions A made on the behaviour of the mechanism which has generated the data (a random sample of n Bernoulli

trials), and any relevant knowledge K on the values of θ which might be available. For example, after a political survey in which 720 citizens out of a random sample of 1500 have declared their support to a particular political measure, one may conclude that $\Pr(\theta < 0.5|720, 1500, A, K) = 0.933$, indicating a probability of about 93% that a referendum on that issue would be lost. Similarly, after a screening test for an infection where 100 people have been tested, none of which has turned out to be infected, one may conclude that $\Pr(\theta < 0.01|0, 100, A, K) = 0.844$, or a probability of about 84% that the proportion of infected people is smaller than 1%. $\triangleleft$

Measurement of a physical constant. A team of scientists, intending to establish the unknown value of a physical constant μ , obtain data $D = \{x_1, \dots, x_n\}$ which are considered to be measurements of μ subject to error. The probabilities of interest are then typically of the form $\Pr(a < \mu < b|x_1, \dots, x_n, A, K)$, the *probability* that the unknown value of μ (fixed in nature, but unknown to the scientists) lies within an interval (a, b) given the information provided by the data D , the assumptions A made on the behaviour of the measurement mechanism, and whatever knowledge K might be available on the value of the constant μ . Again, those probabilities are conditional measures of uncertainty which describe the (necessarily probabilistic) conclusions of the scientists on the true value of μ , given available information and accepted assumptions. For example, after a classroom experiment to measure the gravitational field with a pendulum, a student may report (in m/sec^2) something like $\Pr(9.788 < g < 9.829|D, A, K) = 0.95$, meaning that, under accepted knowledge K and assumptions A , the *observed* data D indicate that the true value of g lies within 9.788 and 9.829 with probability 0.95, a conditional uncertainty measure on a $[0, 1]$ scale. This is naturally compatible with the fact that the value of the gravitational field at the laboratory may well be known with high precision from available literature or from precise previous experiments, but the student may have been instructed *not* to use that information as part of the accepted knowledge K . Under some conditions, it is also true that if the same *procedure* were actually used by many other students with similarly obtained data sets, their reported intervals would actually cover the true value of g in approximately 95% of the cases, thus providing a frequentist *calibration* of the student's probability statement. $\triangleleft$

Prediction. An experiment is made to count the number r of times that an event E takes place in each of n replications of a well defined situation; it is observed that E does take place r_i times in replication i , and it is desired to forecast the number of times r that E will take place in a similar future situation. This is a *prediction* problem on the value of an *observable* (discrete) quantity r , given the information provided by data D , accepted assumptions A on the probability mechanism which generates the r_i 's, and any relevant available knowledge K . Computation of the probabilities $\{\Pr(r|r_1, \dots, r_n, A, K)\}$, for $r = 0, 1, \dots$, is thus

required. For example, the quality assurance engineer of a firm which produces automobile restraint systems may report something like $\Pr(r = 0 | r_1 = \dots = r_{10} = 0, A, K) = 0.953$, after observing that the entire production of airbags in each of $n = 10$ consecutive months has yielded no complaints from their clients. This should be regarded as a measure, on a $[0, 1]$ scale, of the conditional uncertainty, given observed data, accepted assumptions and contextual knowledge, associated with the event that no airbag complaint will come from next month's production and, if conditions remain constant, this is also an estimate of the proportion of months expected to share this desirable property.

A similar problem may naturally be posed with continuous observables. For instance, after measuring some continuous magnitude in each of n randomly chosen elements within a population, it may be desired to forecast the proportion of items in the whole population whose magnitude satisfies some precise specifications. As an example, after measuring the breaking strengths $\{x_1, \dots, x_{10}\}$ of 10 randomly chosen safety belt webbings to verify whether or not they satisfy the requirement of remaining above 26 kN, the quality assurance engineer may report something like $\Pr(x > 26 | x_1, \dots, x_{10}, A, K) = 0.9987$. This should be regarded as a measure, on a $[0, 1]$ scale, of the conditional uncertainty (given observed data, accepted assumptions and contextual knowledge) associated with the event that a randomly chosen safety belt webbing will support no less than 26 kN. If production conditions remain constant, it will also be an estimate of the proportion of safety belts which will conform to this particular specification.

Often, additional information of future observations is provided by related co-variates. For instance, after observing the outputs $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ which correspond to a sequence $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of different production conditions, it may be desired to forecast the output $\mathbf{y}$ which would correspond to a particular set $\mathbf{x}$ of production conditions. For instance, the viscosity of commercial condensed milk is required to be within specified values a and b ; after measuring the viscosities $\{y_1, \dots, y_n\}$ which correspond to samples of condensed milk produced under different physical conditions $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, production engineers will require probabilities of the form $\Pr(a < y < b | \mathbf{x}, (y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n), A, K)$. This is a conditional measure of the uncertainty (always given observed data, accepted assumptions and contextual knowledge) associated with the event that condensed milk produced under conditions $\mathbf{x}$ will actually satisfy the required viscosity specifications. $\triangleleft$

2.2 Statistical Inference and Decision Theory

Decision theory not only provides a precise methodology to deal with decision problems under uncertainty, but its solid axiomatic basis also provides a powerful reinforcement to the logical force of the Bayesian approach. We now summarize the basic argument.

A decision problem exists whenever there are two or more possible courses of action; let $\mathcal{A}$ be the class of possible actions. Moreover, for each $a \in \mathcal{A}$, let Θ_a be the set of *relevant events* which may affect the result of choosing a , and let

$c(a, \theta) \in \mathcal{C}_a$, $\theta \in \Theta_a$, be the *consequence* of having chosen action a when event θ takes place. The class of pairs $\{(\Theta_a, \mathcal{C}_a), a \in \mathcal{A}\}$ describes the *structure* of the decision problem. Without loss of generality, it may be assumed that the possible actions are mutually exclusive, for otherwise one would work with the appropriate Cartesian product.

Different sets of principles have been proposed to capture a minimum collection of logical rules that could sensibly be required for “rational” decision-making. These all consist of axioms with a strong intuitive appeal; examples include the *transitivity* of preferences (if $a_1 > a_2$ given C , and $a_2 > a_3$ given C , then $a_1 > a_3$ given C), and the *sure-thing principle* (if $a_1 > a_2$ given C and E , and $a_1 > a_2$ given C and not E , then $a_1 > a_2$ given C). Notice that these rules are not intended as a description of actual human decision-making, but as a *normative* set of principles to be followed by someone who aspires to achieve coherent decision-making.

There are naturally different options for the set of acceptable principles (see e.g. Ramsey 1926; Savage, 1954; DeGroot, 1970; Bernardo and Smith, 1994, Ch. 2 and references therein), but all of them lead basically to the same conclusions, namely:

- (i) Preferences among consequences should be measured with a real-valued bounded *utility* function $U(c) = U(a, \theta)$ which specifies, on some numerical scale, their desirability.
- (ii) The uncertainty of relevant events should be measured with a set of *probability* distributions $\{\pi(\theta|C, a), \theta \in \Theta_a, a \in \mathcal{A}\}$ describing their plausibility given the conditions C under which the decision must be taken.
- (iii) The desirability of the available actions is measured by their corresponding *expected utility*

$$(1) \quad \bar{U}(a|C) = \int_{\Theta_a} U(a, \theta) \pi(\theta|C, a) d\theta, a \in \mathcal{A}.$$

It is often convenient to work in terms of the non-negative *loss* function defined by

$$(2) \quad L(a, \theta) = \sup_{a \in \mathcal{A}} \{U(a, \theta)\} - U(a, \theta),$$

which directly measures, as a function of θ , the “penalty” for choosing a wrong action. The relative undesirability of available actions $a \in \mathcal{A}$ is then measured by their *expected loss*

$$(3) \quad \bar{L}(a|C) = \int_{\Theta_a} L(a, \theta) \pi(\theta|C, a) d\theta, \quad a \in \mathcal{A}.$$

Notice that, in particular, the argument described above establishes the need to quantify the uncertainty about all relevant unknown quantities (the actual values of the θ 's), and specifies that this quantification *must* have the mathematical structure of probability distributions. These probabilities are conditional on the

circumstances C under which the decision is to be taken, which typically, but not necessarily, include the results D of some relevant experimental or observational data.

It has been argued that the development described above (which is not questioned when decisions have to be made) does not apply to problems of statistical inference, where no specific decision making is envisaged. However, there are two powerful counterarguments to this. Indeed, (i) a problem of statistical inference is typically considered worth analyzing because it *may* eventually help make sensible decisions; a lump of arsenic is poisonous because it *may* kill someone, not because it has actually killed someone [Ramsey, 1926], and (ii) it has been shown [Bernardo, 1979a] that statistical inference on θ actually *has* the mathematical structure of a decision problem, where the class of alternatives is the functional space

$$(4) \quad \mathcal{A} = \left\{ \pi(\theta|D); \quad \pi(\theta|D) > 0, \int_{\Theta} \pi(\theta|D) d\theta = 1 \right\}$$

of the conditional probability distributions of θ given the data, and the utility function is a measure of the amount of information about θ which the data may be expected to provide.

2.3 Exchangeability and Representation Theorem

Available data often take the form of a set $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of “homogeneous” (possibly multidimensional) observations, in the precise sense that only their *values* matter and not the *order* in which they appear. Formally, this is captured by the notion of *exchangeability*. The set of random vectors $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is exchangeable if their joint distribution is invariant under permutations. An infinite sequence $\{\mathbf{x}_j\}$ of random vectors is exchangeable if all its finite subsequences are exchangeable. Notice that, in particular, any random sample from any model is exchangeable in this sense. The concept of exchangeability, introduced by de Finetti [1937], is central to modern statistical thinking. Indeed, the general *representation theorem* implies that if a set of observations is assumed to be a subset of an exchangeable sequence, then it constitutes a *random sample* from some probability model $\{p(\mathbf{x}|\omega), \omega \in \Omega\}$, $\mathbf{x} \in \mathcal{X}$, described in terms of (labeled by) some *parameter vector* ω ; furthermore this parameter ω is *defined* as the limit (as $n \rightarrow \infty$) of some function of the observations. Available information about the value of ω in prevailing conditions C is *necessarily* described by *some* probability distribution $\pi(\omega|C)$.

For example, in the case of a sequence $\{x_1, x_2, \dots\}$ of dichotomous exchangeable random quantities $x_j \in \{0, 1\}$, de Finetti’s representation theorem establishes that the joint distribution of $(x_1, \dots, x_n)$ has an *integral representation* of the form

$$(5) \quad p(x_1, \dots, x_n|C) = \int_0^1 \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} \pi(\theta|C) d\theta, \quad \theta = \lim_{n \rightarrow \infty} \frac{r}{n},$$

where $r = \sum x_j$ is the number of positive trials. This is nothing but the joint distribution of a set of (conditionally) independent Bernoulli trials with parameter θ , over which some probability distribution $\pi(\theta|C)$ is therefore proven to exist. More generally, for sequences of arbitrary random quantities $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$, exchangeability leads to integral representations of the form

$$(6) \quad p(\mathbf{x}_1, \dots, \mathbf{x}_n|C) = \int_{\Omega} \prod_{i=1}^n p(\mathbf{x}_i|\omega) \pi(\omega|C) d\omega,$$

where $\{p(\mathbf{x}|\omega), \omega \in \Omega\}$ denotes some probability *model*, ω is the limit as $n \rightarrow \infty$ of some function $f(\mathbf{x}_1, \dots, \mathbf{x}_n)$ of the observations, and $\pi(\omega|C)$ is some probability distribution over Ω . This formulation includes “nonparametric” (distribution free) modelling, where ω may index, for instance, all continuous probability distributions on $\mathcal{X}$. Notice that $\pi(\omega|C)$ does *not* describe a possible variability of ω (since ω will typically be a fixed *unknown* vector), but a description on the uncertainty associated with its actual value.

Under appropriate conditioning, exchangeability is a very general assumption, a powerful extension of the traditional concept of a *random sample*. Indeed, many statistical analyses directly assume data (or subsets of the data) to be a random sample of conditionally independent observations from some probability model, so that $p(\mathbf{x}_1, \dots, \mathbf{x}_n|\omega) = \prod_{i=1}^n p(\mathbf{x}_i|\omega)$; but *any* random sample is exchangeable, since $\prod_{i=1}^n p(\mathbf{x}_i|\omega)$ is obviously invariant under permutations. Notice that the observations in a random sample are only independent *conditional* on the parameter value ω ; as nicely put by Lindley, the mantra that the observations $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ in a random sample are independent is ridiculous when they are used to infer $\mathbf{x}_{n+1}$. Notice also that, under exchangeability, the general representation theorem provides an *existence theorem* for a probability distribution $\pi(\omega|C)$ on the parameter space Ω , and that this is an argument which only depends on mathematical probability theory.

Another important consequence of exchangeability is that it provides a formal *definition* of the parameter ω which labels the model as the limit, as $n \rightarrow \infty$, of *some* function $f(\mathbf{x}_1, \dots, \mathbf{x}_n)$ of the observations; the function f obviously depends both on the assumed model and the chosen parametrization. For instance, in the case of a sequence of Bernoulli trials, the parameter θ is *defined* as the limit, as $n \rightarrow \infty$, of the relative frequency r/n . It follows that, under exchangeability, the sentence “the true value of ω ” has a well-defined meaning, if only asymptotically verifiable. Moreover, if two different models have parameters which are functionally related by their definition, then the corresponding posterior distributions may be meaningfully compared, for they refer to functionally related quantities. For instance, if a finite subset $\{x_1, \dots, x_n\}$ of an exchangeable sequence of integer observations is assumed to be a random sample from a Poisson distribution $\text{Po}(x|\lambda)$, so that $E[x|\lambda] = \lambda$, then λ is *defined* as $\lim_{n \rightarrow \infty} \{\bar{x}_n\}$, where $\bar{x}_n = \sum_j x_j/n$; similarly, if for some fixed non-zero integer r , the same data are assumed to be a random sample for a negative binomial $\text{Nb}(x|r, \theta)$, so that $E[x|\theta, r] = r(1 - \theta)/\theta$, then θ is *defined* as $\lim_{n \rightarrow \infty} \{r/(\bar{x}_n + r)\}$. It follows that $\theta \equiv r/(\lambda + r)$ and, hence,

θ and $r/(\lambda + r)$ may be treated as the *same* (unknown) quantity whenever this might be needed as, for example, when comparing the relative merits of these alternative probability models.

3 THE BAYESIAN PARADIGM

The statistical analysis of some observed data D typically begins with some informal *descriptive* evaluation, which is used to suggest a tentative, formal *probability model* $\{p(D|\omega), \omega \in \Omega\}$ assumed to represent, for some (unknown) value of ω , the probabilistic mechanism which has generated the observed data D . The arguments outlined in Section 2 establish the logical need to assess a *prior* probability distribution $\pi(\omega|K)$ over the parameter space Ω , describing the available knowledge K about the value of ω prior to the data being observed. It then follows from standard probability theory that, if the probability model is correct, all available information about the value of ω after the data D have been observed is contained in the corresponding *posterior* distribution whose probability density, $\pi(\omega|D, A, K)$, is immediately obtained from Bayes' theorem,

$$(7) \quad \pi(\omega|D, A, K) = \frac{p(D|\omega) \pi(\omega|K)}{\int_{\Omega} p(D|\omega) \pi(\omega|K) d\omega},$$

where A stands for the assumptions made on the probability model. It is this systematic use of Bayes' theorem to incorporate the information provided by the data that justifies the adjective *Bayesian* by which the paradigm is usually known. It is obvious from Bayes' theorem that any value of ω with zero prior density will have zero posterior density. Thus, it is typically assumed (by appropriate restriction, if necessary, of the *parameter space* Ω) that prior distributions are *strictly positive* (as Savage put it, keep the mind open, or at least ajar). To simplify the presentation, the accepted assumptions A and the available knowledge K are often omitted from the notation, but the fact that *all* statements about ω given D are *also* conditional to A and K should always be kept in mind.

EXAMPLE 1 Bayesian inference with a finite parameter space. Let $p(D|\theta)$, $\theta \in \{\theta_1, \dots, \theta_m\}$, be the probability mechanism which is assumed to have generated the observed data D , so that θ may only take a *finite* number of values. Using the finite form of Bayes' theorem, and omitting the prevailing conditions from the notation, the posterior probability of θ_i after data D have been observed is

$$(8) \quad \Pr(\theta_i|D) = \frac{p(D|\theta_i) \Pr(\theta_i)}{\sum_{j=1}^m p(D|\theta_j) \Pr(\theta_j)}, \quad i = 1, \dots, m.$$

For any prior distribution $p(\theta) = \{\Pr(\theta_1), \dots, \Pr(\theta_m)\}$ describing available knowledge about the value of θ , $\Pr(\theta_i|D)$ measures how likely should θ_i be judged, given both the initial knowledge described by the prior distribution, and the information provided by the data D .

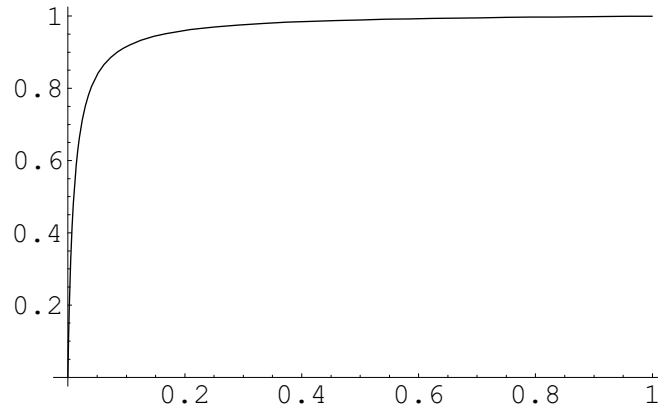


Figure 1. Posterior probability of infection $\Pr(V|+)$ given a positive test, as a function of the prior probability of infection $\Pr(V)$

An important, frequent application of this simple technique is provided by probabilistic diagnosis. For example, consider the simple situation where a particular test designed to detect a virus is known from laboratory research to give a positive result in 98% of infected people and in 1% of non-infected. Then, the posterior probability that a person who tested positive is infected is given by $\Pr(V|+) = (0.98p) / \{0.98p + 0.01(1-p)\}$ as a function of $p = \Pr(V)$, the prior probability of a person being infected (the *prevalence* of the infection in the population under study). Figure 1 shows $\Pr(V|+)$ as a function of $\Pr(V)$.

As one would expect, the posterior probability is only zero if the prior probability is only zero (so that it is *known* that the population is free of infection) and it is only one if the prior probability is one (so that it is *known* that the population is universally infected). Notice that if the infection is rare, then the posterior probability of a randomly chosen person being infected will be relatively low even if the test is positive. Indeed, for say $\Pr(V) = 0.002$, one finds $\Pr(V|+) = 0.164$, so that in a population where only 0.2% of individuals are infected, only 16.4% of those testing positive within a random sample will actually prove to be infected: most positives would actually be *false* positives.

In this section, we describe in some detail the learning process described by Bayes' theorem, discuss its implementation in the presence of nuisance parameters, show how it can be used to forecast the value of future observations, and analyze its large sample behaviour.

3.1 The Learning Process

In the Bayesian paradigm, the process of learning from the data is systematically implemented by making use of Bayes' theorem to combine the available prior

information with the information provided by the data to produce the required posterior distribution. Computation of posterior densities is often facilitated by noting that Bayes' theorem may be simply expressed as

$$(9) \quad \pi(\omega|D) \propto p(D|\omega) \pi(\omega),$$

(where $\propto$ stands for 'proportional to' and where, for simplicity, the accepted assumptions A and the available knowledge K have been omitted from the notation), since the missing proportionality constant $[\int_{\Omega} p(D|\omega) \pi(\omega) d\omega]^{-1}$ may always be deduced from the fact that $\pi(\omega|D)$, a probability density, must integrate to one. Hence, to identify the form of a posterior distribution it suffices to identify a *kernel* of the corresponding probability density, that is a function $k(\omega)$ such that $\pi(\omega|D) = c(D)k(\omega)$ for some $c(D)$ which does not involve ω . In the examples which follow, this technique will often be used.

An *improper prior function* is defined as a positive function $\pi(\omega)$ such that $\int_{\Omega} \pi(\omega) d\omega$ is not finite. Equation (9), the formal expression of Bayes' theorem, remains technically valid if $\pi(\omega)$ is an improper prior function provided that $\int_{\Omega} p(D|\omega) \pi(\omega) d\omega < \infty$, thus leading to a well defined *proper* posterior density $\pi(\omega|D) \propto p(D|\omega) \pi(\omega)$. In particular, as will later be justified (Section 4) it also remains philosophically valid if $\pi(\omega)$ is an appropriately chosen *reference* (typically improper) prior function.

Considered as a function of ω , $l(\omega, D) = p(D|\omega)$ is often referred to as the *likelihood function*. Thus, Bayes' theorem is simply expressed in words by the statement that *the posterior is proportional to the likelihood times the prior*. It follows from equation (9) that, provided the *same* prior $\pi(\omega)$ is used, two different data sets D_1 and D_2 , with possibly different probability models $p_1(D_1|\omega)$ and $p_2(D_2|\omega)$ but yielding *proportional* likelihood functions, will produce identical posterior distributions for ω . This immediate consequence of Bayes' theorem has been proposed as a principle on its own, the *likelihood principle*, and it is seen by many as an obvious requirement for reasonable statistical inference. In particular, for any given prior $\pi(\omega)$, the posterior distribution does not depend on the set of possible data values, or the *sample space*. Notice, however, that the likelihood principle only applies to inferences about the parameter vector ω once the data have been obtained. Consideration of the sample space is essential, for instance, in model criticism, in the design of experiments, in the derivation of predictive distributions, and in the construction of objective Bayesian procedures.

Naturally, the terms prior and posterior are only *relative* to a particular set of data. As one would expect from the coherence induced by probability theory, if data $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ are sequentially presented, the final result will be the same whether data are globally or sequentially processed. Indeed, $\pi(\omega|\mathbf{x}_1, \dots, \mathbf{x}_{i+1}) \propto p(\mathbf{x}_{i+1}|\omega) \pi(\omega|\mathbf{x}_1, \dots, \mathbf{x}_i)$, for $i = 1, \dots, n-1$, so that the "posterior" at a given stage becomes the "prior" at the next.

In most situations, the posterior distribution is "sharper" than the prior so that, in most cases, the density $\pi(\omega|\mathbf{x}_1, \dots, \mathbf{x}_{i+1})$ will be more concentrated around the true value of ω than $\pi(\omega|\mathbf{x}_1, \dots, \mathbf{x}_i)$. However, this is not always the case: oc-

casionaly, a “surprising” observation will increase, rather than decrease, the uncertainty about the value of ω . For instance, in probabilistic diagnosis, a sharp posterior probability distribution (over the possible causes $\{\omega_1, \dots, \omega_k\}$ of a syndrome) describing, a “clear” diagnosis of disease ω_i (that is, a posterior with a large probability for ω_i) would typically update to a less concentrated posterior probability distribution over $\{\omega_1, \dots, \omega_k\}$ if a new clinical analysis yielded data which were unlikely under ω_i .

For a given probability model, one may find that a particular function of the data $\mathbf{t} = \mathbf{t}(D)$ is a *sufficient* statistic in the sense that, given the model, $\mathbf{t}(D)$ contains all information about ω which is available in D . Formally, $\mathbf{t} = \mathbf{t}(D)$ is sufficient if (and only if) there exist nonnegative functions f and g such that the likelihood function may be factorized in the form $p(D|\omega) = f(\omega, \mathbf{t})g(D)$. A sufficient statistic always exists, for $\mathbf{t}(D) = D$ is obviously sufficient; however, a much simpler sufficient statistic, with a fixed dimensionality which is independent of the sample size, often exists. In fact this is known to be the case whenever the probability model belongs to the *generalized exponential family*, which includes many of the more frequently used probability models. It is easily established that if $\mathbf{t}$ is sufficient, the posterior distribution of ω only depends on the data D through $\mathbf{t}(D)$, and may be directly computed in terms of $p(\mathbf{t}|\omega)$, so that, $\pi(\omega|D) = p(\omega|\mathbf{t}) \propto p(\mathbf{t}|\omega) \pi(\omega)$.

Naturally, for fixed data and model assumptions, different priors lead to different posteriors. Indeed, Bayes’ theorem may be described as a data-driven probability transformation machine which maps prior distributions (describing prior knowledge) into posterior distributions (representing combined prior and data knowledge). It is important to analyze whether or not sensible changes in the prior would induce noticeable changes in the posterior. Posterior distributions based on reference “noninformative” priors play a central role in this *sensitivity analysis* context. Investigation of the sensitivity of the posterior to changes in the prior is an important ingredient of the comprehensive analysis of the sensitivity of the final results to *all* accepted assumptions which any responsible statistical study should contain.

EXAMPLE 2 Inference on a binomial parameter. If the data D consist of n Bernoulli observations with parameter θ which contain r positive trials, then $p(D|\theta, n) = \theta^r(1 - \theta)^{n-r}$, so that $\mathbf{t}(D) = \{r, n\}$ is sufficient. Suppose that prior knowledge about θ is described by a Beta distribution $\text{Be}(\theta|\alpha, \beta)$, so that $\pi(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1}$. Using Bayes’ theorem, the posterior density of θ is $\pi(\theta|r, n, \alpha, \beta) \propto \theta^r(1 - \theta)^{n-r} \theta^{\alpha-1}(1 - \theta)^{\beta-1} \propto \theta^{r+\alpha-1}(1 - \theta)^{n-r+\beta-1}$, the Beta distribution $\text{Be}(\theta|r + \alpha, n - r + \beta)$.

Suppose, for example, that in the light of precedent surveys, available information on the proportion θ of citizens who would vote for a particular political measure in a referendum is described by a Beta distribution $\text{Be}(\theta|50, 50)$, so that it is judged to be equally likely that the referendum would be won or lost, and it is judged that the probability that either side wins less than 60% of the vote is 0.95.

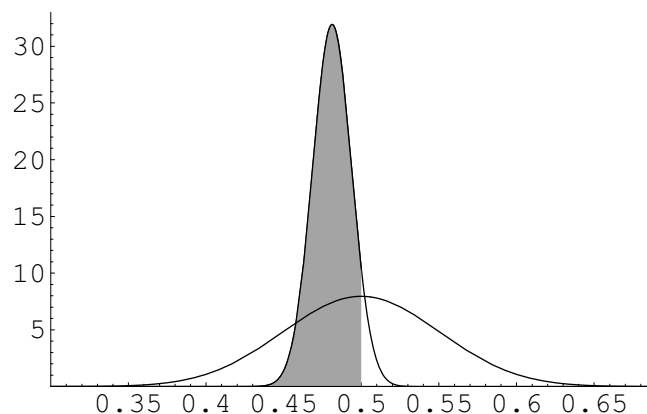


Figure 2. Prior and posterior densities of the proportion θ of citizens that would vote in favour of a referendum

A random survey of size 1500 is then conducted, where only 720 citizens declare to be in favour of the proposed measure. Using the results above, the corresponding posterior distribution is then $\text{Be}(\theta|770, 830)$. These prior and posterior densities are plotted in Figure 2; it may be appreciated that, as one would expect, the effect of the data is to drastically reduce the initial uncertainty on the value of θ and, hence, on the referendum outcome. More precisely, $\Pr(\theta < 0.5|720, 1500, H, K) = 0.933$ (shaded region in Figure 2) so that, after the information from the survey has been included, the probability that the referendum will be lost should be judged to be about 93%.

The general situation where the vector of interest is not the whole parameter vector ω , but some function $\theta = \theta(\omega)$ of possibly lower dimension than ω , will now be considered. Let D be some observed data, let $\{p(D|\omega), \omega \in \Omega\}$ be a probability model assumed to describe the probability mechanism which has generated D , let $\pi(\omega)$ be a probability distribution describing any available information on the value of ω , and let $\theta = \theta(\omega) \in \Theta$ be a function of the original parameters over whose value inferences based on the data D are required. Any valid conclusion on the value of the *vector of interest* θ will then be contained in its posterior probability distribution $\pi(\theta|D)$ which is conditional on the observed data D and will naturally also depend, although not explicitly shown in the notation, on the assumed model $\{p(D|\omega), \omega \in \Omega\}$, and on the available prior information encapsulated by $\pi(\omega)$. The required posterior distribution $p(\theta|D)$ is found by standard use of probability calculus. Indeed, by Bayes' theorem, $\pi(\omega|D) \propto p(D|\omega) \pi(\omega)$. Moreover, let $\lambda = \lambda(\omega) \in \Lambda$ be some other function of the original parameters such that $\psi = \{\theta, \lambda\}$ is a one-to-one transformation of ω , and let $J(\omega) = (\partial\psi/\partial\omega)$ be the corresponding Jacobian matrix. Naturally, the introduction of λ is not necessary if $\theta(\omega)$ is a one-to-one transformation of ω . Using standard change-of-variable probability

techniques, the posterior density of ψ is

$$(10) \quad \pi(\psi|D) = \pi(\theta, \lambda|D) = \left[\frac{\pi(\omega|D)}{|J(\omega)|} \right]_{\omega=\omega(\psi)}$$

and the required posterior of θ is the appropriate *marginal* density, obtained by integration over the *nuisance parameter* λ ,

$$(11) \quad \pi(\theta|D) = \int_{\Lambda} \pi(\theta, \lambda|D) d\lambda.$$

Notice that elimination of unwanted nuisance parameters, a simple integration within the Bayesian paradigm is, however, a difficult (often polemic) problem for frequentist statistics.

Sometimes, the range of possible values of ω is effectively restricted by contextual considerations. If ω is known to belong to $\Omega_c \subset \Omega$, the prior distribution is only positive in Ω_c and, using Bayes' theorem, it is immediately found that the restricted posterior is

$$(12) \quad \pi(\omega|D, \omega \in \Omega_c) = \frac{\pi(\omega|D)}{\int_{\Omega_c} \pi(\omega|D)}, \quad \omega \in \Omega_c,$$

and obviously vanishes if $\omega \notin \Omega_c$. Thus, to incorporate a restriction on the possible values of the parameters, it suffices to *renormalize* the unrestricted posterior distribution to the set $\Omega_c \subset \Omega$ of parameter values which satisfy the required condition. Incorporation of known constraints on the parameter values, a simple renormalization within the Bayesian paradigm, is another very difficult problem for conventional statistics. For further details on the elimination of nuisance parameters see [Liseo, 2005].

EXAMPLE 3 Inference on normal parameters. Let $D = \{x_1, \dots, x_n\}$ be a random sample from a normal distribution $N(x|\mu, \sigma)$. The corresponding likelihood function is immediately found to be proportional to $\sigma^{-n} \exp[-n\{s^2 + (\bar{x} - \mu)^2/(2\sigma^2)\}]$, with $n\bar{x} = \sum_i x_i$, and $ns^2 = \sum_i (x_i - \bar{x})^2$. It may be shown (see Section 4) that absence of initial information on the value of both μ and σ may formally be described by a joint prior function which is uniform in both μ and $\log(\sigma)$, that is, by the (improper) prior function $\pi(\mu, \sigma) = \sigma^{-1}$. Using Bayes' theorem, the corresponding joint posterior is

$$(13) \quad \pi(\mu, \sigma|D) \propto \sigma^{-(n+1)} \exp[-n\{s^2 + (\bar{x} - \mu)^2/(2\sigma^2)\}].$$

Thus, using the Gamma integral in terms of $\lambda = \sigma^{-2}$ to integrate out σ ,

$$(14) \quad \pi(\mu|D) \propto \int_0^\infty \sigma^{-(n+1)} \exp\left[-\frac{n}{2\sigma^2}[s^2 + (\bar{x} - \mu)^2]\right] d\sigma \propto [s^2 + (\bar{x} - \mu)^2]^{-n/2},$$

which is recognized as a kernel of the Student density $\text{St}(\mu|\bar{x}, s/\sqrt{n-1}, n-1)$. Similarly, integrating out μ ,

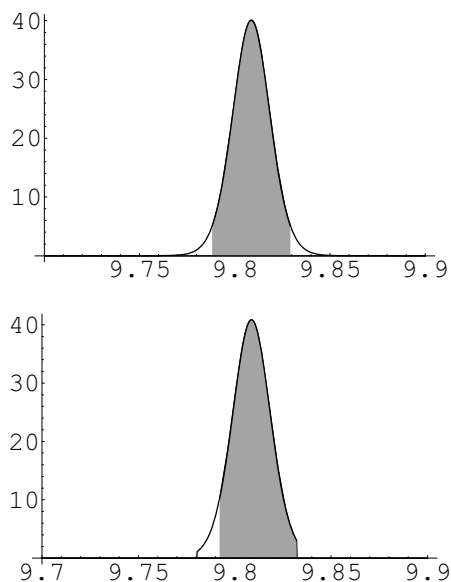


Figure 3. Posterior density $\pi(g|m, s, n)$ of the value g of the gravitational field, given $n = 20$ normal measurements with mean $m = 9.8087$ and standard deviation $s = 0.0428$, (a) with no additional information, and (b) with g restricted to $G_c = \{g; 9.7803 < g < 9.8322\}$. Shaded areas represent 95%-credible regions of g

$$(15) \quad \pi(\sigma|D) \propto \int_{-\infty}^{\infty} \sigma^{-(n+1)} \exp \left[-\frac{n}{2\sigma^2} [s^2 + (\bar{x} - \mu)^2] \right] d\mu \propto \sigma^{-n} \exp \left[-\frac{ns^2}{2\sigma^2} \right].$$

Changing variables to the precision $\lambda = \sigma^{-2}$ results in $\pi(\lambda|D) \propto \lambda^{(n-3)/2} e^{ns^2\lambda/2}$, a kernel of the Gamma density $\text{Ga}(\lambda|(n-1)/2, ns^2/2)$. In terms of the standard deviation σ this becomes $\pi(\sigma|D) = p(\lambda|D)|\partial\lambda/\partial\sigma| = 2\sigma^{-3}\text{Ga}(\sigma^{-2}|(n-1)/2, ns^2/2)$, a square-root inverted gamma density.

A frequent example of this scenario is provided by laboratory measurements made in conditions where central limit conditions apply, so that (assuming no experimental bias) those measurements may be treated as a random sample from a normal distribution centered at the quantity μ which is being measured, and with some (unknown) standard deviation σ . Suppose, for example, that in an elementary physics classroom experiment to measure the gravitational field g with a pendulum, a student has obtained $n = 20$ measurements of g yielding (in m/sec²) a mean $\bar{x} = 9.8087$, and a standard deviation $s = 0.0428$. Using no other information, the corresponding posterior distribution is $\pi(g|D) = \text{St}(g|9.8087, 0.0098, 19)$ represented in Figure 3(a). In particular, $\text{Pr}(9.788 < g < 9.829|D) = 0.95$, so that, with the information provided by this experiment, the gravitational field at the location of the laboratory provided may be expected to lie between 9.788 and 9.829 with

probability 0.95.

Formally, the posterior distribution of g should be restricted to $g > 0$; however, as immediately obvious from Figure 3a, this would not have any appreciable effect, due to the fact that the likelihood function is actually concentrated on positive g values.

Suppose now that the student is further instructed to incorporate into the analysis the fact that the value of the gravitational field g at the laboratory is known to lie between 9.7803 m/sec² (average value at the Equator) and 9.8322 m/sec² (average value at the poles). The updated posterior distribution will be

$$(16) \quad \pi(g|D, g \in G_c) = \frac{\text{St}(g|m, s/\sqrt{n-1}, n)}{\int_{g \in G_c} \text{St}(g|m, s/\sqrt{n-1}, n)}, \quad g \in G_c,$$

represented in Figure 3(b), where $G_c = \{g; 9.7803 < g < 9.8322\}$. One-dimensional numerical integration may be used to verify that $\Pr(g > 9.792|D, g \in G_c) = 0.95$. Moreover, if inferences about the standard deviation σ of the measurement procedure are also requested, the corresponding posterior distribution is found to be $\pi(\sigma|D) = 2\sigma^{-3}\text{Ga}(\sigma^{-2}|9.5, 0.0183)$. This has a mean $E[\sigma|D] = 0.0458$ and yields $\Pr(0.0334 < \sigma < 0.0642|D) = 0.95$.

3.2 Predictive Distributions

Let $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathcal{X}$, be a set of exchangeable observations, and consider now a situation where it is desired to predict the value of a future observation $\mathbf{x} \in \mathcal{X}$ generated by the same random mechanism that has generated the data D . It follows from the foundations arguments discussed in Section 2 that the solution to this prediction problem is simply encapsulated by the *predictive* distribution $p(\mathbf{x}|D)$ describing the uncertainty on the value that $\mathbf{x}$ will take, given the information provided by D and any other available knowledge. Suppose that contextual information suggests the assumption that data D may be considered to be a random sample from a distribution in the family $\{p(\mathbf{x}|\omega), \omega \in \Omega\}$, and let $\pi(\omega)$ be a prior distribution describing available information on the value of ω . Since $p(\mathbf{x}|\omega, D) = p(\mathbf{x}|\omega)$, it then follows from standard probability theory that

$$(17) \quad p(\mathbf{x}|D) = \int_{\Omega} p(\mathbf{x}|\omega) \pi(\omega|D) d\omega,$$

which is an average of the probability distributions of $\mathbf{x}$ conditional on the (unknown) value of ω , weighted with the posterior distribution of ω given D .

If the assumptions on the probability model are correct, the posterior predictive distribution $p(\mathbf{x}|D)$ will converge, as the sample size increases, to the distribution $p(\mathbf{x}|\omega)$ which has generated the data. Indeed, the best technique to assess the quality of the inferences about ω encapsulated in $\pi(\omega|D)$ is to check against the observed data the predictive distribution $p(\mathbf{x}|D)$ generated by $\pi(\omega|D)$. For a good introduction to Bayesian predictive inference, see Geisser [1993].

EXAMPLE 4 Prediction in a Poisson process. Let $D = \{r_1, \dots, r_n\}$ be a random sample from a Poisson distribution $\text{Pn}(r|\lambda)$ with parameter λ , so that $p(D|\lambda) \propto \lambda^t e^{-\lambda n}$, where $t = \sum r_i$. It may be shown (see Section 4) that absence of initial information on the value of λ may be formally described by the (improper) prior function $\pi(\lambda) = \lambda^{-1/2}$. Using Bayes' theorem, the corresponding posterior is

$$(18) \quad \pi(\lambda|D) \propto \lambda^t e^{-\lambda n} \lambda^{-1/2} \propto \lambda^{t-1/2} e^{-\lambda n},$$

the kernel of a Gamma density $\text{Ga}(\lambda|, t + 1/2, n)$, with mean $(t + 1/2)/n$. The corresponding predictive distribution is the Poisson-Gamma mixture

$$(19) \quad p(r|D) = \int_0^\infty \text{Pn}(r|\lambda) \text{Ga}(\lambda|, t + 1/2, n) d\lambda = \frac{n^{t+1/2}}{\Gamma(t + 1/2)} \frac{1}{r!} \frac{\Gamma(r + t + 1/2)}{(1 + n)^{r+t+1/2}}.$$

Suppose, for example, that in a firm producing automobile restraint systems, the entire production in each of 10 consecutive months has yielded no complaint from their clients. With no additional information on the average number λ of complaints per month, the quality assurance department of the firm may report that the probabilities that r complaints will be received in the next month of production are given by equation (19), with $t = 0$ and $n = 10$. In particular, $p(r = 0|D) = 0.953$, $p(r = 1|D) = 0.043$, and $p(r = 2|D) = 0.003$. Many other situations may be described with the same model. For instance, if meteorological conditions remain similar in a given area, $p(r = 0|D) = 0.953$ would describe the chances of no flash flood next year, given 10 years without flash floods in the area.

EXAMPLE 5 Prediction in a Normal process. Consider now prediction of a continuous variable. Let $D = \{x_1, \dots, x_n\}$ be a random sample from a normal distribution $N(x|\mu, \sigma)$. As mentioned in Example 3, absence of initial information on the values of both μ and σ is formally described by the *improper* prior function $\pi(\mu, \sigma) = \sigma^{-1}$, and this leads to the joint posterior density (13). The corresponding (posterior) predictive distribution is

$$(20) \quad p(x|D) = \int_0^\infty \int_{-\infty}^\infty \mathbf{N}(x|\mu, \sigma) \pi(\mu, \sigma|D) d\mu d\sigma = \text{St}(x|\bar{x}, s\sqrt{\frac{n+1}{n-1}}, n-1).$$

If μ is known to be positive, the appropriate prior function will be the restricted function

$$(21) \quad \pi(\mu, \sigma) = \begin{cases} \sigma^{-1} & \text{if } \mu > 0 \\ 0 & \text{otherwise.} \end{cases}$$

However, the result in equation (19) will still hold, provided the likelihood function $p(D|\mu, \sigma)$ is concentrated on positive μ values. Suppose, for example, that in the firm producing automobile restraint systems, the observed breaking strengths of $n = 10$ randomly chosen safety belt webbings have mean $\bar{x} = 28.011$ kN and standard deviation $s = 0.443$ kN, and that the relevant engineering specification requires breaking strengths to be larger than 26 kN. If data may truly be assumed to be a random sample from a normal distribution, the likelihood function is only

appreciable for positive μ values, and only the information provided by this small sample is to be used, then the quality engineer may claim that the probability that a safety belt randomly chosen from the same batch as the sample tested would satisfy the required specification is $\Pr(x > 26|D) = 0.9987$. Besides, if production conditions remain constant, 99.87% of the safety belt webbings may be expected to have acceptable breaking strengths.

3.3 Asymptotic Behaviour

The behaviour of posterior distributions when the sample size is large is now considered. This is important for, at least, two different reasons: (i) asymptotic results provide useful first-order approximations when actual samples are relatively large, and (ii) objective Bayesian methods typically depend on the asymptotic properties of the assumed model. Let $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, $\mathbf{x} \in \mathcal{X}$, be a random sample of size n from $\{p(\mathbf{x}|\omega), \omega \in \Omega\}$. It may be shown that, as $n \rightarrow \infty$, the posterior distribution of a *discrete* parameter ω typically converges to a degenerate distribution which gives probability one to the true value of ω , and that the posterior distribution of a *continuous* parameter ω typically converges to a normal distribution centered at its *maximum likelihood estimate* $\hat{\omega}$ (MLE), with a variance matrix which decreases with n as $1/n$.

Consider first the situation where $\Omega = \{\omega_1, \omega_2, \dots\}$ consists of a *countable* (possibly infinite) set of values, such that the probability model which corresponds to the true parameter value ω_t is *distinguishable* from the others in the sense that the logarithmic divergence $\kappa\{p(\mathbf{x}|\omega_i)|p(\mathbf{x}|\omega_t)\}$ of each of the $p(\mathbf{x}|\omega_i)$ from $p(\mathbf{x}|\omega_t)$ is strictly positive. Taking logarithms in Bayes' theorem, defining $z_j = \log[p(\mathbf{x}_j|\omega_i)/p(\mathbf{x}_j|\omega_t)]$, $j = 1, \dots, n$, and using the strong law of large numbers on the n conditionally independent and identically distributed random quantities $z_1, \dots, z_n$, it may be shown that

$$(22) \quad \lim_{n \rightarrow \infty} \Pr(\omega_t|\mathbf{x}_1, \dots, \mathbf{x}_n) = 1, \quad \lim_{n \rightarrow \infty} \Pr(\omega_i|\mathbf{x}_1, \dots, \mathbf{x}_n) = 0, \quad i \neq t.$$

Thus, under appropriate regularity conditions, the posterior probability of the true parameter value converges to one as the sample size grows.

Consider now the situation where ω is a k -dimensional *continuous* parameter. Expressing Bayes' theorem as $\pi(\omega|\mathbf{x}_1, \dots, \mathbf{x}_n) \propto \exp\{\log[\pi(\omega)] + \sum_{j=1}^n \log[p(\mathbf{x}_j|\omega)]\}$, expanding $\sum_j \log[p(\mathbf{x}_j|\omega)]$ about its maximum (the MLE $\hat{\omega}$), and assuming regularity conditions (to ensure that terms of order higher than quadratic may be ignored and that the sum of the terms from the likelihood will dominate the term from the prior) it is found that the posterior density of ω is the approximate k -variate normal

$$(23) \quad \pi(\omega|\mathbf{x}_1, \dots, \mathbf{x}_n) \approx \mathbf{N}_k\{\hat{\omega}, \mathbf{S}(D, \hat{\omega})\}, \quad \mathbf{S}^{-1}(D, \omega) = \left(- \sum_{l=1}^n \frac{\partial^2 \log[p(\mathbf{x}_l|\omega)]}{\partial \omega_i \partial \omega_j} \right).$$

A simpler, but somewhat poorer, approximation may be obtained by using the

strong law of large numbers on the sums in (22) to establish that $\mathbf{S}^{-1}(D, \hat{\omega}) \approx n \mathbf{F}(\hat{\omega})$, where $\mathbf{F}(\omega)$ is *Fisher's information matrix*, with general element

$$(24) \quad \mathbf{F}_{ij}(\omega) = - \int_X p(\mathbf{x}|\omega) \frac{\partial^2 \log[p(\mathbf{x}|\omega)]}{\partial \omega_i \partial \omega_j} d\mathbf{x},$$

so that

$$(25) \quad \pi(\omega|\mathbf{x}_1, \dots, \mathbf{x}_n) \approx N_k(\omega|\hat{\omega}, n^{-1} \mathbf{F}^{-1}(\hat{\omega})).$$

Thus, under appropriate regularity conditions, the posterior probability density of the parameter vector ω approaches, as the sample size grows, a multivariate normal density centered at the MLE $\hat{\omega}$, with a variance matrix which decreases with n as n^{-1} .

EXAMPLE 2, continued. Asymptotic approximation with binomial data. Let $D = (x_1, \dots, x_n)$ consist of n independent Bernoulli trials with parameter θ , so that $p(D|\theta, n) = \theta^r(1 - \theta)^{n-r}$. This likelihood function is maximized at $\hat{\theta} = r/n$, and Fisher's information function is $F(\theta) = \theta^{-1}(1 - \theta)^{-1}$. Thus, using the results above, the posterior distribution of θ will be the approximate normal,

$$(26) \quad \pi(\theta|r, n) \approx \mathbf{N}(\theta|\hat{\theta}, s(\hat{\theta})/\sqrt{n}), \quad s(\theta) = \{\theta(1 - \theta)\}^{1/2}$$

with mean $\hat{\theta} = r/n$ and variance $\hat{\theta}(1 - \hat{\theta})/n$. This will provide a reasonable approximation to the exact posterior if (i) the prior $\pi(\theta)$ is relatively "flat" in the region where the likelihood function matters, and (ii) both r and n are moderately large. If, say, $n = 1500$ and $r = 720$, this leads to $\pi(\theta|D) \approx \mathbf{N}(\theta|0.480, 0.013)$, and to $\Pr(\theta > 0.5|D) \approx 0.940$, which may be compared with the exact value $\Pr(\theta > 0.5|D) = 0.933$ obtained from the posterior distribution which corresponds to the prior $\text{Be}(\theta|50, 50)$. $\triangleleft$

It follows from the *joint* posterior asymptotic behaviour of ω and from the properties of the multivariate normal distribution that, if the parameter vector is decomposed into $\omega = (\theta, \lambda)$, and Fisher's information matrix is correspondingly partitioned, so that

$$(27) \quad \mathbf{F}(\omega) = \mathbf{F}(\theta, \lambda) = \begin{pmatrix} \mathbf{F}_{\theta\theta}(\theta, \lambda) & \mathbf{F}_{\theta\lambda}(\theta, \lambda) \\ \mathbf{F}_{\lambda\theta}(\theta, \lambda) & \mathbf{F}_{\lambda\lambda}(\theta, \lambda) \end{pmatrix}$$

and

$$(28) \quad \mathbf{S}(\theta, \lambda) = \mathbf{F}^{-1}(\theta, \lambda) = \begin{pmatrix} \mathbf{S}_{\theta\theta}(\theta, \lambda) & \mathbf{S}_{\theta\lambda}(\theta, \lambda) \\ \mathbf{S}_{\lambda\theta}(\theta, \lambda) & \mathbf{S}_{\lambda\lambda}(\theta, \lambda) \end{pmatrix},$$

then the *marginal* posterior distribution of θ will be

$$(29) \quad \pi(\theta|D) \approx \mathbf{N}\{\theta|\hat{\theta}, n^{-1} \mathbf{S}_{\theta\theta}(\hat{\theta}, \hat{\lambda})\},$$

while the *conditional* posterior distribution of λ given θ will be

$$(30) \quad \pi(\lambda|\theta, D) \approx \mathbf{N}\{\lambda|\hat{\lambda} - \mathbf{F}_{\lambda\lambda}^{-1}(\theta, \hat{\lambda})\mathbf{F}_{\lambda\theta}(\theta, \hat{\lambda})(\hat{\theta} - \theta), n^{-1} \mathbf{F}_{\lambda\lambda}^{-1}(\theta, \hat{\lambda})\}.$$

Notice that $\mathbf{F}_{\lambda\lambda}^{-1} = \mathbf{S}_{\lambda\lambda}$ if (and only if) $\mathbf{F}$ is block diagonal, i.e. if (and only if) θ and λ are asymptotically independent.

EXAMPLE 3, continued. Asymptotic approximation with normal data. Let $D = (x_1, \dots, x_n)$ be a random sample from a normal distribution $\mathbf{N}(x|\mu, \sigma)$. The corresponding likelihood function $p(D|\mu, \sigma)$ is maximized at $(\hat{\mu}, \hat{\sigma}) = (\bar{x}, s)$, and Fisher's information matrix is diagonal, with $F_{\mu\mu} = \sigma^{-2}$. Hence, the posterior distribution of μ is *approximately* $\mathbf{N}(\mu|\bar{x}, s/\sqrt{n})$; this may be compared with the *exact* result $\pi(\mu|D) = \text{St}(\mu|\bar{x}, s/\sqrt{n-1}, n-1)$ obtained previously under the assumption of no prior knowledge. $\triangleleft$

4 REFERENCE ANALYSIS

Under the Bayesian paradigm, the outcome of any inference problem (the posterior distribution of the quantity of interest) combines the information provided by the data with relevant available prior information. In many situations, however, either the available prior information on the quantity of interest is too vague to warrant the effort required to have it formalized in the form of a probability distribution, or it is too subjective to be useful in scientific communication or public decision making. It is therefore important to be able to identify the mathematical form of a “noninformative” prior, a prior that would have a minimal effect, relative to the data, on the posterior inference. More formally, suppose that the probability mechanism which has generated the available data D is assumed to be $p(D|\omega)$, for some $\omega \in \Omega$, and that the quantity of interest is some real-valued function $\theta = \theta(\omega)$ of the model parameter ω . Without loss of generality, it may be assumed that the probability model is of the form

$$(31) \quad \mathcal{M} = \{p(D|\theta, \lambda), D \in \mathcal{D}, \theta \in \Theta, \lambda \in \Lambda\}$$

$p(D|\theta, \lambda)$, where λ is some appropriately chosen nuisance parameter vector. As described in Section 3, to obtain the required posterior distribution of the quantity of interest $\pi(\theta|D)$ it is necessary to specify a *joint* prior $\pi(\theta, \lambda)$. It is now required to identify the form of that joint prior $\pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P})$, the *θ -reference prior*, which would have a *minimal effect* on the corresponding posterior distribution of θ ,

$$(32) \quad \pi(\theta|D) \propto \int_{\Lambda} p(D|\theta, \lambda) \pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P}) d\lambda,$$

within the class $\mathcal{P}$ of all the prior distributions compatible with whatever information about (θ, λ) one is prepared to assume, which may just be the class $\mathcal{P}_0$ of *all* strictly positive priors. To simplify the notation, when there is no danger of confusion the reference prior $\pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P})$ is often simply denoted by $\pi(\theta, \lambda)$, but its dependence on the quantity of interest θ , the assumed model $\mathcal{M}$ and the class $\mathcal{P}$ of priors compatible with assumed knowledge, should always be kept in mind.

To use a conventional expression, the reference prior “would let the data speak for themselves” about the likely value of θ . Properly defined, reference *posterior*

distributions have an important role to play in scientific communication, for they provide the answer to a central question in the sciences: conditional on the assumed model $p(D|\theta, \lambda)$, and on any further assumptions of the value of θ on which there might be universal agreement, the reference posterior $\pi(\theta|D)$ should specify what *could* be said about θ *if* the only available information about θ were some well-documented data D and whatever information (if any) one is prepared to assume by restricting the prior to belong to an appropriate class $\mathcal{P}$.

Much work has been done to formulate “reference” priors which would make the idea described above mathematically precise. For historical details, see [Bernardo and Smith, 1994, Sec. 5.6.2; Kass and Wasserman, 1996; Bernardo, 2005a] and references therein. This section concentrates on an approach that is based on information theory to derive reference distributions which may be argued to provide the most advanced general procedure available; this was initiated by Bernardo [1979b; 1981] and further developed by Berger and Bernardo [1989; 1992a; 1982b; 1982c; 1997; 2005a; Bernardo and Ramón, 1998; Berger *et al.*, 2009], and references therein. In the formulation described below, far from ignoring prior knowledge, the reference posterior exploits certain well-defined features of a *possible* prior, namely those describing a situation where relevant knowledge about the quantity of interest (beyond that universally accepted, as specified by the choice of $\mathcal{P}$) may be held to be negligible compared to the information about that quantity which repeated experimentation (from a specific data generating mechanism $\mathcal{M}$) might possibly provide. Reference analysis is appropriate in contexts where the set of inferences which could be drawn in this *possible* situation is considered to be pertinent.

Any statistical analysis contains a fair number of subjective elements; these include (among others) the data selected, the model assumptions, and the choice of the quantities of interest. Reference analysis may be argued to provide an “objective” Bayesian solution to statistical inference problems in just the same sense that conventional statistical methods claim to be “objective”: in that the solutions only depend on model assumptions and observed data.

4.1 Reference Distributions

One parameter. Consider the experiment which consists of the observation of data D , generated by a random mechanism $p(D|\theta)$ which only depends on a real-valued parameter $\theta \in \Theta$, and let $\mathbf{t} = \mathbf{t}(D) \in T$ be *any* sufficient statistic (which may well be the complete data set D). In Shannon’s general information theory, the *amount of information* $I^\theta\{T, \pi(\theta)\}$ which may be expected to be provided by D , or (equivalently) by $\mathbf{t}(D)$, about the value of θ is defined by

$$(33) \quad I^\theta\{T, \pi(\theta)\} = \kappa \{p(\mathbf{t})\pi(\theta)|p(\mathbf{t}|\theta)\pi(\theta)\} = \mathbb{E}_{\mathbf{t}} \left[\int_{\Theta} \pi(\theta|\mathbf{t}) \log \frac{\pi(\theta|\mathbf{t})}{\pi(\theta)} d\theta \right],$$

the expected logarithmic divergence of the prior from the posterior. This is naturally a *functional* of the prior $\pi(\theta)$: the larger the prior information, the smaller the information which the data may be expected to provide. The functional

$I^\theta\{T, \pi(\theta)\}$ is concave, non-negative, and invariant under one-to-one transformations of θ . Consider now the amount of information $I^\theta\{T^k, \pi(\theta)\}$ about θ which may be expected from the experiment which consists of k conditionally independent replications $\{\mathbf{t}_1, \dots, \mathbf{t}_k\}$ of the original experiment. As $k \rightarrow \infty$, such an experiment would provide any *missing information* about θ which could possibly be obtained within this framework; thus, as $k \rightarrow \infty$, the functional $I^\theta\{T^k, \pi(\theta)\}$ will approach the missing information about θ associated with the prior $p(\theta)$. Intuitively, a θ -“noninformative” prior is one which *maximizes the missing information* about θ . Formally, if $\pi_k(\theta)$ denotes the prior density which maximizes $I^\theta\{T^k, \pi(\theta)\}$ in the class $\mathcal{P}$ of prior distributions which are compatible with accepted assumptions on the value of θ (which may well be the class $\mathcal{P}_0$ of *all* strictly positive proper priors) then the θ -reference prior $\pi(\theta|\mathcal{M}, \mathcal{P})$ is the limit as $k \rightarrow \infty$ (in a sense to be made precise) of the sequence of priors $\{\pi_k(\theta), k = 1, 2, \dots\}$.

Notice that this limiting procedure is *not* some kind of asymptotic approximation, but an essential element of the *definition* of a reference prior. In particular, this definition implies that reference distributions only depend on the *asymptotic* behaviour of the assumed probability model, a feature which actually simplifies their actual derivation.

EXAMPLE 6 Maximum entropy. If θ may only take a *finite* number of values, so that the parameter space is $\Theta = \{\theta_1, \dots, \theta_m\}$ and $\pi(\theta) = \{p_1, \dots, p_m\}$, with $p_i = \Pr(\theta = \theta_i)$, and there is no topology associated to the parameter space Θ , so that the θ_i ’s are just labels with no quantitative meaning, then the missing information associated to $\{p_1, \dots, p_m\}$ reduces to

$$(34) \quad \lim_{k \rightarrow \infty} I^\theta\{T^k, \pi(\theta)\} = H(p_1, \dots, p_m) = - \sum_{i=1}^m p_i \log(p_i),$$

that is, the *entropy* of the prior distribution $\{p_1, \dots, p_m\}$.

Thus, in the non-quantitative finite case, the reference prior $\pi(\theta|\mathcal{M}, \mathcal{P})$ is that with *maximum entropy* in the class $\mathcal{P}$ of priors compatible with accepted assumptions. Consequently, the reference prior algorithm contains “maximum entropy” priors as the particular case which obtains when the parameter space is a *finite set of labels*, the *only* case where the original concept of entropy as a measure of uncertainty is unambiguous and well-behaved. In particular, if $\mathcal{P}$ is the class $\mathcal{P}_0$ of *all* priors over $\{\theta_1, \dots, \theta_m\}$, then the reference prior is the uniform prior over the set of possible θ values, $\pi(\theta|\mathcal{M}, \mathcal{P}_0) = \{1/m, \dots, 1/m\}$.

Formally, the *reference prior function* $\pi(\theta|\mathcal{M}, \mathcal{P})$ of a univariate parameter θ is defined to be the limit of the sequence of the proper priors $\pi_k(\theta)$ which maximize $I^\theta\{T^k, \pi(\theta)\}$ in the precise sense that, for any value of the sufficient statistic $\mathbf{t} = \mathbf{t}(D)$, the *reference posterior*, the intrinsic¹ limit $\pi(\theta|\mathbf{t})$ of the corresponding sequence of posteriors $\{\pi_k(\theta|\mathbf{t})\}$, may be obtained from $\pi(\theta|\mathcal{M}, \mathcal{P})$ by formal use of Bayes theorem, so that $\pi(\theta|\mathbf{t}) \propto p(\mathbf{t}|\theta) \pi(\theta|\mathcal{M}, \mathcal{P})$.

¹A sequence $\{\pi_k(\theta|\mathbf{t})\}$ of posterior distributions converges intrinsically to a limit $\pi(\theta|\mathbf{t})$ if the sequence of expected intrinsic discrepancies $E_{\mathbf{t}}[\delta\{\pi_k(\theta|\mathbf{t}), \pi(\theta|\mathbf{t})\}]$ converges to 0, where $\delta\{p, q\} = \min\{k(p|q), k(q|p)\}$, and $k(p|q) = \int_{\Theta} q(\theta) \log[q(\theta)/p(\theta)] d\theta$. For details, see [Berger *et al.*, 2009].

Reference prior *functions* are often simply called reference priors, even though they are usually *not* probability distributions. They should *not* be considered as expressions of belief, but technical devices to obtain (proper) posterior distributions which are a limiting form of the posteriors which could have been obtained from possible prior beliefs which were relatively uninformative with respect to the quantity of interest when compared with the information which data could provide.

If (i) the sufficient statistic $\mathbf{t} = \mathbf{t}(D)$ is a consistent estimator $\tilde{\theta}$ of a continuous parameter θ , and (ii) the class $\mathcal{P}$ contains *all* strictly positive priors, then the reference prior may be shown to have a simple form in terms of any *asymptotic* approximation to the posterior distribution of θ . Notice that, by construction, an *asymptotic* approximation to the posterior does *not* depend on the prior. Specifically, if the posterior density $\pi(\theta|D)$ has an asymptotic approximation of the form $\pi(\theta|\tilde{\theta}, n)$, the (unrestricted) reference prior is simply

$$(35) \quad \pi(\theta|\mathcal{M}, \mathcal{P}_0) \propto \pi(\theta|\tilde{\theta}, n) \Big|_{\tilde{\theta}=\theta}.$$

One-parameter reference priors are *invariant* under reparametrization; thus, if $\psi = \psi(\theta)$ is a piecewise one-to-one function of θ , then the ψ -reference prior is simply the appropriate probability transformation of the θ -reference prior.

EXAMPLE 7 Jeffreys' prior. If θ is univariate and continuous, and the posterior distribution of θ given $\{x_1, \dots, x_n\}$ is asymptotically normal with standard deviation $s(\hat{\theta})/\sqrt{n}$, then, using (34), the reference prior function is $\pi(\theta) \propto s(\theta)^{-1}$. Under regularity conditions (often satisfied in practice, see Section 3.3), the posterior distribution of θ is asymptotically normal with variance $n^{-1} F^{-1}(\hat{\theta})$, where $F(\theta)$ is Fisher's information function and $\hat{\theta}$ is the MLE of θ . Hence, the reference prior function in these conditions is $\pi(\theta|\mathcal{M}, \mathcal{P}_0) \propto F(\theta)^{1/2}$, which is known as Jeffreys' prior. It follows that the reference prior algorithm contains Jeffreys' priors as the particular case which obtains when the probability model only depends on a single continuous univariate parameter, there are regularity conditions to guarantee asymptotic normality, and there is no additional information, so that the class of possible priors is the set $\mathcal{P}_0$ of all strictly positive priors over Θ . These are precisely the conditions under which there is general agreement on the use of Jeffreys' prior as a "noninformative" prior.

EXAMPLE 2, continued. Reference prior for a binomial parameter. Let data $D = \{x_1, \dots, x_n\}$ consist of a sequence of n independent Bernoulli trials, so that $p(x|\theta) = \theta^x(1-\theta)^{1-x}$, $x \in \{0, 1\}$; this is a regular, one-parameter continuous model, whose Fisher's information function is $F(\theta) = \theta^{-1}(1-\theta)^{-1}$. Thus, the reference prior $\pi(\theta)$ is proportional to $\theta^{-1/2}(1-\theta)^{-1/2}$, so that the reference prior is the (proper) Beta distribution $\text{Be}(\theta|1/2, 1/2)$. Since the reference algorithm is invariant under reparametrization, the reference prior of $\phi(\theta) = 2 \arcsin \sqrt{\theta}$ is $\pi(\phi) = \pi(\theta)/|\partial\phi/\partial\theta| = 1$; thus, the reference prior is *uniform on the variance-stabilizing transformation* $\phi(\theta) = 2 \arcsin \sqrt{\theta}$, a feature generally true under reg-

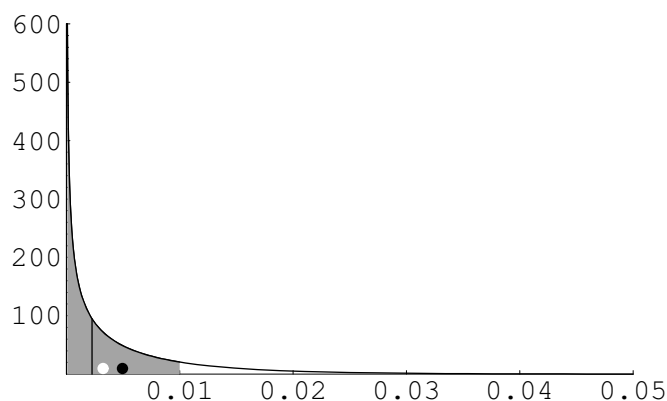


Figure 4. Posterior distribution of the proportion of infected people in the population, given the results of $n = 100$ tests, none of which were positive

ularity conditions. In terms of θ , the reference posterior is $\pi(\theta|D) = \pi(\theta|r, n) = \text{Be}(\theta|r + 1/2, n - r + 1/2)$, where $r = \sum x_j$ is the number of positive trials.

Suppose, for example, that $n = 100$ randomly selected people have been tested for an infection and that all tested negative, so that $r = 0$. The reference posterior distribution of the proportion θ of people infected is then the Beta distribution $\text{Be}(\theta|0.5, 100.5)$, represented in Figure 4. It may well be known that the infection was rare, leading to the assumption that $\theta < \theta_0$, for some upper bound θ_0 ; the (restricted) reference prior would then be of the form $\pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$ if $\theta < \theta_0$, and zero otherwise. However, provided the likelihood is concentrated in the region $\theta < \theta_0$, the corresponding posterior would virtually be identical to $\text{Be}(\theta|0.5, 100.5)$. Thus, just on the basis of the observed experimental results, one may claim that the proportion of infected people is surely smaller than 5% (for the reference posterior probability of the event $\theta > 0.05$ is 0.001), that θ is smaller than 0.01 with probability 0.844 (area of the shaded region in Figure 4), that it is equally likely to be over or below 0.23% (for the median, represented by a vertical line, is 0.0023), and that the probability that a person randomly chosen from the population is infected is 0.005 (the posterior mean, represented in the figure by a black circle), since $\Pr(x = 1|r, n) = E[\theta|r, n] = 0.005$. If a particular point estimate of θ is required (say a number to be quoted in the summary headline) the *intrinsic* estimator suggests itself (see Section 5); this is found to be $\theta^* = 0.0032$ (represented in the figure with a white circle). Notice that the traditional solution to this problem, based on the asymptotic behaviour of the MLE, here $\hat{\theta} = r/n = 0$ for any n , makes absolutely no sense in this scenario. $\triangleleft$

One nuisance parameter. The extension of the reference prior algorithm to the case of two parameters follows the usual mathematical procedure of reducing the problem to a sequential application of the established procedure for the single

parameter case. Thus, if the probability model is $p(\mathbf{t}|\theta, \lambda)$, $\theta \in \Theta$, $\lambda \in \Lambda$ and a θ -reference prior $\pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P})$ is required, the reference algorithm proceeds in two steps:

- (i) Conditional on θ , $p(\mathbf{t}|\theta, \lambda)$ only depends on the nuisance parameter λ and, hence, the one-parameter algorithm may be used to obtain the *conditional* reference prior $\pi(\lambda|\theta, \mathcal{M}, \mathcal{P})$.
- (ii) If $\pi(\lambda|\theta, \mathcal{M}, \mathcal{P})$ is proper, this may be used to integrate out the nuisance parameter thus obtaining the one-parameter integrated model $p(\mathbf{t}|\theta) = \int_\Lambda p(\mathbf{t}|\theta, \lambda) \pi(\lambda|\theta, \mathcal{M}, \mathcal{P}) d\lambda$, to which the one-parameter algorithm may be applied again to obtain $\pi(\theta|\mathcal{M}, \mathcal{P})$. The θ -reference prior is then $\pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P}) = \pi(\lambda|\theta, \mathcal{M}, \mathcal{P}) \pi(\theta|\mathcal{M}, \mathcal{P})$, and the required reference posterior is $\pi(\theta|\mathbf{t}) \propto p(\mathbf{t}|\theta) \pi(\theta|\mathcal{M}, \mathcal{P})$.

If the conditional reference prior is *not* proper, then the procedure is performed within an increasing sequence $\{\Lambda_i\}$ of subsets converging to Λ over which $\pi(\lambda|\theta)$ is integrable. This makes it possible to obtain a corresponding sequence of θ -reference posteriors $\{\pi_i(\theta|\mathbf{t})\}$ for the quantity of interest θ , and the required reference posterior is the corresponding intrinsic limit $\pi(\theta|\mathbf{t}) = \lim_i \pi_i(\theta|\mathbf{t})$.

A θ -reference prior is then defined as a positive function $\pi_\theta(\theta, \lambda)$ which may be formally used in Bayes' theorem as a prior to obtain the reference posterior, i.e. such that, for any sufficient $\mathbf{t} \in T$ (which may well be the whole data set D) $\pi(\theta|\mathbf{t}) \propto \int_\Lambda p(\mathbf{t}|\theta, \lambda) \pi_\theta(\theta, \lambda) d\lambda$. The approximating sequences should be *consistently* chosen within a given model. Thus, given a probability model $\{p(\mathbf{x}|\omega), \omega \in \Omega\}$ an appropriate approximating sequence $\{\Omega_i\}$ should be chosen for the whole parameter space Ω . Thus, if the analysis is done in terms of, say, $\psi = \{\psi_1, \psi_2\} \in \Psi(\Omega)$, the approximating sequence should be chosen such that $\Psi_i = \psi(\Omega_i)$. A natural approximating sequence in location-scale problems is $\{\mu, \log \sigma\} \in [-i, i]^2$.

The θ -reference prior does *not* depend on the choice of the nuisance parameter λ ; thus, for any $\psi = \psi(\theta, \lambda)$ such that (θ, ψ) is a one-to-one function of (θ, λ) , the θ -reference prior in terms of (θ, ψ) is simply $\pi_\theta(\theta, \psi) = \pi_\theta(\theta, \lambda)/|\partial(\theta, \psi)/\partial(\theta, \lambda)|$, the appropriate probability transformation of the θ -reference prior in terms of (θ, λ) . Notice, however, that the reference prior *may* depend on the parameter of interest; thus, the θ -reference prior may differ from the ϕ -reference prior unless either ϕ is a piecewise one-to-one transformation of θ , or ϕ is asymptotically independent of θ . This is an expected consequence of the fact that the conditions under which the missing information about θ is maximized are not generally the same as the conditions which maximize the missing information about an arbitrary function $\phi = \phi(\theta, \lambda)$.

The *non-existence* of a unique “noninformative prior” which would be appropriate for any inference problem within a given model was established by Dawid, Stone and Zidek [1973], when they showed that this is incompatible with *consistent marginalization*. Indeed, if given the model $p(D|\theta, \lambda)$, the reference posterior

of the quantity of interest θ , $\pi(\theta|D) = \pi(\theta|\mathbf{t})$, only depends on the data through a statistic $\mathbf{t}$ whose sampling distribution, $p(\mathbf{t}|\theta, \lambda) = p(\mathbf{t}|\theta)$, only depends on θ , one would expect the reference posterior to be of the form $\pi(\theta|\mathbf{t}) \propto \pi(\theta) p(\mathbf{t}|\theta)$ for some prior $\pi(\theta)$. However, examples were found where this cannot be the case if a *unique* joint “noninformative” prior were to be used for all possible quantities of interest.

EXAMPLE 8 Regular two dimensional continuous reference prior functions. If the joint posterior distribution of (θ, λ) is asymptotically normal, then the θ -reference prior may be derived in terms of the corresponding Fisher’s information matrix, $\mathbf{F}(\theta, \lambda)$. Indeed, if

$$(36) \quad \mathbf{F}(\theta, \lambda) = \begin{pmatrix} F_{\theta\theta}(\theta, \lambda) & F_{\theta\lambda}(\theta, \lambda) \\ F_{\theta\lambda}(\theta, \lambda) & F_{\lambda\lambda}(\theta, \lambda) \end{pmatrix}, \text{ and } \mathbf{S}(\theta, \lambda) = \mathbf{F}^{-1}(\theta, \lambda),$$

then the unrestricted θ -reference prior is $\pi_\theta(\theta, \lambda|\mathcal{M}, \mathcal{P}_0) = \pi(\lambda|\theta) \pi(\theta)$, where

$$(37) \quad \pi(\lambda|\theta) \propto F_{\lambda\lambda}^{1/2}(\theta, \lambda), \quad \lambda \in \Lambda.$$

If $\pi(\lambda|\theta)$ is proper,

$$(38) \quad \pi(\theta) \propto \exp \left\{ \int_{\Lambda} \pi(\lambda|\theta) \log[S_{\theta\theta}^{-1/2}(\theta, \lambda)] d\lambda \right\}, \quad \theta \in \Theta.$$

If $\pi(\lambda|\theta)$ is not proper, integrations are performed on an approximating sequence $\{\Lambda_i\}$ to obtain a sequence $\{\pi_i(\lambda|\theta) \pi_i(\theta)\}$, (where $\pi_i(\lambda|\theta)$ is the proper renormalization of $\pi(\lambda|\theta)$ to Λ_i) and the θ -reference prior $\pi_\theta(\theta, \lambda)$ is defined as its appropriate limit. Moreover, if (i) both $F_{\lambda\lambda}^{1/2}(\theta, \lambda)$ and $S_{\theta\theta}^{-1/2}(\theta, \lambda)$ *factorize*, so that

$$(39) \quad S_{\theta\theta}^{-1/2}(\theta, \lambda) \propto f_\theta(\theta) g_\theta(\lambda), \quad F_{\lambda\lambda}^{1/2}(\theta, \lambda) \propto f_\lambda(\theta) g_\lambda(\lambda),$$

and (ii) the parameters θ and λ are *variation independent*, so that Λ does not depend on θ , then the θ -reference prior is simply $\pi_\theta(\theta, \lambda) = f_\theta(\theta) g_\lambda(\lambda)$, even if the conditional reference prior $\pi(\lambda|\theta) = \pi(\lambda) \propto g_\lambda(\lambda)$ (which will not depend on θ) is actually improper.

EXAMPLE 3, continued. Reference priors for the normal model. The information matrix which corresponds to a normal model $\mathbf{N}(x|\mu, \sigma)$ is

$$(40) \quad \mathbf{F}(\mu, \sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix}, \quad \mathbf{S}(\mu, \sigma) = \mathbf{F}^{-1}(\mu, \sigma) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{pmatrix};$$

hence $F_{\sigma\sigma}^{1/2}(\mu, \sigma) = \sqrt{2} \sigma^{-1} = f_\sigma(\mu) g_\sigma(\sigma)$, with $g_\sigma(\sigma) = \sigma^{-1}$, and thus $\pi(\sigma|\mu) = \sigma^{-1}$. Similarly, $S_{\mu\mu}^{-1/2}(\mu, \sigma) = \sigma^{-1} = f_\mu(\mu) g_\mu(\sigma)$, with $f_\mu(\mu) = 1$, and thus $\pi(\mu) = 1$. Therefore, the μ -reference prior is $\pi_\mu(\mu, \sigma|\mathcal{M}, \mathcal{P}_0) = \pi(\sigma|\mu) \pi(\mu) = \sigma^{-1}$, as already anticipated. Moreover, as one would expect from the fact that $\mathbf{F}(\mu, \sigma)$ is diagonal and also anticipated, it is similarly found that the σ -reference prior is $\pi_\sigma(\mu, \sigma|\mathcal{M}, \mathcal{P}_0) = \sigma^{-1}$, the same as before.

Suppose, however, that the quantity of interest is *not* the mean μ or the standard deviation σ , but the *standardized* mean $\phi = \mu/\sigma$. Fisher's information matrix in terms of the parameters ϕ and σ is $\mathbf{F}(\phi, \sigma) = J^t \mathbf{F}(\mu, \sigma) J$, where $J = (\partial(\mu, \sigma)/\partial(\phi, \sigma))$ is the Jacobian of the inverse transformation; this yields

$$(41) \quad \mathbf{F}(\phi, \sigma) = \begin{pmatrix} 1 & \phi\sigma^{-1} \\ \phi\sigma^{-1} & \sigma^{-2}(2 + \phi^2) \end{pmatrix}, \quad \mathbf{S}(\phi, \sigma) = \begin{pmatrix} 1 + \frac{1}{2}\phi^2 & -\frac{1}{2}\phi\sigma \\ -\frac{1}{2}\phi\sigma & \frac{1}{2}\sigma^2 \end{pmatrix}.$$

Thus, $S_{\phi\phi}^{-1/2}(\phi, \sigma) \propto (1 + \frac{1}{2}\phi^2)^{-1/2}$ and $F_{\sigma\sigma}^{1/2}(\phi, \sigma) \propto \sigma^{-1}(2 + \phi^2)^{1/2}$. Hence, using again the results in Example 8, $\pi_{\phi}(\phi, \sigma | \mathcal{M}, \mathcal{P}_0) = (1 + \frac{1}{2}\phi^2)^{-1/2}\sigma^{-1}$. In the original parametrization, this is $\pi_{\phi}(\mu, \sigma | \mathcal{M}, \mathcal{P}_0) = (1 + \frac{1}{2}(\mu/\sigma)^2)^{-1/2}\sigma^{-2}$, which is *very* different from $\pi_{\mu}(\mu, \sigma | \mathcal{M}, \mathcal{P}_0) = \pi_{\sigma}(\mu, \sigma | \mathcal{M}, \mathcal{P}_0) = \sigma^{-1}$. The corresponding reference posterior of ϕ is $\pi(\phi | x_1, \dots, x_n) \propto (1 + \frac{1}{2}\phi^2)^{-1/2} p(t | \phi)$ where $t = (\sum x_j)/(\sum x_j^2)^{1/2}$, a one-dimensional (marginally sufficient) statistic whose sampling distribution, $p(t | \mu, \sigma) = p(t | \phi)$, only depends on ϕ . Thus, the reference prior algorithm is seen to be consistent under marginalization. $\triangleleft$

Many parameters. The reference algorithm is easily generalized to an arbitrary number of parameters. If the model is $p(\mathbf{t} | \omega_1, \dots, \omega_m)$, a joint reference prior

$$(42) \quad \pi(\theta_m | \theta_{m-1}, \dots, \theta_1) \times \dots \times \pi(\theta_2 | \theta_1) \times \pi(\theta_1)$$

may sequentially be obtained for each *ordered* parametrization $\{\theta_1(\omega), \dots, \theta_m(\omega)\}$ of interest, and these are invariant under reparametrization of any of the $\theta_i(\omega)$'s. The choice of the ordered parametrization $\{\theta_1, \dots, \theta_m\}$ precisely describes the particular prior required, namely that which *sequentially* maximizes the missing information about each of the θ_i 's, conditional on $\{\theta_1, \dots, \theta_{i-1}\}$, for $i = m, m-1, \dots, 1$.

EXAMPLE 9 Stein's paradox. Let D be a random sample from a m -variate normal distribution with mean $\mu = \{\mu_1, \dots, \mu_m\}$ and unitary variance matrix. The reference prior which corresponds to any permutation of the μ_i 's is uniform, and this prior leads indeed to appropriate reference posterior distributions for any of the μ_i 's, namely $\pi(\mu_i | D) = N(\mu_i | \bar{x}_i, 1/\sqrt{n})$. Suppose, however, that the quantity of interest is $\theta = \sum_i \mu_i^2$, the distance of μ to the origin. As showed by Stein [1959], the posterior distribution of θ based on that uniform prior (or in any "flat" *proper* approximation) has very undesirable properties; this is due to the fact that a uniform (or nearly uniform) prior, although "noninformative" with respect to each of the individual μ_i 's, is actually highly informative on the sum of their squares, introducing a severe positive bias (Stein's paradox). However, the reference prior which corresponds to a parametrization of the form $\{\theta, \lambda_1, \dots, \lambda_{m-1}\}$ produces, for any choice of the nuisance parameters $\lambda_i = \lambda_i(\mu)$, the reference posterior $\pi(\theta | D) = \pi(\theta | t) \propto \theta^{-1/2} \chi^2(nt | m, n\theta)$, where $t = \sum_i \bar{x}_i^2$, and this posterior is shown to have the appropriate consistency properties.

Far from being specific to Stein's example, the inappropriate behaviour in problems with many parameters of specific marginal posterior distributions derived

from multivariate “flat” priors (proper or improper) is indeed very frequent. Hence, sloppy, uncontrolled use of “flat” priors (rather than the relevant reference priors), is very strongly discouraged.

Limited information Although often used in contexts where no universally agreed prior knowledge about the quantity of interest is available, the reference algorithm may be used to specify a prior which incorporates any acceptable prior knowledge; it suffices to maximize the missing information within the class $\mathcal{P}$ of priors which is compatible with such accepted knowledge. Indeed, by progressive incorporation of further restrictions into $\mathcal{P}$, the reference prior algorithm becomes a method of (prior) *probability assessment*. As described below, the problem has a fairly simple analytical solution when those restrictions take the form of known expected values. The incorporation of other type of restrictions usually involves numerical computations.

EXAMPLE 10 Univariate restricted reference priors. If the probability mechanism which is assumed to have generated the available data only depends on a univariate continuous parameter $\theta \in \Theta \subset \mathbb{R}$, and the class $\mathcal{P}$ of acceptable priors is a class of proper priors which satisfies some expected value restrictions, so that

$$(43) \quad \mathcal{P} = \left\{ \pi(\theta); \quad \pi(\theta) > 0, \int_{\Theta} \pi(\theta) d\theta = 1, \quad \int_{\Theta} g_i(\theta) \pi(\theta) d\theta = \beta_i, i = 1, \dots, m \right\}$$

then the (restricted) reference prior is

$$(44) \quad \pi(\theta|\mathcal{M}, \mathcal{P}) \propto \pi(\theta|\mathcal{M}, \mathcal{P}_0) \exp \left[\sum_{j=1}^m \gamma_j g_j(\theta) \right]$$

where $\pi(\theta|\mathcal{M}, \mathcal{P}_0)$ is the unrestricted reference prior and the γ_i 's are constants (the corresponding Lagrange multipliers), to be determined by the restrictions which define $\mathcal{P}$. Suppose, for instance, that data are considered to be a random sample from a location model centered at θ , and that it is further assumed that $E[\theta] = \mu_0$ and that $\text{Var}[\theta] = \sigma_0^2$. The unrestricted reference prior for any regular location problem may be shown to be uniform, so that here $\pi(\theta|\mathcal{M}, \mathcal{P}_0) = 1$. Thus, the restricted reference prior must be of the form $\pi(\theta|\mathcal{M}, \mathcal{P}) \propto \exp\{\gamma_1\theta + \gamma_2(\theta - \mu_0)^2\}$, with $\int_{\Theta} \theta \pi(\theta|\mathcal{M}, \mathcal{P}) d\theta = \mu_0$ and $\int_{\Theta} (\theta - \mu_0)^2 \pi(\theta|\mathcal{M}, \mathcal{P}) d\theta = \sigma_0^2$. Hence, $\pi(\theta|\mathcal{M}, \mathcal{P})$ is the *normal* distribution with the specified mean and variance, $\mathbf{N}(\theta|\mu_0, \sigma_0)$.

4.2 Frequentist Properties

Bayesian methods provide a *direct* solution to the problems typically posed in statistical inference; indeed, posterior distributions precisely state what can be said about unknown quantities of interest *given* available data and prior knowledge. In particular, unrestricted reference posterior distributions state what could be said if no prior knowledge about the quantities of interest were available.

A frequentist analysis of the behaviour of Bayesian procedures under repeated sampling may, however, be illuminating, for this provides some interesting connections between frequentist and Bayesian inference. It is found that the frequentist properties of Bayesian reference procedures are typically excellent, and may be used to provide a form of calibration for reference posterior probabilities.

Point Estimation It is generally accepted that, as the sample size increases, a “good” estimator $\tilde{\theta}$ of θ ought to get the correct value of θ eventually, that is to be *consistent*. Under appropriate regularity conditions, any Bayes estimator ϕ^* of any function $\phi(\theta)$ converges in probability to $\phi(\theta)$, so that sequences of Bayes estimators are typically *consistent*. Indeed, it is known that if there is a consistent sequence of estimators, then Bayes estimators are consistent. The rate of convergence is often best for reference Bayes estimators.

It is also generally accepted that a “good” estimator should be *admissible*, that is, *not dominated* by any other estimator in the sense that its expected loss under sampling (conditional to θ) cannot be larger for all θ values than that corresponding to another estimator. Any *proper* Bayes estimator is admissible; moreover, as established by Wald [1950], a procedure *must* be Bayesian (proper or improper) to be admissible. Most published admissibility results refer to quadratic loss functions, but they often extend to more general loss functions. Reference Bayes estimators are typically admissible with respect to appropriate loss functions.

Notice, however, that many other apparently intuitive frequentist ideas on estimation have been proved to be potentially misleading. For example, given a sequence of n Bernoulli observations with parameter θ resulting in r positive trials, the *best unbiased* estimate of θ^2 is found to be $r(r-1)/\{n(n-1)\}$, which yields $\hat{\theta}^2 = 0$ when $r = 1$; but to estimate the probability of two positive trials as zero, when one positive trial has been observed, is less than sensible. In marked contrast, any Bayes reference estimator provides a reasonable answer. For example, the intrinsic estimator of θ^2 is simply $(\theta^*)^2$, where θ^* is the intrinsic estimator of θ described in Section 5.1. In particular, if $r = 1$ and $n = 2$ the intrinsic estimator of θ^2 is (as one would naturally expect) $(\theta^*)^2 = 1/4$.

Interval Estimation As the sample size increases, the frequentist coverage probability of a posterior q -credible region typically converges to q so that, for *large samples*, Bayesian credible intervals may (under regularity conditions) be interpreted as *approximate* frequentist confidence regions: under repeated sampling, a Bayesian q -credible region of θ based on a large sample will cover the true value of θ approximately $100q\%$ of times. Detailed results are readily available for univariate problems. For instance, consider the probability model $\{p(D|\omega), \omega \in \Omega\}$, let $\theta = \theta(\omega)$ be any univariate quantity of interest, and let $\mathbf{t} = \mathbf{t}(D) \in T$ be any sufficient statistic. If $\theta_q(\mathbf{t})$ denotes the $100q\%$ quantile of the posterior distribution of θ which corresponds to some unspecified prior, so that

$$(45) \quad \Pr[\theta \leq \theta_q(\mathbf{t})|\mathbf{t}] = \int_{\theta \leq \theta_q(\mathbf{t})} \pi(\theta|\mathbf{t}) d\theta = q,$$

then the coverage probability of the q -credible interval $\{\theta; \theta \leq \theta_q(\mathbf{t})\}$,

$$(46) \quad \Pr[\theta_q(\mathbf{t}) \geq \theta | \omega] = \int_{\theta_q(\mathbf{t}) \geq \theta} p(\mathbf{t} | \omega) d\mathbf{t},$$

is such that

$$(47) \quad \Pr[\theta_q(\mathbf{t}) \geq \theta | \omega] = \Pr[\theta \leq \theta_q(\mathbf{t}) | \mathbf{t}] + O(n^{-1/2}).$$

This *asymptotic* approximation is true for *all* (sufficiently regular) positive priors. However, the approximation is better, actually $O(n^{-1})$, for a particular class of priors known as (first-order) *probability matching* priors. For details on probability matching priors see Datta and Sweeting [2005] and references therein. Reference priors are typically found to be probability matching priors, so that they provide this improved asymptotic agreement. As a matter of fact, the agreement (in regular problems) is typically quite good even for relatively small samples.

EXAMPLE 11 Product of normal means. Consider the case where independent random samples $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_m\}$ have respectively been taken from the normal densities $N(x | \omega_1, 1)$ and $N(y | \omega_2, 1)$, and suppose that the quantity of interest is the product of their means, $\phi = \omega_1 \omega_2$ (for instance, one may be interested in inferences about the area ϕ of a rectangular piece of land, given measurements $\{x_i\}$ and $\{y_j\}$ of its sides). Notice that this is a simplified version of a problem that it is often encountered in the sciences, where one is interested in the product of several magnitudes, all of which have been measured with error. Using the procedure described in Example 8, with the natural approximating sequence induced by $(\omega_1, \omega_2) \in [-i, i]^2$, the ϕ -reference prior is found to be

$$(48) \quad \pi_\phi(\omega_1, \omega_2 | \mathcal{M}, \mathcal{P}_0) \propto (n\omega_1^2 + m\omega_2^2)^{-1/2},$$

very different from the uniform prior $\pi_{\omega_1}(\omega_1, \omega_2 | \mathcal{M}, \mathcal{P}_0) = \pi_{\omega_2}(\omega_1, \omega_2 | \mathcal{M}, \mathcal{P}_0) = 1$ which should be used to make objective inferences about either ω_1 or ω_2 . The prior $\pi_\phi(\omega_1, \omega_2)$ may be shown to provide approximate agreement between Bayesian credible regions and frequentist confidence intervals for ϕ ; indeed, this prior (with $m = n$) was originally suggested by Stein in the 1980's to obtain such approximate agreement. The same example was later used by Efron [1986] to stress the fact that, even within a fixed probability model $\{p(D | \omega), \omega \in \Omega\}$, the prior required to make objective inferences about some function of the parameters $\phi = \phi(\omega)$ must generally depend on the function ϕ . For further details on the reference analysis of this problem, see [Berger and Bernardo, 1989].

The numerical agreement between reference Bayesian credible regions and frequentist confidence intervals is actually perfect in special circumstances. Indeed, as Lindley [1958] pointed out, this is the case in those problems of inference which may be transformed to location-scale problems.

EXAMPLE 3, continued. Inference on normal parameters. Let $D = \{x_1, \dots, x_n\}$ be a random sample from a normal distribution $N(x | \mu, \sigma)$. As mentioned before,

the reference posterior of the quantity of interest μ is the Student distribution $\text{St}(\mu|\bar{x}, s/\sqrt{n-1}, n-1)$. Thus, normalizing μ , the *posterior* distribution of $t(\mu) = \sqrt{n-1}(\bar{x}-\mu)/s$, as a function of μ given D , is the standard Student $\text{St}(t|0, 1, n-1)$ with $n-1$ degrees of freedom. On the other hand, this function t is recognized to be precisely the conventional t statistic, whose *sampling distribution* is well known to *also* be standard Student with $n-1$ degrees of freedom. It follows that, *for all sample sizes*, posterior *reference* credible intervals for μ given the data will be *numerically identical* to frequentist confidence intervals based on the sampling distribution of t .

A similar result is obtained in inferences about the variance. Thus, the reference *posterior* distribution of $\lambda = \sigma^{-2}$ is the Gamma distribution $\text{Ga}(\lambda|(n-1)/2, ns^2/2)$ and, hence, the *posterior* distribution of $r = ns^2/\sigma^2$, as a function of σ^2 given D , is a (central) χ^2 with $n-1$ degrees of freedom. But the function r is recognized to be a conventional statistic for this problem, whose *sampling distribution* is well known to *also* be χ^2 with $n-1$ degrees of freedom. It follows that, *for all sample sizes*, posterior *reference* credible intervals for σ^2 (or any one-to-one function of σ^2) given the data will be *numerically identical* to frequentist confidence intervals based on the sampling distribution of r . $\triangleleft$

5 INFERENCE SUMMARIES

From a Bayesian viewpoint, the final outcome of a problem of inference about *any* unknown quantity is nothing but the corresponding posterior distribution. Thus, given some data D and conditions C , *all* that can be said about any function ω of the parameters which govern the model is contained in the posterior distribution $\pi(\omega|D, C)$, and *all* that can be said about some function $\mathbf{y}$ of future observations from the same model is contained in its posterior predictive distribution $p(\mathbf{y}|D, C)$. Indeed, Bayesian inference may technically be described as a decision problem where the space of available actions is the class of those posterior probability distributions of the quantity of interest which are compatible with accepted assumptions.

However, to make it easier for the user to assimilate the appropriate conclusions, it is often convenient to *summarize* the information contained in the posterior distribution by (i) providing values of the quantity of interest which, in the light of the data, are likely to be “close” to its true value and by (ii) measuring the compatibility of the results with hypothetical values of the quantity of interest which might have been suggested in the context of the investigation. In this section, those Bayesian counterparts of traditional *estimation* and *hypothesis testing* problems are briefly considered.

5.1 Estimation

In one or two dimensions, a graph of the posterior probability density of the quantity of interest (or the probability mass function in the discrete case) immediately

conveys an intuitive, “impressionist” summary of the main conclusions which may possibly be drawn on its value. Indeed, this is greatly appreciated by users, and may be quoted as an important asset of Bayesian methods. From a plot of its posterior density, the region where (given the data) a univariate quantity of interest is likely to lie is easily distinguished. For instance, all important conclusions about the value of the gravitational field in Example 3 are qualitatively available from Figure 3. However, this does not easily extend to more than two dimensions and, besides, *quantitative* conclusions (in a simpler form than that provided by the mathematical expression of the posterior distribution) are often required.

Point Estimation Let D be the available data, which are assumed to have been generated by a probability model $\{p(D|\omega), \omega \in \Omega\}$, and let $\theta = \theta(\omega) \in \Theta$ be the quantity of interest. A *point estimator* of θ is some function of the data $\tilde{\theta} = \tilde{\theta}(D)$ which could be regarded as an appropriate proxy for the actual, unknown value of θ . Formally, to choose a point estimate for θ is a *decision problem*, where the action space is the class Θ of possible θ values. From a decision-theoretic perspective, to choose a point estimate $\tilde{\theta}$ of some quantity θ is a *decision* to act as though $\tilde{\theta}$ were θ , not to assert something about the value of θ (although desire to assert something simple may well be the reason to obtain an estimate). As prescribed by the foundations of decision theory (Section 2), to solve this decision problem it is necessary to specify a *loss function* $L(\tilde{\theta}, \theta)$ measuring the consequences of acting *as if* the true value of the quantity of interest were $\tilde{\theta}$, when it is actually θ . The expected posterior loss if $\tilde{\theta}$ were used is

$$(49) \quad \bar{L}[\tilde{\theta}|D] = \int_{\Theta} L(\tilde{\theta}, \theta) \pi(\theta|D) d\theta,$$

and the corresponding *Bayes estimator* θ^* is that function of the data, $\theta^* = \theta^*(D)$, which minimizes this expectation.

EXAMPLE 12 Conventional Bayes estimators. For any given model and data, the Bayes estimator obviously depends on the chosen loss function. The loss function is context specific, and should be chosen in terms of the anticipated uses of the estimate; however, a number of conventional loss functions have been suggested for those situations where no particular uses are envisaged. These loss functions produce estimates which may be regarded as simple descriptions of the *location* of the posterior distribution. For example, if the loss function is quadratic, so that $L(\tilde{\theta}, \theta) = (\tilde{\theta} - \theta)^t(\tilde{\theta} - \theta)$, then the Bayes estimator is the *posterior mean* $\theta^* = E[\theta|D]$, assuming that the mean exists. Similarly, if the loss function is a zero-one function, so that $L(\tilde{\theta}, \theta) = 0$ if $\tilde{\theta}$ belongs to a ball or radius ϵ centered in θ and $L(\tilde{\theta}, \theta) = 1$ otherwise, then the Bayes estimator θ^* tends to the *posterior mode* as the ball radius ϵ tends to zero, assuming that a unique mode exists. If θ is univariate and the loss function is linear, so that $L(\tilde{\theta}, \theta) = c_1(\tilde{\theta} - \theta)$ if $\tilde{\theta} \geq \theta$, and $L(\tilde{\theta}, \theta) = c_2(\theta - \tilde{\theta})$ otherwise, then the Bayes estimator is the *posterior quantile* of order $c_2/(c_1 + c_2)$, so that $\Pr[\theta < \theta^*] = c_2/(c_1 + c_2)$. In particular, if $c_1 = c_2$, the Bayes estimator is the *posterior median*. The results derived for linear loss functions

clearly illustrate the fact that *any* possible parameter value may turn out be the Bayes estimator: it all depends on the loss function describing the consequences of the anticipated uses of the estimate.

EXAMPLE 13 Intrinsic estimation. Conventional loss functions are typically non-invariant under reparametrization. It follows that the Bayes estimator ϕ^* of a one-to-one transformation $\phi = \phi(\theta)$ of the original parameter θ is not necessarily $\phi(\theta^*)$ (the *univariate* posterior median, which *is* invariant, is an interesting exception). Moreover, conventional loss functions focus on the “distance” between the estimate $\tilde{\theta}$ and the true value θ , rather than on the “distance” between the probability models they label. Inference-oriented loss functions directly focus on how different the probability model $p(D|\theta, \lambda)$ is from its closest approximation within the family $\{p(D|\tilde{\theta}, \lambda_i), \lambda_i \in \Lambda\}$, and typically produce invariant solutions. An attractive example is the *intrinsic discrepancy*, $\delta(\tilde{\theta}, \theta)$ defined as the minimum logarithmic divergence between a probability model labeled by θ and a probability model labeled by $\tilde{\theta}$. When there are no nuisance parameters, this is given by

$$(50) \quad \delta(\tilde{\theta}, \theta) = \min\{\kappa(\tilde{\theta}|\theta), \kappa(\theta|\tilde{\theta})\}, \quad \kappa(\theta_i|\theta) = \int_T p(\mathbf{t}|\theta) \log \frac{p(\mathbf{t}|\theta)}{p(\mathbf{t}|\theta_i)} d\mathbf{t},$$

where $\mathbf{t} = \mathbf{t}(D) \in T$ is *any* sufficient statistic (which may well be the whole data set D). The definition is easily extended to problems with nuisance parameters; in this case,

$$(51) \quad \delta(\tilde{\theta}, \theta, \lambda) = \min_{\lambda_i \in \Lambda} \delta(\tilde{\theta}, \lambda_i, \theta, \lambda)$$

measures the logarithmic divergence from $p(\mathbf{t}|\theta, \lambda)$ of its closest approximation with $\theta = \tilde{\theta}$, and the loss function now depends on the complete parameter vector (θ, λ) . Although not explicitly shown in the notation, the intrinsic discrepancy function typically depends on the sample size n ; indeed, when the data consist of a random sample $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ from some model $p(\mathbf{x}|\theta)$ then $\kappa(\theta_i|\theta) = n \int_X p(\mathbf{x}|\theta) \log[p(\mathbf{x}|\theta)/p(\mathbf{x}|\theta_i)] d\mathbf{x}$ so that the discrepancy associated with the full model is simply n times the discrepancy which corresponds to a single observation. The intrinsic discrepancy is a symmetric, non-negative loss function with a direct interpretation in information-theoretic terms as the minimum amount of information which is expected to be necessary to distinguish between the model $p(D|\theta, \lambda)$ and its closest approximation within the class $\{p(D|\tilde{\theta}, \lambda_i), \lambda_i \in \Lambda\}$. Moreover, it is invariant under one-to-one reparametrization of the parameter of interest θ , and does not depend on the choice of the nuisance parameter λ . The *intrinsic estimator* is naturally obtained by minimizing the reference posterior expected intrinsic discrepancy

$$(52) \quad d(\tilde{\theta}|D) = \int_{\Lambda} \int_{\Theta} \delta(\tilde{\theta}, \theta, \lambda) \pi(\theta, \lambda|D) d\theta d\lambda.$$

Since the intrinsic discrepancy is invariant under reparametrization, minimizing its posterior expectation produces invariant estimators. For further details on intrinsic point estimation see [Bernardo and Juárez, 2003; Bernardo, 2006].

EXAMPLE 2, continued. Intrinsic estimation of a binomial parameter. In the estimation of a binomial proportion θ , given data $D = (n, r)$, the Bayes reference estimator associated with the quadratic loss (the corresponding posterior mean) is $E[\theta|D] = (r + \frac{1}{2})/(n + 1)$, while the quadratic loss based estimator of, say, the log-odds $\phi(\theta) = \log[\theta/(1 - \theta)]$, is found to be $E[\phi|D] = \psi(r + \frac{1}{2}) - \psi(n - r + \frac{1}{2})$ (where $\psi(x) = d \log[\Gamma(x)]/dx$ is the *digamma* function), which is *not* equal to $\phi(E[\theta|D])$. The intrinsic loss function in this problem is

$$(53) \quad \delta(\tilde{\theta}, \theta) = n \min\{\kappa(\tilde{\theta}|\theta), \kappa(\theta|\tilde{\theta})\}, \quad \kappa(\theta_i|\theta) = \theta \log \frac{\theta}{\theta_i} + (1 - \theta) \log \frac{1 - \theta}{1 - \theta_i},$$

and the corresponding intrinsic estimator θ^* is obtained by minimizing the expected posterior loss $d(\tilde{\theta}|D) = \int \delta(\tilde{\theta}, \theta) \pi(\theta|D) d\theta$. The exact value of θ^* may be obtained by numerical minimization, but a very good approximation is given by $\theta^* \approx (r + \frac{1}{3})/(n + \frac{2}{3})$.

Since intrinsic estimation is an invariant procedure, the intrinsic estimator of the log-odds will simply be the log-odds of the intrinsic estimator of θ . As one would expect, when r and $n - r$ are both large, all Bayes estimators of any well-behaved function $\phi(\theta)$ will cluster around $\phi(E[\theta|D])$. $\triangleleft$

Interval Estimation To describe the inferential content of the posterior distribution of the quantity of interest $\pi(\theta|D)$ it is often convenient to quote regions $R \subset \Theta$ of given probability under $\pi(\theta|D)$. For example, the identification of regions containing 50%, 90%, 95%, or 99% of the probability under the posterior may be sufficient to convey the general quantitative messages implicit in $\pi(\theta|D)$; indeed, this is the intuitive basis of graphical representations of univariate distributions like those provided by boxplots.

Any region $R \subset \Theta$ such that $\int_R \pi(\theta|D) d\theta = q$ (so that, given data D , the true value of θ belongs to R with probability q), is said to be a posterior *q-credible region* of θ . Notice that this provides immediately a direct intuitive statement about the unknown quantity of interest θ in probability terms, in marked contrast to the circumlocutory statements provided by frequentist confidence intervals. Clearly, for any given q there are generally infinitely many credible regions. A credible region is invariant under reparametrization; thus, for any q -credible region R of θ , $\phi(R)$ is a q -credible region of $\phi = \phi(\theta)$. Sometimes, credible regions are selected to have minimum size (length, area, volume), resulting in *highest probability density* (HPD) regions, where all points in the region have larger probability density than all points outside. However, HPD regions are *not* invariant under reparametrization: the image $\phi(R)$ of an HPD region R will be a credible region for ϕ , but will not generally be HPD; indeed, there is no compelling reason to restrict attention to HPD credible regions. In one-dimensional problems, posterior quantiles are often used to derive credible regions. Thus, if $\theta_q = \theta_q(D)$ is the 100 q % posterior quantile of θ , then $R = \{\theta; \theta \leq \theta_q\}$ is a one-sided, typically unique q -credible region, and it is invariant under reparametrization. Indeed, *probability centered q-credible*

regions of the form $R = \{\theta; \theta_{(1-q)/2} \leq \theta \leq \theta_{(1+q)/2}\}$ are easier to compute, and are often quoted in preference to HPD regions.

EXAMPLE 3. Inference on normal parameters, continued. In the numerical example about the value of the gravitational field described in Figure 3a, the interval $[9.788, 9.829]$ in the unrestricted posterior density of g is a HPD, 95%-credible region for g . Similarly, the interval $[9.7803, 9.8322]$ in Figure 3b is also a 95%-credible region for g , but it is not HPD. $\triangleleft$

Decision theory may also be used to select credible regions. Thus, *lowest posterior loss* (LPL) regions, are defined as those where all points in the region have smaller posterior expected loss than all points outside. Using the intrinsic discrepancy as a loss function yields *intrinsic credible regions* which, as one would expect from an invariant loss function, are coherent under one-to-one transformations. For details, see [Bernardo, 2005b; 2007].

The concept of a credible region for a function $\theta = \theta(\omega)$ of the parameter vector is trivially extended to prediction problems. Thus, a posterior q -credible region for $\mathbf{x} \in \mathcal{X}$ is a subset R of the sample space $\mathcal{X}$ with posterior predictive probability q , so that $\int_R p(\mathbf{x}|D) d\mathbf{x} = q$.

5.2 Hypothesis Testing

The reference posterior distribution $\pi(\theta|D)$ of the quantity of interest θ conveys immediate intuitive information on those values of θ which, given the assumed model, may be taken to be *compatible* with the observed data D , namely, those with a relatively high probability density. Sometimes, a *restriction* $\theta \in \Theta_0 \subset \Theta$ of the possible values of the quantity of interest (where Θ_0 may possibly consists of a single value θ_0) is suggested in the course of the investigation as deserving special consideration, either because restricting θ to Θ_0 would greatly simplify the model, or because there are additional, context specific arguments suggesting that $\theta \in \Theta_0$. Intuitively, the *hypothesis* $H_0 \equiv \{\theta \in \Theta_0\}$ should be judged to be *compatible* with the observed data D if there are elements in Θ_0 with a relatively high posterior density. However, a more precise conclusion is often required and, once again, this is made possible by adopting a decision-oriented approach. Formally, testing the hypothesis $H_0 \equiv \{\theta \in \Theta_0\}$ is a *decision problem* where the action space has only two elements, namely to accept (a_0) or to reject (a_1) the proposed restriction. To solve this decision problem, it is necessary to specify an appropriate loss function, $L(a_i, \theta)$, measuring the consequences of accepting or rejecting H_0 as a function of the actual value θ of the vector of interest. Notice that this requires the statement of an *alternative* a_1 to accepting H_0 ; this is only to be expected, for an action is taken not because it is good, but because it is better than anything else that has been imagined.

Given data D , the optimal action will be to reject H_0 if (and only if) the expected posterior loss of accepting, $\int_{\Theta} L(a_0, \theta) \pi(\theta|D) d\theta$, is larger than the expected posterior loss of rejecting, $\int_{\Theta} L(a_1, \theta) \pi(\theta|D) d\theta$, that is, if (and only if)

$$(54) \quad \int_{\Theta} [L(a_0, \theta) - L(a_1, \theta)] \pi(\theta|D) d\theta = \int_{\Theta} \Delta L(\theta) \pi(\theta|D) d\theta > 0.$$

Therefore, only the loss difference $\Delta L(\theta) = L(a_0, \theta) - L(a_1, \theta)$, which measures the *advantage* of rejecting H_0 as a function of θ , has to be specified. Thus, as common sense dictates, the hypothesis H_0 should be rejected whenever the expected advantage of rejecting H_0 is positive.

A crucial element in the specification of the loss function is a description of what is actually meant by rejecting H_0 . By assumption a_0 means to act *as if* H_0 were true, i.e. as if $\theta \in \Theta_0$, but there are at least two options for the alternative action a_1 . This may either mean (i) the *negation* of H_0 , that is to act as if $\theta \notin \Theta_0$ or, alternatively, it may rather mean (ii) to reject the simplification implied by H_0 and to keep the unrestricted model, $\theta \in \Theta$, which is true by assumption. Both options have been analyzed in the literature, although it may be argued that the problems of scientific data analysis where hypothesis testing procedures are typically used are better described by the second alternative. Indeed, an established model, identified by $H_0 \equiv \{\theta \in \Theta_0\}$, is often embedded into a more general model, $\{\theta \in \Theta, \Theta_0 \subset \Theta\}$, constructed to include possibly promising departures from H_0 , and it is required to verify whether presently available data D are still compatible with $\theta \in \Theta_0$, or whether the extension to $\theta \in \Theta$ is really required.

EXAMPLE 14 Conventional hypothesis testing. Let $\pi(\theta|D)$, $\theta \in \Theta$, be the posterior distribution of the quantity of interest, let a_0 be the decision to work under the restriction $\theta \in \Theta_0$ and let a_1 be the decision to work under the complementary restriction $\theta \notin \Theta_0$. Suppose, moreover, that the loss structure has the simple, zero-one form given by $\{L(a_0, \theta) = 0, L(a_1, \theta) = 1\}$ if $\theta \in \Theta_0$ and, similarly, $\{L(a_0, \theta) = 1, L(a_1, \theta) = 0\}$ if $\theta \notin \Theta_0$, so that the *advantage* $\Delta L(\theta)$ of rejecting H_0 is 1 if $\theta \notin \Theta_0$ and it is -1 otherwise. With this loss function it is immediately found that the optimal action is to reject H_0 if (and only if) $\Pr(\theta \notin \Theta_0|D) > \Pr(\theta \in \Theta_0|D)$. Notice that this formulation requires that $\Pr(\theta \in \Theta_0) > 0$, that is, that the hypothesis H_0 has a strictly positive prior probability. If θ is a continuous parameter and Θ_0 has zero measure (for instance if H_0 consists of a single point θ_0), this requires the use of a non-regular “sharp” prior concentrating a positive probability mass on Θ_0 . For details see [Kaas and Rafetery, 1995] and references therein.

EXAMPLE 15 Intrinsic hypothesis testing. Again, let $\pi(\theta|D)$, $\theta \in \Theta$, be the posterior distribution of the quantity of interest, and let a_0 be the decision to work under the restriction $\theta \in \Theta_0$, but let a_1 now be the decision to keep the general, unrestricted model $\theta \in \Theta$. In this case, the advantage $\Delta L(\theta)$ of rejecting H_0 as a function of θ may safely be assumed to have the form $\Delta L(\theta) = \delta(\Theta_0, \theta) - \delta^*$, for some $\delta^* > 0$, where (i) $\delta(\Theta_0, \theta)$ is some measure of the discrepancy between the assumed model $p(D|\theta)$ and its closest approximation within the class $\{p(D|\theta_0), \theta_0 \in \Theta_0\}$, such that $\delta(\Theta_0, \theta) = 0$ whenever $\theta \in \Theta_0$, and (ii) δ^* is a context dependent *utility constant* which measures the (necessarily positive) advantage of being able to work with the simpler model when it is true. Choices for both $\delta(\Theta_0, \theta)$ and δ^* which

may be appropriate for general use will now be described.

For reasons similar to those supporting its use in point estimation, an attractive choice for the function $d(\Theta_0, \theta)$ is an appropriate extension of the intrinsic discrepancy; when there are no nuisance parameters, this is given by

$$(55) \quad \delta(\Theta_0, \theta) = \inf_{\theta_0 \in \Theta_0} \min\{\kappa(\theta_0|\theta), \kappa(\theta|\theta_0)\}$$

where $\kappa(\theta_0|\theta) = \int_T p(\mathbf{t}|\theta) \log\{p(\mathbf{t}|\theta)/p(\mathbf{t}|\theta_0)\} d\mathbf{t}$, and $\mathbf{t} = \mathbf{t}(D) \in T$ is *any* sufficient statistic, which may well be the whole dataset D . As before, if the data $D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ consist of a random sample from $p(\mathbf{x}|\theta)$, then

$$(56) \quad \kappa(\theta_0|\theta) = n \int_X p(\mathbf{x}|\theta) \log \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\theta_0)} d\mathbf{x}.$$

Naturally, the loss function $\delta(\Theta_0, \theta)$ reduces to the intrinsic discrepancy $\delta(\theta_0, \theta)$ of Example 13 when Θ_0 contains a single element θ_0 . Besides, as in the case of estimation, the definition is easily extended to problems with nuisance parameters, with

$$(57) \quad \delta(\Theta_0, \theta, \lambda) = \inf_{\theta_0 \in \Theta_0, \lambda_0 \in \Lambda} \delta(\theta_0, \lambda_0, \theta, \lambda).$$

The hypothesis H_0 should be rejected if the posterior expected advantage of rejecting is

$$(58) \quad d(\Theta_0|D) = \int_{\Lambda} \int_{\Theta} \delta(\Theta_0, \theta, \lambda) \pi(\theta, \lambda|D) d\theta d\lambda > \delta^*,$$

for some $\delta^* > 0$. As an expectation of a non-negative quantity, $d(\Theta_0, D)$ is obviously nonnegative. Moreover, if $\phi = \phi(\theta)$ is a one-to-one transformation of θ , then $d(\phi(\Theta_0), D) = d(\Theta_0, D)$ so that, as one should clearly require, the expected intrinsic loss of rejecting H_0 is invariant under reparametrization.

It may be shown that, as the sample size increases, the expected value of $d(\Theta_0, D)$ under sampling tends to one when H_0 is true, and tends to infinity otherwise; thus $d(\Theta_0, D)$ may be regarded as a continuous, positive measure of how inappropriate (in loss of information units) it would be to simplify the model by accepting H_0 . In traditional language, $d(\Theta_0, D)$ is a *test statistic* for H_0 and the hypothesis should be rejected if the value of $d(\Theta_0, D)$ exceeds some *critical value* δ^* . In sharp contrast to conventional hypothesis testing, this critical value δ^* is found to be a context specific, positive utility constant δ^* , which may precisely be described as the number of *information units* which the decision maker is prepared to lose in order to be able to work with the simpler model H_0 , and does not depend on the sampling properties of the probability model. The procedure may be used with standard, continuous regular priors even in *sharp* hypothesis testing, when Θ_0 is a zero-measure set (as would be the case if θ is continuous and Θ_0 contains a single point θ_0).

Naturally, to implement the test, the utility constant δ^* which defines the rejection region must be chosen. Values of $d(\Theta_0, D)$ of about 1 should be regarded as

an indication of no evidence against H_0 , since this is precisely the expected value of the test statistic $d(\Theta_0, D)$ under repeated sampling from the null. It follows from its definition that $d(\Theta_0, D)$ is the reference posterior expectation of the log-likelihood ratio against the null. Hence, values of $d(\Theta_0, D)$ of about $\log[12] \approx 2.5$, and $\log[150] \approx 5$ should be respectively regarded as an indication of mild evidence against H_0 , and significant evidence against H_0 . In the canonical problem of testing a value $\mu = \mu_0$ for the mean of a normal distribution with known variance (see below), these values correspond to the observed sample mean $\bar{x}$ respectively lying 2 or 3 posterior standard deviations from the null value μ_0 . Notice that, in sharp contrast to frequentist hypothesis testing, where it is hazily recommended to adjust the significance level for dimensionality and sample size, this provides an absolute scale (in information units) which remains valid for any sample size and any dimensionality.

For further details on intrinsic hypothesis testing see [Bernardo and Rueda, 2003; Bernardo and Pérez, 2007].

EXAMPLE 16 Testing the value of a normal mean. Let the data $D = \{x_1, \dots, x_n\}$ be a random sample from a normal distribution $\mathbf{N}(x|\mu, \sigma)$, where σ is assumed to be known, and consider the problem of testing whether these data are or are not compatible with some specific sharp hypothesis $H_0 \equiv \{\mu = \mu_0\}$ on the value of the mean.

The conventional approach to this problem requires a non-regular prior which places a probability mass, say p_0 , on the value μ_0 to be tested, with the remaining $1 - p_0$ probability continuously distributed over $\mathfrak{R}$. If this prior is chosen to be $\pi(\mu|\mu \neq \mu_0) = N(\mu|\mu_0, \sigma_0)$, Bayes theorem may be used to obtain the corresponding posterior probability,

$$(59) \quad \Pr[\mu_0|D, \lambda] = \frac{B_{01}(D, \lambda) p_0}{(1 - p_0) + p_0 B_{01}(D, \lambda)},$$

$$(60) \quad B_{01}(D, \lambda) = \left(1 + \frac{n}{\lambda}\right)^{1/2} \exp\left[-\frac{1}{2} \frac{n}{n + \lambda} z^2\right],$$

where $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ measures, in standard deviations, the distance between $\bar{x}$ and μ_0 and $\lambda = \sigma^2/\sigma_0^2$ is the ratio of model to prior variance. The function $B_{01}(D, \lambda)$, a ratio of (integrated) likelihood functions, is called the *Bayes factor* in favour of H_0 . With a conventional zero-one loss function, H_0 should be rejected if $\Pr[\mu_0|D, \lambda] < 1/2$. The choices $p_0 = 1/2$ and $\lambda = 1$ or $\lambda = 1/2$, describing particular forms of *sharp* prior knowledge, have been suggested in the literature for routine use. The conventional approach to sharp hypothesis testing deals with situations of *concentrated* prior probability; it *assumes* important prior knowledge about the value of μ and, hence, should *not* be used unless this is an appropriate assumption. Moreover [Bartlett, 1957], the resulting posterior probability is extremely sensitive to the specific prior specification. In most applications, H_0 is really a hazily defined small region rather than a point. For moderate sample sizes, the posterior probability $\Pr[\mu_0|D, \lambda]$ is an *approximation* to the posterior

probability $\Pr[\mu_0 - \epsilon < \mu < \mu_0 + \epsilon | D, \lambda]$ for some small interval around μ_0 which would have been obtained from a regular, continuous prior heavily concentrated around μ_0 ; however, this approximation *always* breaks down for sufficiently large sample sizes. One consequence (which is immediately apparent from the last two equations) is that for any *fixed* value of the pertinent statistic z , the posterior probability of the null, $\Pr[\mu_0 | D, \lambda]$, tends to one as $n \rightarrow \infty$. Far from being specific to this example, this unappealing behaviour of posterior probabilities based on sharp, non-regular priors generally known as *Lindley's paradox* [Lindley, 1957] is *always* present in the conventional Bayesian approach to *sharp* hypothesis testing.

The intrinsic approach may be used without assuming any sharp prior knowledge. The intrinsic discrepancy is $\delta(\mu_0, \mu) = n(\mu - \mu_0)^2 / (2\sigma^2)$, a simple transformation of the standardized distance between μ and μ_0 . The reference prior is uniform and the corresponding (proper) posterior distribution is $\pi(\mu | D) = N(\mu | \bar{x}, \sigma / \sqrt{n})$. The expected value of $\delta(\mu_0, \mu)$ with respect to this posterior is $d(\mu_0, D) = (1 + z^2)/2$, where $z = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$ is the standardized distance between $\bar{x}$ and μ_0 . As foretold by the general theory, the expected value of $d(\mu_0, D)$ under repeated sampling is one if $\mu = \mu_0$, and increases linearly with n if $\mu \neq \mu_0$. Moreover, in this canonical example, to reject H_0 whenever $|z| > 2$ or $|z| > 3$, that is whenever μ_0 is 2 or 3 posterior standard deviations away from $\bar{x}$, respectively corresponds to rejecting H_0 whenever $d(\mu_0, D)$ is larger than 2.5, or larger than 5.

If σ is unknown, the reference prior is $\pi(\mu, \sigma) = \sigma^{-1}$, and the intrinsic discrepancy becomes

$$(61) \quad \delta(\mu_0, \mu, \sigma) = \frac{n}{2} \log \left[1 + \left(\frac{\mu - \mu_0}{\sigma} \right)^2 \right].$$

The intrinsic test statistic $d(\mu_0, D)$ is found as the expected value of $\delta(\mu_0, \mu, \sigma)$ under the corresponding joint referenceposterior distribution; this may be exactly expressed in terms of hypergeometric functions, and is well approximated by

$$(62) \quad d(\mu_0, D) \approx \frac{1}{2} + \frac{n}{2} \log \left(1 + \frac{t^2}{n} \right),$$

where t is the traditional statistic $t = \sqrt{n-1}(\bar{x} - \mu_0)/s$, $ns^2 = \sum_j (x_j - \bar{x})^2$. For instance, for samples sizes 5, 30 and 1000, and using the utility constant $\delta^* = 5$, the hypothesis H_0 would be rejected whenever $|t|$ is respectively larger than 5.025, 3.240, and 3.007.

6 DISCUSSION

This article focuses on the basic concepts of the Bayesian paradigm, with special emphasis on the derivation of “objective” methods, where the results only depend on the data obtained and the model assumed. Many technical aspects have been spared; the interested reader is referred to the bibliography for further information. This final section briefly reviews the main arguments for an objective Bayesian approach.

6.1 Coherence

By using probability distributions to characterize *all* uncertainties in the problem, the Bayesian paradigm reduces statistical inference to applied probability, thereby ensuring the coherence of the proposed solutions. There is no need to investigate, on a case by case basis, whether or not the solution to a particular problem is logically correct: a Bayesian result is only a *mathematical consequence of explicitly stated assumptions* and hence, unless a logical mistake has been committed in its derivation, it cannot be formally wrong. In marked contrast, conventional statistical methods are plagued with counterexamples. These include, among many others, negative estimators of positive quantities, q -confidence regions ($q < 1$) which consist of the whole parameter space, empty sets of “appropriate” solutions, and incompatible answers from alternative methodologies simultaneously supported by the theory.

The Bayesian approach does require, however, the specification of a (prior) probability distribution over the parameter space. The sentence “a prior distribution does not exist for this problem” is often stated to justify the use of non-Bayesian methods. However, the general representation theorem *proves the existence* of such a distribution whenever the observations are assumed to be exchangeable (and, if they are assumed to be a random sample then, *a fortiori*, they are assumed to be exchangeable). To ignore this fact, and to proceed as if a prior distribution did not exist, just because it is not easy to specify, is mathematically untenable.

6.2 Objectivity

It is generally accepted that any statistical analysis is subjective, in the sense that it is always conditional on accepted assumptions (on the structure of the data, on the probability model, and on the outcome space) and those assumptions, although possibly well founded, are definitely *subjective* choices. It is, therefore, mandatory to make all assumptions very explicit.

Users of conventional statistical methods rarely dispute the mathematical foundations of the Bayesian approach, but claim to be able to produce “objective” answers in contrast to the possibly subjective elements involved in the choice of the prior distribution.

Bayesian methods do indeed require the choice of a prior distribution, and critics of the Bayesian approach systematically point out that in many important situations, including scientific reporting and public decision making, the results must exclusively depend on documented data which might be subject to independent scrutiny. This is of course true, but those critics choose to ignore the fact that this particular case is covered within the Bayesian approach by the use of *reference* prior distributions which (i) are mathematically derived from the accepted probability model (and, hence, they are “objective” insofar as the choice of that model might be objective) and, (ii) by construction, they produce posterior probability distributions which, given the accepted probability model, *only* contain the information about their values which data may provide and, *optionally*, any further

contextual information over which there might be universal agreement.

6.3 Operational meaning

An issue related to objectivity is that of the operational meaning of reference posterior probabilities; it is found that the analysis of their behaviour under repeated sampling provides a suggestive form of calibration. Indeed, $\Pr[\theta \in R|D] = \int_R \pi(\theta|D) d\theta$, the reference posterior probability that $\theta \in R$, is *both* a measure of the conditional uncertainty (given the assumed model and the observed data D) about the event that the unknown value of θ belongs to $R \subset \Theta$, and the limiting proportion of the regions which would cover θ under repeated sampling conditional on data “sufficiently similar” to D . Under broad conditions (to guarantee regular asymptotic behaviour), all large data sets from the same model are “sufficiently similar” among themselves in this sense and hence, given those conditions, reference posterior credible regions are *approximate* frequentist confidence regions.

The conditions for this approximate equivalence to hold exclude, however, important special cases, like those involving “extreme” or “relevant” observations. In very special situations, when probability models may be transformed to location-scale models, there is an exact equivalence; in those cases reference posterior credible intervals are, for any sample size, exact frequentist confidence intervals.

6.4 Generality

In sharp contrast to most conventional statistical methods, which may only be exactly applied to a handful of relatively simple stylized situations, Bayesian methods are defined to be totally general. Indeed, for a given probability model and prior distribution over its parameters, the derivation of posterior distributions is a well-defined mathematical exercise. In particular, Bayesian methods do not require any particular regularity conditions on the probability model, do not depend on the existence of sufficient statistics of finite dimension, do not rely on asymptotic relations, and do not require the derivation of any sampling distribution, nor (a fortiori) the existence of a “pivotal” statistic whose sampling distribution is independent of the parameters.

However, when used in complex models with many parameters, Bayesian methods often require the computation of multidimensional definite integrals and, for a long time in the past, this requirement effectively placed practical limits on the complexity of the problems which could be handled. This has dramatically changed in recent years with the general availability of large computing power, and the parallel development of simulation-based numerical integration techniques like *importance sampling* or *Markov chain Monte Carlo* (MCMC). These methods provide a structure within which many complex models may be analyzed using generic software. MCMC is numerical integration using Markov chains. Monte Carlo integration proceeds by drawing samples from the required distributions, and computing sample averages to approximate expectations. MCMC methods

draw the required samples by running appropriately defined Markov chains for a long time; specific methods to construct those chains include the Gibbs sampler and the Metropolis algorithm, originated in the 1950's in the literature of statistical physics. The development of improved algorithms and appropriate diagnostic tools to establish their convergence, remains a very active research area.

For an introduction to MCMC methods in Bayesian inference, see [Gilks *et al.*, 1996; Mira, 2005], and references therein.

BIBLIOGRAPHY

- [Bartlett, 1957] M. Bartlett. A comment on D. V. Lindley's statistical paradox. 44, 533–534, 1957.
- [Berger and Bernardo, 1989] J. O. Berger and J. M. Bernardo. Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.*, 84, 200–207, 1989.
- [Berger and Bernardo, 1992a] J. O. Berger and J. M. Bernardo. Ordered group reference priors with applications to a multinomial problem. 79, 25–37, 1992.
- [Berger and Bernardo, 1992b] J. O. Berger and J. M. Bernardo. Reference priors in a variance components problem. *Bayesian Analysis in Statistics and Econometrics*, 323–340, 1992.
- [Berger and Bernardo, 1992c] J. O. Berger and J. M. Bernardo. On the development of reference priors. 4, 35–60, 1992 (with discussion).
- [Berger *et al.*, 2009] J. O. Berger, J. M. Bernardo, and D. Sun. The formal definition of reference priors. *Ann Statist.*, 37, 905–938, 2009.
- [Berger, 1985] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Berlin: Springer., 1985.
- [Bernardo, 1979a] J. M. Bernardo. Expected information as expected utility. *Ann Statist.*, 7, 686–690, 1979.
- [Bernardo, 1979b] J. M. Bernardo. Reference posterior distributions for Bayesian inference. *J. R. Statist. Soc. B* **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference* **1** (G. C. Tiao and N. G. Polson, eds). Oxford: Edward Elgar, 229–263, 1979.
- [Bernardo, 1981] J. M. Bernardo. Reference decisions. *Symposia Mathematica* **25**, 85–94, 1981.
- [Bernardo, 1997] J. M. Bernardo. Noninformative priors do not exist. *J. Statist. Planning and Inference*, 65, 159–189, 1997 (with discussion).
- [Bernardo, 2005a] J. M. Bernardo. Reference analysis. *Handbook of Statistics* **25** (D. K. Dey and C. R. Rao eds.) Amsterdam: Elsevier, 17–90, 2005.
- [Bernardo, 2005b] J. M. Bernardo. Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test*, 14, 317–384, 2005 (with discussion).
- [Bernardo, 2006] J. M. Bernardo. Intrinsic point estimation of the normal variance. *Bayesian Statistics and its Applications*. (S. K. Upadhyay, U. Singh and D. K. Dey, eds.) New Delhi: Anamaya Pub, 110–121, 2006.
- [Bernardo, 2007] J. M. Bernardo. Objective Bayesian point and region estimation in location-scale models. *Sort* **14**, 3–44, 2007.
- [Bernardo and Juárez, 2003] J. M. Bernardo and M. A. Juárez. Intrinsic Estimation. 7, 465–476, 2003.
- [Bernardo and Pérez, 2007] J. M. Bernardo and S. Pérez. Comparing normal means: New methods for an old problem. *Bayesian Analysis* **2**, 45–58, 2007.
- [Bernardo and Ramón, 1998] J. M. Bernardo and J. M. Ramón. An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician*, 47, 1–35, 1998.
- [Bernardo and Rueda, 2002] J. M. Bernardo and R. Rueda. Bayesian hypothesis testing: A reference approach. *International Statistical Review* **70**, 351–372, 2002.
- [Bernardo and Smith, 1994] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*, Chichester: Wiley, 1994. Second edition forthcoming.
- [Box and Tiao, 1973] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley, 1973.

- [Datta and Sweeting, 2005] G. S. Datta and T. J. Sweeting. Probability matching priors. *Handbook of Statistics* **25** (D. K. Dey and C. R. Rao eds.) Amsterdam: Elsevier, 91–114, 2005.
- [Dawid et al., 1973] A. P. Dawid, M. Stone, and J. V. Zidek. Marginalization paradoxes in Bayesian and structural inference. *J. R. Statist. Soc. B* **35**, 189–233, 1973 (with discussion).
- [de Finetti, 1937] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* **7**, 1–68, 1937. Reprinted in 1980 as ‘Foresight; its logical laws, its subjective sources’ in *Studies in Subjective Probability*, 93–158.
- [de Finetti, 1970] B. de Finetti *Teoria delle Probabilità* Turin: Einaudi, 1970. English translation as *Theory of Probability* Chichester: Wiley, 1975.
- [DeGroot, 1970] M. H. DeGroot. *Optimal Statistical Decisions*, New York: McGraw Hill, 1970.
- [Efron, 1986] B. Efron. Why isn’t everyone a Bayesian? *Amer. Statist.*, 40, 1–11, 198 (with discussion).
- [Geisser, 1993] S. Geisser. *Predictive Inference: an Introduction*. London: Chapman and Hall, 1993.
- [Gilks et al., 1996] W. R. Gilks, S. Y. Richardson, and D. J. Spiegelhalter, eds. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall, 1996.
- [Jaynes, 1976] E. T. Jaynes. Confidence intervals vs. Bayesian intervals. *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science* **2** (W. L. Harper and C. A. Hooker, eds). Dordrecht: Reidel, 175–257, 1976 (with discussion).
- [Jeffreys, 1939] H. Jeffreys. *Theory of Probability*. Oxford: Oxford University Press. Third edition in 1961, Oxford: Oxford University Press, 1939.
- [Kass and Raftery, 1995] R. E. Kass and A. E. Raftery. Bayes factors. *J. Amer. Statist. Assoc.*, 90, 773–795, 1995.
- [Kass and Wasserman, 1996] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.*, 91, 1343–1370, 1996.
- [Laplace, 1812] P. S. Laplace. *Théorie Analytique des Probabilités*. Paris: Courcier, 1812. Reprinted as *Oeuvres Complètes de Laplace* **7**, 1878–1912. Paris: Gauthier-Villars.
- [Lindley, 1957] D. V. Lindley. A statistical paradox. 44, 187–192, 1957.
- [Lindley, 1958] D. V. Lindley. Fiducial distribution and Bayes’ Theorem. *J. Roy. Statist. Soc.*, 20, 102–107, 1958.
- [Lindley, 1965] D. V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge: Cambridge University Press, 1965.
- [Lindley, 1972] D. V. Lindley. *Bayesian Statistics, a Review*. Philadelphia, PA: SIAM, 1972.
- [Liseo, 2005] B. Liseo. The elimination of nuisance parameters. *Handbook of Statistics* **25** (D. K. Dey and C. R. Rao eds.) Amsterdam: Elsevier, 193–219, 2005.
- [Mira, 2005] A. Mira. MCMC methods to estimate Bayesian parametric models. *Handbook of Statistics* **25** (D. K. Dey and C. R. Rao eds.) Amsterdam: Elsevier, 415–436, 2005.
- [Ramsey, 1926] F. P. Ramsey. Truth and probability. *The Foundations of Mathematics and Other Logical Essays* (R. B. Braithwaite, ed.). London: Kegan Paul 1926 (1931), 156–198. Reprinted in 1980 in *Studies in Subjective Probability*, 61–92.
- [Savage, 1954] L. J. Savage. *The Foundations of Statistics*. New York: Wiley, 1954. Second edition in 1972, New York: Dover.
- [Stein, 1959] C. Stein. An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.*, 30, 877–880, 1959.
- [Zellner, 1971] A. Zellner. *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley, 1971. Reprinted in 1987, Melbourne, FL: Kreiger.

EVIDENTIAL PROBABILITY AND OBJECTIVE BAYESIAN EPISTEMOLOGY

Gregory Wheeler and Jon Williamson

1 INTRODUCTION

Evidential probability (EP), developed by Henry Kyburg, offers an account of the impact of statistical evidence on single-case probability. According to this theory, observed frequencies of repeatable outcomes determine a probability interval that can be associated with a proposition. After giving a comprehensive introduction to EP in §2, in §3 we describe a recent variant of this approach, *second-order evidential probability* (2oEP). This variant, introduced in [Haenni *et al.*, 2010], interprets a probability interval of EP as bounds on the sharp probability of the corresponding proposition. In turn, this sharp probability can itself be interpreted as the degree to which one ought to believe the proposition in question.

At this stage we introduce objective Bayesian epistemology (OBE), a theory of how evidence helps determine appropriate degrees of belief (§4). OBE might be thought of as a rival to the evidential probability approaches. However, we show in §5 that they can be viewed as complimentary: one can use the rules of EP to narrow down the degree to which one should believe a proposition to an interval, and then use the rules of OBE to help determine an appropriate degree of belief from within this interval. Hence bridges can be built between evidential probability and objective Bayesian epistemology.

2 EVIDENTIAL PROBABILITY

2.1 *Motivation*

Rudolf Carnap [Carnap, 1962] drew a distinction between probability₁, which concerned rational degrees of belief, and probability₂, which concerned statistical regularities. Although he claimed that both notions of probability were crucial to scientific inference, Carnap practically ignored probability₂ in the development of his systems of inductive logic. Evidential probability (EP) [Kyburg, 1961; Kyburg and Teng, 2001], by contrast, is a theory that gives primacy to probability₂, and Kyburg's philosophical program was an uncompromising approach to see how far he could go with relative frequencies. Whereas Bayesianism springs from the view

that probability₁ is all the probability needed for scientific inference, EP arose from the view that probability₂ is all that we really have.

The theory of evidential probability is motivated by two basic ideas: probability assessments should be based upon relative frequencies, to the extent that we know them, and the assignment of probability to specific individuals should be determined by everything that is known about that individual. Evidential probability is conditional probability in the sense that the probability of a sentence χ is evaluated given a set of sentences Γ_δ . But the evidential probability of χ given Γ_δ , written $\text{Prob}(\chi, \Gamma_\delta)$, is a meta-linguistic operation similar in kind to the relation of provability within deductive systems.

The semantics governing the operator $\text{Prob}(\cdot, \cdot)$ is markedly dissimilar to axiomatic theories of probability that take conditional probability as primitive, such as the system developed by Lester Dubbins [Dubbins, 1975; Arló-Costa and Parikh, 2005], and it also resists reduction to linear [de Finetti, 1974] as well as lower previsions [Walley, 1991]. One difference between EP and the first two theories is that EP is interval-valued rather than point-valued, because the relative frequencies that underpin assignment of evidential probability are typically incomplete and approximate. But more generally, EP assignments may violate coherence. For example, suppose that χ and φ are sentences in the object language of evidential probability. The evidential probability of $\chi \wedge \varphi$ given Γ_δ might fail to be less than or equal to the evidential probability that χ given Γ_δ .¹ A point to stress from the start is that evidential probability is a logic of statistical probability statements, and there is nothing in the activity of observing and recording statistical regularities that guarantees that a set of statistical probability statements will comport to the axioms of probability. So, EP is neither a species of Carnapian logical probability nor a kind of Bayesian probabilistic logic.^{2,3} EP is instead a logic for approximate reasoning, thus it is more similar in kind to the theory of rough sets [Pawlak, 1991] and to systems of fuzzy logic [Dubois and Prade, 1980] than to probabilistic logic.

The operator $\text{Prob}(\cdot, \cdot)$ takes as arguments a sentence χ in the first coordinate and a set of statements Γ_δ in the second. The statements in Γ_δ represent a knowledge base, which includes categorical statements as well as statistical generalities. Theorems of logic and mathematics are examples of categorical statements, but so too are contingent generalities. One example of a contingent categorical statement is the ideal gas law. EP views the propositions “ $2 + 2 = 4$ ” and “ $PV = nRT$ ”

¹Specifically, the lower bound of $\text{Prob}(\chi \wedge \varphi, \Gamma_\delta)$ may be strictly greater than the lower bound of $\text{Prob}(\chi, \Gamma_\delta)$.

²See the essays by Levi and by Seidenfeld in [Harper and Wheeler, 2007] for a discussion of the sharp differences between EP and Bayesian approaches, particularly on the issue of conditionalization. A point sometimes overlooked by critics is that there are different *systems* of evidential probability corresponding to different conditions we assume to hold. Results pertaining to a qualitative representation of EP inference, for instance, assume that Γ_δ is consistent. A version of conditionalization holds in EP given that there is specific statistical statement pertaining to the relevant joint distribution. See [Kyburg, 2007] and [Teng, 2007].

³EP does inherit some notions from Keynes’s [Keynes, 1921], however, including that probabilities are interval-valued and not necessarily comparable.

within a chemistry knowledge base as indistinguishable analytic truths that are built into a particular language adopted for handling statistical statements to do with gasses. In light of EP's expansive view of analyticity, the theory represents all categorical statements as universally quantified sentences within a guarded fragment of first-order logic [Andréka *et al.*, 1998].⁴

Statistical generalities within Γ_δ , by contrast, are viewed as direct inference statements and are represented by syntax that is unique to evidential probability. *Direct inference*, recall, is the probability assigned a target subclass given known frequency information about a reference population, and is often contrasted to *indirect inference*, which is the assignment of probability to a population given observed frequencies in a sample. Kyburg's ingenious idea was to solve the problem of indirect inference by viewing it as a form of direct inference. Since the philosophical problems concerning direct inference are much less contentious than those raised by indirect inference, the unusual properties and behavior of evidential probability should be weighed against this achievement [Levi, 2007].

Direct inference statements are statements that record the observed frequency of items satisfying a specified reference class that also satisfy a particular target class, and take the form of

$$\% \vec{x}(\tau(\vec{x}), \rho(\vec{x}), [l, u]).$$

This schematic statement says that given a sequence of propositional variables $\vec{x}$ that satisfies the reference class predicate ρ , the proportion of ρ that also satisfies the target class predicate τ is between l and u .

Syntactically, ' $\tau(\vec{x}), \rho(\vec{x}), [l, u]$ ' is an open formula schema, where ' $\tau(\cdot)$ ' and ' $\rho(\cdot)$ ' are replaced by open first-order formulas, ' $\vec{x}$ ' is replaced by a sequence of propositional variables, and ' $[l, u]$ ' is replaced by a specific sub-interval of $[0, 1]$. The binding operator '%' is similar to the ordinary binding operators ($\forall, \exists$) of first-order logic, except that '%' is a 3-place binding operator over the propositional variables appearing the *target formula* $\tau(\vec{x})$ and the *reference formula* $\rho(\vec{x})$, and binding those formulas to an interval.⁵ The language $\mathcal{L}^{ep}$ of evidential probability then is a guarded first-order language augmented to include direct inference statements. There are additional formation rules for direct inference statements that are designed to block spurious inference, but we shall pass over these details of the theory.⁶ An example of a direct inference statement that might appear in Γ_δ is

$$\% x(B(x), A(x), [.71, .83]),$$

which expresses that the proportion of As that are also Bs lies between 0.71 and 0.83.

As for semantics, a model M of $\mathcal{L}^{ep}$ is a pair, $\langle \mathcal{D}, \mathcal{I} \rangle$, where $\mathcal{D}$ is a two-sorted domain consisting of mathematical objects, $\mathcal{D}_m$, and a *finite* set of empirical objects, $\mathcal{D}_e$. EP assumes that there is a first giraffe and a last carbon molecule. $\mathcal{I}$ is

⁴A guarded fragment of first-order logic is a decidable fragment of first-order logic.

⁵Hereafter we relax notation and simply use an arbitrary variable ' x ' for ' $\vec{x}$ '.

⁶See [Kyburg and Teng, 2001].

an interpretation function that is the union of two partial functions, one defined on $\mathcal{D}_m$ and the other on $\mathcal{D}_e$. Otherwise M behaves like a first-order model: the interpretation function $\mathcal{I}$ maps (empirical/mathematical) terms into the (empirical/mathematical) elements of $\mathcal{D}$, monadic predicates into subsets of $\mathcal{D}$, n -arity relation symbols into $\mathcal{D}^n$, and so forth. Variable assignments also behave as one would expect, with the only difference being the procedure for assigning truth to direct inference statements.

The basic idea behind the semantics for direct inference statements is that the statistical quantifier ‘ $\%$ ’ ranges over the finite empirical domain $\mathcal{D}_e$, not the field terms l, u that denote real numbers in $\mathcal{D}_m$. This means that the only free variables in a direct inference statement range over a finite domain, which will allow us to look at proportions of models in which a sentence is true. A *satisfaction set* of an open formula φ whose only free n variables are empirical in the subset of $\mathcal{D}^n$ that satisfies φ .

A direct inference statement $\%x(\tau(x), \rho(x), [l, u])$ is true in M under variable assignment v iff the cardinality of the satisfaction sets for the open formula ρ under v is greater than 0 and the ratio of the cardinality of satisfaction sets for $\tau(x^*) \wedge \rho(x^*)$ over the cardinality of the satisfaction sets for $\rho(x)$ (under v) is in the closed interval $[l, u]$, where all variables of x occur in ρ , all variables of τ occur in ρ , and x^* is the sequence of variables free in ρ but not bound by $\%x$ [Kyburg and Teng, 2001].

The operator $\text{Prob}(\cdot, \cdot)$ then provides a semantics for a nonmonotonic consequence operator [Wheeler, 2004; Kyburg *et al.*, 2007]. The structural properties enjoyed by this consequence operator are as follows:⁷

Properties of EP Entailment: Let $\models$ denote classical consequence and let $\equiv$ denote classical logical equivalence. Whenever $\mu \wedge \xi, \nu \wedge \xi$ are sentences of $\mathcal{L}^{ep}$,

Right Weakening: if $\mu \approx \nu$ and $\nu \models \xi$ then $\mu \approx \xi$.

Left Classical Equivalence: if $\mu \approx \nu$ and $\mu \equiv \xi$ then $\xi \approx \nu$.

(KTW) Cautious Monotony: if $\mu \models \nu$ and $\mu \approx \xi$ then $\mu \wedge \xi \approx \nu$.

(KTW) Premise Disjunction: if $\mu \models \nu$ and $\xi \approx \nu$ then $\mu \vee \xi \approx \nu$.

(KTW) Conclusion Conjunction: if $\mu \models \nu$ and $\mu \approx \xi$ then $\mu \approx \nu \wedge \xi$.

As an aside, this qualitative EP-entailment relation presents challenges in handling disjunction in the premises since the KTW disjunction property admits a novel reversal effect similar to, but distinct from, Simpson’s paradox [Kyburg *et al.*, 2007; Wheeler, 2007]. This raises a question over how best to axiomatize EP. One approach, which is followed by [Hawthorne and Makinson, 2007] and considered in

⁷Note that these properties are similar to, but strictly weaker than, the properties of the class of cumulative consequence relations specified by System P [Kraus *et al.*, 1990]. To yield the axioms of System P, replace the nonmonotonic consequence operator $\approx$ for $\models$ in the premise position of [And*], [Or*], and [Cautious Monotonicity*].

[Kyburg *et al.*, 2007], is to replace Boolean disjunction by ‘exclusive-or’. While this route ensures nice properties for $\approx$, it does so at the expense of introducing a dubious connective into the object language that is neither associative nor compositional.⁸ Another approach explored in [Kyburg *et al.*, 2007] is a weakened disjunction axiom (KTW Or) that yields a sub-System P nonmonotonic logic and preserves the standard properties of the positive Boolean connectives.

Now that we have a picture of what EP is, we turn to consider the inferential behavior of the theory. We propose to do this with a simple ball-draw experiment before considering the specifics of the theory in more detail in the next section.

EXAMPLE 1. Suppose the proportion of white balls (W) in an urn (U) is known to be within $[\cdot 33, 4]$, and that ball t is drawn from U . These facts are represented in Γ_δ by the sentences, $\%x(W(x), U(x), [\cdot 33, 4])$ and $U(t)$.

- (i) If these two statements are all that we know about t , i.e., they are the only statements in Γ_δ pertaining to t , then $\text{Prob}(W(t), \Gamma_\delta) = [\cdot 33, 4]$.
- (ii) Suppose additionally that the proportion of plastic balls (P) that are white is observed to be between $[\cdot 31, \cdot 36]$, t is plastic, and that every plastic ball is a white ball. That means that $\%x(P(x), U(x), [\cdot 31, \cdot 36])$, $P(t)$, and $\forall x.P(x) \rightarrow W(x)$ are added to Γ_δ as well. Then there is conflicting statistical knowledge about t , since either:
 1. the probability that ball t is white is between $[\cdot 33, 4]$, by reason of $\%x(W(x), U(x), [\cdot 33, 4])$, or
 2. the probability that ball t is white is between $[\cdot 31, \cdot 36]$, by reason of $\%x(W(x), P(x), [\cdot 31, \cdot 36])$,

may apply. There are several ways that statistical statements may conflict and there are rules for handling each type, which we will discuss in the next section. But in this particular case, because it is known that the class of plastic balls is more *specific* than the class of balls in U and we have statistics for the proportion of plastic balls that are also white balls, the statistical statement in (2) dominates the statement in (1). So, the probability that t is white is in $[\cdot 31, \cdot 36]$.

- (iii) Adapting an example from [Kyburg and Teng, 2001, 216], suppose U is partitioned into three cells, u_1 , u_2 , and u_3 , and that the following compound experiment is performed. First, a cell of U is selected at random. Then a ball is drawn at random from that cell. To simplify matters, suppose that there are 25 balls in U and 9 are white such that 3 of 5 balls from u_1 are white, but only 3 of 10 balls in u_2 and 3 of 10 in u_3 are white. The following table summarizes this information.

⁸Example: ‘ $A \text{ xor } B \text{ xor } C$ ’ is true if A, B, C are; and ‘ $(A \text{ xor } B) \text{ xor } C$ ’ is not equivalent to ‘ $A \text{ xor } (B \text{ xor } C)$ ’ when A is false but B and C both true.

Table 1. Compound Experiment

	u_1	u_2	u_3	
W	3	3	3	9
$\overline{W}$	2	7	7	16
	5	10	10	25

We are interested in the probability that t is white, but we have a conflict. Given these over all precise values, we would have $\text{Prob}(W(t), \Gamma_\delta) = \frac{9}{25}$. However, since we know that t was selected by performing this compound experiment, then we also have the conflicting direct inference statement $\%x, y(W^*(x, y), U^*(x, y), [.4, .4])$, where U^* is the set of compound two stage experiments, and W^* is the set of outcomes in which the ball selected is white.⁹ We should prefer the statistics from the compound experiment because they are *richer* in information. So, the probability that t is white is .4.

- (iv) Finally, if there happens to be *no* statistical knowledge in Γ_δ pertaining to t , then we would be completely ignorant of the probability that t is white. So in the case of total ignorance, $\text{Prob}(W(t), \Gamma_\delta) = [0, 1]$.

We now turn to a more detailed account of how EP calculates probabilities.

2.2 Calculating Evidential Probability

In practice an individual may belong to several reference classes with known statistics. Selecting the appropriate statistical distribution among the class of potential probability statements is the *problem of the reference class*. The task of assigning evidential probability to a statement χ relative to a set of evidential certainties relies upon a procedure for eliminating excess candidates from the set of potential candidates. This procedure is described in terms of the following definitions.

Potential Probability Statement: A *potential probability statement* for χ with respect to Γ_δ is a tuple $\langle t, \tau(t), \rho(t), [l, u] \rangle$, such that instances of $\chi \leftrightarrow \tau(t)$, $\rho(t)$, and $\%x(\tau(x), \rho(x), [l, u])$ are each in Γ_δ .

Given χ , there are possibly many target statements of form $\tau(t)$ in Γ_δ that have the same truth value as χ . If it is known that individual t satisfies ρ , and known that between .7 and .8 of ρ 's are also τ 's, then $\langle t, \tau(t), \rho(t), [.7, .8] \rangle$ represents a potential probability statement for χ based on the knowledge base Γ_δ . Our focus

⁹ Γ_δ should also include the categorical statements $\forall x, y(U^*(x, y) \rightarrow W(y))$, which says that the second stage of U concerns the proportion of balls that are white, and three statements of the form $[W^*(\mu, t) \leftrightarrow W(t)]$, where μ is replaced by u_1, u_2, u_3 , respectively. This statement tells us that everything that's true of W^* is true of W , which is what ensures that this conflict is detected.

will be on the statistical statements $\%x(\tau(x), \rho(x), [l, u])$ in Γ_δ that are the basis for each potential probability statement.

Selecting the appropriate probability interval for χ from the set of potential probability statements reduces to identifying and resolving conflicts among the statistical statements that are the basis for each potential probability statement.

Conflict: Two intervals $[l, u]$ and $[l', u']$ *conflict* iff neither $[l, u] \subset [l', u']$ nor $[l, u] \supset [l', u']$. Two statistical statements conflict iff their intervals conflict.

Note that conflicting intervals may be disjoint or intersect. For technical reasons an interval is said to conflict with itself.

Cover: Let X be a set of intervals. An interval $[l, u]$ *covers* X iff for every $[l', u'] \in X$, $l \leq l'$ and $u' \leq u$. A cover $[l, u]$ of X is the *smallest cover*, $Cov(X)$, iff for all covers $[l^*, u^*]$ of X , $l^* \leq l$ and $u \leq u^*$.

Difference Set: (i) Let X be a non-empty set of intervals and $\mathcal{P}(X)$ be the powerset of X . A non-empty $Y \in \mathcal{P}(X)$ is a *difference set of X* iff Y includes every $x \in X$ that conflicts with some $y \in Y$. (ii) Let X be the set of intervals associated with a set Γ of statistical statements, and Y be the set of intervals associated with a set Λ of statistical statements. Λ is a difference set to Γ iff Y is closed under difference with respect to X .

EXAMPLE 2. An example might help. Let X be the set of intervals $[\cdot 30, \cdot 40]$, $[\cdot 35, \cdot 45]$, $[\cdot 325, \cdot 475]$, $[\cdot 50, \cdot 55]$, $[\cdot 30, \cdot 70]$, $[\cdot 20, \cdot 60]$, $[\cdot 10, \cdot 90]$. There are three sets closed under difference with respect to X :

- (i) $\{[\cdot 30, \cdot 40], [\cdot 35, \cdot 45], [\cdot 325, \cdot 475], [\cdot 50, \cdot 55]\}$,
- (ii) $\{[\cdot 30, \cdot 70], [\cdot 20, \cdot 60]\}$,
- (iii) $\{[\cdot 10, \cdot 90]\}$.

The intuitive idea behind a difference set is to eliminate intervals from a set that are broad enough to include all other intervals in that set. The interval $[\cdot 10, \cdot 90]$ is the broadest interval in X . So, it only appears as a singleton difference set and is not included in any other difference set of X . It is not necessary that all intervals in a difference set X be pairwise conflicting intervals. Difference sets identify the set of all possible conflicts for each potential probability statement in order to find that conflicting set with the shortest cover.

Minimal Cover Under Difference: (i) Let X be a non-empty set of intervals and $\mathcal{Y} = \{Y_1, \dots, Y_n\}$ the set of all difference sets of X . The *minimal cover under difference* of X is the smallest cover of the elements of $\mathcal{Y}$, i.e., the shortest cover in $\{Cov(Y_1), \dots, Cov(Y_n)\}$.

(ii) Let X be the set of intervals associated with a set Γ of statistical statements, and $\mathcal{Y}$ be the set of all difference sets of X associated with a set Λ of statistical statements. Then the minimal cover under difference of Γ is the minimal cover under difference of X .

EP resolves conflicting statistical data concerning χ by applying two principles to the set of potential probability assignments, *Richness* and *Specificity*, to yield

a class of *relevant statements*. The (controversial) principle of *Strength* is then applied to this set of relevant statistical statements, yielding a unique probability interval for χ . For discussion of these principles, see [Teng, 2007].

We illustrate these principles in terms of a pair (φ, ϑ) of conflicting statistical statements for χ , and represent their respective reference formulas by ρ_φ and ρ_ϑ . The probability interval assigned to χ is the shortest cover of the relevant statistics remaining after applying these principles.

1. **[Richness]** If φ and ϑ conflict and ϑ is based on a marginal distribution while φ is based on the full joint distribution, eliminate ϑ .
2. **[Specificity]** If φ and ϑ both survive the principle of richness, and if $\rho_\varphi \subset \rho_\vartheta$, then eliminate $\langle \tau, \rho_\vartheta, [l, u] \rangle$ from *all* difference sets.

The principle of specificity says that if it is known that the reference class ρ_φ is included in the reference class ρ_ϑ , then eliminate the statement ϑ . The statistical statements that survive the sequential application of the principle of richness followed by the principle of specificity are called *relevant statistics*.

3. **[Strength]** Let Γ^{RS} be the set of relevant statistical statements for χ with respect to Γ_δ , and let the set $\{\Lambda_1, \dots, \Lambda_n\}$ be the set of difference sets of Γ^{RS} . The principle of strength is the choosing of the minimal cover under difference of Γ^{RS} , i.e., the selection of the shortest cover in $\{Cov(\Lambda_1), \dots, Cov(\Lambda_n)\}$.

The evidential probability of χ is the minimal cover under difference of Γ^{RS} .

We may define Γ_ϵ , the set of **practical certainties**, in terms of a body of evidence Γ_δ :

$$\Gamma_\epsilon = \{\chi : \exists l, u \text{ (Prob}(\neg\chi, \Gamma_\delta) = [l, u] \wedge u \leq \epsilon)\},$$

or alternatively,

$$\Gamma_\epsilon = \{\chi : \exists l, u \text{ (Prob}(\chi, \Gamma_\delta) = [l, u] \wedge l \geq 1 - \epsilon)\}.$$

The set Γ_ϵ is the set of statements that the evidence Γ_δ warrants accepting; we say a sentence χ is ϵ -*accepted* if $\chi \in \Gamma_\epsilon$. Thus we may add to our knowledge base statements that are nonmonotonic consequences of Γ_δ with respect to a threshold point of acceptance.

Finally, we may view the evidence Γ_δ to provide real-valued bounds on ‘degrees of belief’ owing to the logical structure of sentences accepted into Γ_δ . However, the probability interval $[l, u]$ associated with χ does not specify a range of equally rational degrees of belief between l and u : the interval $[l, u]$ itself is not a quantity, only l and u are quantities, which are used to specify bounds. On this view, no degree of belief within $[l, u]$ is defensible, which is in marked contrast to the view offered by Objective Bayesianism.

3 SECOND-ORDER EVIDENTIAL PROBABILITY

3.1 *Motivation*

Second-order evidential probability—developed in [Haenni *et al.*, 2010]—goes beyond Kyburg’s evidential probability in two ways. First, it treats an EP interval as bounds on sharp probability. Second, it disentangles reasoning under uncertainty from questions of acceptance and rejection. Here we explain both moves in more detail.

3.1.0.1 Bounds on Degrees of Belief. Kyburg maintained that one can interpret an evidential probability interval for proposition χ as providing bounds on the degree to which an agent should believe χ , but he had reservations about this move:

Should we speak of partial beliefs as ‘degrees of belief’? Although probabilities are intervals, we could still do so. Or we could say that any ‘degree of belief’ satisfying the probability bounds was ‘rational’. But what would be the point of doing so? We agree with Ramsey that logic cannot determine a real-valued a priori degree of belief in pulling a black ball from an urn. This seems a case where degrees of belief are not appropriate. No particular degree of belief is defensible. We deny that there are any appropriate a priori degrees of belief, though there is a fine a priori probability: $[0, 1]$. There are real valued *bounds* on degrees of belief, determined by the logical structure of our evidence. [Kyburg, 2003, p. 147]

Kyburg is making the following points here. Evidence rarely determines a unique value for an agent’s degree of belief—rather, it narrows down rational belief to an interval. One can view this interval as providing bounds on rational degree of belief, but since evidence can not be used to justify the choice of one point over another in this interval, there seems to be little reason to talk of the individual points and one can instead simply treat the interval itself as a partial belief.

This view fits very well with the interpretation of evidential probability as some kind of measure of weight of evidence. (And evidential probability provides a natural measure of weight of evidence: the narrower the interval, the weightier the evidence.) Hence if evidence only narrows down probability to an interval, then there does indeed seem to be little need to talk of anything but the interval when measuring features of the evidence. But the view does not address how to fix a sharp degree of belief—intentionally so, since Kyburg’s program was designed in part to show us how far one can go with relative frequency information alone. Even so, we may ask whether there is a way to use the resources of evidential probability to fix sharp degrees of belief. In other words, we might return to Carnap’s original distinction between probability₁ and probability₂ and ask how a theory of the latter can be used to constrain the former. If we want to talk

Step 1	Evidence	$\{P(\varphi) \in [l_\varphi, u_\varphi]\}$
Step 2	Acceptance	$\Gamma_\delta = \{\varphi : l_\varphi \geq 1 - \delta\}$
Step 3	Uncertain reasoning	$\{P(\chi) \in [l_\chi, u_\chi]\}$
Step 4	Acceptance	$\Gamma_\varepsilon = \{\chi : l_\chi \geq 1 - \varepsilon\}$

Figure 1. The structure of (first-order) EP inferences.

not only of the quality of our evidence but also of our disposition to act on that evidence, then it would appear that we need a richer language than that provided by EP alone: while evidence—and hence EP—cannot provide grounds to prefer one point in the interval over another as one’s degree of belief, there may be other, non-evidential grounds for some such preference, and formalising this move would require going beyond EP.

Reconciling EP with a Bayesian approach has been considered to be highly problematic [Levi, 1977; Levi, 1980; Seidenfeld, 2007], and was vigorously resisted by Kyburg throughout his life. On the other hand, Kyburg’s own search for an EP-compatible decision theory was rather limited [Kyburg, 1990]. It is natural then to explore how to modify evidential probability in order that it might handle point-valued degrees of belief and thereby fit with Bayesian decision theory. Accordingly second-order EP departs from Kyburg’s EP by viewing evidential probability intervals as bounds on rational degree of belief, $P(\chi) \in \text{Prob}(\chi, \Gamma_\delta)$. In §5 we will go further still by viewing the results of EP as feeding into objective Bayesian epistemology.

3.1.0.2 Acceptance and Rejection. If we allow ourselves the language of point-valued degrees of belief, (first-order) EP can be seen to work like this. An agent has evidence which consists of some propositions $\varphi_1, \dots, \varphi_n$ and information about their risk levels. He then accepts those propositions whose risk levels are below the agent’s threshold δ . This leaves him with the evidential certainties, $\Gamma_\delta = \{\varphi_i : P(\varphi_i) \geq 1 - \delta\}$. From Γ_δ the agent infers propositions ψ of the form $P(\chi) \in [l, u]$. In turn, from these propositions the agent infers the practical certainties $\Gamma_\varepsilon = \{\chi : l \geq 1 - \varepsilon\}$. This sequence of steps is depicted in Figure 1.

There are two modes of reasoning that are intermeshed here: on the one hand the agent is using evidence to reason under uncertainty about the conclusion proposition ψ , and on the other he is deciding which propositions to accept and reject. The acceptance mode appears in two places: deciding which evidential propositions to accept and deciding whether to accept the proposition χ to which the conclusion ψ refers.

With second-order EP, on the other hand, *acceptance is delayed until all reasoning under uncertainty is completed*. Then we treat acceptance as a decision problem requiring a decision-theoretic solution—e.g., accept those propositions

Step 1	Evidence	$\Phi = \{P(\varphi) \in [l_\varphi, u_\varphi]\}$
Step 2	Uncertain reasoning	$\Psi = \{P(\chi) \in [l_\chi, u_\chi]\}$
Step 3	Acceptance	$\{\chi : \text{decision-theoretically optimal}\}$

Figure 2. The structure of 2oEP inferences.

whose acceptance maximises expected utility.¹⁰ Coupling this solution with the use of point-valued probabilities we have second-order evidential probability (2oEP), whose inferential steps are represented in Figure 2.

There are two considerations that motivate this more strict separation of uncertain reasoning and acceptance.

First, such a separation allows one to chain inferences—something which is not possible in 1oEP. By ‘chaining inferences’ we mean that the results of step 2 of 2oEP can be treated as an input for a new round of uncertain reasoning, to be recombined with evidence and to yield further inferences. Only once the chain of uncertain reasoning is complete will the acceptance phase kick in. Chaining of inferences is explained in further detail in §3.2.

Second, such a separation allows one to keep track of the uncertainties that attach to the evidence. To each item of evidence φ attaches an interval $[l_\varphi, u_\varphi]$ representing the risk or reliability of that evidence. In 1oEP, step 2 ensures that one works just with those propositions φ whose risk levels meet the threshold of acceptance. But in 2oEP there is no acceptance phase before uncertain reasoning is initiated, so one works with the entirety of the evidence, including the risk intervals themselves. While the use of this extra information makes inference rather more complicated, it also makes inference more accurate since the extra information can matter—the results of 2oEP can differ from the results of 1oEP.

We adopt a decision-theoretic account of acceptance for the following reason. In 1oEP, each act of acceptance uniformly accepts those propositions whose associated risk is less than some fixed threshold: δ in step 2 and ε in step 4. (This allows statements to detach from their risk levels and play a role as logical constraints in inference.) But in practice thresholds of acceptance depend not so much on the step in the chain of reasoning as on the proposition concerned, and, indeed, the whole inferential set-up. To take a favourite example of Kyburg’s, consider a lottery. The threshold of acceptance of the proposition *the lottery ticket that the seller is offering me will lose* may be higher than that of *the coin with a bias in favour of heads that I am about to toss will land heads* and lower than that of *the moon is made of blue cheese*. This is because nothing may hang on the

¹⁰Note that maximising expected utility is not straightforward in this case since bounds on probabilities, rather than the probabilities themselves, are input into the decision problem. EP-calibrated objective Bayesianism (§5) goes a step further by determining point-valued probabilities from these bounds, thereby making maximisation of expected utility straightforward. See [Williamson, 2009] for more on the combining objective Bayesianism with a decision-theoretic account of acceptance.

coin toss (in which case a 60% bias in favour of heads may be quite adequate for acceptance), while rather a lot hangs on accepting that the moon is made of blue cheese—many other propositions that I have hitherto granted will have to be revisited if I were to accept this proposition. Moreover, if I am going to use the judgement to decide whether to buy a ticket then the threshold of acceptance of the lottery proposition should plausibly depend on the extent to which I value the prize. Given these considerations, acceptance of a proposition can fruitfully be viewed as a decision problem, depending on the decision set-up including associated utilities [Williamson, 2009]. Again, while this is more complicated than the 1oEP solution of modelling acceptance using a fixed threshold, the subtleties of a full-blown decision-theoretic account can matter to the resulting inferences.

3.2 Calculating Second-order EP

In this section we will be concerned with developing some machinery to perform uncertain reasoning in second-order evidential probability (step 2 in Figure 2). See [Haenni *et al.*, 2010] for further details of this approach.

3.2.1 Entailment

Let $\mathcal{L}^\sharp$ be a propositional language whose propositional variables are of the form $\varphi^{[a,b]}$ for atomic propositions $\varphi \in \mathcal{L}$.¹¹ Here $\mathcal{L}$ is the language of (first-order) EP extended to include statements of the form $P(\chi) \in [l, u]$, and, for proposition φ of $\mathcal{L}$, $\varphi^{[a,b]}$ is short for $P(\varphi) \in [a, b]$. Hence in $\mathcal{L}^\sharp$ we can express propositions about higher-order probabilities, e.g., $P(\chi) \in [l, u]^{[a,b]}$ which is short for $P(P(\chi) \in [l, u]) \in [a, b]$. We write φ^a as an abbreviation of $\varphi^{[a,a]}$.

For $\mu, \nu \in \mathcal{L}^\sharp$ write $\mu \approx_{2o} \nu$ if ν deductively follows from μ by appealing to the axioms of probability and the following EP-motivated axioms:

A1: Given $\varphi_1^1, \dots, \varphi_n^1$, if $\text{Prob}(\chi, \{\varphi_1, \dots, \varphi_n\}) = [l, u]$ is derivable by (first-order) EP then infer ψ^1 , where $\psi \in \mathcal{L}$ is the statement $P(\chi) \in [l, u]$.

A2: Given ψ^1 then infer $\chi^{[l,u]}$, where $\psi \in \mathcal{L}$ is the statement $P(\chi) \in [l, u]$.

Axiom A1 ensures that EP inferences carry over to 2oEP, while axiom A2 ensures that probabilities at the first-order level can constrain those at the second-order level.

The entailment relation $\approx_{2o}$ will be taken to constitute *core second-order EP*. The idea is that when input evidence Φ consisting of a set of sentences of $\mathcal{L}^\sharp$, one infers a set Ψ of further such sentences using the above consequence relation. Note that although $\approx_{2o}$ is essentially classical consequence with extra axioms, it is a nonmonotonic consequence relation since 1oEP is nonmonotonic. But 2oEP yields a strong logic inasmuch as it combines the axioms of probability with the rules of

¹¹As it stands $\mathcal{L}^\sharp$ contains uncountably many propositional variables, but restrictions can be placed on a, b to circumscribe the language if need be.

EP, and so questions of consistency arise. Will there always be some probability function that satisfies the constraints imposed by 1oEP consequences of evidence? Not always: see [Seidenfeld, 2007] for some counterexamples. Consequently, some consistency-maintenance procedure needs to be invoked to cope with such cases. (Of course, some consistency maintenance procedure will in any case be required to handle certain inconsistent sets of evidential propositions, so there may be no extra burden here.) One option is to consider probability functions satisfying (EP consequences of) maximal satisfiable subsets of evidential statements, for example. In this paper we will not commit to a particular consistency-maintenance procedure; we leave this interesting question as a topic for further research.

3.2.2 Credal Networks

This entailment relation can be implemented using probabilistic networks, as we shall now explain. For efficiency reasons, we make the following further assumptions. First we assume that P is distributed uniformly over the EP interval unless there is evidence otherwise:

A3: If $\Phi \models_{2o} \chi^{[l,u]}$ then $P(\chi^{[l',u']}|\Phi) = \frac{|[l,u] \cap [l',u']|}{|[l,u]|}$, as long as this is consistent with other consequences of Φ .

Second, we assume that items of evidence are independent unless there is evidence of dependence:

A4: If $\varphi_1^{[a_1,b_1]}, \dots, \varphi_k^{[a_k,b_k]} \in \Phi$ then $P(\varphi_1^{[a_1,b_1]}, \dots, \varphi_k^{[a_k,b_k]}) = P(\varphi_1^{[a_1,b_1]}) \dots P(\varphi_k^{[a_k,b_k]})$, as long as this is consistent with other consequences of Φ .

These assumptions are not essential to second-order EP, but they make the probabilistic network implementation particularly straightforward.¹² Note that these assumptions are default rules; when determined by A1-4, the consequence relation $\models_{2o}$ is nonmonotonic.

A *credal network* can be used to represent and reason with a set of probability functions [Cozman, 2000]. A credal network consists of (i) a directed acyclic graph whose nodes are variables $A_1, \dots, A_n$ and (ii) constraints on conditional probabilities of the form $P(a_i | par_i) \in [l, u]$ where a_i is an assignment of a value to a variable and par_i is an assignment of values to its parents in the graph. It is assumed that each variable is probabilistically independent of its non-descendants conditional on its parents in the graph, written $A_i \perp\!\!\!\perp ND_i | Par_i$; this assumption is known as the *Markov Condition*.

¹²If items of evidence are known to be dependent then the corresponding nodes will be connected by arrows in the credal network representation outlined below. Any information that helps to quantify the dependence will help determine the conditional probability distributions associated with these arrows. If P is known to be distributed non-uniformly over the EP intervals then information about its distribution will need to be used to determine conditional probability distributions in the credal net.

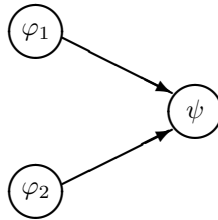
Credal networks are of fundamental importance for inference in probabilistic logic [Haenni *et al.*, 2010]. A logic is a *probabilistic logic* if its semantic interpretations are probability functions; the entailment relation of first-order EP does not constitute a probabilistic logic in this sense, but the entailment relation $\models_{2o}$ of second-order EP does. In a probabilistic logic we are typically faced with the following sort of question: given premiss propositions $\varphi_1, \dots, \varphi_n$ and their respective probabilities $X_1, \dots, X_n$, what probability should we attach to a conclusion proposition ψ ? This question can be written in the form

$$\varphi_1^{X_1}, \dots, \varphi_n^{X_n} \approx \psi?$$

where $\approx$ is the entailment relation of the probabilistic logic. For example, in second-order evidential probability we might be faced with the following question

$$\%x(Fx, Rx, [.2, .4])^{[.9, 1]}, Rt \approx_{2o} P(Ft) \in [.2, .4]?$$

This asks, given evidence that (i) the proposition that the frequency of attribute F in reference class R is between .2 and .4 has probability at least .9, and (ii) t falls in reference class R , what probability interval should attach to the proposition that the probability that t has attribute F is between .2 and .4? In first-order EP, if $1 - \delta \geq .9$ then $\text{Prob}(Ft, \Gamma_\delta) = [.2, .4]$ would be conclusively inferred (and hence treated as if it had probability 1). Clearly this disregards the uncertainty that attaches to the statistical evidence; the question is, what uncertainty should attach to the conclusion as a consequence? (This is a second-order uncertainty; hence the name *second-order* evidential probability.) One can construct a credal network to answer this question as follows. Let φ_1 be the proposition $\%x(Fx, Rx, [.2, .4])$, φ_2 be Rt and ψ be $P(Ft) \in [.2, .4]$. These can all be thought of as variables that take possible values True and False. The structure of 1oEP calculations determines the structure of the directed acyclic graph in the credal net:



The conditional probability constraints involving the premiss propositions are simply their given risk levels:

$$P(\varphi_1) \in [.9, 1],$$

$$P(\varphi_2) = 1.$$

Turning to the conditional probability constraints involving the conclusion proposition, these are determined by 1oEP inferences via axioms A1-3:

$$P(\psi | \varphi_1 \wedge \varphi_2) = 1,$$

$$P(\psi|\neg\varphi_1 \wedge \varphi_2) = P(\psi|\varphi_1 \wedge \neg\varphi_2) = P(\psi|\neg\varphi_1 \wedge \neg\varphi_2) = .2.$$

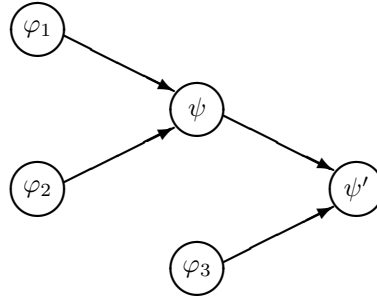
Finally, the Markov condition holds in virtue of A4, which implies that $\varphi_1 \perp\!\!\!\perp \varphi_2$. Inference algorithms for credal networks can then be used to infer the uncertainty that should attach to the conclusion, $P(\psi) \in [.92, 1]$. Hence we have:

$$\%x(Fx, Rx, [.2, .4])^{[.9, 1]}, Rt \approx_{2o} P(Ft) \in [.2, .4]^{[.92, 1]}$$

3.2.2.1 Chaining Inferences. While it is not possible to chain inferences in 1oEP, this is possible in 2oEP, and the credal network representation can just as readily be applied to this more complex case. Consider the following question:

$$\%x(Fx, Rx, [.2, .4])^{[.9, 1]}, Rt, \%x(Gx, Fx, [.2, .4])^{[.6, .7]} \approx_{2o} P(Gt) \in [0, .25]^?$$

As we have just seen, the first two premisses can be used to infer something about Ft , namely $P(Ft) \in [.2, .4]^{[.92, 1]}$. But now this inference can then be used in conjunction with the third premiss to infer something about Gt . To work out the probability bounds that should attach to an inference to $P(Gt) \in [0, .25]$, we can apply the credal network procedure. Again, the structure of the graph in the network is given by the structure of EP inferences:



Here φ_3 is $\%x(Gx, Fx, [.2, .4])$ and ψ' is $P(Gt) \in [0, .25]$; other variables are as before. The conditional probability bounds of the previous example simply carry over

$$P(\varphi_1) \in [.9, 1], P(\varphi_2) = 1,$$

$$P(\psi|\varphi_1 \wedge \varphi_2) = 1, P(\psi|\neg\varphi_1 \wedge \varphi_2) = .2 = P(\psi|\varphi_1 \wedge \neg\varphi_2) = P(\psi|\neg\varphi_1 \wedge \neg\varphi_2).$$

But we need to provide further bounds. As before, the risk level associated with the third premiss φ_3 provides one of these:

$$P(\varphi_3) \in [.6, .7],$$

and the constraints involving the new conclusion ψ' are generated by A3:

$$P(\psi'|\psi \wedge \varphi_3) = \frac{|[.2 \times .6 + .8 \times .1, .4 \times .7 + .6 \times .1] \cap [0, .25]|}{| [.2 \times .6 + .8 \times .1, .4 \times .7 + .6 \times .1] |} = .31,$$

$$P(\psi'|\neg\psi \wedge \varphi_3) = .27, P(\psi'|\psi \wedge \neg\varphi_3) = P(\psi'|\neg\psi \wedge \neg\varphi_3) = .25.$$

The Markov Condition holds in virtue of A4 and the structure of EP inferences. Performing inference in the credal network yields $P(\psi') \in [.28, .29]$. Hence

$$\%x(Fx, Rx, [.2, .4])^{[.9, 1]}, Rt, \%x(Gx, Fx, [.2, .4])^{[.6, .7]} \approx_{20} P(Gt) \in [0, .25]^{[.28, .29]}.$$

This example shows how general inference in 2oEP can be: we are not asking which probability bounds attach to a 1oEP inference in this example, but rather which probability bounds attach to an inference that cannot be drawn by 1oEP. The example also shows that the probability interval attaching to the conclusion can be narrower than intervals attaching to the premisses.

4 OBJECTIVE BAYESIAN EPISTEMOLOGY

4.1 *Motivation*

We saw above that evidential probability concerns the impact of evidence upon a conclusion. It does not on its own say how strongly one should *believe* the conclusion. Kyburg was explicit about this, arguing that evidential probabilities can at most be thought of as ‘real-valued *bounds* on degrees of belief, determined by the logical structure of our evidence’ [Kyburg, 2003, p. 147]. To determine rational degrees of belief themselves, one needs to go beyond EP, to a normative theory of partial belief.

Objective Bayesian epistemology is just such a normative theory [Rosenkrantz, 1977; Jaynes, 2003; Williamson, 2005]. According to the version of objective Bayesianism presented in [Williamson, 2005], one’s beliefs should adhere to three norms:

Probability: The strengths of one’s beliefs should be representable by probabilities. Thus they should be measurable on a scale between 0 and 1, and should be additive.

Calibration: These degrees of belief should fit one’s evidence. For example, degrees of belief should be calibrated with frequency: if all one knows about the truth of a proposition is an appropriate frequency, one should believe the proposition to the extent of that frequency.

Equivocation: One should not believe a proposition more strongly than the evidence demands. One should equivocate between the basic possibilities as far as the evidence permits.

These norms are imprecisely stated: some formalism is needed to flesh them out.

4.1.0.2 Probability. In the case of the Probability norm, the mathematical calculus of probability provides the required formalism. Of course mathematical probabilities attach to abstract events while degrees of belief attach to propositions, so the mathematical calculus needs to be tailored to apply to propositions. It is usual to proceed as follows — see, e.g., [Paris, 1994]. Given a predicate language $\mathcal{L}$ with constants t_i that pick out all the members of the domain, and sentences θ, φ of $\mathcal{L}$, a function P is a *probability function* if it satisfies the following axioms:

P1: If $\models \theta$ then $P(\theta) = 1$;

P2: If $\models \neg(\theta \wedge \varphi)$ then $P(\theta \vee \varphi) = P(\theta) + P(\varphi)$;

P3: $P(\exists x \theta(x)) = \lim_{n \rightarrow \infty} P(\bigvee_{i=1}^n \theta(t_i))$.

P1 sets the scale, P2 ensures that probability is additive, and P3, called *Gaifman's condition*, sets the probability of ' θ holds of something' to be the limit of the probability of ' θ holds of one or more of $t_1, \dots, t_n$ ', as n tends to infinity. The Probability norm then requires that the strengths of one's beliefs be representable by a probability function P over (a suitable formalisation of) one's language. Writing $\mathbb{P}$ for the set of probability functions over $\mathcal{L}$, the Probability norm requires that one's beliefs be representable by some $P \in \mathbb{P}$.

4.1.0.3 Calibration. The Calibration norm says that the strengths of one's beliefs should be appropriately constrained by one's evidence $\mathcal{E}$. (By evidence we just mean everything taken for granted in the current operating context—observations, theory, background knowledge etc.) This norm can be explicated by supposing that there is some set $\mathbb{E} \subseteq \mathbb{P}$ of probability functions that satisfy constraints imposed by evidence and that one's degrees of belief should be representable by some $P_{\mathcal{E}} \in \mathbb{E}$. Now typically one has two kinds of evidence: quantitative evidence that tells one something about physical probability (frequency, chance etc.), and qualitative evidence that tells one something about how one's beliefs should be structured. In [Williamson, 2005] it is argued that these kinds of evidence should be taken into account in the following way. First, quantitative evidence (e.g., evidence of frequencies) tells us that the physical probability function P^* must lie in some set $\mathbb{P}^*$ of probability functions. One's degrees of belief ought to be similarly constrained by evidence of physical probabilities, subject to a few provisos:

C1: $\mathbb{E} \neq \emptyset$.

If evidence is inconsistent this tells us something about our evidence rather than about physical probability, so one cannot conclude that $\mathbb{P}^* = \emptyset$ and one can hardly insist that $P_{\mathcal{E}} \in \emptyset$. Instead $\mathbb{P}^*$ must be determined by some consistency maintenance procedure—one might, for example, take $\mathbb{P}^*$ to be determined by maximal consistent subsets of one's evidence—and neither $\mathbb{P}^*$ nor $\mathbb{E}$ can ever be empty.

C2: If $\mathcal{E}$ is consistent and implies proposition θ that does not mention physical probability P^* , then $P(\theta) = 1$ for all $P \in \mathbb{E}$.

This condition merely asserts that categorical evidence be respected—it prevents $\mathbb{E}$ from being too inclusive. The qualification that θ must not mention physical probability is required because in some cases evidence of physical probability should be treated more pliantly:

C3: If $P, Q \in \mathbb{P}^*$ and $R = \lambda P + (1 - \lambda)Q$ for $\lambda \in [0, 1]$ then, other things being equal, one should be permitted to take R as one's belief function $P_{\mathcal{E}}$.

Note in particular that C3 implies that, other things being equal, if $P \in \mathbb{P}^*$ then $P \in \mathbb{E}$; it also implies C1 (under the understanding that $\mathbb{P}^* \neq \emptyset$). C3 is required to handle the following kind of scenario. Suppose for example that you have evidence just that an experiment with two possible outcomes, a and $\neg a$, has taken place. As far as you are aware, the physical probability of a is now 1 or 0 and no value in between. But this does not imply that your degree of belief in a should be 1 or 0 and no value in between—a value of $\frac{1}{2}$, for instance, is quite reasonable in this case. C3 says that, in the absence of other overriding evidence, $\langle \mathbb{P}^* \rangle \subseteq \mathbb{E}$ where $\langle \mathbb{P}^* \rangle$ is the convex hull of $\mathbb{P}^*$. The following condition imposes the converse relation:

C4: $\mathbb{E} \subseteq \langle \mathbb{P}^* \rangle$.

Suppose for example that evidence implies that either $P^*(a) = 0.91$ or $P^*(a) = 0.92$. While C3 permits any element of the interval $[0.91, 0.92]$ as a value for one's degree of belief $P_{\mathcal{E}}(a)$, C4 confines $P_{\mathcal{E}}(a)$ to this interval—indeed a value outside this interval is unwarranted by this particular evidence. Note that C4 implies C2: θ being true implies that its physical probability is 1, so $P(\theta) = 1$ for all $P \in \mathbb{P}^*$, hence for all $P \in \langle \mathbb{P}^* \rangle$, hence for all $P \in \mathbb{E}$.

In the absence of overriding evidence the conditions C1–4 set $\mathbb{E} = \langle \mathbb{P}^* \rangle$. This sheds light on how *quantitative* evidence constrains degrees of belief, but one may also have qualitative evidence which constrains degrees of belief in ways that are not mediated by physical probability. For example, one may know about causal influence relationships involving variables in one's language: this may tell one something about physical probability, but it also tells one other things—e.g., that if one extends one's language to include a new variable that is not a cause of the current variables, then that does not on its own provide any reason to change one's beliefs about the current variables. These constraints imposed by evidence of influence relationships, discussed in detail in [Williamson, 2005], motivate a further principle:

C5: $\mathbb{E} \subseteq \mathbb{S}$ where $\mathbb{S}$ is the set of probability functions satisfying structural constraints.

We will not dwell on C5 here since structural constraints are peripheral to the theme of this paper, namely to connections between objective Bayesian epistemology and evidential probability. It turns out that the set $\mathbb{S}$ is always non-empty, hence C1–5 yield:

Calibration: One's degrees of belief should be representable by $P_{\mathcal{E}} \in \mathbb{E} = \langle \mathbb{P}^* \rangle \cap \mathbb{S}$.

4.1.0.4 Equivocation. The third norm, Equivocation, can be fleshed out by requiring that $P_{\mathcal{E}}$ be a probability function, from all those that are calibrated with evidence, that is as close as possible to a totally equivocal probability function $P_{=}$ called the *equivocator* on $\mathcal{L}$. But we need to specify the equivocator and also what we mean by 'as close as possible'. To specify the equivocator, first create an ordering $a_1, a_2, \dots$ of the atomic sentences of $\mathcal{L}$ —sentences of the form Ut where U is a predicate or relation and t is a tuple of constants of corresponding arity—such that those atomic sentences involving constants $t_1, \dots, t_{n-1}$ occur earlier in the ordering than those involving t_n . Then we can define the equivocator $P_{=}$ by $P_{=}(a_j^{e_j} \mid a_1^{e_1} \wedge \dots \wedge a_{j-1}^{e_{j-1}}) = 1/2$ for all j and all $e_1, \dots, e_j \in \{0, 1\}$, where a_i^1 is just a_i and a_i^0 is $\neg a_i$. Clearly $P_{=}$ equivocates between each atomic sentence of $\mathcal{L}$ and its negation. In order to explicate 'as close as possible' to $P_{=}$ we shall appeal to the standard notion of distance between probability functions, the *n-divergence* of P from Q :

$$d_n(P, Q) \stackrel{\text{df}}{=} \sum_{e_1, \dots, e_{r_n}=0}^1 P(a_1^{e_1} \wedge \dots \wedge a_{r_n}^{e_{r_n}}) \log \frac{P(a_1^{e_1} \wedge \dots \wedge a_{r_n}^{e_{r_n}})}{Q(a_1^{e_1} \wedge \dots \wedge a_{r_n}^{e_{r_n}})}.$$

Here $a_1, \dots, a_{r_n}$ are the atomic sentences involving constants $t_1, \dots, t_n$; we follow the usual convention of taking $0 \log 0$ to be 0, and note that the *n-divergence* is not a distance function in the usual mathematical sense because it is not symmetric and does not satisfy the triangle inequality—rather, it is a measure of the amount of information that is encapsulated in P but not in Q . We then say that P is closer to the equivocator than Q if there is some N such that for $n \geq N$, $d_n(P, P_{=}) < d_n(Q, P_{=})$. Now we can state the Equivocation norm as follows. For a set $\mathbb{Q}$ of probability functions, denote by $\downarrow \mathbb{Q}$ the members of $\mathbb{Q}$ that are closest to the equivocator $P_{=}$.¹³ Then,

E1: $P_{\mathcal{E}} \in \downarrow \mathbb{E}$.

This principle is discussed at more length in [Williamson, 2008]. It can be construed as a version of the maximum entropy principle championed by Edwin Jaynes. Note that while some versions of objective Bayesianism assume that an agent's degrees of belief are uniquely determined by her evidence and language, we make no such assumption here: $\downarrow \mathbb{E}$ may not be a singleton.

¹³If there are no closest members (i.e., if chains are all infinitely descending: for any member P of $\mathbb{Q}$ there is some P' in $\mathbb{Q}$ that is closer to the equivocator than P) the context may yet determine an appropriate subset $\downarrow \mathbb{Q} \subseteq \mathbb{Q}$ of probability functions that are *sufficiently close* to the equivocator; for simplicity of exposition we shall ignore this case in what follows.



Figure 3. Constraint graph.

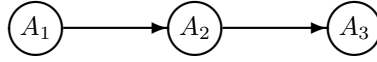


Figure 4. Graph satisfying the Markov Condition.

4.2 Calculating Objective Bayesian Degrees of Belief

Just as credal nets can be used for inference in 2oEP, so too can they be used for inference in OBE. The basic idea is to use a credal net to represent $\downarrow\mathbb{E}$, the set of rational belief functions, and then to perform inference to calculate the range of probability values these functions ascribe to some proposition of interest. These methods are explained in detail in [Williamson, 2008]; here we shall just give the gist.

For simplicity we shall describe the approach in the base case in which the evidence consists of interval bounds on the probabilities of sentences of the agent's language $\mathcal{L}$, $\mathcal{E} = \{P^*(\varphi_i) \in [l_i, u_i] : i = 1, \dots, k\}$, $\mathcal{E}$ is consistent and does not admit infinite descending chains; but these assumptions can all be relaxed. In this case $\mathbb{E} = \langle \mathbb{P}^* \rangle \cap \mathbb{S} = \mathbb{P}^*$. Moreover, the evidence can be written in the language $\mathcal{L}^\sharp$ introduced earlier: $\mathcal{E} = \{\varphi_1^{[l_1, u_1]}, \dots, \varphi_k^{[l_k, u_k]}\}$, and the question facing objective Bayesian epistemology takes the form

$$\varphi_1^{[l_1, u_1]}, \dots, \varphi_k^{[l_k, u_k]} \vDash_{\text{OBE}} \psi?$$

where $\vDash_{\text{OBE}}$ is the entailment relation defined by objective Bayesian epistemology as outlined above. As explained in [Williamson, 2008], this entailment relation is nonmonotonic but it is well-behaved in the sense that it satisfies all the System-P properties of nonmonotonic logic.

The method is essentially this. First construct an undirected graph, the *constraint graph*, by linking with an edge those atomic sentences that appear in the same item of evidence. One can read off this graph a list of probabilistic independencies that any function in $\downarrow\mathbb{E}$ must satisfy: if node A separates nodes B and C in this graph then $B \perp\!\!\!\perp C \mid A$ for each probability function in $\downarrow\mathbb{E}$. This constraint graph can then be transformed into a directed acyclic graph for which the Markov Condition captures many or all of these independencies. Finally one can calculate bounds on the probability of each node conditional on its parents in the graph by using entropy maximisation methods: each probability function in $\downarrow\mathbb{E}$ maximises entropy subject to the constraints imposed by $\mathcal{E}$, and one can identify the probability it gives to one variable conditional on its parents using numerical optimisation methods [Williamson, 2008].

To take a simple example, suppose we have the following question:

$$\forall x(Ux \rightarrow Vx)^{3/5}, \forall x(Vx \rightarrow Wx)^{3/4}, Ut_1^{[0.8,1]} \approx_{\text{OBE}} Wt_1^?$$

A credal net can be constructed to answer this question. There is only one constant symbol t_1 , and so the atomic sentences of interest are Ut_1, Vt_1, Wt_1 . Let A_1 be Ut_1 , A_2 be Vt_1 and A_3 be Wt_1 . Then the constraint graph $\mathcal{G}$ is depicted in Fig. 3 and the corresponding directed acyclic graph $\mathcal{H}$ is depicted in Fig. 4. It is not hard to see that $P(A_1) = 4/5, P(A_2|A_1) = 3/4, P(A_2|\neg A_1) = 1/2, P(A_3|A_2) = 5/6, P(A_3|\neg A_2) = 1/2$; together with $\mathcal{H}$, these probabilities yield a credal network. (In fact, since the conditional probabilities are precisely determined rather than bounded, we have a special case of a credal net called a *Bayesian net*.) The Markov Condition holds since separation in the constraint graph implies probabilistic independence. Standard inference methods then give us $P(A_3) = 11/15$ as an answer to our question.

5 EP-CALIBRATED OBJECTIVE BAYESIANISM

5.1 Motivation

At face value, evidential probability and objective Bayesian epistemology are very different theories. The former concerns the impact of evidence of physical probability, Carnap's probability₂, and concerns acceptance and rejection; it appeals to interval-valued probabilities. The latter theory concerns rational degree of belief, probability₁, and invokes the usual point-valued mathematical notion of probability. Nevertheless the core of these two theories can be reconciled, by appealing to second-order EP as developed above.

2oEP concerns the impact of evidence on rational degree of belief. Given statistical evidence, 2oEP will infer statements about rational degrees of belief. These statements can be viewed as constraints that should be satisfied by the degrees of belief of a rational agent with just that evidence. So 2oEP can be thought of as mapping statistical evidence $\mathcal{E}$ to a set $\mathbb{E}$ of rational belief functions that are compatible with that evidence. (This is a non-trivial mapping because frequencies attach to a sequence of outcomes or experimental conditions that admit repeated instantiations, while degrees of belief attach to propositions. Hence the epistemological reference-class problem arises: how can one determine appropriate single-case probabilities from information about generic probabilities? Evidential probability is a theory that tackles this reference-class problem head on: it determines a probability interval that attaches to a sentence from statistical evidence about repetitions.)

But this mapping from $\mathcal{E}$ to $\mathbb{E}$ is just what is required by the Calibration norm of OBE. We saw in §4 that OBE maps evidence $\mathcal{E}$ to $\mathbb{E} = \langle \mathbb{P}^* \rangle \cap \mathbb{S}$, a set of probability functions calibrated with that evidence. But no precise details were given as to how $\langle \mathbb{P}^* \rangle$, nor indeed $\mathbb{P}^*$, is to be determined. In special cases this is

straightforward. For example, if one's evidence is just that the chance of a is $\frac{1}{2}$, $P^*(a) = 1/2$, then $\langle \mathbb{P}^* \rangle = \mathbb{P}^* = \{P \in \mathbb{P} : P(a) = 1/2\}$. But in general, determining $\langle \mathbb{P}^* \rangle$ is not a trivial enterprise. In particular, statistical evidence takes the form of information about generic frequencies rather than single-case chances, and so the reference-class problem arises. It is here that 2oEP can be plugged in: if $\mathcal{E}$ consists of propositions of $\mathcal{L}^\#$ —i.e., propositions, including statistical propositions, to which probabilities or closed intervals of probabilities attach—then $\langle \mathbb{P}^* \rangle$ is the set of probability functions that satisfy the $\approx_{2o}$ consequences of $\mathcal{E}$.

C6: If $\mathcal{E}$ is a consistent set of propositions of $\mathcal{L}^\#$ then $\langle \mathbb{P}^* \rangle = \{P : P(\chi) \in [l, u]$ for all χ, l, u such that $\mathcal{E} \approx_{2o} \chi^{[l, u]}$.

We shall call OBE that appeals to calibration principles C1–6 *epistemic-probability-calibrated objective Bayesian epistemology*, or EP-OBE for short. We shall denote the corresponding entailment relation by $\approx_{\text{OBE}}^{\text{EP}}$.

We see then that there is a sense in which EP and OBE can be viewed as complementary rather than in opposition. Of course, this isn't the end of the matter. Questions still arise as to whether EP-OBE is the right way to flesh out OBE. One can, for instance, debate the particular rules that EP uses to handle reference classes (§2.2). One can also ask whether EP tells us everything we need to know about calibration. As mentioned in §4.1, further rules are needed in order to handle structural evidence, fleshing out C5. Moreover, both 1oEP and 2oEP take statistical statements as input; these statements themselves need to be inferred from particular facts—indeed EP, OBE and EP-OBE each presume a certain amount of statistical inference. Consequently we take it as understood that Calibration requires more than just C1–6.

And questions arise as to whether the alterations to EP that are necessary to render it compatible with OBE are computationally practical. Second-order EP replaces the original theory of acceptance with a decision theoretic account which will incur a computational burden. Moreover, some thought must be given as to which consistency maintenance procedure should be employed in practice. Having said this, we conjecture that there will be real inference problems for which the benefits will be worth the necessary extra work.

5.2 Calculating EP-Calibrated Objective Bayesian Probabilities

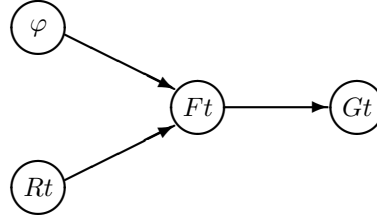
Calculating EP-OBE probabilities can be achieved by combining methods for calculating 2oEP probabilities with methods for calculating OBE probabilities. Since credal nets can be applied to both formalisms independently, they can also be applied to their unification. In fact in order to apply the credal net method to OBE, some means is needed of converting statistical statements, which can be construed as constraints involving generic, repeatably-instantiatable variables, to constraints involving the single-case variables which constitute the nodes of the objective Bayesian credal net; only then can the constraint graph of §4.2 be constructed. The 2oEP credal nets of §3.2 allow one to do this, since this kind

of net incorporates both statistical variables and single-case variables as nodes. Thus 2oEP credal nets are employed first to generate single-case constraints, at which stage the OBE credal net can be constructed to perform inference. This fits with the general view of 2oEP as a theory of how evidence constrains rational degrees of belief and OBE as a theory of how further considerations—especially equivocation—further constrain rational degrees of belief.

Consider the following very simple example:

$$\%x(Fx, Rx, [.2, .5]), Rt, \forall x(Fx \rightarrow Gx)^{3/4} \approx_{\text{OBE}}^{\text{EP}} Gt?$$

Now the first two premisses yield $Ft^{[.2, .5]}$ by EP. This constraint combines with the third premiss to yield an answer to the above question by appealing to OBE. This answer can be calculated by constructing the following credal net:



Here φ is the first premiss. The left-hand side of this net is the 2oEP net, with associated probability constraints

$$P(\varphi) = 1, P(Rt) = 1,$$

$$P(Ft|\varphi \wedge Rt) \in [.2, .5], P(Ft|\neg\varphi \wedge Rt) = 0 = P(Ft|\varphi \wedge \neg Rt) = P(Ft|\neg\varphi \wedge \neg Rt).$$

The right-hand side of this net is the OBE net with associated probabilities

$$P(Gt|Ft) = 7/10, P(Gt|\neg Ft) = 1/2.$$

Standard inference algorithms then yield an answer of $7/12$ to our question:

$$\%x(Fx, Rx, [.2, .5]), Rt, \forall x(Fx \rightarrow Gx)^{3/4} \approx_{\text{OBE}}^{\text{EP}} Gt^{7/12}$$

6 CONCLUSION

While evidential probability and objective Bayesian epistemology might at first sight appear to be chalk and cheese, on closer inspection we have seen that their relationship is more like horse and carriage—together they do a lot of work, covering the interface between statistical inference and normative epistemology.

Along the way we have taken in an interesting array of theories—first-order evidential probability, second-order evidential probability, objective Bayesian epistemology and EP-calibrated OBE—that can be thought of as nonmonotonic logics.

2oEP and OBE are probabilistic logics in the sense that they appeal to the usual mathematical notion of probability. More precisely, their entailment relations are probabilistic: premisses entail the conclusion if every model of the premisses satisfies the conclusion, where models are probability functions. This connection with probability means that credal networks can be applied as inference machinery. Credal nets yield a perspicuous representation and the prospect of more efficient inference [Haenni *et al.*, 2010].

ACKNOWLEDGEMENTS

We are grateful to the Leverhulme Trust for supporting this research, and to Prasanta S. Bandyopadhyay, Teddy Seidenfeld and an anonymous referee for helpful comments.

BIBLIOGRAPHY

- [Andréka *et al.*, 1998] H. Andréka, J. van Benthem, and I. Németi. Modal languages and bounded fragments of predicate logic. *Journal of Philosophical Logic*, 27:217–274, 1998.
- [Arló-Costa and Parikh, 2005] H. Arló-Costa and R. Parikh. Conditional probability and defeasible inference. *Journal of Philosophical Logic*, 34:97–119, 2005.
- [Carnap, 1962] R. Carnap. *The Logical Foundations of Probability*. University of Chicago Press, 2nd edition, 1962.
- [Cozman, 2000] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000.
- [de Finetti, 1974] B. de Finetti. *Theory of Probability: A critical introductory treatment*. Wiley, 1974.
- [Dubbins, 1975] L. E. Dubbins. Finitely additive conditional probability, conglomerability, and disintegrations. *Annals of Probability*, 3:89–99, 1975.
- [Dubois and Prade, 1980] D. Dubois and H. Prade. *Fuzzy Sets and Systems: Theory and Applications*. Kluwer, North Holland, 1980.
- [Haenni *et al.*, 2010] R. Haenni, J.-W. Romeijn, G. Wheeler, and J. Williamson. *Probabilistic Logic and Probabilistic Networks*. The Synthese Library, Springer, 2010.
- [Harper and Wheeler, 2007] W. Harper and G. Wheeler, editors. *Probability and Inference: Essays In Honor of Henry E. Kyburg, Jr.* King’s College Publications, London, 2007.
- [Hawthorne and Makinson, 2007] J. Hawthorne and D. Makinson. The quantitative/qualitative watershed for rules of uncertain inference. *Studia Logica*, 86(2):247–297, 2007.
- [Jaynes, 2003] E. T. Jaynes. *Probability theory: the logic of science*. Cambridge University Press, Cambridge, 2003.
- [Keynes, 1921] J. M. Keynes. *A Treatise on Probability*. Macmillan, London, 1921.
- [Kraus *et al.*, 1990] S. Kraus, D. Lehman, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.
- [Kyburg and Teng, 2001] H. E. Kyburg, Jr. and C. M. Teng. *Uncertain Inference*. Cambridge University Press, Cambridge, 2001.
- [Kyburg *et al.*, 2007] H. E. Kyburg, Jr., C. M. Teng, and G. Wheeler. Conditionals and consequences. *Journal of Applied Logic*, 5(4):638–650, 2007.
- [Kyburg, 2003] H. E. Kyburg Jr. Are there degrees of belief? *Journal of Applied Logic*, 1:139–149, 2003.
- [Kyburg, 1961] H. E. Kyburg, Jr. *Probability and the Logic of Rational Belief*. Wesleyan University Press, Middletown, CT, 1961.

- [Kyburg, 1990] H. E. Kyburg, Jr. *Science and Reason*. Oxford University Press, New York, 1990.
- [Kyburg, 2007] H. E. Kyburg, Jr. Bayesian inference with evidential probability. In William Harper and Gregory Wheeler, editors, *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr.*, pages 281–296. King’s College, London, 2007.
- [Levi, 1977] I. Levi. Direct inference. *Journal of Philosophy*, 74:5–29, 1977.
- [Levi, 1980] I. Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, MA, 1980.
- [Levi, 2007] I. Levi. Probability logic and logical probability. In William Harper and Gregory Wheeler, editors, *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr.*, pages 255–266. College Publications, 2007.
- [Paris, 1994] J. B. Paris. *The uncertain reasoner’s companion*. Cambridge University Press, Cambridge, 1994.
- [Pawlak, 1991] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer, Dordrecht, 1991.
- [Rosenkrantz, 1977] R. D. Rosenkrantz. *Inference, method and decision: towards a Bayesian philosophy of science*. Reidel, Dordrecht, 1977.
- [Seidenfeld, 2007] T. Seidenfeld. Forbidden fruit: When Epistemic Probability may not take a bite of the Bayesian apple. In William Harper and Gregory Wheeler, editors, *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr.* King’s College Publications, London, 2007.
- [Teng, 2007] C. M. Teng. Conflict and consistency. In William L. Harper and Gregory Wheeler, editors, *Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr.* King’s College Publications, London, 2007.
- [Walley, 1991] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
- [Wheeler, 2004] G. Wheeler. A resource bounded default logic. In James Delgrande and Torsten Schaub, editors, *NMR 2004*, pages 416–422, 2004.
- [Wheeler, 2007] G. Wheeler. Two puzzles concerning measures of uncertainty and the positive Boolean connectives. In José Neves, Manuel Santos, and José Machado, editors, *Progress in Artificial Intelligence, 13th Portuguese Conference on Artificial Intelligence*, LNAI 4874, pages 170–180, Berlin, 2007. Springer-Verlag.
- [Williamson, 2005] J. Williamson. *Bayesian nets and causality: philosophical and computational foundations*. Oxford University Press, Oxford, 2005.
- [Williamson, 2008] J. Williamson. Objective Bayesian probabilistic logic. *Journal of Algorithms in Cognition, Informatics and Logic*, 63:167–183, 2008.
- [Williamson, 2009] J. Williamson. Aggregating judgements by merging evidence. *Journal of Logic and Computation*, 19(3): 461–473, 2009.

This page intentionally left blank

CONFIRMATION THEORY

James Hawthorne

INTRODUCTION

Confirmation theory is the study of the logic by which scientific hypotheses may be confirmed or disconfirmed (or supported or refuted) by evidence. A specific theory of confirmation is a proposal for such a logic. Presumably the epistemic evaluation of scientific hypotheses should largely depend on their empirical content — on what they *say* the evidentially accessible parts of the world are like, and on the extent to which they turn out to be right about that. Thus, all theories of confirmation rely on measures of how well various alternative hypotheses account for the evidence.¹ Most contemporary confirmation theories employ probability functions to provide such a measure. They measure how well the evidence fits what the hypothesis *says* about the world in terms of *how likely* it is that the evidence would occur if the hypothesis were true. Such hypothesis-based probabilities of evidence claims are called *likelihoods*. Clearly, when the evidence is more likely according to one hypothesis than according to an alternative, that should redound to the credit of the former hypothesis and the discredit of the later. But various theories of confirmation diverge on precisely how this credit is to be measured.

A natural approach is to also employ a probabilistic measure to directly represent the degree to which the hypothesis is confirmed or disconfirmed on the evidence. The idea is to rate the degree to which a hypothesis is confirmed on a scale from 0 to 1, where tautologies are always assigned maximal confirmation (degree 1), and where the degree of confirmation of the disjunction of mutually incompatible hypotheses sum to the degrees of confirmation of each taken separately. This way of rating confirmation just recapitulates the standard axioms

¹I make no distinction here between scientific hypotheses and scientific theories. For our purposes a scientific theory is just a large, complex hypothesis. We will suppose (as a formal logic must) that scientific hypotheses are expressible as sentences of a language — e.g. a mathematically sophisticated dialect of English. This supposition need not be in opposition to a *semantic view of theories*. Presumably, if scientists can express a theory well enough to agree about what it says about the world (or at least about its testable empirical content), it must be expressible in some bit of language. For instance, if a theory is taken to be a set of models, presumably that set of models may be described in mathematical English, perhaps drawing on set theory. In that case the empirical content of the theory will consist of various hypotheses about what parts of the world are modeled (to what degree of approximation) by some particular model or set of models. Such hypotheses should be subject to empirical evaluation. So a theory of confirmation should apply to them.

of probability theory, but applies them as a measure of evidential support. Any theory of confirmation that employs such a measure is a *probabilistic confirmation theory*. However, confirmation functions of this sort will be of little value unless it can be shown that under reasonable conditions the accumulation of evidence tends to drive the confirmation values given by these functions towards 0 for false hypotheses and towards 1 for true hypotheses.

How should confirmation values be related to what hypotheses imply about evidence claims via the likelihoods? The most straightforward idea would be to have the confirmation function assign to a hypothesis whatever numerical value is had by the *likelihood* the hypothesis assigns to the evidence. However, this idea won't work. For one thing, in cases where the hypothesis logically entails the evidence, the likelihood is 1. But we cannot require the confirmational probability of each hypothesis that entails the evidence to be 1. For, several alternative hypotheses may entail the evidence, and a probabilistic confirmation measure cannot assign probability 1 to two (or more) alternative hypotheses based on the same evidence.²

If both *likelihoods* and the *degrees of confirmation* for hypotheses are to be measured probabilistically, it seems natural to represent both by a common probability function. In that case what relationship does the *degree to which a hypothesis is confirmed on evidence* have to the *likelihood of the evidence according to the hypothesis*? This is where Bayes' Theorem comes into play. Bayes' Theorem follows from the standard axioms for probabilities; and it explicitly shows how the probabilistic degree of confirmation for a hypothesis depends on the likelihoods of evidence claims. Thus, any confirmation measure that satisfies the standard axioms of probability theory and employs the same probability function to represent likelihoods will have to be a *Bayesian confirmation measure*. Any theory of confirmation that employs such a probabilistic measure of confirmation will thus be a *Bayesian confirmation theory*.

Various Bayesian approaches to confirmation primarily diverge with regard to how they understand the concept of probabilistic confirmation — i.e. with regard to how they interpret the notion of *probability* that is supposed to be captured by confirmational probability functions. Is the confirmation function supposed to represent the warranted *degree of confidence* or *degree of belief* an agent should have in hypotheses (based on the evidence), as the subjectivists and personalists would have it? Or, is the confirmation function some kind of *objective logical relationship* by which the evidence is supposed to *probabilistically entail* hypotheses, as some logical objectivists maintain? Or, might the confirmation function represent some other coherent conception of evidential support?

These days the most prominent strain of Bayesian confirmation theory takes probabilistic confirmation functions to represent the rationally ideal agent's subjective *degrees of confidence* or *degrees of belief* in the truth of statements or propositions. This view has become so influential that the term 'Bayesian confir-

²The same argument applies whenever two alternative hypotheses assign the evidence a likelihood greater than $1/2$ — and similarly, whenever n alternative hypotheses each assign the evidence likelihoods greater than $1/n$.

mation theory' is often taken to be synonymous with it. But to identify Bayesian confirmation theory with the subjectivist view is a mistake. It either mischaracterizes or entirely disregards a host of non-subjectivist Bayesian accounts. Properly speaking, it is not the subjective interpretation of the confirmation function that makes a confirmation theory *Bayesian*. Rather, any confirmation theory that gives Bayes' Theorem a central role in the representation of how a hypothesis is confirmed or refuted by evidence is a *Bayesian confirmation theory*. And any account that employs the same probability function to represent the hypothesis-based *likelihoods* of evidence claims and to provide a probabilistic measure of the *support strength* for those hypotheses by that evidence must give Bayes' Theorem this central role.

Historically a number of proposals for objective Bayesian confirmation theories have been developed — e.g., by Keynes [1921], Carnap [1950; 1952; 1971; 1980], Jeffreys [1939], Jaynes [1968]. More recently, Rosenkrantz [1981], Maher [1996; 2006], Fitelson [2005], and Williamson [2007] defend objectivist Bayesian accounts of confirmation. The founding proponents of subjectivist Bayesianism include Ramsey [1926], de Finetti [1937], and Savage [1954]. Prominent contemporary subjectivist treatments include those of Jeffrey [1965; 1992], Levi [1967; 1980], Lewis [1980], Skyrms [1984], Earman [1992], Howson and Urbach [1993], Howson [1997] and Joyce [1999; 2003; 2004; 2005]. Although all these subjectivist Bayesian approaches draw on a belief-strength notion of probability, there remain important differences among them. For instance, all agree that belief strengths should be *probabilistically coherent* (i.e., consistent with the probability axioms), but some maintain that belief strengths should also satisfy additional epistemic constraints.

In this article I will explicate the probabilistic logic underlying most probabilistic accounts of evidential support. I'll also discuss some difficulties faced by objectivist and subjectivist readings of confirmation functions, and suggest a conception that I think overcomes the interpretative problems. Here is how I'll proceed.

In section 1 I will discuss the most basic features of this probabilistic logic. I'll set down axioms that characterize these functions, but without yet describing how the logic is supposed to apply to the evaluation of scientific hypotheses.

Section 2 will briefly describe two accounts of the nature of confirmation functions — views about what the confirmation functions are supposed to be, or what they are supposed to represent. We won't be able to delve too far into this issue until after we see how the logic of confirmation functions represents the evidential support of hypotheses, which I'll address in section 3. But the logic of confirmation described in section 3 will be more easily comprehended if we have some preliminary idea about what the confirmation functions represent.

Section 3 shows how the logic of confirmation functions represents the evidential support of hypotheses. I will spell out several forms of Bayes' Theorem, showing how the Bayesian formulae represent the role of the likelihoods and another important factor, *prior probabilities* of hypotheses, in the logic of evidential support.

In section 4 I will return to the issue of what confirmation functions are supposed

to be. I'll describe major problems with the two most prominent interpretations, and suggest an alternative view. In particular I'll address the issue of how confirmation functions are supposed to inform our beliefs about the truth or falsity of hypotheses.

Is there any good reason to think that on a suitable amount of evidence, false hypothesis will become strongly disconfirmed and true hypotheses will become highly confirm? In section 5 I will explicate a *Bayesian Convergence Theorem* that provides such assurances. It shows that under reasonable conditions the accumulation of evidence should result in the near refutation of false hypotheses — in confirmational probability approaching 0 — and should lead to a high degree of confirmation for the true alternative — to confirmational probability near 1.³

The discussion throughout sections 3 through 5 will suppose that the likelihoods for evidence statements on the various hypotheses, which express the empirical content of hypotheses, are either objectively determinate or intersubjectively agreed to by the relevant scientific community — so all confirmation functions employed by a scientific community will agree on the numerical values of the likelihoods. In section 6 I will weaken this supposition, but argue that the important truth-acquiring features of probabilistic confirmation theory are nevertheless maintained.

1 THE PROBABILISTIC LOGIC OF CONFIRMATION FUNCTIONS

Confirmation functions represent a logic by which scientific theories are refuted or supported by evidence. So they should be defined on an *object language* rich enough to fully express any scientific theory. A language with the expressive strength of first-order predicate logic should suffice.⁴ All current scientific theories are expressible in such a language. Notation for the standard logical connectives

³This *Bayesian Convergence Theorem* avoids many of the usual objections to such theorems. It depends only on *likelihoods*; so those suspicious of *prior probabilities* may still find it edifying. And this theorem provides explicit bounds on the likely rate of convergence.

⁴Everything I'll say here also applies to languages for second-order and higher-order logics. Some confirmation theorists take confirmation functions to apply directly to "propositions" rather than to sentences of a language, where propositions are supposed to be sets of possible worlds, or some other sort of non-linguistic entities. I prefer to define confirmation functions directly on a language. There are a number of reasons for preferring this approach. Among them are: (1) it avoids the sticky metaphysical issue of what a proposition is (indeed, on many accounts, all logical truths represent the same proposition — so a logic that directly employs such propositions may hide important logical structure); (2) it makes the logic more directly comparable to deductive logics; (3) although there is widespread agreement about how formal languages and their logics work, this may not be true for propositions; (4) all real scientific theories are expressed in language (including the language of set theory and mathematics), so if the results of a propositional treatment of confirmation are to be applicable to real scientific theories, they will have to be translated into results that apply to expressions in a language; (4) to the extent that the propositional approach is richer, and permits the proof of stronger results than the sentential approach, it isn't at all clear that these stronger results are relevant to the confirmation of real scientific theories, which are always expressed in a language. So, when important results are stated and proved in terms of propositions, one may be left wondering what sort assumptions about the nature of propositions play a crucial role, and whether these results can really be translated into results about the confirmation of real scientific theories.

and quantifiers for this language are as follows: “not”, ‘ $\sim$ ’; “and”, ‘ $\cdot$ ’; “or”, ‘ $\vee$ ’; truth-functional “if-then”, ‘ $\supset$ ’; “if and only if”, ‘ $\equiv$ ’; the quantifiers “all”, ‘ $\forall$ ’, and “some”, ‘ $\exists$ ’; and the identity relation, “is the same thing as”, ‘ $=$ ’. These are the *logical terms* of the language. Standard deductive logic depends only on the meanings of these logical terms. The other terms of a language (e.g., names and predicate and relation expressions) are called *non-logical terms*. The central logical notion of standard deductive logic, the *logical entailment relation*, neither depends on their meanings nor on the actual truth-values of sentences composed of these terms. It only supposes that the non-logical terms are meaningful, and that sentences containing them must be either true or false.⁵

A *degree of confirmation function* represents a relationship between statements (i.e. declarative sentences) that is somewhat analogous to the deductive *logical entailment relation*. However, the logic of confirmation will have to deviate from the deductive paradigm in several significant ways. For one thing, deductive logical entailment is an absolute relationship between sentences, while confirmation comes in degrees. Deductive logical entailment is monotonic: when B *logically entails* A , adding a premise C cannot undermine that logical entailment — i.e., $(C \cdot B)$ must *logically entail* A as well. But *confirmation* is *nonmonotonic*. Adding a new premise C to B may substantially raise the degree to which A is confirmed, or may substantially lower it, or may leave it unchanged — i.e., for a confirmation function P_α , the value of $P_\alpha[A|C \cdot B]$ may be much larger than $P_\alpha[A|B]$ (for some statements C), while it may be much smaller (for some other C), and it may have the same value, or nearly the same value (for yet other statements C).

Another very significant difference is this. A given deductive logic specifies a unique *logical entailment relation*, and that relation depends *only* on the meanings of the logical connectives and quantifiers. Is there, similarly, a uniquely good confirmation function? And does it depend only on the meanings of logical terms, and not on the specifics of what the individual sentences mean (due to the meanings of the names and predicates they contain)? Most confirmation theorists would answer “no” to both questions. (I’ll discuss some reasons for this later.) Rather, from the logician’s point of view, confirmation functions are technically somewhat analogous to *truth-value assignments* to sentences of a formal language. That is, holding the meanings of the logical terms (connectives and quantifiers) fixed, there are lots of ways to assign meanings to names and predicate terms of a language; and for each such meaning assignment, each of the various ways the world might turn out to be assigns a corresponding truth-value assignment to each sentence of the language. Similarly, keeping the meanings of logical terms fixed, each way of

⁵Any language with this syntax permits the expression of set theory and all of mathematics, so it should be sufficiently rich to express all statements made by real scientific theories. But if you doubt the adequacy of this language for the expression of real scientific theories, then please think of the logic of confirmation presented here as a logician’s model of the real thing. Viewed this way, the logician’s model is, like any model employed by the sciences, an approximation that is intended to capture the essential features of its subject matter, which in this case is the nature of correct scientific inferences concerning hypotheses expressible in a scientific dialect of a natural language such as English.

assigning meanings to names and predicate expressions may give rise to a distinct confirmation function; and more than one confirmation function may apply to each fully meaningful language. So there are many possible confirmation functions.

In deductive logic the possible truth-value assignments to sentences of a formal language L are constrained by certain *semantic rules*, which are axioms regarding the meanings of the logical terms ('not', 'and', 'or', etc., the quantifiers, and the identity relation). The rules, or axioms, for confirmation functions play a similar role. They constrain each member of the family of possible confirmation functions, $\{P_\beta, P_\gamma, \dots, P_\delta, \dots\}$, to respect the meanings of the logical terms, but without regard for what the other terms of the language may mean. Although each confirmation function satisfies the same basic axioms, the further issue of which among them provides an appropriate measure of confirmation is not settled by these axioms alone. It presumably depends on additional factors, including the meanings of the non-logical terms of the language. Here are the semantic rules (or axioms) that constrain probabilistic confirmation functions.⁶

Let L be a language for predicate logic with identity, and let ' $\models$ ' be the standard logical entailment relation (where ' $B \models A$ ' abbreviates ' B logically entails A ', and ' $\models A$ ' abbreviates ' A is a logical truth'). A confirmation function is a function P_α from pairs of sentences of L to real numbers between 0 and 1 that satisfies the following rules:

1. $P_\alpha[D|E] < 1$ for at least one pair of sentences D and E .

For all sentence A, B, C ,

2. if $B \models A$, then $P_\alpha[A|B] = 1$;
3. If $\models (B \equiv C)$, then $P_\alpha[A|B] = P_\alpha[A|C]$;
4. If $C \models \sim(A \cdot B)$, then either $P_\alpha[(A \vee B)|C] = P_\alpha[A|C] + P_\alpha[B|C]$,
or for every sentence D , $P_\alpha[D|C] = 1$;
5. $P_\alpha[(A \cdot B)|C] = P_\alpha[A|(B \cdot C)] \times P_\alpha[B|C]$.

This axiomatization takes conditional probability as basic. The conditional probability functions it characterizes are identical to the usual unconditional probability functions when the latter are defined: just let $P_\alpha[A] = P_\alpha[A|(D \vee \sim D)]$. But, these axioms permit conditional probabilities $P_\alpha[A|C]$ to remain defined even when a condition statement C has probability 0 (e.g., even when $P_\alpha[C|(D \vee \sim D)] = 0$). Though useful, this feature is not the primary reason for taking conditional probability to be basic when specifying confirmation functions.

⁶The language L in which scientific hypotheses and evidence claims are expressed is what logicians call the *object-language*. As logicians see it, confirmation functions are not themselves part of the object-language L . Rather, we take confirmation functions, along with other semantic notions like *truth* and *logical entailment*, to reside in the metalanguage, which is the language where properties of object-language expressions are treated. Logicians distinguish object-language from metalanguage in order to avoid certain kinds of paradoxes and logical incoherencies that can arise from applying semantic concepts to themselves.

The primary reason is this. On the usual account, where unconditional probability is basic, conditional probability is defined as follows: $P[A|B] = P[(A \cdot B)]/P[B]$ if $P[B] > 0$, and is undefined otherwise. (This is closely related to axiom 5.) But if one takes conditional probability as defined in this way, the likelihood a hypothesis H assigns to some evidence statement E under experimental or observation conditions C , $P[E|H \cdot C]$, must be *defined* as follows: $P[E|H \cdot C] = P[E \cdot H \cdot C]/P[H \cdot C]$. However, in the context of confirmation functions it seems really unnatural to take such likelihoods as *defined* like this. Likelihoods often have very well-defined, well-known values all on their own, whereas the values of the probabilities in the numerator and denominator of ' $P[E \cdot H \cdot C]/P[H \cdot C]$ ' are often unknown or only vaguely specified.

For example, let H say "the coin is fair" (i.e. it has a propensity to come up *heads* half the time when tossed in the usual way), let C say "the coin is tossed at present in the usual way", and let E say "the coin lands *heads* on the present toss". We may not be at all clear about the values of $P[E \cdot H \cdot C]$ (the probability that "the coin is fair, is presently tossed in the usual way, and lands *heads*") or of $P[H \cdot C]$ (the probability that "the coin is fair, and is presently tossed in the usual way"). Nevertheless, the value of $P[E|H \cdot C]$ seems to be perfectly well-defined and well-known (it should clearly equal $1/2$), due to what $H \cdot C$ says about E .⁷ Thus, because confirmation functions are supposed to represent such relationships between hypotheses and evidence statements, and because such *likelihoods* are often better defined and better known than the probabilities that would define them via the usual ratio definition, it seems more natural to axiomatize confirmational probabilities in a way that takes conditional probabilities as basic.⁸

One important context where the present approach is especially useful is this. Consider a statistical hypothesis that says that the chance (or measure) of an attribute X among systems in a state Y is 1. Formally, we might express such a hypothesis this way: ' $Ch(X, Y) = 1$ '. Suppose it's known that a physical system g is in state Y (i.e., $g \in Y$). This gives rise to a likelihood of the following form: $P_\alpha[g \in X | Ch(X, Y) = 1 \cdot g \in Y] = 1$. Adding additional information to the premise may decrease this likelihood — e.g., $P_\alpha[g \in X | Ch(X, Y) = 1 \cdot g \in Y \cdot \sim g \in X] = 0$. The present approach to confirmation functions permits this, whereas on the usual approach to conditional probability this "lowering of the likelihood" cannot happen when the likelihood has value 1.⁹

⁷The idea that confirmation functions should agree on the values of such likelihoods is an additional supposition, not captured by the rules (axioms) for confirmational probabilities given above. I'll say more about the nature of these *direct inference likelihoods* a bit later.

⁸See [Hájek, 2003a] for more reasons to take conditional probability as basic. Although an account of confirmation functions that takes conditional probability as basic makes good conceptual sense, the account of Bayesian confirmation in this article will not rely on this approach in any crucial way. The more usual axiomatization will suffice for most purposes. It goes like this: a confirmation function is a function P_α from sentences of L to real numbers between 0 and 1 that satisfies the following rules: (1) if $\models A$, then $P_\alpha[A] = 1$; (2) if $\models \sim(A \cdot B)$, then $P_\alpha[A \vee B] = P_\alpha[A] + P_\alpha[B]$; by definition $P_\alpha[A|B] = P_\alpha[A \cdot B]/P_\alpha[B]$ whenever $P_\alpha[B] > 0$.

⁹For, if by definition $P_\alpha[g \in X | Ch(X, Y) = r \cdot g \in Y] = P_\alpha[g \in X \cdot Ch(X, Y) = 1 \cdot g \in Y]$, $P_\alpha[Ch(X, Y) = r \cdot g \in Y]$, then whenever $P_\alpha[g \in X | Ch(X, Y) = 1 \cdot g \in Y] = 1$ we must have

Let us now briefly consider each axiom, to see what sort of constraint it places on a measure of confirmation. First, notice that adopting a scale between 0 and 1 is merely a convenience. This scale is usual for probabilities, but another scale might do as well.

Rule (1) is a non-triviality requirement. It says that some sentences must be supported by others to degree less than 1. We might instead have required that $P_\alpha[(A \cdot \sim A)|(A \vee \sim A)] < 1$; but this turns out to be derivable from Rule (1) together with the other rules.

Rule (2) says that if B logically entails A , then B must maximally confirm A . This makes each probabilistic confirmation function a kind of generalization of the deductive logical entailment relation. Rule (3) adds the plausible constraint that whenever statements B and C are logically equivalent, each must provide precisely the same confirmational support to all other statements.

Rule (4) says that confirmational support adds up in a plausible way. When C logically entails the incompatibility of A and B , the support C provides each separately must sum to the support it provides for their disjunction. The only exception is the case where C acts like a contradiction and supports every sentence to degree 1.

To understand what Rule (5) says, think of a confirmation function P_α as describing a measure of the proportion of possible states of affairs (or possible worlds) in which statements are true: an expression of form ' $P_\alpha[C|D] = r$ ' says that the proportion of states in which C is true among those where D is true is r . On this reading, Rule (5) says the following: suppose that B (together with C) is true in proportion q of all states where C is true; and suppose that A is true in fraction r of all those states where B and C are true together; then A and B (and C) will be true in fraction r of the proportion q (i.e. in $r \times q$) of all states where C is true.

All of the usual theorems of probability theory are easily derived from these axioms. For example, logically equivalent sentences are always supported to the same degree: if $C \models (B \equiv A)$, then $P_\alpha[A|C] = P_\alpha[B|C]$. And the following generalization of the Addition Rule (4) is derivable: $P_\alpha[(A \vee B)|C] = P_\alpha[A|C] + P_\alpha[B|C] - P_\alpha[(A \cdot B)|C]$.

It also follows that if $\{B_1, \dots, B_n, \dots\}$ is any countable set of sentences that are mutually exclusive, given C (i.e., for each pair B_i and B_j , $C \models \sim(B_i \cdot B_j)$), then $\lim_n P_\alpha[(B_1 \vee B_2 \vee \dots \vee B_n)|C] = \sum_{i=1}^{\infty} P_\alpha[B_i|C]$ (unless $P_\alpha[D|C] = 1$ for every sentence D).¹⁰

$P_\alpha[g \in X \cdot Ch(X, Y) = 1 \cdot g \in Y] = P_\alpha[Ch(X, Y) = 1 \cdot g \in Y] = P_\alpha[(g \in X \vee \sim g \in X) \cdot Ch(X, Y) = 1 \cdot g \in Y] = P_\alpha[g \in X \cdot Ch(X, Y) = 1 \cdot g \in Y] + P_\alpha[\sim g \in X \cdot Ch(X, Y) = 1 \cdot g \in Y]$. So $P_\alpha[\sim g \in X \cdot Ch(X, Y) = 1 \cdot g \in Y] = 0$. Thus, $P_\alpha[g \in X | Ch(X, Y) = 1 \cdot g \in Y \cdot \sim g \in X]$ cannot be 0, but must instead be undefined, because the term $P_\alpha[\sim g \in X \cdot Ch(X, Y) = 1 \cdot g \in Y] = 0$ in the denominator of its definition has value 0.

¹⁰This should not be confused with *countable additivity*. In this context countable additivity would require the object language in which scientific theories are expressed to possess a means of expressing infinitely long disjunctions of sentences. If our object language had such a connective, ' $\bigvee_{\{B_k : B_k \in S\}} B_k$ ' for countable sets S of sentences, then an axiom for countable additivity need only say that whenever, for all distinct B_i and B_j in a countable set $S = \{B_1, B_2, \dots\}$, $C \models \sim(B_i \cdot B_j)$, then $P_\alpha[\bigvee_{\{B_k : B_k \in S\}} B_k | C] = \lim_n P_\alpha[(B_1 \vee B_2 \vee \dots \vee B_n) | C]$. It then follows from the derived rule stated above that $P_\alpha[\bigvee_{\{B_k : B_k \in S\}} B_k | C] = \sum_{i=1}^{\infty} P_\alpha[B_i | C]$. Here is a

In the context of the logic of confirmational support it makes good sense to supplement the above rules with two more. Here's the first one:

6. If A is an axiom of set theory or any other piece of pure mathematics employed by the sciences, or if A is analytically true (given the meanings that P_α presupposes for the terms in L), then, for all C , $P_\alpha[A|C] = 1$.

The idea is that the logic of confirmation is about evidential support for contingent claims. Nothing can count as empirical evidence against or for non-contingent

derivation of the above rule (proved without appeal to countable additivity):

Suppose that for each distinct B_i and B_j , $C \models \sim(B_i \cdot B_j)$, and suppose $P_\alpha[D|C] < 1$ for at least one sentence D . Notice that, for each i , $C \models (\sim(B_i \cdot B_{i+1}) \cdot \dots \cdot \sim(B_i \cdot B_n))$. This implies $C \models \sim(B_i \cdot (B_{i+1} \vee \dots \vee B_n))$. So for each finite list of the B_i , $P_\alpha[(B_1 \vee B_2 \vee \dots \vee B_n)|C] = P_\alpha[B_1|C] + P_\alpha[(B_2 \vee \dots \vee B_n)|C] = \dots = \sum_{i=1}^n P_\alpha[B_i|C]$. By definition, $\sum_{i=1}^\infty P_\alpha[B_i|C] = \lim_n \sum_{i=1}^n P_\alpha[B_i|C]$. Thus, $\lim_n P_\alpha[(B_1 \vee B_2 \vee \dots \vee B_n)|C] = \sum_{i=1}^\infty P_\alpha[B_i|C]$.

I have no objection to countable additivity in general. But infinite disjunction doesn't appear to be a natural part of the language in which scientific theories are expressed. That is, although there are well worked out extensions of predicate logic to infinitary languages, which include a means of expressing infinite disjunctions (see [Scott and Krauss, 1966]), no scientific theory I know of draws on the resources of such a language. So I won't add such devices to the object language for the confirmation functions under consideration here.

Regular first-order predicate logic does have a limited means of expressing infinite disjunctions via existential quantification. So one might consider adding a kind of countable additivity axiom (sometimes called the *Gaifman Condition*, after [Gaifman, 1962]) as follows:

for each open expression Fx , $P_\alpha[\exists xFx|B] = \lim_n P[Fc_1 \vee \dots \vee Fc_n|B]$, where the individual constants $c_1, \dots, c_n, \dots$, exhaust the countably infinite list of L 's individual constants.

From this axiom the following form of countable additivity follows: if for each distinct c_i and c_j , $B \models \sim(Fc_i \cdot Fc_j)$, then $P_\alpha[\exists xFx|B] = \sum_{i=1}^\infty P_\alpha[Fc_i|C]$. However, the proposed axiom seems overly strong, since it effectively assumes that every individual object gets named, or at least that enough exemplars are named that the probability of the disjunction approaches the probability of the existential claim. This seems implausible in contexts where the number of individuals is uncountably infinite.

If we don't assume that the *Gaifman Condition* holds, the strongest claim we should want is this: $P_\alpha[\exists xFx|B] \geq \lim_n P_\alpha[Fc_1 \vee \dots \vee Fc_n|B]$. But the other axioms already imply this, as follows:

$$\begin{aligned} P_\alpha[\exists xFx|B] &\geq P_\alpha[\exists xFx|B] \times P_\alpha[(Fc_1 \vee \dots \vee Fc_n)|B \cdot \exists xFx] = P_\alpha[(Fc_1 \vee \dots \vee Fc_n) \cdot \\ &\quad \exists xFx|B] = P_\alpha[\exists xFx|B \cdot (Fc_1 \vee \dots \vee Fc_n)] \times P_\alpha[(Fc_1 \vee \dots \vee Fc_n)|B] = P_\alpha[(Fc_1 \vee \dots \vee \\ &\quad Fc_n)|B] \text{ (since } P_\alpha[\exists xFx|B \cdot (Fc_1 \vee \dots \vee Fc_n)] = 1, \text{ because } B \cdot (Fc_1 \vee \dots \vee Fc_n) \models \exists xFx). \\ \text{So } P_\alpha[\exists xFx|B] &\geq \lim_n P_\alpha[Fc_1 \vee \dots \vee Fc_n|B]. \end{aligned}$$

If, in addition, for each pair of distinct c_i and c_j , $B \models \sim(Fc_i \cdot Fc_j)$, then $P_\alpha[(Fc_1 \vee \dots \vee Fc_n)|B] = \sum_{i=1}^n P_\alpha[Fc_i|C]$, so $\lim_n P_\alpha[Fc_1 \vee \dots \vee Fc_n|B] = \sum_{i=1}^\infty P_\alpha[Fc_i|C]$. Thus, if for each pair of distinct c_i and c_j , $B \models \sim(Fc_i \cdot Fc_j)$, then $P_\alpha[\exists xFx|B] \geq \sum_{i=1}^\infty P_\alpha[Fc_i|C]$.

One more point: Confirmation functions reside in the meta-language, where logical relationships usually reside, rather than in the object-language L , where scientific hypotheses live. So, even though countable additivity may not be a feature of confirmation functions themselves, scientific hypotheses (expressed in the object-language) may employ object-language probability functions defined on sets to represent chance processes in nature. These object-language probability functions may well satisfy a countable additivity rule for countable unions of disjoint sets. The existence of such countably additive object-language probabilities is perfectly compatible with the present approach to confirmational probability.

truths. They should be maximally confirmed relative to each possible statement.

An important respect in which the logic of confirmation functions should follow the deductive paradigm is in not presupposing the truth-values of contingent sentences. For, the whole idea of a logic of confirmation is to provide a measure of the extent to which contingent premise sentences indicate the likely truth-values of contingent conclusion sentences. But this idea won't work properly if the truth-values of some contingent sentences are presupposed by the confirmation function. Such presuppositions may hide significant premises, making the logic of confirmation enthymematic. Thus, for example, no confirmation function P_α should permit a tautological premise to assign degree of confirmation 1 to a contingent claim: $P_\alpha[C|B \vee \sim B]$ should always be less than 1 when C is contingent.

However, it is common practice for probabilistic logicians to sweep provisionally accepted contingent claims under the rug by assigning them probability 1. This saves the trouble of repeatedly writing a given contingent sentence B as a premise, since $P_\gamma[A|B \cdot C]$ will just equal $P_\gamma[A|C]$ whenever $P_\gamma[B|C] = 1$. Although this device is useful, such functions should be considered mere abbreviations of proper, logically explicit, non-enthymematic, confirmational relationships. Thus, properly speaking, a confirmation function P_α should assign probability 1 to a sentence *on every possible premise* only if that sentence is either (i) logically true, or (ii) an axiom of set theory or some other piece of pure mathematics employed by the sciences, or (iii) that sentence is *analytic* according to the meanings of terms in the language presupposed by confirmation function P_α , and so outside the realm of evidential support. Thus, it is natural to adopt the following version of the so-called "axiom of regularity".

7. If A is not a consequence of set theory or some other piece of pure mathematics employed by the sciences, and is neither a logically nor an analytically true statement (given the meanings of the terms of L as represented in P_α), then $P_\alpha[A|\sim A] < 1$.¹¹

Taken together with axiom 6, this axiom tells us that a confirmation function P_α counts as non-contingently true just those sentences that it assigns probability 1 on every possible premise.¹²

Some logicist Bayesians have proposed that the logic of confirmation might be made to depend solely on the logical form of sentences, just like deductive logic. The idea is, effectively, to supplement axioms 1–7 with additional axioms that depend only on the logical structures of sentences, and to add enough such axioms to reduce the number of possible confirmation functions to a single unique function. It is now widely agreed that this project cannot be carried out in a plausible way. Perhaps there are additional rules that should be added to 1–7. But it is doubtful

¹¹It follows from the other axioms that when $P_\alpha[A|\sim A] < 1$, $P_\alpha[A|\sim A] = 0$: for, $\sim A \models (A \vee \sim A)$, so $1 = P_\alpha[A \vee \sim A|\sim A] = P_\alpha[A|\sim A] + P_\alpha[\sim A|\sim A] = P_\alpha[A|\sim A] + 1$.

¹²Because, $P_\alpha[A|C] = 1$ for all C if and only if $P_\alpha[A|\sim A] = 1$. So, taken together with axiom 7, ' $P_\alpha[A|C] = 1$ for all C ' implies that A must be either a consequence of set theory or some other piece of pure mathematics, or it must be logically or analytically true.

that such rules can suffice to specify a single, uniquely qualified confirmation function based only on the formal structure of sentences. I'll say more about why this is doubtful a bit later, after we first see how confirmational probabilities capture the important relationships between hypotheses and evidence.

2 TWO CONCEPTIONS OF CONFIRMATIONAL PROBABILITY

Axioms 1–7 merely place formal constraints on what may properly count as a probabilistic confirmation function. Each function P_α that satisfies these rules may be viewed as a possible way of specifying a confirmation function that respects the meanings of the logical terms, much as each possible *truth-value assignment* for a language represents a possible way of assigning truth-values to its sentences in a way that respects the semantic rules expressing the meanings of the logical terms 'not', 'and', 'or', etc. The issue of which of the possible truth-value assignments to sentences of a language represents the *actual* truth or falsehood of its sentences depends on more than this — it depends on the meanings of the non-logical terms and on the state of the actual world. Similarly, the degree to which some sentences *actually* support others in a fully meaningful language must rely on more than the mere satisfaction of these axioms for confirmation functions. It must at least rely on what the sentences of the language mean, and perhaps on much more besides. What more? Perhaps an *interpretation of probability* appropriate to the confirmation functions may shed some light on this issue by filling out our conception of what confirmational probability is really about. I'll briefly describe two prominent views.¹³

One reading, a kind of *logicist* reading of confirmation functions, is to take each function P_α to be associated with a meaning assignment to sentences, and to take these interpreted sentences to represent possible states of affairs, or ways the world might turn out to be, or possible worlds. Then a confirmation measure P_α is effectively a measure on (expressible) sets of logically possible states of affairs. The idea is that, given a fully meaningful language, ' $P_\alpha[A|B] = r$ ' says that among the states in which B is true, A is true in "proportion" r of them. The relevant sets of possible states will usually contain infinitely many states, so the measure on sets of states supplied by P_α is not "intrinsic to the sets of possibilities themselves", but rather is imposed by a particular way of measuring (by a particular measure function on) the space of logical possibilities. Thus, each confirmation function imposes its own way of measuring possible states, and generally there may be any number of respectable ways to assign meanings to sentences and measures on possible states. So, when an agent chooses to employ a particular confirmation function (or a particular collection of them), she is effectively choosing a way of measuring possibilities — a way that reflects how she understands the meanings and inferential implications of sentences in her language.

If we read confirmation functions as measures on possible states, then axioms

¹³[Hájek, 2003b] provides a helpful discussion of the various interpretations of probability.

1-7 turn out to be plausible constraints on such measures. Each axiom is a rule that is automatically satisfied by finite proportions — so these axioms extend the logic of finite proportions to “measures of proportionality” on infinite sets. However, perhaps not every function that satisfies these axioms represents a good way to measure confirmation. That is, these axioms may not be jointly sufficient to pick out “proper” confirmation functions. What further restrictions should confirmation functions satisfy? As we see how the confirmation of scientific hypotheses is supposed to work in a Bayesian confirmation theory, perhaps some plausible additional restrictions will become apparent. So let’s put that question aside for now. In any case, I hope that the basic idea behind this kind of *logicist reading* of the confirmation functions is clear enough. I invite you to see how well this reading fits with how Bayesian confirmation theory is supposed to work, as explicated in subsequent sections of this article.

Subjectivist Bayesians offer an alternative reading of the confirmation functions. First, they usually take unconditional probabilities as basic, and they take conditional probabilities as defined in terms of them. Furthermore, subjectivist Bayesians take each unconditional probability function P_α to represent the *belief strengths* or *confidence strengths* of an ideally rational agent, α . On this understanding ‘ $P_\alpha[A] = r$ ’ says, “the strength of α ’s belief (or confidence) that A is true is r .” Subjectivist Bayesians usually tie such belief strengths to what the agent would be willing to bet on A being true. Roughly, the idea is this. Suppose that an ideally rational agent α is willing to accept a wager that would yield (at least) $\$u$ if A turns out to be true and would lose him $\$1$ if A turns out to be false. Then, under reasonable assumptions about his desires for small amounts of money, it can be shown that his belief strength that A is true should be $P_\alpha[A] = 1/(u + 1)$. And it can further be shown that any function P_α that expresses such betting-related belief strengths on all statements in agent α ’s language must satisfy the usual axioms for unconditional probabilities.¹⁴ Moreover, it can be shown that any function P_β that satisfies these axioms is a possible rational belief function for some ideally rational agent β . Such relationships between belief strengths and the desirability of outcomes (e.g., gains in money or goods on bets) are at the core of *Bayesian decision theory*.¹⁵ Subjectivist Bayesians usually take confirmational probability to just *be* this notion of probabilistic belief strength.¹⁶

Undoubtedly real agents do believe some claims more strongly than others. And, arguably, the belief strengths of real agents can be measured on a probabilistic scale between 0 and 1, at least approximately. And clearly the confirmational support of evidence for a hypothesis should influence the strength of an agent’s

¹⁴Note 8 lists these axioms.

¹⁵An alternative, but conceptually similar approach is to set down intuitively plausible constraints on the notion of *rational preference* among acts (or their outcomes), and then show that any such notion of preference can be represented (uniquely) by a probabilistic belief function together with a utility function, and that *preferred* acts (or their outcomes) are just those that maximize expected utility.

¹⁶For various versions of this approach see [Ramsey, 1926; de Finetti, 1937; Savage, 1954; Jeffrey, 1965; Skyrms, 1984; Joyce, 1999]. [Hájek, 2005] provides a perceptive analysis.

belief that the hypothesis is true. However, there is good reason for caution about taking confirmation functions to *be* Bayesian belief strength functions, as we will see later. So, perhaps an agent's confirmation function is not simply identical to his belief function, and perhaps the relationship between confirmation and belief strength is somewhat more complicated than the subjective Bayesian supposes.

In any case, some account of what confirmation functions are supposed to represent is clearly needed. The belief function account and the possible states (or possible worlds) account are two attempts to provide this. I'll now put this interpretative issue aside until later (section 4). We'll return to trying to get a better handle on what probabilistic confirmation functions *really are* after we take a careful look at how the logic that draws on them is supposed to work.

3 THE LOGICAL STRUCTURE OF EVIDENTIAL SUPPORT AND THE ROLE OF BAYES' THEOREM IN THAT LOGIC

A theory of confirmation should explicate the notion of evidential support for all sorts of scientific hypotheses, ranging from simple diagnostic claims (e.g., the patient has pneumonia) to scientific theories about the fundamental nature of the world, like quantum mechanics or the theory of relativity. We'll now look at how the logic of probabilistic confirmation functions draws on Bayes' Theorem to bring evidence to bear, via the *likelihoods*, on the refutation or support of scientific hypotheses.

To begin with, consider some exhaustive set of mutually incompatible hypotheses or theories about some common subject matter, $\{h_1, h_2, \dots\}$. The set of alternatives may consist of a simple pair of alternatives — e.g., {“the patient has pneumonia”, “the patient doesn't have pneumonia”}. Or it may consist of a long list of possible alternatives, as is the case when the physician tries to determine which among a range of diseases is causing the patient's symptoms. For the cosmologist the alternatives may consist of several alternative theories of the structure and dynamics of space-time, and may include various “versions of the same theory”. Where confirmation theory is concerned, even a slightly different (logically distinct) version of a given theory will count as a distinct theory, especially if it differs from the original in empirical import.

In principle there may be finitely or infinitely many alternative hypotheses under consideration. They may all be considered at once, or they may be constructed and assessed over many centuries. One may even take the set of alternative hypotheses to consist of all logically possible alternatives expressible in a given language about a given subject matter — e.g., all possible theories of the origin and evolution of the universe expressible in English and mathematics. Although testing every logically possible alternative poses practical challenges, the logic works much the same way in this logically ideal case as it does in more practical cases.

The set of alternative hypotheses may contain a *catch-all hypothesis* h_K that says that none of the other hypotheses are true — e.g., “the patient has none of the known diseases”. When only a finite number u of explicit alternative hypotheses

is under consideration, h_K will be equivalent to the sentence that denies each explicitly specified alternative: $(\sim h_1 \cdot \dots \cdot \sim h_u)$

Evidence for scientific hypotheses comes from specific experiments or observations. For a given experiment or observation, let ' c ' represent a description of the relevant experimental or observational *conditions* under which the evidence is obtained, and let ' e ' represent a description of the *evidential outcome* that comes about under conditions c .

Scientific hypotheses often require the mediation of background knowledge and auxiliary hypotheses to help them express claims about evidence. Let ' b ' represent all background and auxiliary hypotheses not at issue in the assessment of the hypotheses h_i , but that mediate their implications about evidence. In cases where a hypothesis is deductively related to evidence, either $h_i \cdot b \cdot c \models e$ or $h_i \cdot b \cdot c \models \sim e$ (i.e., either $h_i \cdot b \cdot c$ *logically entails* e or it *logically entails* $\sim e$).

For example, h_i might be the Newtonian Theory of Gravitation. A test of the theory may involve conditions described by a statement c about how measurements of Jupiter's position are made at various times; the outcome description e states the results of each position measurement; and the background information (or auxiliary hypotheses) b may state some trustworthy (already well confirmed) theory about the workings and accuracy of the devices used to make the position measurements. If outcome e can be calculated from the theory h_i together with b and c , we have that $h_i \cdot b \cdot c \models e$. The so-called *hypothetico-deductive* account of confirmation says that in such cases, if $(c \cdot e)$ actually occurs, this may be considered good evidence for h_i , given b . On the other hand, if from $h_i \cdot b \cdot c$ we calculate some outcome incompatible with e , then we have $h_i \cdot b \cdot c \models \sim e$. In that case, from deductive logic alone we get that $b \cdot c \cdot e \models \sim h_i$, and h_i is said to be *falsified* by $b \cdot c \cdot e$.

Duhem [1906] and Quine [1953] deserve credit for alerting inductive logicians to the importance of auxiliary hypotheses. They point out that scientific hypotheses often make little contact with evidence claims on their own. So, generally speaking, the evidence can only falsify hypotheses relative to background or auxiliary hypotheses that tie them to that evidence. However, the auxiliaries themselves will usually be testable on some separate body of evidence in much the same way that the hypotheses $\{h_1, h_2, \dots\}$ are tested. Furthermore, when we are not simply interested in assessing the hypotheses $\{h_1, h_2, \dots\}$ relative to a specific package of auxiliaries b , but instead want to consider various alternative packages of auxiliary hypotheses, $\{b_1, b_2, \dots\}$, as well, the set of *alternative hypotheses* to which the logic of confirmation applies should be the various possible combinations of original hypotheses in conjunction with the possible alternative auxiliaries, $\{h_1 \cdot b_1, h_1 \cdot b_2, \dots, h_2 \cdot b_1, h_2 \cdot b_2, \dots, h_3 \cdot b_1, h_3 \cdot b_2, \dots\}$. When this is the case, the logic of confirmation will remain the same. The only difference is that the hypotheses ' h_i ' under discussion below should be taken to stand for the complex conjunctive hypotheses of form $(h_j \cdot b_v)$, and ' b ' in our discussion below should stand for whatever remaining, common auxiliary hypotheses are not at issue. In the most extreme case, where each hypothesis at issue includes within itself all relevant auxiliaries, the term ' b ' may be empty — i.e. we may take it to be some simple tautology.

In probabilistic confirmation theory the degree to which a hypothesis h_i is supported or confirmed on evidence $c \cdot e$, relative to background b , is represented by the *posterior probability* of h_i , $P_\alpha[h_i|b \cdot c \cdot e]$. It turns out that the *posterior probability* of a hypothesis depends on two kinds of factors: (1) its *prior probability*, $P_\alpha[h_i|b]$, together with the prior probabilities of its competitors, $P_\alpha[h_j|b]$, etc....; and (2) the *likelihood* of evidential outcomes e according to h_i (give that b and c are true), $P[e|h_i \cdot b \cdot c]$, together with the likelihoods of the outcomes according to h_i 's competitors h_j , $P[e|h_j \cdot b \cdot c]$, etc., ... I'll now examine each of these two kinds of factors more closely. Then I'll discuss how the values of posterior probabilities depend on them.

3.1 Likelihoods

Hypotheses express their empirical import via *likelihoods*, which are confirmation function probabilities of form $P[e|h_i \cdot b \cdot c]$.¹⁷ A likelihood expresses how likely it is that outcome e will occur according hypothesis h_i under conditions c , supposing that auxiliaries b hold.¹⁸ If a hypothesis together with auxiliaries and observation conditions deductively entails an evidence claim, the probability axioms make the corresponding likelihood objective in the sense that every confirmation function must agree on its values: i.e., for all confirmation functions P_α , $P_\alpha[e|h_i \cdot b \cdot c] = 1$ if $h_i \cdot b \cdot c \models e$, and $P_\alpha[e|h_i \cdot b \cdot c] = 0$ if $h_i \cdot b \cdot c \models \sim e$. However, in many cases the hypothesis h_i will not be deductively related to the evidence, but will only imply it probabilistically. There are at least two ways this might happen. Either h_i may itself be an explicitly probabilistic or statistical hypothesis, or there may be an auxiliary statistical hypothesis in the background b that connects h_i to the evidence. For the sake of clarity let's briefly consider examples of each.

A blood test for HIV has a known false-positive rate and a known true-positive rate. Suppose the false positive rate is .05 — i.e., the test may be expected to incorrectly show the blood sample to be positive for HIV in about 5% of all cases where no HIV is present. And suppose the true-positive rate is .99 — i.e., the blood test may be expected to correctly show the blood sample to be positive for HIV in about 99% all cases where HIV really is present. When a particular patient's blood is tested, the hypotheses under consideration are 'the patient is infected with HIV', h , and 'the patient is not infected with HIV', $\sim h$. In this context the known test characteristics play the role of background information, b . The experimental condition c merely states that this patient was subjected to this particular blood test for HIV, which was processed by the lab in the usual way. Let us suppose that

¹⁷Presentations of the logic of confirmation often suppress c and b , and simply write ' $P[e|h]$ '. But c and b are important to logic of the likelihoods. So I'll continue to make them explicit.

¹⁸Bayesians often refer to the probability of an evidence statement on a hypothesis, $P[e|h \cdot b \cdot c]$, as the *likelihood of the hypothesis*. This terminology was introduced by R.A. Fisher [1922], who treated likelihoods as *functions on the space of possible hypotheses*, which he took to be measures of how strongly the evidence supports hypotheses. This can be a somewhat confusing way of talking, since it is clearly the evidence that is made *likely* to whatever degree by a hypothesis. So I'll avoid this usual (but unusual) way of talking about likelihoods.

the outcome e states that the result is positive for HIV. The relevant likelihoods, then, are $P[e|h \cdot b \cdot c] = .99$ and $P[e|\sim h \cdot b \cdot c] = .05$. In this example the values of the likelihoods are entirely due to the statistical characteristics of the accuracy of the test, which is carried by the background information b . The hypothesis h being tested is not itself statistical.

This kind of situation may, of course, arise for much more complex hypotheses. The hypothesis of interest may be some deterministic physical theory, say Newtonian Gravitation Theory. Some of the experiments that test this theory may rely on imprecise measurements that have known statistical error characteristics, which are expressed as part of the background or auxiliary hypotheses b . For example, the auxiliary b may describe the error characteristics of a device that measures the torque imparted to a quartz fiber, used to assess the strength of the gravitational force between test masses. In that case b may say that for this kind of device the measurement errors are normally distributed about whatever value a given gravitational theory predicts, with some specified standard deviation that is characteristic of the device. This results in specific likelihood values, $P[e|h_i \cdot b \cdot c] = r_i$, for each of the various alternative gravitational theories h_i being tested.

Alternatively, the hypotheses being tested may themselves be statistical in nature. One of the simplest examples of statistical hypotheses and their role in likelihoods consists of hypotheses about the chance-characteristics of coin-tossing. Let $h_{[r]}$ be a hypothesis that says a specific coin has a propensity r for coming up *heads* on normal tosses, and that all such tosses are probabilistically independent of one another. Let c state that the coin is tossed n times in the usual way; and let e state that on this specific run of n tosses the coin comes up heads m times. In cases like this the value of the likelihood of the outcome e under conditions c on hypothesis $h_{[r]}$ is well-known: $P[e|h_{[r]} \cdot b \cdot c] = [n!/(m!(n-m)!)] r^m(1-r)^{n-m}$.

There are, of course, more complex cases of likelihoods involving statistical hypotheses. Consider the hypothesis that plutonium 233 nuclei have a half-life of 20 minutes — i.e., the propensity for a Pu-233 nucleus to decay within a 20 minute period is $1/2$. This hypothesis, h , together with background b about decay products and the efficiency of the equipment used to detect them (which may itself be an auxiliary statistical hypothesis), yields precisely calculable values for likelihoods $P[e|h \cdot b \cdot c]$ of possible outcomes of the experimental arrangement.

Likelihoods that arise from explicit statistical claims — either within the hypotheses being tested, or from statistical background claims that tie the hypotheses to the evidence — are sometimes called *direct inference likelihoods*. Such likelihoods are generally taken to be completely objective.¹⁹ So all *reasonable confirma-*

¹⁹If you have doubts about the objectivity of direct inference likelihoods, I refer you to David Lewis's Questionnaire near the beginning of "A Subjectivist's Guide to Objective Chance" (1980). In that paper Lewis argues for the objectivity of direct inference likelihoods based on chance statements. Lewis's *Principal Principle* expresses his version of the principle governing these likelihoods. Indeed, Lewis thinks that this kind of principle about direct inference likelihoods captures "all we know about chance." (p. 266) He takes such likelihoods to be correct, regardless of what the right view might be about the metaphysical basis of chance and of confirmation probability function (which he calls *credence function*). "I shall not attempt to decide between

tion functions should be required to agree on their values, just as all confirmation functions agree on likelihoods when evidence is logically entailed. In other words, those probability functions P_α that satisfy our axioms but get the direct inference likelihoods wrong should be regarded as failing to represent proper confirmation functions. Direct inference likelihoods may be thought of as *logical* in an extended, non-deductive sense. Indeed, some logicians have attempted to spell out the logic of *direct inferences* in terms of the logical form of the sentences involved.²⁰ If that project can be made to work (for Bayesian confirmation functions), then the axioms for probabilistic confirmation functions (in section 1) should be supplemented with additional axioms that capture the logic of the direct inference likelihoods. But regardless of whether that formal project can be made to work, it seems reasonable to take likelihoods that derive from explicit statistical claims to have objective or intersubjectively agreed values, and to disregard any probability function that gets this wrong as failing to represent a proper confirmation function.²¹

the Humean and the anti-Humean variants of my approach to credence and chance. The Principal Principle doesn't." (Final two lines of the paper, p. 292.)

²⁰Attempts to spell out the logic of direct inference likelihoods for Bayesian confirmation functions in purely formal terms (i.e. in terms of logical structure alone) have not been wholly satisfactory thus far, but research continues. For an illuminating discussion of the logic of direct inference and the difficulties involved in providing a purely formal account, see the series of papers [Levi, 1977; Kyburg, 1978; Levi, 1978].

²¹In several places I've drawn on coin tossing as an example of a chancy process, and as having *propensities towards heads*. I sometimes encounter the following complaint about this treatment:

It is reasonable to think that the propensities for single outcomes in macro-physical systems are all either 1 or 0. Only in quantum mechanics do there seem to be non-extreme single case propensities. So those antecedents of direct inference likelihoods that attribute propensities to macro-systems, like coin tossing systems, are (most likely) literally false.

This objection is not just nit-picking. The point is that there may be very few real non-deductive cases of direct inference likelihoods (except in microphysics). It seems to me that two responses are in order.

(1) As in the case of deductive logic, the falsity of the premise has nothing to do with the validity of the inference. I take this to be the case for probabilistic direct inferences as well.

(2) But perhaps the issue is that if the hypothesis (e.g. about coin propensities, or about any other macro-system) is literally false, and if we know that it's false, then even though the likelihood based on it may "make sense", we shouldn't want to try to confirm such a hypothesis. So there are few instances of true chance claims that figure in direct inference likelihoods in the context of a confirmation theory. This assessment seems correct, up to a point. But I think there is an important caveat. In the sciences we often use "literally false claims" in models that give good approximations. I think we do exactly that when we treat systems that we take to be *non-chancy* (e.g. coin tossing) as though they are chance processes. We model the system as though it is "truly chancy" (has real non-extreme propensities), and within that framework we test to find which chance model best captures the phenomena at the level of detail we are using to describe it. Understood this way, there is indeed a *literally true* hypothesis in the neighborhood that we may wish to confirm — i.e. a hypothesis of the sort that says:

for coins of *this sort* (e.g. bent in this particular way), tossed in the usual way, coin tossing is *more accurately modeled* as though it is a chance mechanism for which the chances of *heads* is (say) 57/100 *than by any other model of a similar sort* — e.g. than any other chance model that assigns the chances of *heads* (according to

Not all likelihoods of interest in confirmational contexts are warranted deductively or by explicitly stated statistical claims. Nevertheless, the likelihoods that relate hypotheses to evidence in scientific contexts should often have objective or intersubjectively agreed values. So, although we may need to draw on a range of different functions, $P_\alpha, P_\beta, \dots, P_\gamma$, etc., to represent the confirmation functions employed by various members of a scientific community (due to the different values they assign to *prior probabilities*), all should agree, at least approximately, on the values of the likelihoods. For, likelihoods represent the empirical content of a hypothesis, what the hypothesis (together with background b) *probabilistically implies* about the evidence. Indeed, the empirical objectivity of a science relies on a high degree of objectivity or intersubjective agreement among scientists on the values of likelihoods.

To see the point more vividly, imagine what a science would be like if scientists disagreed widely about the values of likelihoods. Each practitioner interprets a theory to *say* quite different things about how likely it is that various possible evidence statements will turn out to be true. Whereas scientist α takes theory h_1 to probabilistically imply that event e is highly likely, his colleague β understands the empirical import of h_1 to say that e is very unlikely. And whereas α takes competing theory h_2 to probabilistically imply that e is quite unlikely, his colleague β reads h_2 as saying that e is very likely. So, for α the outcome e supplies strong support for h_1 over h_2 , because $P_\alpha[e|h_1 \cdot b \cdot c] \gg P_\alpha[e|h_2 \cdot b \cdot c]$. But his colleague β takes outcome e to show just the opposite — that h_2 is strongly supported over h_1 — because $P_\beta[e|h_1 \cdot b \cdot c] \ll P_\beta[e|h_2 \cdot b \cdot c]$. If this kind of thing were to occur often or for significant evidence claims in a scientific domain, it would make a shambles of the empirical objectivity of that science. The empirical testability of its hypotheses and theories would be completely undermined. Under such circumstances, although each scientist employs the same *sentences* to express a given theory h , each understands the *empirical import* of these sentences so differently that h as understood by α is an empirically very different theory than h as understood by β .²² Thus, the empirical objectivity of the sciences requires that experts should be in close agreement about the values of the likelihoods.

that model) to be some other rational number $k/100$ (for another integer k).

One of the 101 alternative hypotheses of this form — about which chance model best captures the phenomena at a given level of descriptive detail — may be literally true. We test such hypotheses against one another in the usual way — by counting heads that result from tosses of coins of the appropriate sort. Isn't this, or something very much like it, what we think we are doing when we assess hypotheses about the chanciness of coin tossing?

²²This idea should not be confused with *logical positivism*. *Positivism* holds that if two theories assign the same likelihood values to each possible evidence claim, then they are essentially the same theory, though they may be couched in different words. The slogan is: *same likelihoods implies same theory*. The point being defended here, however, is not the *positivistic* claim, but its inverse, which should be much less controversial: *different likelihoods implies different theories*. That is, given that all of the relevant background and auxiliaries are made explicit (represented in ' b '), if two scientists disagree significantly about the likelihoods of important evidence claims on a given hypothesis, they must understand the empirical content of that hypothesis quite differently. To that extent, though they may be using the same sentences — the same syntactic expressions — they use them to express empirically distinct hypotheses.

For now we will suppose that the likelihoods have objective or intersubjectively agreed values, common to all agents in a scientific community, and shared by all confirmation functions they employ. We'll mark this agreement by dropping the subscript ' α ', ' β ', etc., from expressions that represent likelihoods. One might worry that this supposition is overly strong. There are many legitimate scientific contexts where, although scientists should have enough of a common understanding of the empirical import of hypotheses to assign quite similar values to likelihoods, precise agreement on the numerical values is not realistic. Later (in section 6) we will see how to relax the supposition that likelihood values precisely agree. But for now the main ideas behind probabilistic confirmation theory will be more easily explained if we focus on those contexts where objective or intersubjectively agreed likelihoods are available. Later we will see how the same logic applies in contexts where the values of likelihoods are somewhat vague, or where members of the scientific community disagree about their values to some extent.

An adequate treatment of the likelihoods requires the introduction of one additional notational device. Scientific hypotheses are generally tested by a sequence of experiments or observations conducted over a period of time. In order to explicitly represent the accumulation of evidence, let the series of sentences $c_1, c_2, \dots, c_n$, describe the conditions under which a sequence of experiments or observations are conducted. And let the corresponding outcomes of these conditions be represented by the respective sentences $e_1, e_2, \dots, e_n$. We will abbreviate the conjunction of the first n descriptions of experimental or observation conditions as ' c^n ', and abbreviate the conjunction of descriptions of their outcomes as ' e^n '. Then, for a stream of n observations or experiments and their outcomes, the likelihoods take the form $P[e^n|h_i \cdot b \cdot c^n] = r$, for appropriate r between 0 and 1. In many cases the likelihood of the evidence stream is equal to the product of the likelihoods of the individual outcomes: $P[e^n|h_i \cdot b \cdot c^n] = P[e_1|h_i \cdot b \cdot c_1] \times \dots \times P[e_n|h_i \cdot b \cdot c_n]$ — that is, the individual bits of evidence are *probabilistically independent given the hypothesis*. I'll discuss this kind of probabilistic independence in more detail later. For present purposes we will *not assume* that it holds.

3.2 Posterior Probabilities and Prior Probabilities

In a probabilistic confirmation theory the degree to which a hypothesis is supported on the evidence is represented by the *posterior probability* of the hypothesis, $P_\alpha[h_i|b \cdot c^n \cdot e^n]$. The posterior probability represents the net plausibility of the hypothesis. It depends on the combined influence of evidence together with relevant “non-evidential” plausibility considerations — i.e. all considerations not representable by *likelihoods*. The likelihoods are the means through which the usual sort of empirical evidence from experiments and carefully specified observations makes its contribution. But other kinds of considerations may also be relevant to a proper assessment of the truth of scientific hypotheses. The *prior probability* of the hypothesis (on background b), $P_\alpha[h_i|b]$, represents the weight of all such additional plausibility considerations on which posterior probabilities may

depend. Bayes' Theorem shows precisely how the values of (ratios of) likelihoods *combine with* the values of (ratios of) prior probabilities of hypotheses determine the posterior probabilities of hypotheses.

To understand the role of prior probabilities, consider the HIV test example described earlier. What the physician and patient want to know is the posterior probability $P_{\alpha}[h|b \cdot c \cdot e]$ of the hypothesis h that the patient has HIV, given the evidence of the positive test, $c \cdot e$, and given the error rates of the test, described by b . The value of this posterior probability depends on the likelihood (due to the error rates) of this patient obtaining a true-positive result, $P[e|h \cdot b \cdot c] = .99$, and of obtaining a false positive result, $P[e|\sim h \cdot b \cdot c] = .05$. In addition, the value of the posterior probability depends on how plausible it is that the patient has HIV before taking the test results into account, $P_{\alpha}[h|b]$. In the context of medical diagnosis this prior probability is called the *base rate*. It is the plausibility that the patient may have contracted HIV based on his risk group (i.e., whether he is an IV drug user, has unprotected sex with multiple partners, etc.). Such information may be explicitly stated in the background, b . To see its importance, consider the following numerical results (which may be calculated using Bayes' Theorem, as presented in the next subsection). If the base rate for the patient's risk group is relatively high, say $P_{\alpha}[h|b] = .10$, then the positive test result yields a probability for his having HIV of $P_{\alpha}[h|b \cdot c \cdot e] = .69$. However, if the patient is in a very low risk group, $P_{\alpha}[h|b] = .001$, then a positive test only raises the plausibility of HIV infection to $P_{\alpha}[h|b \cdot c \cdot e] = .02$. This posterior probability is much higher than the prior probability of .001, but still shouldn't worry the patient too much. The positive result is more likely due to the false-positive rate of the test than to the presence of HIV. (This sort of test, with such a large false-positive rate, .05, is best used as a screening test; a positive result should lead to a second, more expensive test with better error characteristics.)

In more theoretical disciplines prior probabilities may represent assessments of non-evidential, conceptually motivated *plausibility weightings* among hypotheses. However, because such plausibility assessments tend to vary among agents, critics often brand them as *merely subjective*, and take their role in the evaluation of hypotheses to be highly problematic. Bayesian confirmation theorists may counter that such assessments often play an important role in the sciences, especially when there is insufficient evidence to distinguish among some of the alternatives. And they point out that the epithet "*merely subjective*" is unwarranted. Such plausibility assessments are often backed by extensive arguments that may draw on forceful conceptual considerations.

Consider, for example, the kinds of plausibility considerations brought to bear in assessing the various interpretations of quantum theory (e.g., those related to the measurement problem). Many of these considerations go to the heart of conceptual issues that were central to the development of the theory. Indeed, many of these issues were first explored by those scientists who made the greatest contributions to the theory's development, in the attempt to get a conceptual hold on the theory and its implications. Such arguments seem to play a legitimate role

in the assessment of the relative plausibility of alternative views, especially when distinguishing evidence has yet to be found, or is far from definitive.

Scientists often bring plausibility arguments to bear in assessing their views. Although such arguments are seldom decisive, they may bring the scientific community into widely shared agreement, especially with regard to the implausibility of some “logically possible” alternatives. This seems to be the primary epistemic role of the thought experiment. It is arguably a virtue of probabilistic confirmation theory that it provides a place for such assessments to figure into the net evaluation of hypotheses. Prior probabilities are subjective in the sense that agents may continue to disagree on the relative merits of plausibility arguments — and so disagree on the prior plausibilities of various hypotheses. But assessments of priors are far from being *mere subjective whims*. Moreover, it can be shown that when sufficient empirical evidence becomes available, such plausibility assessments may be “washed out” or overridden by the evidence. We’ll see how this works in the next subsection, and in more detail in section 5.

Our discussion of the nature of prior probabilities isn’t over yet. We will return to it a bit later. But let’s now see precisely how the logic of confirmation is supposed to work — how the likelihoods combine with prior probabilities to yield posterior probabilities for hypotheses.

3.3 Bayes’ Theorem and the Logic of Hypothesis Evaluation

Any *probabilistic logic of confirmation* that draws on the usual axioms of probability theory to represent how evidence supports hypotheses must be a *Bayesian inductive logic* in the broad sense. For, Bayes’ Theorem is just a simple theorem of probability theory. Its importance derives from the way it shows how evidence, through the likelihoods, combines with prior probabilities of hypotheses to produce assessments of their posterior probabilities.

We will now examine several forms of Bayes’ Theorem, each strictly derivable from the axioms of probability theory. The simplest is this:

8. Bayes’ Theorem: Simple Form:

$$\begin{aligned} P_{\alpha}[h_i|b \cdot c^n \cdot e^n] &= \frac{P[e^n|h_i \cdot b \cdot c^n] \times P_{\alpha}[h_i|b] \times P_{\alpha}[c^n|h_i \cdot b]}{P_{\alpha}[e^n|b \cdot c^n] \times P_{\alpha}[c^n|b]} \\ &= \frac{P[e^n|h_i \cdot b \cdot c^n] \times P_{\alpha}[h_i|b]}{P_{\alpha}[e^n|b \cdot c^n]} \text{ if } P_{\alpha}[c^n|h_i \cdot b] = P_{\alpha}[c^n|b]. \end{aligned}$$

This equation expresses the posterior probability of h_i , $P_{\alpha}[h_i|b \cdot c^n \cdot e^n]$, in terms of the *likelihood* of the evidence on the hypothesis (together with background and observation conditions), $P[e^n|h_i \cdot b \cdot c^n]$, the *prior probability* of the hypothesis (given background conditions), $P_{\alpha}[h_i|b]$, and the *simple probability* of the evidence (given background and observation conditions), $P_{\alpha}[e^n|b \cdot c^n]$. This latter probability is sometimes called *the marginal probability of the evidence* or *the prior probability*

of the evidence or the expectedness of the evidence. This version of Bayes' Theorem also includes a term, $(P_\alpha[c^n|h_i \cdot b]/P_\alpha[c^n|b])$, that represents the ratio of the likelihood of the experimental conditions on the hypothesis and background, to the "likelihood" of the experimental conditions on the background alone. Bayes' Theorem is usually expressed in a way that suppresses this factor, perhaps by building conditions c^n into the background b . However, if c^n is built into b , then technically b must change as new evidence is accumulated. It is better to make this factor explicit, and see how to deal with it logically.

Arguably the term $(P_\alpha[c^n|h_i \cdot b]/P_\alpha[c^n|b])$ should be 1, or nearly 1, since the truth of the hypothesis at issue should not significantly affect how likely it is that the experimental conditions themselves are satisfied. If various alternative hypotheses assign significantly different likelihoods to the experimental conditions themselves, then such conditions should more properly be included as part of the evidential outcomes e^n .

Both the *prior probability of the hypothesis* and the *marginal probability of the evidence* tend to be more subjective than the likelihoods. That is, various agents from the same scientific community may legitimately disagree on what values these factors should take. Bayesian logicians usually accept the subjectivity of the prior probabilities for hypotheses, but they find the subjectivity of the *marginal probability of the evidence* more troubling. How likely is the evidence supposed to be, given only the background — how are we supposed to assess this value? However, this problem is easily finessed.

The subjective *marginal probability of the evidence* may be circumvented by considering a ratio form of Bayes' Theorem, a form that compares hypotheses one pair at a time:

9. Bayes' Theorem: Ratio Form:

$$\begin{aligned} \frac{P_\alpha[h_j|b \cdot c^n \cdot e^n]}{P_\alpha[h_i|b \cdot c^n \cdot e^n]} &= \frac{P[e^n|h_j \cdot b \cdot c^n]}{P[e^n|h_i \cdot b \cdot c^n]} \times \frac{P_\alpha[h_j|b]}{P_\alpha[h_i|b]} \times \frac{P_\alpha[c^n|h_j \cdot b]}{P_\alpha[c^n|h_i \cdot b]} \\ &= \frac{P[e^n|h_j \cdot b \cdot c^n]}{P[e^n|h_i \cdot b \cdot c^n]} \times \frac{P_\alpha[h_j|b]}{P_\alpha[h_i|b]} \text{ if } P_\alpha[c^n|h_j \cdot b] = P_\alpha[c^n|h_i \cdot b]. \end{aligned}$$

The second line follows if c^n is no more likely on $h_i \cdot b$ than on $h_j \cdot b$ — i.e., if neither hypothesis makes the occurrence of experimental or observation conditions more likely than the other.²³

This ratio form of Bayes' Theorem expresses how much more plausible one hypothesis is than another, based on the evidence and on their non-evidential plausibilities. Notice that the only subjective element affecting the ratio of posterior probabilities is the ratio of prior probabilities for hypotheses. We see from

²³This assumption may be substantially relaxed without affecting the analysis below; we might instead only suppose that the ratios $P_\alpha[c^n|h_j \cdot b]/P_\alpha[c^n|h_i \cdot b]$ are bounded so as not to get exceptionally far from 1. If *that* supposition were to fail, then the mere occurrence of the experimental conditions would count as very strong evidence for or against hypotheses, which is a highly implausible effect. Our analysis could include such bounded condition-ratios, but that would only add inessential complexity.

this equation that the *likelihood ratios* carry the full import of the evidence. The evidence influences the evaluation of hypotheses in no other way.

Let's consider a simple example of how this form of Bayes' Theorem applies. Suppose you possess a warped coin and want to determine its propensity for *heads*. You may compare two hypotheses, $h_{[q]}$ and $h_{[r]}$, that propose the propensity for *heads* is q and r , respectively. Let c^n report that the coin is tossed n times in the usual way, and let e^n report a total m *heads*. Equation (9) then yields:

$$\frac{P_\alpha[h_{[q]} | b \cdot c^n \cdot e^n]}{P_\alpha[h_{[r]} | b \cdot c^n \cdot e^n]} = \frac{q^m(1-q)^{n-m}}{r^m(1-r)^{n-m}} \times \frac{P_\alpha[h_{[q]} | b]}{P_\alpha[h_{[r]} | b]}$$

For example, when the coin is tossed $n = 100$ times and comes up *heads* $m = 72$ times, the evidence for hypothesis $h_{[1/2]}$ as compared to $h_{[3/4]}$ is given by the likelihood ratio $[(1/2)^{72}(1/2)^{28}]/[(3/4)^{72}(1/4)^{28}] = .000056269$. So, even if prior to taking account of the evidence, one considers it 100 times more plausible that the coin is "fair" than that it is "warped towards heads with propensity 3/4", i.e., even if $P_\alpha[h_{[1/2]} | b]/P_\alpha[h_{[3/4]} | b] = 100$, the evidence provided by these tosses makes the posterior plausibility that the coin is "fair" only about $56/10,000^{th}$ as plausible as the hypothesis that it is "warped towards heads with propensity 3/4" — i.e., $P_\alpha[h_{[1/2]} | b \cdot c^n \cdot e^n]/P_\alpha[h_{[3/4]} | b \cdot c^n \cdot e^n] = .0056269$. Thus, such evidence *strongly refutes* the "fairness hypothesis" relative to the "3/4-*heads*-propensity hypothesis", provided the assessment of prior plausibilities doesn't make the latter hypothesis *too extremely implausible* to begin with. Notice, however, that *strong refutation* is not *absolute refutation*. Additional evidence could reverse the trend against the fairness hypothesis.

This example employs repetitions of the same kind of experiment — repeated tosses of a coin. But the point holds more generally. If, as the evidence increases, the *likelihood ratios* $P[e^n | h_j \cdot b \cdot c^n]/P[e^n | h_i \cdot b \cdot c^n]$ approach 0, then the Ratio Form of Bayes' Theorem (Equation 9) shows that the posterior probability of h_j must approach 0 as well. The evidence comes to strongly refute h_j with little regard for its prior plausibility value. Indeed, Bayesian confirmation turns out to be a version of *eliminative induction*, and Equation 9 begins to illustrate this. For, suppose that h_i is the true hypothesis, and consider what happens to *each* of its false competitors, h_j . If enough evidence becomes available to drive each of the likelihood ratios $P[e^n | h_j \cdot b \cdot c^n]/P[e^n | h_i \cdot b \cdot c^n]$ toward 0 (as n increases), then Equation 9 says that each false h_j will become effectively refuted — each of their posterior probabilities approaches 0. As a result, the posterior probability of h_i must approach 1, as shown by the next two equations.

Summing the ratio versions of Bayes' Theorem over all alternatives to hypothesis h_i (including the catch-all h_K) produces the Odds Form of Bayes' Theorem:

10. Bayes' Theorem: Odds Form

$$\begin{aligned} \Omega_\alpha[\sim h_i | b \cdot c^n \cdot e^n] &= \sum_{j \neq i} \frac{P[e^n | h_j \cdot b \cdot c^n]}{P[e^n | h_i \cdot b \cdot c^n]} \times \frac{P_\alpha[h_j | b]}{P_\alpha[h_i | b]} \\ &+ \frac{P_\alpha[e^n | h_K \cdot b \cdot c^n]}{P_\alpha[e^n | h_i \cdot b \cdot c^n]} \times \frac{P_\alpha[h_K | b]}{P_\alpha[h_i | b]} \end{aligned}$$

(By definition the *odds against* A given B is $\Omega_\alpha[\sim A|B] = P_\alpha[\sim A|B]/P_\alpha[A|B]$.) If $\{h_1, h_2, \dots, h_n, \dots\}$ is an exhaustive collection of alternative hypotheses, given b , then the *catch-all alternative* isn't needed and the expression after the '+' sign should be discarded. We represent the term for the catch-all hypothesis separately because the likelihood of evidence relative to the catch-all hypothesis will not usually enjoy the kind of objectivity possessed by the likelihoods for *explicit* hypotheses. We indicate this by leaving the subscript α on its likelihood.

Although the catch-all hypothesis may lack objective likelihoods, the influence of the catch-all term in Bayes' theorem diminishes as additional *explicit* hypotheses become articulated. That is, when a new hypothesis h_{u+1} is formulated and made explicit, the old catch-all hypothesis h_K is replaced by a new catch-all, h_{K*} , of form $(\sim h_1 \dots \sim h_u \cdot \sim h_{u+1})$. The prior probability for the new catch-all hypothesis comes from diminishing the prior of the old catch-all: $P_\alpha[h_{K*}|b] = P_\alpha[h_K|b] - P_\alpha[h_{u+1}|b]$. Thus, the influence of the catch-all term diminishes towards 0 as new alternative hypotheses are made explicit and "peeled off" of previous catch-all terms.²⁴

If increasing evidence drives the likelihood ratios comparing h_i with each competitor towards 0, then the odds against h_i , $\Omega_\alpha[\sim h_i|b \cdot c^n \cdot e^n]$, will approach 0 (provided that priors of catch-all terms, if needed, approach 0 too, as new alternative hypotheses are made explicit and peeled off). And, as $\Omega_\alpha[\sim h_i|b \cdot c^n \cdot e^n]$ approaches 0, the posterior probability of h_i goes to 1. The relationship between the odds against h_i and its posterior probability is this:

11. Bayes' Theorem: From Posterior Odds to Posterior Probability

$$P_\alpha[h_i|b \cdot c^n \cdot e^n] = 1/(1 + \Omega_\alpha[\sim h_i|b \cdot c^n \cdot e^n]).$$

A confirmation function provides a probabilistic index of the net support a hypothesis receives *all-things-considered*. It explicitly divides considerations that bear on the evaluation of hypotheses into two kinds of components: non-evidential comparative plausibility considerations, represented by ratios of prior probabilities; and evidential support that derives from what hypotheses imply about the evidence claims, represented by ratios of likelihoods. Equation 9 shows precisely how these two kinds of considerations are brought to bear. In particular, it makes clear that if the series of likelihood ratios $P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n]$ approaches 0 on increasing evidence (i.e., as n increases), the posterior probability of hypothesis h_j must approach 0 as well, with little regard for the value of its prior probability $P_\alpha[h_j|b]$. As this happens to each of h_i 's competitors, Equations 10 and 11 show that the posterior probability of hypothesis h_i approaches 1, as evidence increases.

²⁴For example, when a new disease is discovered, a new hypothesis h_{u+1} about that disease being a possible cause of patients' symptoms is made explicit. The old catch-all was, "the symptoms are caused by some unknown disease — some disease other than $h_1, \dots, h_u$ ". The new catch-all hypothesis states that "the symptoms are caused by one of the remaining unknown diseases — some disease other than $h_1, \dots, h_u, h_{u+1}$ ". Then, $P_\alpha[h_K|b] = P_\alpha[\sim h_1 \dots \sim h_u|b] = P_\alpha[\sim h_1 \dots \sim h_u \cdot (h_{u+1} \vee \sim h_{u+1})|b] = P_\alpha[\sim h_1 \dots \sim h_u \cdot h_{u+1}|b] + P_\alpha[\sim h_1 \dots \sim h_u \cdot \sim h_{u+1}|b] = P_\alpha[h_{K*}|b] + P_\alpha[h_{u+1}|b]$.

Is there any good reason for thinking that when h_i is true, the series of likelihood ratios $P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n]$ will indeed favor it by heading towards 0 as evidence accumulates? There is a result, a kind of *Bayesian Convergence Theorem*, that speaks directly to this issue. It says: given that h_i (together with $b \cdot c^n$) is true and that h_j is empirically distinct from h_i (in that h_j assigns different likelihoods to some possible evidential outcomes than does h_i), it's *very likely* that the evidence stream will produce outcomes for which the likelihood ratios approach 0 as the evidence accumulates. The theorem expresses this in terms of a likelihood, based on $(h_i \cdot b \cdot c^n)$. The theorem itself supposes nothing about the values of the prior probabilities. Let's call this result the *Likelihood Ratio Convergence Theorem*. When it applies, putting it together with Equation 9 shows that the posterior probability of false competitor h_j is *very likely* to approach 0 as evidence accumulates, with little regard for its prior probability. And as this happens to each of h_i 's false competitors, Equations 10 and 11 show that the posterior probability of the true hypothesis, h_i , is *very likely* to approach 1.²⁵ Thus, Bayesian confirmation is at bottom a version of *induction by elimination*, where the elimination of false alternatives comes by way of likelihood ratios approaching 0 as evidence accumulates. We will examine the *Likelihood Ratio Convergence Theorem* in detail in section 5.

3.4 Likelihood Ratios, Likelihoodism, and the Law of Likelihood

The versions of Bayes' Theorem expressed by Equations 9-11 show that for probabilistic confirmation theory the influence of empirical evidence on posterior probabilities of hypotheses is completely captured by the ratios of likelihoods, $P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n]$. The evidence ($c^n \cdot e^n$) influences the posterior probabilities in no other way.²⁶ So, the following "Law" is implied by the logic of confirmation functions.

General Law of Likelihood: Given any pair of incompatible hypotheses h_i and h_j , whenever the likelihoods $P_\alpha[e^n|h_j \cdot b \cdot c^n]$ and $P_\alpha[e^n|h_i \cdot b \cdot c^n]$

²⁵This claim depends, of course, on h_i being empirically distinct from each alternative h_j . I.e., there must be conditions c_k with possible outcomes o_{ku} on which the likelihoods differ: $P[o_{ku}|h_i \cdot b \cdot c_k] \neq P[o_{ku}|h_j \cdot b \cdot c_k]$. Otherwise h_i and h_j are empirically equivalent, and no amount of evidence can distinguish one from the other. If the true hypothesis has empirically equivalent rivals, then convergence just implies that the odds against the *disjunction* of the true hypothesis with these rivals very probably goes to 0, and so the posterior probability of this *disjunction* goes to 1. Among empirically equivalent hypotheses the ratio of their posterior probabilities equals the ratio of their priors: $P_\alpha[h_j|b \cdot c^n \cdot e^n]/P_\alpha[h_i|b \cdot c^n \cdot e^n] = P_\alpha[h_j|b]/P_\alpha[h_i|b]$. So the true hypothesis will have a posterior probability near 1 (after evidence drives the posteriors of empirically distinct rivals near 0) *just in case* considerations that go beyond the kind of evidence engendered by likelihoods make the true hypothesis much more plausible than the empirically equivalent rivals. Furthermore, additional plausibility consideration may be brought to bear whenever they are developed. Nothing about the logic Bayesian confirmation requires that the so-called "prior" probabilities be set in advance of the evidence; the so-called "priors" may themselves be updated at any time due to additional plausibility considerations.

²⁶However, there may be other useful ways to measure evidential import. Fitelson [1999] provides a penetrating comparison of a number of such measures.

are defined, *the evidence* ($c^n \cdot e^n$) favors h_i over h_j , given b , if and only if $P_\alpha[e^n|h_i \cdot b \cdot c^n] > P_\alpha[e^n|h_j \cdot b \cdot c^n]$. The ratio of likelihoods $P_\alpha[e^n|h_i \cdot b \cdot c^n]/P_\alpha[e^n|h_j \cdot b \cdot c^n]$ measures the *strength of the evidence* for h_i over h_j given b .²⁷

The Law of Likelihood says that the likelihood ratios represent the *total impact of the evidence*. Bayesians agree with this, but take prior probabilities to also play a role in the net assessment of confirmation, as represented by the posterior probabilities. So, for Bayesians, even when the *strength of the evidence*, $P_\alpha[e^n|h_i \cdot b \cdot c^n]/P_\alpha[e^n|h_j \cdot b \cdot c^n]$, is be very high, strongly favoring h_i over h_j , the net degree of confirmation of h_i may be much smaller than that of h_j if h_i is taken to be much less plausible than h_j on grounds not captured by this evidence (where the weight of these additional considerations is represented by the confirmational prior probabilities of hypotheses).

Two features of the way the General Law of Likelihood is stated here need some explanation. As stated, this law does not presuppose that likelihoods of form $P_\alpha[e^n|h_j \cdot b \cdot c^n]$ and $P_\alpha[e^n|h_i \cdot b \cdot c^n]$ are always defined. This qualification is introduced to accommodate a conception of evidential support called *Likelihoodism*, which I'll say more about in a moment. Also, the likelihoods in the law are expressed with the subscript α attached, to indicate that the law holds for each confirmation function P_α , even if the values of the likelihoods are not completely objective or agreed on by a given scientific community. These two features of the law both involve issues concerning the objectivity of the likelihoods.

Each confirmation function (each function that satisfies the axioms of section 1) is defined on every pair of sentences. So, the likelihoods are always defined for a given confirmation function. Thus, for a Bayesian confirmation theory the qualifying clause about the likelihoods being defined is automatically satisfied. Furthermore, for confirmation functions the versions of Bayes' theorem (Equations 8-11) hold even when the likelihoods are not objective or intersubjectively agreed. When intersubjective agreement on likelihoods may fail, we leave the subscripts α , β , etc. attached to the likelihoods to indicate this possible lack of objective agreement. Even so, the General Law of Likelihood applies to the confirmation function likelihoods taken one confirmation function at a time. For each confirmation function, the impact of the evidence in distinguishing between hypotheses is completely captured by the likelihood ratios.

A view (or family of views) called *likelihoodism* maintains that confirmation theory should only concern itself with how much *the evidence* supports one hypothesis over another, and maintains that evidential support should only involve ratios of *completely objective* likelihoods. When the likelihoods are objective, their ratios provide an objective measure of how strongly the evidence supports h_i as compared to h_j , one that is "untainted" by such subjective elements as prior plausibility considerations. According to *likelihoodists*, objective likelihood ratios

²⁷The General Law is sometimes presented (by *likelihoodists*) in a stronger form — a form that adds that *nothing but likelihood ratios* is relevant to the scientific evaluation of hypotheses. In that form it is an explicitly anti-Bayesian thesis.

are the only scientifically appropriate way to assess *what the evidence says* about hypotheses.

Likelihoodists need not reject Bayesian confirmation theory altogether. Many are statisticians and logicians who hold that the logical assessment of the *evidential impact* should be kept separate from other considerations. They often add that the only job of the statistician/logician is to evaluate the objective strength of the evidence. Some concede that the way in which these objective likelihoods should influence the agents' posterior confidence in the truth of a hypothesis may depend on additional considerations — and that perhaps these considerations may be represented by individual *subjective* prior probabilities for agents in the way Bayesians suggest. But such considerations go beyond *the impact of the evidence*. So it's not the place of the statistician/logician to compute *recommended values of posterior probabilities* for the scientific community.²⁸

For most pairs of sentences conditional probabilities fail to be *objectively defined* in a way that suits likelihoodists. So, by their lights the *logic of confirmation functions* (captured by the axioms of section 1) cannot represent an *objective logic* of evidential support. Because of this, *likelihoodists* do not have Bayes' theorem available (except in special cases where an objective probability measure on the hypothesis space is available), and so cannot extract the Law of Likelihood from it (as do Bayesians via Equations 9-11). Rather, *likelihoodists* must state the Law of Likelihood as an axiom of their logic of evidential support, an axiom that (for them) applies only when likelihoods have well-defined objective values.

Likelihoodists tend to have a very strict conception of what it takes for likelihoods to be *well-defined*. They consider a likelihood well-defined only when it has the form of what we referred to earlier as a *direct inference likelihood* — i.e., only when either, (1) the hypothesis (together with background and experimental conditions) logically entails the evidence claim, or (2) the hypothesis (together with background conditions) logically entails an explicit *simple statistical hypothesis* that (together with experimental conditions) specifies precise probabilities for each type of event that makes up the evidence.

Likelihoodists make a point of contrasting *simple statistical hypotheses* with *composite statistical hypotheses*, which only entail imprecise, or disjunctive, or directional claims about the statistical probabilities of evidential events. A *simple statistical hypothesis* might say, for example, “the chance of *heads* on tosses of the coin is precisely .65”; a composite statistical hypothesis might say, “the chance of *heads* on tosses is either .65 or .75”, or it may be a *directional hypothesis* that says, “the chance of *heads* on tosses is greater than .65”. Likelihoodists maintain that *composite hypotheses* are not an appropriate basis for well-defined likelihoods, because such hypotheses represent a kind of disjunction of simple statistical hypotheses, and so must depend on non-objective factors — i.e. they must depend on the prior probabilities of the various hypotheses in the disjunction. For example, “the chance of heads on tosses is either .65 or .75”, is a disjunction of the two *simple statistical hypotheses* $h_{.65}$ and $h_{.75}$. From the axioms of probability

²⁸Royall [1997] expresses a view of this sort.

theory it follows that the likelihood of any specific sequence of outcomes e from appropriate tosses c is given by

$$P_{\alpha}[e|c \cdot (h_{.65} \vee h_{.75})] = \frac{(P[e|c \cdot h_{.65}]P_{\alpha}[h_{.65}|c] + P[e|c \cdot h_{.75}]P_{\alpha}[h_{.75}|c])}{(P_{\alpha}[h_{.65}|c] + P_{\alpha}[h_{.75}|c])}$$

where only the likelihoods based on simple hypotheses (those from which I have dropped the ‘ α ’) are completely objective. Thus, likelihoods based on disjunctive hypotheses depend (at least implicitly) on the prior probabilities of the simple statistical hypotheses involved; and likelihoodists consider such factors to be too subjective to be permitted a role in a logic that is supposed to represent only the impact of the evidence.

Taking all of this into account, the version of the Law of Likelihood appropriate to likelihoodists may be stated as follows.

Special Law of Likelihood:

Given a pair of incompatible hypotheses h_i and h_j that imply statistical models regarding outcomes e^n given $b \cdot c^n$, the likelihoods $P[e^n|h_j \cdot b \cdot c^n]$ and $P[e^n|h_i \cdot b \cdot c^n]$ are well defined. For such likelihoods, *the evidence* ($c^n \cdot e^n$) *supports* h_i *over* h_j , given b , *if and only if* $P[e^n|h_i \cdot b \cdot c^n] > P[e^n|h_j \cdot b \cdot c^n]$; the ratio of likelihoods $P[e^n|h_i \cdot b \cdot c^n] / P[e^n|h_j \cdot b \cdot c^n]$ measures the *strength of the evidence* for h_i over h_j given b .

Notice that when either version of the Law of Likelihood holds, the absolute size of any particular likelihood is irrelevant to the strength of the evidence. All that matters is the relative size of the likelihoods — i.e., the size of their ratio. Here is a way to see the point. Let c_1 and c_2 be the conditions for two different experiments having outcomes e_1 and e_2 , respectively. Suppose that e_1 is 1000 times more likely on hypothesis h_i (given $b \cdot c_1$) than is e_2 on h_i (given $b \cdot c_2$); and suppose that e_1 is also 1000 times more likely on h_j (given $b \cdot c_1$) than is e_2 on h_j (given $b \cdot c_2$) — i.e., suppose that $P_{\alpha}[e_1|h_i \cdot b \cdot c_1] = 1000 \times P_{\alpha}[e_2|h_i \cdot b \cdot c_2]$, and $P_{\alpha}[e_1|h_j \cdot b \cdot c_1] = 1000 \times P_{\alpha}[e_2|h_j \cdot b \cdot c_2]$. Which piece of evidence, $(c_1 \cdot e_1)$ or $(c_2 \cdot e_2)$, is stronger evidence with regard to the comparison of h_i to h_j ? The Law of Likelihood implies both are equally strong. All that matters evidentially are the ratios of the likelihoods, and they are the same in this case:

$$P_{\alpha}[e_1|h_i \cdot b \cdot c_1]/P_{\alpha}[e_1|h_j \cdot b \cdot c_1] = P_{\alpha}[e_2|h_i \cdot b \cdot c_2]/P_{\alpha}[e_2|h_j \cdot b \cdot c_2].$$

Thus, the General Law of Likelihood implies the following principle.

General Likelihood Principle:

Suppose two different experiments or observations (or two sequences of them) c_1 and c_2 produce outcomes e_1 and e_2 , respectively. Let $\{h_1, h_2, \dots\}$ be any set of alternative hypotheses. If there is a constant r such that for each hypothesis h_j from the set, $P_{\alpha}[e_1|h_j \cdot b \cdot c_1] = r \times P_{\alpha}[e_2|h_j \cdot b \cdot c_2]$, then the *evidential import* of $(c_1 \cdot e_1)$ for distinguishing

among hypotheses in the set (given b) is precisely the same as the *evidential import* of $(c_2 \cdot e_2)$.

Similarly, the Special Law of Likelihood implies a corresponding Special Likelihood Principle that applies only to hypotheses that express simple statistical models.²⁹

Bayesians agree with likelihoodists that likelihood ratios completely characterize the extent to which *the evidence* favors one hypothesis over another (as shown by Equations 9-11). So they agree with the letter of the Law of Likelihood and the Likelihood Principle. Furthermore, Bayesian confirmationists may agree that it's important to keep likelihoods separate from other factors, such as prior probabilities, in scientific reports about the evidence. However, Bayesians go further than most likelihoodists in finding a legitimate role for prior plausibility assessments to play in the full evaluation of scientific hypotheses. They propose to combine a measure of the impact of evidence (couched in terms of ratios of likelihoods) with a measure of the plausibility of hypotheses based on all other relevant factors (couched in terms of ratios of prior probabilities) to yield a probabilistic measure of the net confirmation of each hypothesis (its posterior probability).

Throughout the remainder of this article I will not assume that likelihoods must be based on simple statistical hypotheses, as *likelihoodist* would have them. However, most of what will be said about likelihoods, including the convergence results in section 5 (which only involved likelihoods), applies to the *likelihoodist* conception of likelihoods as well. We'll continue for now to take the likelihoods with which we are dealing to be *objective* in the sense that all members of the scientific community agree on their numerical values. In section 6 we'll see how to extend this approach to cases where the likelihoods are less objectively determinate.

3.5 The Representation of Vague and/or Diverse Prior Probabilities

Given that a scientific community should largely agree on the values of the likelihoods for hypotheses, any significant disagreement regarding the posterior probabilities of hypotheses should derive from disagreements over prior probabilities. The community is *diverse* with respect to the range of confirmational probability functions they employ. Furthermore, individual agents may not be able to specify precisely how plausible they consider a hypothesis to be; so their prior probabilities for hypotheses may be *vague*. Both *diversity* due to disagreement among agents and *vagueness* for each individual agent can be represented by sets of confirmation functions, $\{P_\beta, P_\delta, \dots\}$, that agree on the likelihoods, but encompass a range of values for the prior plausibilities of hypotheses. *Diversity* and *vagueness* are different issues, but they may be represented in much the same way. We consider each in turn.

²⁹The Law of Likelihood and the Likelihood Principle have been formulated in somewhat different ways by various logicians and statisticians. R.A. Fisher [1922] argued for the Likelihood Principle early in the 20th century, though he didn't call it that. One of the first places it is discussed under that name is in [Savage *et al.*, 1962]. The Law of Likelihood was first called by that name by Hacking [1965], and has been invoked more recently by the *likelihoodist* statisticians A.F.W. Edwards [1972] and R. Royall [1997].

An individual's assessments of the non-evidential plausibilities of hypotheses will often be vague — not subject to the kind of precise quantitative treatment that a probabilistic logic of confirmation may seem to require for prior probabilities. So it is sometimes objected that the kind of assessment of prior probabilities required to get the Bayesian appraisal going cannot be had in practice. Bayesian confirmation theory has a way of addressing this worry. An agent's vague assessments of prior plausibilities may be represented by a collection of probability functions, a *vagueness set*, which covers the range of plausibility values that the agent finds acceptable. Notice that if accumulating evidence drives the likelihood ratios to extremes, the range of functions in the agent's *vagueness set* will be driven to near agreement, near 0 or 1, for values for posterior probabilities of hypotheses. Thus, as evidence accumulates, the agent's vague initial plausibility assessments may transform into quite sharp posterior probabilities that indicate the strong refutation or support of the various hypotheses. Intuitively this seems like quite a reasonable way for the logic to work.

The various agents in a community may widely disagree over the non-evidential plausibilities of hypotheses. Bayesian confirmation theory may represent this kind of *diversity* across the community of agents as a collection of all functions in the agents' *vagueness sets*. Let's call such a collection a *diversity set*. So, while there may well be disagreement among agents regarding the prior plausibilities of hypotheses, and while individual agents may only have vague priors, the logic of probabilistic confirmation may readily represent this feature. Furthermore, if accumulating evidence drives the likelihood ratios to extremes, the range of functions in a *diversity set* will come to near agreement on sharp values, near 0 or 1, for posterior probabilities of hypotheses. So, not only can such evidence *firm up* each agent's vague initial plausibilities, it also brings the whole community into agreement on the near refutation of some alternative hypotheses and on the strong support of others.

Under what conditions might the likelihood ratios go to such extremes as evidence accumulates, effectively washing out vagueness and diversity? The *Likelihood Ratio Convergence Theorem* (discussed in detail in section 5) implies that if a true hypothesis disagrees with false alternatives on the likelihoods of possible outcomes for a long enough (or a strong enough) stream of experiments or observations, then that evidence stream will very probably produce *actual outcomes* that drive the likelihood ratios of false alternatives as compared to the true hypothesis to approach 0. As this happens, almost any range of prior plausibility assessments will be driven to agreement on the posterior plausibilities for hypotheses. Thus, the accumulating evidence will very probably bring all confirmation functions in the *vagueness* and *diversity sets* for a community of agents to near agreement on posterior plausibility values — near 0 for the false competitors, and near 1 for the true hypothesis.

One more point about prior probabilities and Bayesian convergence is worth noting. Some subjectivist versions of Bayesianism seem to suggest that an agent's prior plausibility assessments for hypotheses should stay fixed once and for all, and

that the only plausibility updating should be brought about via the likelihoods in accord with Bayes' Theorem. Critics argue that this is unreasonable. The members of a scientific community may quite legitimately revise their "prior", non-evidential plausibility assessments for hypotheses from time to time as they rethink plausibility arguments and bring new considerations to bear. This seems a natural part of the conceptual development of a science. It turns out that such reassessments of priors poses no difficulty for Bayesian confirmation theory. Reassessments may sometimes come about by the addition of explicit statements that supplement or modify the background information b . Or they may take the form of (non-Bayesian) transitions to new *vagueness sets* for individual agents and to new *diversity sets* for the community. The *logic* of Bayesian confirmation theory places no restrictions on how values for prior plausibility assessments might change. Provided that the series of reassessments of prior plausibilities doesn't push the non-evidential plausibility assessments of the true hypothesis ever nearer to zero, the *Likelihood Ratio Convergence Theorem* implies that the evidence will very probably bring the posterior probabilities of the empirically distinct rivals to approach 0 via decreasing likelihood ratios; and as this happens, the posterior probability of the true hypothesis will head towards 1.

4 WHAT IS CONFIRMATIONAL PROBABILITY ANYWAY?

If confirmation functions aren't some sort of normative guide to what we may legitimately believe, then they are useless, and probabilistic confirmation theory is a pointless enterprise. Ideally a confirmation function should be a kind of truth-indicating index. When things are working right, evidence should induce a confirmation function to indicate the falsehood of false hypotheses by tagging them with *confirmational probability numbers* near 0, and it should eventually induce the confirmation function to indicate a true hypothesis by tagging it with *confirmation numbers* that approach 1. Provided a confirmation function has this truth-indicating feature, it makes good epistemic sense for *degree of confirmation* to influence *belief strength*. But exactly how is this supposed to work. Precisely how is the *degree of confirmation* of a hypothesis supposed to hook up with one's *level of confidence* or *degree of belief* in its truth or falsehood? Views about the nature of confirmation functions, about what they really are, should be sensitive to this question. A theory of confirmation that cannot reasonably tie confirmation to appropriate belief provides only a useless contrivance.

4.1 *Against Syntactic-Structural Versions of a Logical Interpretation: Grue-Hypotheses*

Some Bayesian logicians maintain that confirmation is *logical* in the same way that deductive logic is *logical*, and that it should play an analogous role in informing the beliefs of agents. Without going into a lot of technical detail, let's get a basic handle on what these logical confirmation functions are supposed to be like.

A leading idea is that the posterior probabilities of hypotheses should be determined by logical structure alone. The idea of basing probabilities on syntactic structure may seem plausible enough in the case of *likelihoods* that deductively or statistically relate hypotheses to the evidence. So, if logical form could also be made to determine the values of prior probabilities, then the logic of confirmation would be *fully formal* in the same way that deductive logical entailment is formal — i.e., it would be based only on the syntactic structure of the sentences involved. Such confirmation functions would be *logical probabilities* in the sense that their values would be uniquely specified by the syntactic structures of the sentences of the language. A sufficiently rigorous version of this approach would specify a uniquely best way of assigning logically appropriate priors to hypotheses, resulting in a single uniquely best logical confirmation function. This confirmation function would be completely objective in that it would not be influenced by anyone's subjective assessments of plausibilities of various hypotheses.

Keynes and Carnap each tried to implement this kind of approach through syntactic versions of the principle of indifference. The idea is that hypotheses that share the same syntactic structure should be assigned the same prior probability values. Carnap showed how to carry out this project in detail, but only for extremely simple formal languages. Most logicians now take this project to have failed because of a fatal flaw in the whole idea that reasonable prior probabilities can be made to depend on logical form alone. Semantic content should matter. Goodmanian *grue*-predicates provide one way to illustrate this point.³⁰

Call an object *grue* at a given time *just in case* “either the time is earlier than the first second of the year 2030 and the object is green, or the time not earlier than the first second of 2030 and the object is blue”. Now the statement ‘All emeralds are *grue* (at all times)’ has the same syntactic structure as ‘All emeralds are green (at all times)’. So, if syntactic structure determines priors, then these two hypotheses should have the same prior probabilities. Indeed, both should have prior probabilities approaching 0. For, there are an infinite number of competitors of these two hypotheses, each sharing the same syntactic structure: consider the hypotheses ‘All emeralds are *grue_n* (at all times)’, where an object is *grue_n* at a given time *just in case* “either the time is earlier than the first second of the *nth* day after January 1, 2030, and the object is green, or the time is not earlier than the first second of the *nth* day after January 1, 2030, and the object is blue.” A purely syntactic specification of the priors should assign all of these hypotheses the same prior probability. But these are mutually exclusive hypotheses; so their prior probabilities must sum to a value no greater than 1. The only way this can happen is for ‘All emeralds are green’ and each of its *grue_n* competitors to have prior probability values equal to 0. In that case the green hypothesis can never receive a posterior probability above 0.

One might object that the predicate ‘*grue*’ is defined in terms of ‘green’, and so hides the extra syntactic complexity. But from a purely formal, syntactic point of

³⁰ Goodman [1955] introduced predicates of the following sort as a challenge to inductive logic. However, the details of my example and the use to which I’ll put it differs from Goodman’s.

view (which is all this view is entitled to), the predicates we happen to actually employ are only an accident of the language we happen to speak. We could have spoken the *grue*-language, where ‘*grue*’ is the more primitive predicate, where the predicate ‘*green*’ is defined, and hides the extra syntactic complexity. Here’s how to spell out this point in detail. Suppose the *grue*-language also contains a predicate ‘*bleen*’ which, translated to our usual language works like this: an object is *bleen* at a given time *just in case* “either the time is earlier than the first second of the year 2030 and the object is blue, or the time not earlier than the first second of 2030 and the object is green”. Now, it is easy to show that from the perspective of the *grue*-language our predicate ‘*green*’ is *defined* as follows: an object is green at a given time *just in case* “either the time is earlier than the first second of the year 2030 and the object is *grue*, or the time not earlier than the first second of 2030 and the object is *bleen*”. (‘*Blue*’ may be similarly defined). The point is that from a purely logical perspective there is no reason to prefer one set of primitive predicates over another. Presumably part of the mission of confirmation theory is to discover what hypotheses, couched in terms of what primitive predicates, best describe the world. The syntactic-structural view attempts to avoid prejudicing the confirmatory process by assigning prior probabilities in a logically/syntactically unbiased way. The *grue* example shows that this can’t work. If you pick a preferred set of predicates, you build in a bias. If you don’t pick a preferred set, then all of the *grue*-like hypotheses must be given equal footing to the green hypothesis. But then all prior probabilities must be 0 (or so close to 0 that no significant amount of confirmation can occur).

Even if some version of the syntactic-structural approach could be made to work, its advocates would still owe us an account of how, and why, such confirmation functions should inform our belief strengths for various hypotheses. In particular, for cases where the evidence is not yet sufficient to strongly favor one specific hypothesis over an alternative (i.e. where the likelihood ratio is near 1), why should an agent’s *belief strength* (or *level of confidence*) be governed by the *syntactic structure* of these hypotheses, rather than by their (*semantic*) *meanings* together with whatever plausibility considerations make the most sense to the scientific community? The defenders of the syntactic-structural view owe us credible reasons to conform belief to their confirmation functions.

4.2 Against the Subjective Belief-Function Interpretation: the Problem of Old Evidence

The *subjectivist* or *personalist* Bayesian view solves the problem of how confirmation is supposed to influence belief in the most direct way possible. It says that the agent’s confirmation function P_α should just *be* his belief function, Bel_α , where Bel_α is a probability function that measures how confident the agent is (or should be) in the truth of various statements. Belief is, of course, dynamic. We learn new truths, including evidence claims. On the subjectivist account, upon learning new evidence e , an agent α is supposed to update his belief strengths

via Bayesian conditionalization: for all sentences S (including the hypotheses h_j), $Bel_{\alpha:new}[S] = Bel_{\alpha:old}[S|e]$. This is where Bayes' Theorem comes in. When sentence S is a hypothesis h_i , we have (from combining Equations 10 and 11, but suppressing 'c' and 'b', as subjectivists usually do):

$$Bel_{\alpha:new}[h_i] = Bel_{\alpha:old}[h_i|e] = 1 / \left(1 + \sum_{j \neq i} \frac{Bel_{\alpha:old}[e|h_j]}{Bel_{\alpha:old}[e|h_i]} \times \frac{Bel_{\alpha:old}[h_j]}{Bel_{\alpha:old}[h_i]} \right)$$

(where the catch-all hypothesis, if needed, is included among the hypotheses h_j). This is how subjectivist Bayesians employ Bayes' Theorem to represent the updating of belief strengths on new evidence.

Formally this account works just fine. However there are reasons for thinking that *confirmation functions* must be distinct from subjectivist or personalist *degree-of-belief functions*. One such reason is the *problem of old evidence*.³¹ To understand the problem we need to first consider more carefully what *belief functions* are supposed to represent.

Bayesian belief functions are supposed to provide an idealized model of belief strengths for agents. They extend the notion of ideally consistent belief to a probabilistic notion of ideally coherent belief strengths. I have no objection to this kind of idealization as a normative guide for real decision making. An agent is supposed to make decisions based on her belief strengths about the state of the world, her belief strengths about possible consequences of her actions, and her assessment of the desirability (or utility) of these consequences. But the very role that *belief functions* are supposed to play in decision making makes them ill-suited to hypothesis confirmation, where the likelihoods are often supposed to be objective, or at least possess intersubjectively agreed values that represent the empirical import of hypotheses. That is, for the purposes of decision making, degree-of-belief functions *should* represent the agent's belief strengths *based on everything she presently knows*. But then the degree-of-belief likelihoods must represent how strongly the agent would believe the evidence if a hypothesis h_i were added to *everything else she presently knows*. This makes them quite different than confirmation function likelihoods, which represent what the hypothesis (together with explicit background and experimental conditions) *says* or *implies* about the evidence. In particular, *degree-of-belief* likelihoods are saddled with a version of the *problem of old evidence*, a problem not shared by confirmation function likelihoods.

Here is the problem. An evidence statement e may be well-known far in advance of the time when we first attempt to account for it with some new hypothesis or theory. For example, the rate of advance in Mercury's perihelion was known well before Einstein developed the theory of General Relativity, and then figured out how the theory could account for that phenomenon. If the agent is already certain of an evidence statement e before using e to test a hypothesis, then her *belief-function likelihoods* for e must have value 1, so the *belief-function likelihood* must

³¹Glymour (1980) first raised this problem. Eells (1985) extends the problem. For a more extensive version of the following treatment see (Hawthorne 2005).

also be 1 on each alternative hypothesis. That is, if $Bel_{\alpha:old}$ is α 's *belief function* and she already knows that e , then $Bel_{\alpha:old}[e] = 1$. It then follows from the axioms of probability theory that $Bel_{\alpha:old}[e|h_i] = 1$ as well, regardless of what h_i says, indeed even if h_i says that e is very unlikely.

This problem runs even deeper. It not only applies to evidence that the agent *knows with certainty*. It turns out that almost anything the agent learns that could influence how strongly she believes e will also influence the value of her *belief-function likelihood* for e , because a belief function $Bel_{\alpha:old}[e|h_i]$ represents the agent's belief strength *given everything she knows*.

To see the difficulty with less-than-certain evidence, consider the following example (where I'll continue to suppress the 'b' and 'c' terms.) A physician intends to use a treadmill test to find evidence about whether her patient has heart disease, h . She knows from medical studies that there is a 10% false negative rate for this test; so her belief strength for a negative result, e , given heart disease is present, h , is $Bel_{\alpha:old}[e|h] = .10$. Now, her nurse is very professional and is usually unaffected by patients' test results. So, if asked, the physician would say her belief strength that her nurse will "feel awful about it", s , "if the test is positive" (i.e. if $\sim e$) is around $Bel_{\alpha:old}[s|\sim e] = .05$. Let us suppose, as seems reasonable, that this belief strength is independent of whether h is in fact true — i.e. $Bel_{\alpha:old}[s|\sim e \cdot h] = Bel_{\alpha:old}[s|\sim e]$. The nurse then tells the physician, in a completely convincing way, "if *his* test comes out positive, I'll feel just awful about it." The physician's new belief function likelihood for a false negative must then become $Bel_{\alpha:new}[e|h] = Bel_{\alpha:old}[e|h \cdot (\sim e \supset s)] = .69$.³² Now, if a negative test result comes back from the lab, what likelihood is the physician supposed to use in her evaluation of the patient's prospects for having heart disease, her present personal belief-function likelihood, $Bel_{\alpha:new}[e|h] = .69$, or the "real" false-negative rate likelihood, $P[e|h] = .10$?

The main point is that even the most trivial knowledge of conditional (or disjunctive) claims involving e may completely upset the objective values of likelihoods for an agent's belief function. And an agent will almost always have some such trivial knowledge. E.g., the physician in the previous example may also learn that if the treadmill test is negative for heart disease, *then*, (1) the patient's worried mother will throw a party, (2) the patient's insurance company won't cover additional tests, (3) it will be the thirty-seventh negative treadmill test result she has received for a patient this year, ..., etc. Updating on such conditionals can force the physician's *belief function likelihoods* to deviate widely from the evidentially objective, textbook values for likelihoods.

More generally, it can be shown that the incorporation into Bel_{α} of almost any kind of evidence for or against the truth of a prospective evidence claim e — even uncertain evidence for e , as may come through Jeffrey updating³³ — completely

³²Since $Bel_{\alpha:old}[e|h \cdot (\sim e \supset s)] = Bel_{\alpha:old}[\sim e \supset s|h \cdot e] \times Bel_{\alpha:old}[e|h] / (Bel_{\alpha:old}[\sim e \supset s|h \cdot e] \times Bel_{\alpha:old}[e|h] + Bel_{\alpha:old}[\sim e \supset s|h \cdot \sim e] \times Bel_{\alpha:old}[\sim e|h]) = Bel_{\alpha:old}[e|h] / (Bel_{\alpha:old}[e|h] + Bel_{\alpha:old}[s|\sim e \cdot h] \times Bel_{\alpha:old}[\sim e|h]) = .1 / (.1 + (.05)(.9)) = .69$.

³³See Jeffrey [1965; 1987; 1992].

undermines the objective or intersubjectively agreed likelihood values that a belief function might have otherwise expressed.³⁴ This should be no surprise. The agent's belief function likelihoods reflect her *total degree of belief* in e , based on h together with *everything else she knows* about e . So the agent's present belief function may capture objective likelihoods for e only if the possibility of the truth of e can be completely isolated from all of the agent's other beliefs. And this will rarely be the case.

One Bayesian subjectivist response to this kind of problem is that the belief functions employed in scientific inferences should often be "counterfactual belief functions", which represent what the agent *would believe* if e were subtracted (in some suitable way) from everything else she knows (see, e.g. [Howson and Urbach, 1993]). However, our example shows that merely subtracting e won't do. One must also subtract all conditional and disjunctive statements containing e . And one must subtract any uncertain evidence for or against e as well. So the counterfactual belief function idea needs a lot of working out if it is to rescue the idea that subjectivist Bayesian belief functions can provide a viable account of the likelihoods employed by the sciences.

There is important work for the *degree of belief* notion to do as part of our best formal account of belief and decision. But degree-of-confirmation functions, associated with objective or public likelihoods, do different work. It seems that confirmation functions should help guide changes in belief, but they should not themselves be agents' belief functions. Taking probabilistic confirmation functions to be degree-of-belief functions, even counterfactual ones, forces the *degree of belief* conception into a mold that doesn't suit it given the other work it does. Better to keep these two notions distinct, and connect them with an account of how *degree-of-confirmation* should inform *degree-of-belief*.

4.3 *How Confirmational Support should influence Belief Strength: the Truth-Index Interpretation*

Rather than ask what a confirmation function *is*, perhaps it's more fruitful to ask what a confirmation function is supposed to do. That is, I want to suggest a kind of *functionalist* view of the nature of confirmation functions. You might call this the *they-are-what-they-do* interpretation. But what is a confirmation function designed to do? What is its functional role supposed to be?

As I see it, a confirmation function is supposed to be a kind of *truth-indicating index*. It can be expected to successfully perform this function when things are working right. That is, when things are working right, an increasing stream of evidence will induce a confirmation function to indicate the *falsehood* of false hypotheses by tagging them with *confirmational probability numbers* approaching 0, and the evidence stream will eventually induce the confirmation function to indicate the *truth* of a true hypothesis by tagging with it *confirmational probability*

³⁴See [Hawthorne, 2005] for more details.

numbers that approach 1. What does it take for “things to work right”? Although spelling this out is not completely trivial, it’s not as daunting as one might think.

If, among the alternative hypotheses proposed to account for a given subject-matter we are fortunate enough to think up a hypothesis that happens to in fact be true, and if we find ways to empirically test it against rivals, then all that’s needed for success is persistence and not too much bad luck with how the evidence actually turns out. For, according to the Likelihood Ratio Convergence Theorem (section 5), the true hypothesis itself *says*, via its likelihoods, that a long enough (but finite), stream of observations or experiments is very likely to produce outcomes that will drive the likelihood ratios that compare empirically distinct false competitors to the true hypothesis to approach 0. As this happens, the confirmation index of these competitors, as measured by their posterior probabilities, also approaches 0, and the confirmation index of the true hypothesis (or at least its disjunction with empirically equivalent rivals) will approach 1.

One must read this result carefully, however. The result does not imply that whatever hypothesis has index near 1 at a given moment is likely to be the true alternative. The convergence theorem doesn’t say that. Rather, the result suggests the pragmatic strategy of continually testing hypotheses, and taking whichever of them has an index near 1 (if there is one) as the *best current candidate* for being true. The convergence theorem implies that maintaining this strategy of continual testing is very likely to eventually promote the true hypothesis (or its disjunction with empirically indistinguishable rivals) to the status of *best current candidate*, and maintain it there. Thus, this strategy is very likely to eventually produce the truth for us. But notice, the theorem doesn’t imply that we’ll ever be in a position to justly be certain that our *best current candidate* is the true alternative. Thus, this eliminative strategy promises to work *only if* we continue to look for rivals and continue to test the best alternative candidates against them. This strategy shouldn’t seem novel or surprising. It’s merely a rigorously justified version of scientific common sense.

When the kind of empirical evidence that’s related to hypotheses via likelihoods is too meager to distinguish between a particular pair of hypotheses, the confirmation index must rely on whatever our most probative “non-evidential” considerations can tell us. We often have good reasons besides the evidence from likelihoods to discount some logically possible alternatives as *just too implausible*, or as significantly less plausible than some better conceived competitors. Indeed, we always bring some such considerations to bear, at least implicitly. For, given any specific hypothesis, logicians know how to easily cook up numerous alternatives that agree with it on all the evidence gathered thus far. Any reasonable scientist will reject most of these inventions immediately, because they look *ad hoc*, contrived, or just foolish. Such reasons for rejection appeal to neither purely logical characteristics of these hypotheses, nor to the usual sorts of evidential considerations. Such reasons bring into play *plausibility consideration*, some purely conceptual, some *broadly empirical*. I refer to plausibility considerations as *broadly empirical* when they reflect our experience, but their import cannot be properly captured by

observation conditions c and evidential outcomes e in evidential likelihoods that express what the alternative hypotheses at issue *say* about such evidence. On a Bayesian account, whatever cannot be represented by likelihoods may only be introduced via the “prior” probabilities. They are the conduit through which any legitimate considerations not expressed by likelihoods may be brought to bear in the net evaluation of scientific hypotheses.

This all suggests that the normative connection between confirmation and belief should go something like this:

The Belief-Confirmation Alignment Condition:

Each agent should bring her belief strengths for hypotheses into alignment with their degrees of confirmation due to all the relevant evidence of which she is aware, and where the confirmation function she employs draws on prior probabilities that represent her best estimates of the comparative plausibilities of alternative hypotheses based on all relevant non-evidential considerations of which she is aware. That is, if P_α is her confirmation function, and she is certain of background and auxiliaries b and evidence $c^n \cdot e^n$, and this is the totality of her evidence that is relevant to h_i , then her belief strength Bel_α should be (or become) $Bel_\alpha[h_i] = P_\alpha[h_i|b \cdot c^n \cdot e^n]$.

Furthermore, if she has partial or uncertain evidence that's relevant to h_i , then her belief strength should be the weighted sum of the degrees of confirmation of the hypothesis on *each possible alternative evidence sequence* $c^n \cdot e^n$, weighted by her belief strengths for each of those possible evidence sequences (and similarly for possible alternative auxiliaries b , if they are uncertain), as follows:³⁵

$$Bel_\alpha[h_i] = \sum_{\{b \cdot c^n \cdot e^n\}} P_\alpha[h_i|b \cdot c^n \cdot e^n] \times Bel_\alpha[b \cdot c^n \cdot e^n].$$

The *Alignment Condition* may be difficult for real agents to follow precisely. But it should be a normative guide for real agents (much as Bayesian decision theory is supposed to be a normative guide).³⁶ The *Alignment Condition* merely recommends that a real agent's confidence in scientific hypotheses should conform to the

³⁵If the agent is *certain* of some particular bit of evidence $c_k \cdot e_k$ in the evidence stream, her belief function will assign belief strength 0 to each possible evidence sequence $c^n \cdot e^n$ that fails to contain $c_k \cdot e_k$ — i.e., $Bel_\alpha[b \cdot c^n \cdot e^n] = 0$ for all $c^n \cdot e^n$ that don't contain $c_k \cdot e_k$.

³⁶The idea that Bayesian epistemology should draw on two distinct probability functions in roughly this way was suggested by Carnap (1971). He calls the *degree of belief* notion ‘rational credence’, and he calls the *degree of confirmation* notion ‘credibility’. Carnap takes *initial credence functions* to derive from *credibility functions*, which themselves are taken to be *logical probability functions*. Brian Skyrms largely adopts this Carnapian idea in the third edition of *Choice and Chance* (1986, Ch. 1, Ch. 6, Sects. 7 and 8), though he doesn't identify his version of *credibility functions* with Carnapian *logical probabilities*. Skyrms calls the *degree of belief* notion ‘epistemic probability’, and calls the *degree of confirmation* notion ‘inductive probability’. More recently Marc Lange [1999] also argues for a two function Bayesian model. See [Hawthorne, 2005] for more about the alignment of belief with confirmational support.

level indicated by her confirmation function, moderated by how confident she is in the truth of the evidence claims. It shouldn't be overly difficult for real agents to approximately align belief to confirmation in this way. Furthermore, supposing (as argued earlier) that probabilistic confirmation functions should not *just be* belief functions, the *Alignment Condition* shows how probabilistic confirmation can plausibly be made to mesh with the usual Bayesian account of belief and decision. The *Alignment Condition* is recommended as a norm by the following fact: if the agent comes to strongly believe true evidence statements, then *alignment* takes advantage of *Likelihood Ratio Convergence* to very probably bring the agent to strongly doubt false hypotheses and strongly believe true ones. What better recommendation for the formation of belief strengths about scientific hypotheses could one reasonably expect to have?

5 THE LIKELIHOOD RATIO CONVERGENCE THEOREM

The *Likelihood Ratio Convergence Theorem* shows that when h_i is true and h_j is empirically distinct from h_i on a sequence of experiments and observations c^n , then (provided b is also true) it's *very likely* that a sequence of outcomes e^n will result that yields a sequence of likelihood ratios $P[e^n | h_j \cdot b \cdot c^n] / P[e^n | h_i \cdot b \cdot c^n]$ that approach 0 as the evidence accumulates (i.e., as n increases). The theorem places an explicit bound on the *rate of probable convergence*. That is, it puts a lower bound that approaches 1 *on how likely it is* that, when h_i is true, some stream of outcomes will occur that yields a likelihood ratio within a specified small region of 0 (counting heavily against the truth of alternative h_j).

This convergence theorem draws only on likelihoods. Neither the statement of the theorem nor its proof employs prior probabilities of any kind. Likelihoodists and Bayesian confirmationists agree that when the sequence of likelihood ratios $P[e^n | h_j \cdot b \cdot c^n] / P[e^n | h_i \cdot b \cdot c^n]$ approach 0 for increasing n , the evidence goes strongly against h_j as compared to h_i . So, even likelihoodists, who eschew the use of prior probabilities, may embrace this result.

For Bayesians, the *Likelihood Ratio Convergence Theorem* has the additional implication that the posterior probabilities of empirically distinct false competitors of a true hypothesis are very likely to converge to 0. That's because whenever the ratios $P[e^n | h_j \cdot b \cdot c^n] / P[e^n | h_i \cdot b \cdot c^n]$ approach 0 for increasing n , the Ratio Form of Bayes' Theorem, Equation 9, says that the posterior probability of h_j will also approach 0. The values of prior probabilities only accelerate or retard this process of convergence. This also implies that all confirmation functions in a collection that constitutes a *vagueness set* (that represent the range of vagueness in an agent's assessments of the prior plausibilities of hypotheses) will very likely come to near agreement, all coming to agree that the posterior probability of false alternatives approach 0.³⁷ And as the posterior probabilities of false competitors approach 0,

³⁷The same goes for *diversity sets*, which represent the range of plausibility assessments among members of a scientific community.

the posterior probability of the true hypothesis heads towards 1.

The *Likelihood Ratio Convergence Theorem* avoids or overcomes the usual objections raised against Bayesian convergence results:

- The theorem does not employ *second-order probabilities* — it doesn't rely on assessing the probability of a probability. The theorem only concerns the probability of particular disjunctive sentences that represent possible sequences of outcomes.
- The theorem does not rely on countable additivity (to which some commentators have objected with regard to other convergence results).
- The theorem does not require that evidence consist of sequences of outcomes that, according to the hypotheses, are *identically distributed* (like repeated tosses of a die). The version of the theorem I'll present does, however, suppose that the evidential outcomes in the sequence of experiments or observations are *probabilistically independent* of one another given each hypothesis (or at least that the outcomes can be grouped into clusters that are probabilistically independent of one another). Another version of this theorem (not presented here) applies without supposing probabilistic independence. Nevertheless, I will argue that the sort of probabilistic independence that the present version of the theorem draws on should almost always be present in real scientific contexts.
- The rate of likely convergence of the likelihood ratios is explicitly calculable from the likelihoods specified by individual hypotheses. These convergence rates depend only on finite sequences of experiments or observations. So this theorem overcomes the often repeated objection that Bayesian convergence results may only apply in the infinite long run (when we'll all be long dead).
- The values of prior probabilities for hypotheses need not be "locked in" permanently for the theorem to apply. Indeed, the theorem itself doesn't draw on prior probabilities at all. It only employs likelihoods. However, the convergence of the likelihoods leads directly to the convergence of posterior probabilities; and this convergence of posteriors occurs even if agents reassess the non-evidential plausibilities of hypotheses from time to time, and assign new prior probabilities accordingly.

This last point needs some explanation. It is sometimes objected that Bayesian convergence results only work when prior probabilities are held fixed — that the theorems fall through if an agent is permitted to change her evidence-independent assessments of prior plausibilities from time to time. Critics point out that real agents may quite legitimately change their assessments of the evidence-independent plausibilities of hypotheses, perhaps due to newly developed plausibility arguments, or due to the reassessment of old ones. A Bayesian confirmation theory has to represent such reassessments as non-Bayesian shifts from one confirmation

function (or from one *vagueness* or *diversity set* of confirmation functions) to another. But, critics object, Bayesian convergence theorems always assume that the only dynamic element in the confirmational process is due to the addition of new evidence, brought to bear by the associated likelihoods. This kind of updating of posterior probabilities via Bayes' Theorem is supposed to be the only means of updating available to the Bayesian. Thus, it looks like Bayesian confirmation is severely handicapped as an account of scientific hypothesis evaluation.

However, the *Likelihood Ratio Convergence Theorem* is not subject to this kind of objection. It applies even if agents revise their evidence-independent priors from time to time. The theorem itself only involves the values of likelihoods. Thus, provided that reassessments of prior plausibilities do not push the non-evidential plausibility of the true hypothesis down towards 0 *too rapidly*, the theorem shows that posterior probabilities of the empirically distinct false competitors of a true hypothesis will *very probably* approach 0 as evidence increases.³⁸

I raise these points in advance so that the reader may be on the look-out to see that the theorem really does avoid these challenges.³⁹ Let's now turn to the details.⁴⁰

³⁸That is, for each confirmation function P_α , the posterior $P_\alpha[h_j|b \cdot c^n \cdot e^n]$ must go to 0 if the ratio $P_\alpha[h_j|b \cdot c^n \cdot e^n]/P_\alpha[h_i|b \cdot c^n \cdot e^n]$ goes to 0; and that will occur if the likelihood ratios $P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n]$ approach 0 and the prior $P_\alpha[h_i|b]$ is a bit greater than 0. The Likelihood Ratio Convergence Theorem will show that when $h_i \cdot b$ is true, it is very likely that the evidence will indeed be such as to drive the likelihood ratios as near to 0 as you please (given a long enough evidence stream). As that happens, the only way a Bayesian agent can avoid having his confirmation function yield *posterior probabilities* for h_j that approach 0 (as n gets large) is to continually switch among confirmation functions (moving from P_α to P_β to ... to P_γ to ...) in a way that revises the prior probability of the true hypothesis, h_i , downward towards 0. And even then he can only avoid having the posterior probability for alternative h_j approach 0 by continually switching to new confirmation functions at a rate that keeps the new priors for h_i diminishing towards 0 at least as quickly as the likelihood ratios that disfavor h_j (as compared to h_i) diminish towards 0. To see this, suppose, *to the contrary*, that $P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n]$ approaches 0 *faster* than does sequence $P_\gamma[h_i|b]$, for changing P_γ and increasing n — i.e. *approaches 0 faster* in the sense that $(P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n])/P_\gamma[h_i|b]$ goes to 0, for changing P_γ and increasing n . Then, we'd have $(P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n])/P_\gamma[h_i|b] > (P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n]) \cdot (P_\gamma[h_j|b]/P_\gamma[h_i|b]) = P_\gamma[h_j|b \cdot c^n \cdot e^n]/P_\gamma[h_i|b \cdot c^n \cdot e^n]$. So, the ratios of posterior probabilities $P_\gamma[h_j|b \cdot c^n \cdot e^n]/P_\gamma[h_i|b \cdot c^n \cdot e^n]$ must still go to 0, for changing P_γ and increasing n ; and thus, so must $P_\gamma[h_j|b \cdot c^n \cdot e^n]$.

³⁹My version of the theorem is related to L. J. Savage's (1954) convergence theorem, but generalizes that result considerably. In particular, Savage's theorem supposes that the outcomes are both *independent* and *identically distributed*, whereas the Likelihood Ratio Convergence Theorem (hereafter 'LRCT') only supposes independent outcomes, not identical distribution. Savage's version does not provide bounds on the rate of convergence, while the LRCT will provide explicit lower bounds. And Savage's theorem is stated in terms of the convergence of *posterior probabilities* to 0 or 1 for each of a pair of alternative hypotheses, whereas the LRCT deals directly with *likelihoods* and *likelihood ratios*, and does not involve prior probabilities.

⁴⁰For a nice presentation of the most prominent Bayesian convergence results and a discussion of their weaknesses see [Earman, 1992, Ch. 6]. Earman was not aware of the Likelihood Ratio Convergence Theorem I'll be presenting here.

Among the convergence results discussed by Earman is a well-known result due to Gaifman and Snir [1982] (hereafter 'G&S'). The G&S theorem result is a *strong law of large numbers* result, and so may at first blush appear to be a stronger result than the Likelihood Ratio Convergence Theorem, which is a *weak law of large numbers* result. However, in important respects the LRCT

5.1 The Space of Possible Outcomes of Experimental and Observational Conditions

To spell out the details of the *Likelihood Ratio Convergence Theorem* we'll need a few additional notational conventions and definitions. Here they are.

For a sequence of n experiments or observations c^n , consider the set of those possible sequences of outcomes that would result in likelihood ratios for h_j over h_i that are less than some chosen small number $\varepsilon > 0$. This set is represented by the following expression:

$$\{e^n : P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n] < \varepsilon\}$$

One may choose any small value of ε that seems interesting, and then form the corresponding set. Placing the disjunction symbol ' $\vee$ ' in front of this expression yields an expression

$$\vee\{e^n : P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n] < \varepsilon\},$$

that represents the disjunction of all outcome sequences in this set. So, the expression ' $\vee\{e^n : P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n] < \varepsilon\}$ ' just represents a particular sentence that says, in effect, "one of those sequences of outcomes from the first n experiments or observations will occur that makes the likelihood ratio for h_j over h_i less than ε ."

The *Likelihood Ratio Convergence Theorem* says that, for any specific ε you choose, the likelihood of a disjunctive sentence of this sort, given that ' $h_i \cdot b \cdot c^n$ ' is true,

$$P[\vee\{e^n : P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n] < \varepsilon\} \mid h_i \cdot b \cdot c^n]$$

must have a value of at least $1 - (\psi/n)$, for some explicitly calculable term ψ . And clearly this lower bound, $1 - (\psi/n)$, will approach 0 as n increases. Thus,

is actually the stronger of the two results. For, although the G&S result may arguably have a stronger consequent, it relies on a much stronger antecedent — it supposes that the evidence is *separating* in the sense that, for any pair of possible worlds (or models of the formal language), there is an evidence statement E that is true in one world and false in the other world. This is a very strong assumption, much stronger than anything the LRCT depends on. By contrast, the LRCT supposes nothing stronger than conditions that are commonly present in the testing of real scientific theories.

The consequent of the G&F result shows probabilistic convergence to the truth-value of a hypothesis *almost everywhere* — i.e. on a set of models, or possible worlds, of measure 1. However, as is common with "strong law" results, it says nothing about how fast convergence is likely to be. Thus, the G&F result is open to the "infinite long run" objection mentioned above. By contrast, the LRCT only shows that the likelihood *approaches* 1 that the likelihood ratios comparing false competitors to the true hypothesis *approach* 0 as evidence increases. However, this consequent of the LRCT also provides explicit information about how fast this kind of convergence is likely to be, and it shows exactly how the likely rate of convergence depends on the degree to which likelihood values of possible outcomes differ for the alternative hypotheses. Thus, although the consequent of the LRCT is weaker than the G&S result in one way, it is stronger in another way. See [Earman, 1992] for a detailed critique of the G&S result.

the true hypothesis h_i implies that as the amount of evidence, n , increases, it is highly likely (as close to 1 as you please) that one of the outcome sequences e^n will occur that yields a likelihood ratio $P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n]$ less than ε , for any value of ε you may choose. As this happens, the posterior probability of h_i 's false competitor, h_j , must approach 0, as required by the Ratio Form of Bayes' Theorem.

The term ψ in the theorem depends on a measure of the empirical distinctness of the two hypotheses for the proposed sequence of experiments and observations c^n . To specify this measure we need to contemplate not only the actual outcomes, but the collection of alternative possible outcomes of each experiment or observation. So, consider some sequence of experimental or observation conditions described by sentences $c_1, c_2, \dots, c_n$. Corresponding to each condition c_k there will be some range of possible alternative outcomes; let $O_k = \{o_{k1}, o_{k2}, \dots, o_{kw}\}$ be a set of statements describing the alternative possible outcomes for condition c_k . (The number of alternative outcomes will usually differ for distinct experiments $c_1, \dots, c_n$; so, the value of w depends on each specific c_k). For each hypothesis h_j , the alternative outcomes of c_k in O_k are mutually exclusive and exhaustive — that is, we have:

$$P[o_{ku} \cdot o_{kv} | h_j \cdot b \cdot c_k] = 0 \text{ and } \sum_{u=1}^w P[o_{ku} | h_j \cdot b \cdot c_k] = 1.$$

Expressions like ' e_k ' represent possible outcomes of c_k — i.e., ' e_k ' ranges over the members of O_k . As before, ' c^n ' denotes the conjunction of the first n test conditions, $(c_1 \cdot c_2 \cdot \dots \cdot c_n)$, and ' e^n ' represents possible sequences of corresponding outcomes, $(e_1 e_2 \cdot \dots \cdot e_n)$. We'll take ' E^n ' to represent the set of all possible outcome sequences form conditions c^n . So, for each hypothesis h_j (including h_i), we have $\sum_{e^n \in E^n} P[e^n | h_j \cdot b \cdot c^n] = 1$. There are no substantive assumptions in any of this — only notational conventions.

5.2 About Probabilistic Independence

In almost all scientific contexts the outcomes in a series of experiments or observations are *probabilistically independent* of one another relative to each hypothesis under consideration. We may divide the kind of independence involved into two types.

Definition: Independent Evidence Conditions:

1. A sequence of outcomes e^k is *condition-independent* of a condition for an additional experiment or observation c_{k+1} , given $h \cdot b$ and its own conditions c^k , if and only if $P[e^k | h \cdot b \cdot c^k \cdot c_{k+1}] = P[e^k | h \cdot b \cdot c^k]$.
2. An individual outcome e_k is *result-independent* of a sequence of other observations and their outcomes $(c^{k-1} \cdot e^{k-1})$, given $h \cdot b$ and its own condition c_k , if and only if $P[e_k | h \cdot b \cdot c_k \cdot (c^{k-1} \cdot e^{k-1})] = P[e_k | h \cdot b \cdot c_k]$.

When these two conditions hold, the likelihood for a sequence of experiments or observations may be decomposed into the product of the likelihoods for individual experiments or observations. To see how the two *independence conditions* affect the decomposition, first consider the following formula, which holds even if neither *independence condition* is satisfied:

$$(1) \quad P[e^n | h_j \cdot b \cdot c^n] = \prod_{k=1}^n P[e_k | h_j \cdot b \cdot c^n \cdot e^{k-1}].$$

When *condition-independence* holds, the likelihood of the whole evidence stream parses into a product of likelihoods that *probabilistically depend* on only past observation conditions and their outcomes. They do not depend on the conditions for other experiments whose outcomes are not yet specified. Here is the formula:

$$(2) \quad P[e^n | h_j \cdot b \cdot c^n] = \prod_{k=1}^n P[e_k | h_j \cdot b \cdot c_k \cdot (c^{k-1} \cdot e^{k-1})].$$

Finally, whenever both independence conditions are satisfied we obtain the following relationship between the likelihood of the evidence stream and the likelihoods of individual experiments or observations:⁴¹

$$(3) \quad P[e^n | h_j \cdot b \cdot c^n] = \prod_{k=1}^n P[e_k | h_j \cdot b \cdot c_k].$$

In almost all scientific contexts both clauses of the *Independent Evidence Condition* will be satisfied. To see this, let us consider each independence condition more carefully.

Condition-independence says that the mere addition of a new observation condition c_{k+1} , *without specifying one of its outcomes*, does not alter the likelihood of the outcomes e^k of other experiments c^k . To appreciate the significance of this condition, imagine how the world would be if it were violated. Suppose hypothesis h_j is some statistical theory, say, a quantum theory of superconductivity. The conditions expressed in c^k describe a number of experimental setups, perhaps conducted in numerous labs, that test a variety of aspects of the theory (e.g., experiments that test electrical conductivity in different materials at a range of temperatures). Outcome sequence e^k describes the results of these experiments. The violation of *condition-independence* would mean that merely adding to $h_j \cdot b \cdot c^k$ a statement c_{k+1} describing the set-up of an additional experiment, but with no mention of its outcome, changes how likely the evidence sequence e^k is: i.e., $P[e^k | h \cdot b \cdot c^k \cdot c_{k+1}] \neq P[e^k | h \cdot b \cdot c^k]$. What $(h_j \cdot b)$ *says*, via likelihoods, about the outcomes e^k of experiments c^k differs as a result of merely supplying a description of another experimental arrangement, c_{k+1} . *Condition-independence*, when it holds, rules out such strange effects.

⁴¹For derivations of equations (13) and (14) see [Hawthorne, 2004, supplement 3] at <http://plato.stanford.edu/entries/logic-inductive/supplement3.html>

Result-independence says that the description of previous test conditions *together with their outcomes* is irrelevant to the likelihoods of outcomes for additional experiments. If this condition were widely violated, then in order to specify the most informed likelihoods for a given hypothesis one would need to include information about volumes of past observations and their outcomes. What a hypothesis says about future cases would depend on how past cases have gone. Such *dependence* had better not happen on a large scale. Otherwise, the hypothesis would be fairly useless, since its empirical import in each specific case would depend on taking into account volumes of past observational and experimental results. However, even if such dependencies occur, provided they are not too pervasive, *result-independence* can be accommodated rather easily by packaging each collection of *result-dependent* data together, treating it like a single extended experiment or observation. The *result-independence condition* will then be satisfied by letting each term ' c_k ' in the statement of the independence condition represent a conjunction of test conditions for a collection of *result-dependent* tests, and by letting each term ' e_k ' (and each term ' o_{ku} ') stand for a conjunction of the corresponding *result-dependent* outcomes. Thus, by packaging *result-dependent* data together in this way, the *result-independence* condition is satisfied by those (conjunctive) statements that describe the separate, *result-independent* chunks.⁴²

The version of the *Likelihood Ratio Convergence Theorem* I'll present depends on the usual axioms of probability theory together with the *Independent Evidence Conditions*. It depends on no other assumptions (except those explicitly stated in the antecedent of the theorem itself). Thus, from this point on, let's suppose that the following two assumptions holds.

Independent Evidence Assumptions: For each hypothesis h and background b under consideration, let's assume that the experiments and observations can be packaged into condition statements, $c_1, \dots, c_k, c_{k+1}, \dots$, and possible outcomes in a way that satisfies the following independence conditions:

1. Each sequence of possible outcomes e^k of a sequence of conditions c^k is *condition-independent* of additional conditions c_{k+1} — i.e., $P[e^k | h \cdot b \cdot c^k \cdot c_{k+1}] = P[e^k | h \cdot b \cdot c^k]$.
2. Each possible outcome e_k of condition c_k is *result-independent* of

⁴²In scientific contexts the most prominent kind of case where data may fail to be *result-independent* is where some quantity of past data helps tie down the numerical value of a parameter not completely specified by the hypothesis at issue, where the value of this parameter influences the likelihoods for outcomes of lots of other experiments. Such hypotheses (with their free parameters) are effectively *disjunctions* of more specific hypotheses, where each distinct disjunct is a distinct version of the original hypothesis that has a specific value for the parameter filled in. Evidence that "fills in the value" for the parameter just amounts to evidence that refutes (via likelihood ratios) those alternative more specific, filled-in hypotheses that possess incorrect parameter values. For any specific, filled-in hypotheses, the evidence that bears on whether it has the correct parameter value will be independent of other evidence that relies on the parameter value. So, relative to each of these more specific hypotheses, *result-independence* holds.

sequences of other observations and possible outcomes ($c^{k-1} \cdot e^{k-1}$)
 — i.e., $P[e_k | h \cdot b \cdot c_k \cdot (c^{k-1} \cdot e^{k-1})] = P[e_k | h \cdot b \cdot c_k]$.

We now have all that is needed to begin to state the *Likelihood Ratio Convergence Theorem*. The convergence theorem comes in two parts. The first part applies to only those experiments or observations that have possible outcomes, according to h_i , that alternative h_j says are impossible. The second part of the theorem applies to all other experiments or observations.

5.3 *Likelihood Ratio Convergence under Conditions where Falsifying Outcomes are Possible*

The first part of the *Likelihood Ratio Convergence Theorem* applies whenever some of the experiments or observations in sequence c^n have possible outcomes with non-0 likelihoods on hypothesis h_i , but 0 likelihoods on alternative h_j . Such outcomes are highly desirable. If they occur, the likelihood ratio comparing h_j to h_i will be 0, and h_j will be *falsified*. A *crucial experiment* is a special case of this, the case where, for at least one possible outcome o_{ku} , $P[o_{ku} | h_i \cdot b \cdot c_k] = 1$ and $P[o_{ku} | h_j \cdot b \cdot c_k] = 0$. In the more general case h_i together with b says that one of the outcomes of c_k is at least minimally probable, whereas h_j says that outcome is impossible: $P[o_{ku} | h_i \cdot b \cdot c_k] > 0$ and $P[o_{ku} | h_j \cdot b \cdot c_k] = 0$.

*Likelihood Ratio Convergence Theorem Part 1: The Falsification Theorem:*⁴³ Suppose c^m , a subsequence of the whole evidence sequence c^n , consists of experiments or observations with the following property: there are outcomes o_{ku} of each c_k in c^m *deemed impossible* by $h_j \cdot b$ but *deemed possible* by $h_i \cdot b$ to at least some small degree δ . That is, suppose there is some $\delta > 0$ such that for each c_k in c^m , $P[\vee\{o_{ku} : P[o_{ku} | h_j \cdot b \cdot c_k] = 0\} | h_i \cdot b \cdot c_k] \geq \delta$. Then,

$$P[\vee\{e^n : P[e^n | h_j \cdot b \cdot c^n] / P[e^n | h_i \cdot b \cdot c^n] = 0\} | h_i \cdot b \cdot c^n] =$$

$$P[\vee\{e^n : P[e^n | h_j \cdot b \cdot c^n] = 0\} | h_i \cdot b \cdot c^n] \geq 1 - (1 - \delta)^m$$

which approaches 1 for large m .

In other words, suppose h_i says observation c_k has at least a small likelihood of producing one of the outcomes o_{ku} that h_j says is impossible — that is, $P[\vee\{o_{ku} : P[o_{ku} | h_j \cdot b \cdot c_k] = 0\} | h_i \cdot b \cdot c_k] \geq \delta > 0$. And suppose that some number m of experiments or observations are of this kind. If the number of such observations is large enough, and h_i (together with $b \cdot c^n$) is true, then it is highly likely that one of the outcomes held to be impossible by h_j will occur, and the likelihood ratio of h_j over h_i will then become 0. Bayes' theorem then goes on to

⁴³For proof see [Hawthorne, 2004] <http://plato.stanford.edu/entries/logic-inductive/supplement4.html>.

imply that when this happens, h_j is absolutely refuted — its posterior probability becomes 0.

The Falsification Theorem is very commonsensical. First, notice that when there is a *crucial experiment* in the evidence stream, the theorem is completely obvious. That is, suppose for the specific experiment c_k (in evidence stream c^n) there are two incompatible possible outcomes o_{kv} and o_{ku} such that $P[o_{kv}|h_j \cdot b \cdot c_k] = 1$ and $P[o_{ku}|h_i \cdot b \cdot c_k] = 1$. Then, clearly, $P[\vee\{o_{ku} : P[o_{ku}|h_j \cdot b \cdot c_k] = 0\} | h_i \cdot b \cdot c_k] = 1$, since o_{ku} is “one of the o_{ku} such that $P[o_{ku}|h_j \cdot b \cdot c_k] = 0$ ”. So where there is a crucial experiment available, the theorem applies with $m = 1$ and $\delta = 1$.

The theorem is equally commonsensical when there is no crucial experiment. To see what it says in such cases, consider an example. Let h_i be some theory that implies a specific rate of proton decay, but a rate so low that there is only a very small probability that any particular proton will decay in a given year. Consider an alternative theory h_j that implies that protons *never* decay. If h_i is true, then for a persistent enough sequence of observations (i.e., if proper detectors can be built and billions of protons kept under observation for long enough), eventually a proton decay will almost surely be detected. When this happens, the likelihood ratio becomes 0. Thus, the posterior probability of h_j becomes 0.

It may be instructive to plug some specific numbers into the formula given by the Falsification Theorem, to see what the convergence rate might look like. For example, the theorem tells us that if we compare any pair of hypotheses h_i and h_j on an evidence stream c^n that contains at least $m = 19$ observations or experiments, each having $\delta \geq .10$ for the likelihood of yielding a *falsifying outcome*, then the likelihood (on $h_i \cdot b \cdot c^n$) of obtaining an outcome sequence e^n that will yield a likelihood ratio $P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n] = 0$ must be least $1 - (1 - .1)^{19} = .865$.

A comment about the *need for*, and *usefulness of* such convergence theorems is in order, now that we’ve seen one. Given some specific pair of scientific hypotheses h_i and h_j , one may always directly compute the likelihood, given $(h_i \cdot b \cdot c^n)$, that any specific sequence of experiments or observations c^n will result in one of the specific sequences of outcomes that yields low likelihood ratios. So, given a specific pair of hypotheses and a proposed sequence of experiments, we don’t need a general *Convergence Theorem* to tell us the likelihood of obtaining refuting evidence. The specific hypotheses h_i and h_j tell us this *themselves*. Indeed, they tell us the likelihood of obtaining each specific outcome stream, including those that refute the competitor or produce a very small likelihood ratio for it. Furthermore, after we’ve actually performed an experiment and recorded its outcome, all that matters is the actual ratio of likelihoods for that outcome. Convergence theorems become moot.

The point of Likelihood Ratio Convergence Theorem (both the Falsification Theorem, and the other part of the theorem, still to come) is to assure us *in advance of the consideration of any specific pair of hypotheses* that if the possible evidence streams that test hypotheses have certain characteristics that reflect the empirical distinctness of the hypotheses, then it is highly likely that one of the sequences of outcomes will occur that results in a very small likelihood ratio.

These theorems provide relatively loose, finite lower bounds on how quickly such convergence is likely to be. Thus, the only point of such convergence theorems is to assure us, in advance of our using the logic of confirmation to test specific hypotheses, that this logic is likely to do what we want it to do — i.e., to result in the refutation of empirically distinct false alternatives to the true hypothesis, and to generate a high degree of positive confirmation for the true hypothesis.

5.4 Likelihood Ratio Convergence under Conditions where No Falsifying Outcomes are Possible

The Falsification Theorem shows what happens when the evidence stream includes possible outcomes that may *falsify* the alternative hypothesis. But what if no *possibly falsifying outcomes* are present? That is, what if hypothesis h_j only specifies various non-zero likelihoods for possible outcomes? Or what if h_j does specify 0 likelihoods for some outcomes, but only for those that h_i says are impossible as well? Such evidence streams are undoubtedly much more common in practice than those containing possibly falsifying outcomes. To cover evidence streams of this kind we first need to identify a useful way to measure the *degree* to which hypotheses are empirically distinguishable by such evidence.

Consider some particular sequence of outcomes e^n that results from observations c^n . The likelihood ratio $P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n]$ measures the extent to which that outcome sequence distinguishes between h_i and h_j . But, as a measure of the power of evidence to distinguish among hypotheses, likelihood ratios themselves provide a rather lopsided scale, a scale that ranges from 0 to infinity with the midpoint, the point where e^n doesn't distinguish at all between h_i and h_j , at 1. So, rather than using raw likelihood ratios to measure the ability of e^n to distinguish between hypotheses, it proves more useful to employ a symmetric measure. The logarithm of the likelihood ratio provides just such a measure.

Definition: QI — the Quality of the Information.

For each experiment or observation c_k , define the *quality of the information* provided by possible outcome o_{ku} for distinguishing h_j from h_i , given b , as follows (where we take the log to be base 2):

$$QI[o_{ku}|h_i/h_j|b \cdot c_k] = \log[P[o_{ku}|h_i \cdot b \cdot c_k]/P[o_{ku}|h_j \cdot b \cdot c_k]].$$

Similarly, define $QI[e^n|h_i/h_j|b \cdot c^n] = \log[P[e^n|h_i \cdot b \cdot c^n]/P[e^n|h_j \cdot b \cdot c^n]]$.

We measure the *Quality of the Information* an outcome would yield in distinguishing between two hypotheses as the base-2 logarithm of the likelihood ratio. This is clearly a measure of the outcome's evidential strength at distinguishing between the two hypotheses. By this measure, two hypotheses, h_i and h_j , assign the same likelihood value to a given outcome o_{ku} just in case $QI[o_{ku}|h_i/h_j|b \cdot c_k] = 0$. Furthermore, because the log is base 2, when $P[o_{ku}|h_i \cdot b \cdot c_k]/P[o_{ku}|h_j \cdot b \cdot c_k] = 2^r$, $QI[o_{ku}|h_i/h_j|b \cdot c_k] = r$, for any real number r . So, QI measures information on a

logarithmic scale that is symmetric about the natural no-information midpoint, 0. Positive information ($r > 0$) favors h_i over h_j and negative information ($r < 0$) favors h_j over h_i .

Given the *Independent Evidence Assumptions* it is easy to see that relative to each hypothesis (with background), $h_i \cdot b$ and $h_j \cdot b$, the QI for a sequence of outcomes is just the sum of the QIs of the individual outcomes in the sequence:

$$(4) \quad \text{QI}[e^n|h_i/h_j|b \cdot c^n] = \sum_{k=1}^n \text{QI}[e_k|h_i/h_j|b \cdot c_k].$$

QI only measures the amount by which each specific outcome counts for or against the two hypotheses. But what we want to know is something about how the experiment or observation as a whole tends to produce distinguishing outcomes. The *expected value* of QI turns out to be very helpful in this regard. The *expected value* of a quantity is gotten by first multiplying each of its *possible values* by its probability of occurring, and then summing up these products. Thus, the *expected value* of QI is given by the following formula:

Definition: EQI — the *Expected Quality of the Information*.

Let's call h_j *outcome-compatible* with h_i on evidence stream c^k just when for each possible outcome sequence e^k of c^k , if $P[e^k|h_i \cdot b \cdot c^k] > 0$, then $P[e^k|h_j \cdot b \cdot c^k] > 0$. We also adopt the convention that if $P[o_{ku}|h_j \cdot b \cdot c_k] = 0$, then the term $\text{QI}[o_{ku}|h_i/h_j|b \cdot c_k] \cdot P[o_{ku}|h_i \cdot b \cdot c_k] = 0$, since the outcome o_{ku} has 0 probability of occurring given $h_i \cdot b \cdot c_k$.

For h_j *outcome-compatible* with h_i on c^k , define

$$\text{EQI}[c_k|h_i/h_j|b] = \sum_u \text{QI}[o_{ku}|h_i/h_j|b \cdot c_k] \times P[o_{ku}|h_i \cdot b \cdot c_k].$$

Also, define $\text{EQI}[c^n|h_i/h_j|b] = \sum_{e^n \in E^n} \text{QI}[e^n|h_i/h_j|b] \times P[e^n|h_i \cdot b \cdot c^n]$.

Notice that whenever h_j is not *outcome-compatible* with h_i on evidence stream c^m , Part 1 of the *Likelihood Ratio Convergence Theorem*, the *Falsification Theorem* applies.

The EQI of an experiment or observation is the *Expected Quality of its Information* for distinguishing h_i from h_j when h_i is true. It is a measure of the expected evidential strength of the possible outcomes of an experiment or observation at distinguishing between the hypotheses. Whereas QI measures the ability of each particular outcome or sequence of outcomes to empirically distinguish hypotheses, EQI measures the tendency of experiments or observations to produce distinguishing outcomes. EQI tracks empirical distinctness in a very precise way, as we'll see in a moment.

The EQI for a sequence of observations c^n turns out to be just the sum of the EQIs of the individual observations c_k in the sequence.⁴⁴

⁴⁴For a derivation see [Hawthorne, 2004] <http://plato.stanford.edu/entries/logic-inductive/supplement5.html>.

$$(5) \quad \text{EQI}[c^n|h_i/h_j|b] = \sum_{k=1}^n \text{EQI}[c_k|h_i/h_j|b]$$

This suggests that it may be useful to average the values of the $\text{EQI}[c_k|h_i/h_j|b]$ over the number of observations n . We then obtain a measure of the *average expected quality of the information* from the experiments and observations that make up c^n .

Definition: $\underline{\text{EQI}}$ — The Average Expected Quality of Information.

The average expected quality of information, $\underline{\text{EQI}}$, from c^n for distinguishing h_j from h_i , given $h_i \cdot b$, is defined as:

$$\underline{\text{EQI}}[c^n|h_i/h_j|b] = \text{EQI}[c^n|h_i/h_j|b] / n.$$

This definition together with equation (16) yields the following:

$$(6) \quad \underline{\text{EQI}}[c^n|h_i/h_j|b] = (1/n) \times \sum_{k=1}^n \text{EQI}[c_k|h_i/h_j|b]$$

It turns out that the value of $\text{EQI}[c_k|h_i/h_j|b]$ cannot be less than 0; and it will be greater than 0 just in case h_i is *empirically distinct* from h_j on at least one outcome o_{ku} — i.e., just in case for at least one o_{ku} , $P[o_{ku}|h_i \cdot b \cdot c_k] \neq P[o_{ku}|h_j \cdot b \cdot c_k]$. The same goes for the average, $\underline{\text{EQI}}[c^n|h_i/h_j|b]$.

*Theorem: Nonnegativity of $\underline{\text{EQI}}$.*⁴⁵

$\text{EQI}[c_k|h_i/h_j|b] \geq 0$; and, $\text{EQI}[c_k|h_i/h_j|b] > 0$ if and only if for at least one of its possible outcomes o_{ku} , $P[o_{ku}|h_i \cdot b \cdot c_k] \neq P[o_{ku}|h_j \cdot b \cdot c_k]$.

Also, $\underline{\text{EQI}}[c^n|h_i/h_j|b] \geq 0$; and $\underline{\text{EQI}}[c^n|h_i/h_j|b] > 0$ if and only if at least one experiment or observation c_k has at least one possible outcome o_{ku} such that $P[o_{ku}|h_i \cdot b \cdot c_k] \neq P[o_{ku}|h_j \cdot b \cdot c_k]$.

In fact it can be shown that increasing the *fineness* of the partition of the outcome space $O_k = \{o_{k1}, \dots, o_{kv}, \dots, o_{kw}\}$ by breaking it up into more distinct outcomes (if it can be so divided) always results in a larger value for EQI , provided that at least some of the additional outcomes have distinct likelihood ratio values.⁴⁶ EQI tracks empirical distinctness in a very precise way. The importance of the *Non-negativity of $\underline{\text{EQI}}$* result for the *Likelihood Ratio Convergence Theorem* will become apparent in a moment.

We are now in a position to state the second part of the *Likelihood Ratio Convergence Theorem*. It applies to all evidence streams that do not contain *possibly falsifying outcomes* for h_j when h_i holds — i.e., it applies to all evidence streams for which h_j is *outcome-compatible* with h_i on each c_k in the stream.

⁴⁵For proof see [Hawthorne, 2004] <http://plato.stanford.edu/entries/logic-inductive/supplement6.html>

⁴⁶See [Hawthorne, 2004] <http://plato.stanford.edu/entries/logic-inductive/supplement6.html>

*Likelihood Ratio Convergence Theorem Part 2: The Probabilistic Refutation Theorem.*⁴⁷

Let $\gamma > 0$ be any number smaller than $1/e^2 (\approx .135)$; where this ‘e’ is the base of the natural logarithm). And suppose that for each possible outcome o_{ku} of each observation condition c_k in c^n , either $P[o_{ku}|h_i \cdot b \cdot c_k] = 0$ or $P[o_{ku}|h_j \cdot b \cdot c_k]/P[o_{ku}|h_i \cdot b \cdot c_k] \geq \gamma$. Choose any positive $\varepsilon < 1$, as near to 0 as you like, but large enough that (for the number of observations n being contemplated) the value of $\underline{\text{EQI}}[c^n|h_i/h_j|b] > -(\log \varepsilon)/n$. Then

$$P[\vee \{e^n : P[e^n|h_j \cdot b \cdot c^n]/P[e^n|h_i \cdot b \cdot c^n] < \varepsilon\} \mid h_i \cdot b \cdot c^n] >$$

$$1 - (1/n) \times \frac{(\log \gamma)^2}{(\underline{\text{EQI}}[c^n|h_i/h_j|b] + (\log \varepsilon)/n)^2}$$

which approaches 1 for large n when $\underline{\text{EQI}}[c^n|h_i/h_j|b]$ has a positive lower bound — i.e., when the sequence of observation c^n has an average expected quality of information (for empirically distinguishing h_j from h_i) that doesn’t diminish towards 0 as the evidence sequence increases.

This theorem provides a very reasonable sufficient condition for the *likely refutation* of false alternatives via exceedingly small likelihood ratios. The condition under which this happens draws only on a characterization of the degree to which the hypotheses involved are empirically distinct from each other. The theorem says that when these conditions of empirical distinctness are met, hypothesis h_i (together with $b \cdot c^n$) provides a likelihood that is *at least* within $(1/n) \times (\log \gamma)^2 / (\underline{\text{EQI}}[c^n|h_i/h_j|b] + (\log \varepsilon)/n)^2$ of 1 that some outcome sequence e^n will occur that yields a likelihood ratio smaller than chosen ε . It turns out that in almost every case the actual likelihood of obtaining such evidence will be much closer to 1 than this factor indicates. Thus, this theorem provides a rather loose lower bound on the likelihood of obtaining small likelihood ratios. It shows that the larger the value of $\underline{\text{EQI}}$ for an evidence stream, the more likely it is that the stream will produce a sequence of outcomes that yield very small likelihood ratios. But even if $\underline{\text{EQI}}$ remains quite small, a long enough stream, n , will almost surely do the trick.⁴⁸

⁴⁷For a proof see [Hawthorne, 2004] <http://plato.stanford.edu/entries/logic-inductive/supplement7.html>

⁴⁸It should now be clear why the boundedness of $\underline{\text{EQI}}$ above 0 is important. Convergence Theorem 2 applies only when $\underline{\text{EQI}}[c^n|h_i/h_j|b] > -(\log \varepsilon)/n$. But this requirement is not a strong assumption. For, the *Nonnegativity of $\underline{\text{EQI}}$ Theorem* shows that the empirical distinctness of two hypotheses on a single possible outcome suffices to make the average $\underline{\text{EQI}}$ positive for the whole sequence of experiments. So, given any small fraction $\varepsilon > 0$, the value of $-(\log \varepsilon)/n$ (which is greater than 0) will eventually become smaller than $\underline{\text{EQI}}$, provided that the degree to which the hypotheses are empirical distinct for the various observations c_k does not on average degrade too much as the length n of the evidence stream increases. This seems a reasonable condition on the empirical distinctness of hypotheses.

Notice that the antecedent condition of the theorem, that “either $P[o_{ku}|h_i \cdot b \cdot c_k] = 0$ or $P[o_{ku}|h_j \cdot b \cdot c_k]/P[o_{ku}|h_i \cdot b \cdot c_k] \geq \gamma$, for some $\gamma > 0$ but γ less than $1/e^2 (\approx .135)$ ”, does not favor hypothesis h_i in any way. This condition only rules out the possibility that some outcomes might furnish *extremely strong* evidence *against* h_j relative to h_i . This condition is only needed because our measure of the evidential distinguishability of pairs of hypotheses, QI, blows up when the likelihood ratio $P[o_{ku}|h_j \cdot b \cdot c_k]/P[o_{ku}|h_i \cdot b \cdot c_k]$ is extremely small. Furthermore, this condition is really no restriction at all on the application of the theorem to possible experiments or observations. If c_k has some possible outcome description o_{ku} that would make $P[o_{ku}|h_j \cdot b \cdot c_k]/P[o_{ku}|h_i \cdot b \cdot c_k] < \gamma$ (for a given small γ of interest), one may disjunctively lump o_{ku} together with some other outcome description o_{kv} for c_k . Then, the antecedent condition of the theorem will be satisfied, but with the sentence ‘ $(o_{ku} \vee o_{kv})$ ’ treated as a single outcome in the formula for EQI. It can be proved that the only effect of such “disjunctive lumping” is to make EQI a bit smaller than it would otherwise be. If, when the evidence is actually collected, such a “too refuting” outcome o_{ku} actually occurs, so much the better. We merely failed to take this possibility for refutation into account in computing our *lower bound* on the likelihood that refutation via likelihood ratios will occur.

The point of the two *Convergence Theorems* explored in this section is to assure us, in advance of considering any *specific* pair of hypotheses, that if the possible evidence streams that test pairs of hypotheses against each other have certain characteristics which reflect their evidential distinguishability, it is highly likely that outcomes yielding small likelihood ratios will result. These theorems provide finite lower bounds on how quickly convergence is likely to occur, bounds that show one need not wait for convergence through some infinitely long run. Indeed, for any evidence sequence in which the probability distributions are at all well behaved, the *actual likelihood* of obtaining outcomes that yield small likelihood ratio values will inevitably be *much higher* than the lower bounds given by Parts 1 and 2 of the Theorem.

In sum, according to the Theorem, each hypothesis h_i *says*, via likelihoods, the following: “given enough observations, *I* am very likely to dominate my empirically distinct rivals in a contest of likelihood ratios.” Even a sequence of observations with an extremely low *average expected quality of information* is very likely to do the job, provided that the sequence is long enough. Presumably, in *saying this*, the true hypothesis speaks truthfully, and its false competitors lie. Thus (by Equation 9), as evidence accumulates, the degree of confirmation for false hypotheses will very probably approach 0, which will indicate that they are probably false; and as this happens, (by Equations 10 and 11) the degree of confirmation of the true hypothesis will approach 1, indicating its probable truth.

6 WHEN THE LIKELIHOODS ARE VAGUE AND/OR DIVERSE

Up to this point I’ve been supposing that likelihoods possess objective or agreed numerical values. Although this supposition is often satisfied in scientific contexts,

there are important settings where it is unrealistic, where individuals are pretty vague about the numerical values of likelihoods, even though the evidence seems to weigh strongly against one hypothesis and in support of another. So let's see how the supposition of precise, agreed values for likelihoods may be relaxed in a reasonable way.

Let's first consider an example of evidence for an important scientific hypothesis where the likelihoods are vague. Consider the following *drift hypothesis*: the land masses of Africa and South America were once joined together, then split and have drifted apart over the eons. Let's compare it to an alternative *contractionist hypothesis*: the continents have fixed positions acquired when the earth first formed, cooled and contracted into its present configuration. On each of these hypotheses, how likely is it that: (1) the shape of the east coast of South America should match the shape of the west coast of Africa as closely as it in fact does; (2) the geology of the two coasts should match up so well; (3) the plant and animal species on these distant continents should be as similar as they are. One may not be able to determine anything like precise numerical values for such likelihoods. But experts readily agree that each of these observations is much more likely on the *drift hypothesis* than on the *contractionist hypothesis*. Jointly these observations constitute very strong evidence in favor of *drift* over its *contraction* alternative. On a Bayesian analysis this is due to the fact that experts in the scientific community widely agree (at least implicitly) that the *ratio of the likelihoods* strongly favors *drift* over *contraction*. As equations 9-11 show, this suffices to strongly refute the *contractionist hypothesis* with respect to the *drift hypothesis* (unless the *contractionist hypothesis* is taken to be quite a bit more plausible than the *drift hypothesis* on other grounds).⁴⁹

⁴⁹Historically the case for continental drift is somewhat more complicated. Geologists tended to largely dismiss the evidence referred to above until the 1960s. Although this evidence may seem to be quite strong, it was unconvincing because it was not sufficiently strong to overcome certain non-evidential plausibility considerations that made the *drift hypothesis* seem extremely implausible — much less plausible than the more traditional *contraction* view. The chief problem was that there appeared to be no plausible mechanism by which *drift* might occur. It was argued that no known force or mechanism could push or pull the continents apart, and that the less dense continental material cannot possibly push through the denser material that makes up the ocean floor. These objections were eventually overcome when a plausible mechanism was articulated — i.e. that the continental crust floats atop molten material and moves apart as convection currents in the molten material carry it along. The case was pretty well clinched when evidence for this mechanism was found in the form of *spreading zones* containing alternating strips of magnetized material at regular distances from mid-ocean ridges. The magnetic alignments of materials in these strips correspond closely to the magnetic alignments found in magnetic materials in dateable sedimentary layers at other locations on the earth. These magnetic alignments indicate time periods when the direction of earth's magnetic field has reversed. This gave geologists a way of measuring the rate at which the sea floor might spread, and the continents move apart. Although geologists may not be able to determine anything like *precise values* for the likelihoods of any of this evidence on each of the alternative hypotheses, the evidence is universally agreed to be *much more likely* on the *drift hypothesis* than on the *contractionist* alternative. Also, with the emergence of a plausible mechanism, the *drift hypothesis* no longer seems so overwhelmingly implausible due to non-evidential considerations. Thus, *the weight of a likelihood ratio* may be objective or public enough to strongly support a hypothesis over an alternative even in cases

Recall now the reasons given earlier for the desirability of agreement or near agreement on values for likelihoods in scientific contexts. I argued that to the extent members of a scientific community disagree on the values of the likelihoods, they disagree about the empirical content of their hypotheses — about what each hypothesis *says* the world is like. Such disagreement about empirical import may result in widely disparate assessments regarding which hypotheses are favored or disfavored by a given body of evidence. Similarly, to the extent that the values of likelihoods are vague for an individual agent, he or she may be unable to determine which of several hypotheses is favored or disfavored by a given body of evidence.

Notice, however, that on a Bayesian account of confirmation the values of individual likelihoods are not really crucial to the way evidence sorts among hypotheses. Rather (as Equations 9-11 show), it is *ratios of likelihoods* that do the heavy lifting. So, even if two confirmation functions P_α and P_β disagree on the values of likelihoods, they may, nevertheless, largely agree on the refutation or support that accrues to various rival hypotheses provided that the following Directional Agreement Condition is satisfied:

Directional Agreement Condition:

The likelihood ratios due to each of a pair of confirmation functions P_α and P_β will be said to *agree in direction* (with respect to the possible outcomes of experiments or observations relevant to a pair of hypotheses) *just in case* each of the following conditions hold:

- for each possible outcome e_k of experiments and observations c_k in the evidence stream,
 $P_\alpha[e_k|h_j \cdot b \cdot c_k]/P_\alpha[e_k|h_i \cdot b \cdot c_k] < 1$ just in case
 $P_\beta[e_k|h_j \cdot b \cdot c_k]/P_\beta[e_k|h_i \cdot b \cdot c_k] < 1$; and
 $P_\alpha[e_k|h_j \cdot b \cdot c_k]/P_\alpha[e_k|h_i \cdot b \cdot c_k] > 1$ just in case
 $P_\beta[e_k|h_j \cdot b \cdot c_k]/P_\beta[e_k|h_i \cdot b \cdot c_k] > 1$.
- each of these likelihood ratios is either close to 1 for neither confirmation function or for both functions.

When this condition holds, the evidence will support h_i over h_j according to P_α just in case it does so for P_β as well, although the strength of support may differ somewhat. Furthermore, although the *rate* at which the likelihood ratios increase or decrease as evidence accumulates may differ for these confirmation functions, the total impact of the cumulative evidence will ultimately affect the refutation and support of hypotheses in much the same way for each function.

Thus, when *likelihoods* are vague or diverse, we may take the approach we employed for *vague* and *diverse* prior plausibility assessments. We may represent the vagueness in an agent's assessments of both prior plausibilities and likelihoods in terms of a *vagueness set* — a set of confirmation functions that covers the range of values that are acceptable to the agent. Similarly, we may extend the *diversity sets* for communities of agents to include confirmation functions for both the

where precise values for likelihoods cannot be determined.

range of *likelihoods* and the range of prior plausibilities (from individual *vagueness sets*) that represent the considered views of the members of the relevant scientific community.

The *Likelihood Ratio Convergence Theorem* can still do its work in this context, provided that the *Directional Agreement Condition* is satisfied by all confirmation functions in these extended *vagueness* and *diversity sets*. The proof of the theorem doesn't depend on supposing that likelihoods are objective or have intersubjectively agreed values. The theorem may be applied to each confirmation function P_α individually. The only real difficulty that comes from applying the theorem to a range of confirmation functions that disagree on the values of likelihoods is that the specific outcome sequences that strongly favors h_i according to P_α may instead strongly favor h_j according to P_β . However, when the *Directional Agreement Condition* holds for a family of confirmation functions, this kind of confirmational bifurcation cannot happen. *Directional Agreement* means that the empirical import of hypotheses as represented by P_α and P_β is similar enough that each evidence sequence must favor the same hypotheses for both of functions. Thus, when the *Directional Agreement Condition* holds, if enough empirically distinguishing experiments or observations are forthcoming, all support functions in an *extended vagueness* or *diversity* set will *very probably* come to agree in assigning extremely small likelihood ratios to empirically distinct false competitors of a true hypothesis. As that happens, the community comes to agree on the refutation of these competitors, and the true hypothesis rises to the top of the heap.⁵⁰

What if the true hypothesis has empirically equivalent rivals? Then their posterior probabilities must rise as well. The Likelihood Ratio Convergence Theorem only assures us that the disjunction of the true hypothesis with its empirically equivalent rivals will be driven to 1 as evidence lays low the empirically distinct rivals. The true hypothesis may itself approach 1 only if either it has no empirically equivalent rivals, or whatever equivalent rivals it has are also laid low by non-evidential plausibility considerations.

ACKNOWLEDGEMENTS

Thanks to Prasanta Bandyopadhyay, Mark Gutel, Mary Gwin, Adam Morton, and an anonymous referee for their helpful comments on various drafts of this article.

BIBLIOGRAPHY

- [Carnap, 1950] R. Carnap. *Logical Foundations of Probability*, Chicago: University of Chicago Press, 1950.

⁵⁰Even if some part of the evidence gives rise to *directionally disagreeing likelihood ratios* (e.g., due to minor disagreements about empirical import), these may not interfere too much with agreement about evidential support for hypotheses, provided that the most substantial part of the evidence gives rise to overwhelmingly powerful *directionally agreeing likelihood ratios*.

- [Carnap, 1952] R. Carnap. *The Continuum of Inductive Methods*, Chicago: University of Chicago Press, 1952.
- [Carnap, 1971] R. Carnap. A Basic System of Inductive Logic, Part I, in R. Carnap and R. C. Jeffrey (eds.), *Studies in Inductive Logic and Probability, Vol. 1*, Berkeley: University of California Press, 1971.
- [Carnap, 1980] R. Carnap. A Basic System of Inductive Logic, Part II, in R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability, Vol. 2*, Berkeley: University of California Press, 1980.
- [De Finetti, 1937] B. De Finetti. La Prévision: Ses Lois Logiques, Ses Sources Subjectives, *Annales de l'Institut Henri Poincaré*, 7, 1-68; translated as "Foresight. Its Logical Laws, Its Subjective Sources", in H. E. Kyburg, Jr. and H. E. Smokler (eds.), 1980, *Studies in Subjective Probability*, Robert E. Krieger Publishing.
- [Duhem, 1906] P. Duhem. *La theorie physique. Son objet et sa structure*, Paris: Chevalier et Riviere; translated by P.P. Wiener, 1954, *The Aim and Structure of Physical Theory*, Princeton, NJ: Princeton University Press, 1906.
- [Earman, 1992] J. Earman. *Bayes or Bust?*, Cambridge, MA: MIT Press, 1992.
- [Edwards, 1972] A. W. F. Edwards. *Likelihood*, Cambridge: Cambridge University Press 1972.
- [Eells, 1985] E. Eells. Problems of Old Evidence, *Pacific Philosophical Quarterly*, 66, 283-302, 1985.
- [Fisher, 1922] R. A. Fisher. On the Mathematical Foundations of Theoretical Statistics, *Philosophical Transactions of the Royal Society*, series A, 309-368, 1922.
- [Fitelson, 1999] B. Fitelson. The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity, *Philosophy of Science* 66, S362-S378, 1999.
- [Fitelson, 2005] B. Fitelson. Inductive Logic, in J. Pfeifer and S. Sarkar (eds.), *The Philosophy of Science: An Encyclopedia*, Oxford: Routledge, 2005.
- [Goodman, 1955] N. Goodman. The New Riddle of Induction, Chapter 3 of *Fact, Fiction, and Forecast*, Cambridge, Mass.: Harvard University Press, 1955.
- [Gaifman, 1964] H. Gaifman. Concerning Measures in First Order Calculi, *Israel Journal of Mathematics*, 2, 1-18, 1964.
- [Gaifman and Snir, 1982] H. Gaifman and M. Snir. Probabilities Over Rich Languages, *Journal of Symbolic Logic*, 47, 495-548, 1982.
- [Glymour, 1980] C. Glymour. *Theory and Evidence*, Princeton: Princeton University Press, 1980.
- [Hacking, 1965] I. Hacking. *Logic of Statistical Inference*, Cambridge: Cambridge University Press, 1965.
- [Hájek, 2003a] A. Hájek. What Conditional Probability Could Not Be, *Synthese* 137, 273-323, 2003.
- [Hájek, 2003b] A. Hájek. Interpretations of the Probability Calculus, in the *Stanford Encyclopedia of Philosophy*, (Summer 2003 Edition), Edward N. Zalta (ed.), 2003. <http://plato.stanford.edu/archives/sum2003/entries/probability-interpret/>.
- [Hájek, 2005] A. Hájek. Scotching Dutch Books? *Philosophical Perspectives* 19 (Epistemology), 139-151, 2005.
- [Hawthorne, 2004] J. Hawthorne. Inductive Logic, *The Stanford Encyclopedia of Philosophy* (Winter 2006 Edition), Edward N. Zalta (ed.), 2004. <http://plato.stanford.edu/archives/win2006/entries/logic-inductive/>.
- [Hawthorne, 2005] J. Hawthorne. Degree-of-Belief and Degree-of-Support: Why Bayesians Need Both Notions, *Mind* 114, 277-320, 2005.
- [Howson, 1997] C. Howson. A Logic of Induction, *Philosophy of Science* 64, 268-290, 1997.
- [Howson and Urbach, 1993] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*, 2nd edition, Chicago: Open Court, 1993.
- [Jaynes, 1968] E. T. Jaynes. Prior Probabilities, *I.E.E.E. Transactions on Systems Science and Cybernetics*, SSC-4, 227-241, 1968.
- [Jeffrey, 1965] R. C. Jeffrey. *The Logic of Decision*, McGraw Hill; 1983, 2nd edition, Chicago: University of Chicago Press, 1965.
- [Jeffrey, 1987] R. C. Jeffrey. Alias Smith and Jones: The Testimony of the Senses, *Erkenntnis* 26, 391-399, 1987.
- [Jeffrey, 1992] R. C. Jeffrey. *Probability and the Art of Judgment*, New York: Cambridge University Press, 1992.
- [Jeffreys, 1939] H. Jeffreys. *Theory of Probability*, Oxford: Oxford University Press, 1939.

- [Joyce, 1999] J. Joyce. *The Foundations of Causal Decision Theory*, New York: Cambridge U. Press. 1999.
- [Joyce, 2003] J. Joyce. Bayes' Theorem, in the *Stanford Encyclopedia of Philosophy*, (Summer 2003 Edition), Edward N. Zalta (ed.), 2003. <http://plato.stanford.edu/archives/win2003/entries/bayes-theorem/>
- [Joyce, 2004] J. Joyce. Bayesianism, in A. Mele and P. Rawling (eds.), *The Oxford Handbook of Rationality*, Oxford: Oxford University Press, 2004.
- [Joyce, 2005] J. Joyce. How Degrees of Belief Reflect Evidence, *Philosophical Perspectives* 19, 153-179, 2005.
- [Keynes, 1921] J. M. Keynes. *A Treatise on Probability*, Macmillan and Co. 1921.
- [Kyburg, 1977] H. Kyburg. Randomness and the Right Reference Class, *Journal of Philosophy*, 74, 501-520, 1977.
- [Lange, 1999] M. Lange. Calibration and the Epistemological Role of Bayesian Conditionalization, *Journal of Philosophy*, 96, 294-324, 1999.
- [Levi, 1967] I. Levi. *Gambling with Truth*, New York: Knopf, 1967.
- [Levi, 1977] I. Levi. Direct Inference, *Journal of Philosophy*, 74, 5-29, 1977.
- [Levi, 1978] I. Levi. Confirmational Conditionalization, *Journal of Philosophy*, 75, 730-737, 1978.
- [Levi, 1980] I. Levi. *The Enterprise of Knowledge*, Cambridge, Mass.: MIT Press, 1980.
- [Lewis, 1980] D. Lewis. A Subjectivist's Guide to Objective Chance, in R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, Vol. 2, Berkeley: University of California Press, 263-293, 1980.
- [Maher, 1996] P. Maher. Subjective and Objective Confirmation, *Philosophy of Science* 63, 149-174, 1996.
- [Maher, 2006] P. Maher. The Concept of Inductive Probability, *Erkenntnis* 65, 185-206, 2006.
- [Quine, 1953] W. V. Quine. Two Dogmas of Empiricism, in *From a Logical Point of View*, New York: Harper Torchbooks, 1953.
- [Ramsey, 1926] F. P. Ramsey. Truth and Probability, in *Foundations of Mathematics and other Essays*, R. B. Braithwaite (ed.), Routledge & P. Kegan, 1931, 156-198; reprinted in H. E. Kyburg, Jr. and H. E. Smokler (eds.), *Studies in Subjective Probability*, 2nd ed., R. E. Krieger Publishing Company, 1980, 23-52; reprinted in D. H. Mellor (ed.), *Philosophical Papers*, Cambridge: University Press, Cambridge, 1990.
- [Rosenkrantz, 1981] R. Rosenkrantz. *Foundations and Applications of Inductive Probability*, New York/Atascadero, CA: Ridgeview Publishing, 1981.
- [Royall, 1997] R. Royall. *Statistical Evidence: A Likelihood Paradigm*, New York: Chapman & Hall/CRC. 1997.
- [Savage, 1954] L. J. Savage. *The Foundations of Statistics*, John Wiley, 1954. (2nd ed., New York: Dover 1972).
- [Savage et al., 1962] L. J. Savage et al. *The Foundations of Statistical Inference*, London: Methuen, 1962.
- [Scott and Krauss, 1966] D. Scott and P. Krauss. Assigning Probabilities to Logical Formulas, in J. Hintikka and P. Suppes (eds.), *Aspects of Inductive Logic*, Amsterdam: North Holland, 219-264, 1966.
- [Skyrms, 1984] B. Skyrms. *Pragmatics and Empiricism*, New Haven: Yale University Press, 1984.
- [Skyrms, 1986] B. Skyrms. *Choice and Chance*, 3rd ed., Belmont, CA: Wadsworth 1986.
- [Williamson, 2007] J. Williamson. Inductive Influence, *British Journal for Philosophy of Science* 58, 689-708, 2007.

This page intentionally left blank

CHALLENGES TO BAYESIAN CONFIRMATION THEORY

John D. Norton

1 INTRODUCTION

Proponents of Bayesian confirmation theory believe that they have the solution to a significant, recalcitrant problem in philosophy of science. It is the identification of the logic that governs evidence and its inductive bearing in science. That is the logic that lets us say that our catalog of planetary observations strongly confirms Copernicus' heliocentric hypothesis; or that the fossil record is good evidence for the theory of evolution; or that the 3°K cosmic background radiation supports big bang cosmology. The definitive solution to this problem would be a significant achievement. The problem is of central importance to philosophy of science, for, in the end, what distinguishes science from myth making is that we have good evidence for the content of science, or at least of mature sciences, whereas myths are evidentially ungrounded fictions.

The core ideas shared by all versions of Bayesian confirmation theory are, at a good first approximation, that a scientist's beliefs are or should conform to a probability measure; and that the incorporation of new evidence is through conditionalization using Bayes' theorem. While the burden of this chapter will be to inventory why critics believe this theory may not be the solution after all, it is worthwhile first to summarize here the most appealing virtues of this simple account. There are three. First, the theory reduces the often nebulous notion of a logic of induction to a single, unambiguous calculus, the probability calculus. Second, the theory has proven to be spacious, with a remarkable ability to absorb, systematize and vindicate what elsewhere appear as independent evidential truisms. Third is its most important virtue, an assurance of consistency. The larger our compass, the more we must digest evidence of diverse form and we must do it consistently. Most accounts of evidence provide no assurance of consistency in their treatment of larger bodies of evidence.¹ Bayesian confirmation theory affords us a simple picture: the entire bearing of evidence at any moment is captured by a probability distribution. No matter how large a body of evidence we contemplate,

¹Some even fail in simple cases. Consider an enumerative induction on the white swans of Europe, which lets us conclude that all swans are white; and an enumerative induction on the black swans of Western Australia, which leads us to the contradictory conclusion that all swans are black.

we will not be led to contradictions in our evidential judgments as long as we form and update our beliefs in conformity with the probability calculus.

Perhaps because of these virtues, Bayesian confirmation theory now enjoys the status of the leading account of induction and confirmation in the philosophy of science literature. Those who expect this circumstance to persist should recall that the present success of Bayesianism is relatively momentary. The probability calculus was born over 350 years ago in the seventeenth century in correspondence between Pascal and Fermat. One hundred years later, in the eighteenth century, the Reverend Thomas Bayes published his theorem as part of a proposal that probability theory be used to answer Hume's inductive skepticism. Nonetheless, the idea that the probability calculus afforded the right way to understand inductive inference remained a minority view. The dominant view of induction in the nineteenth century followed in the tradition of Bacon and his tables, with its most influential expression in John Stuart Mills' *System of Logic* (1891). This dominance persisted through to the mid twentieth century.² It competed with the hypothetico-deductive approach to inductive inference. That view's venerable history can be traced through Descartes' method of hypothesis to ancient Athens where astronomers were, as legend has it, asked by Plato to find the geometrical planetary constructions that would "save the phenomena" of astronomy [Duhem, 1969]. As recently as a few decades ago, the philosophy of science literature was filled with proposals for dealing with Hempel's [1945] notorious "paradox of the raven." That paradox arose within a development of Hempel's "satisfaction" criterion of confirmation, that was itself merely the ancient notion of enumerative induction transported from the context of Aristotle's syllogistic logic to modern, first order predicate logic. These differing approaches played out against persistent claims by Popper [1959] and his followers that the very notion of a logic of induction is mistaken. The Bayesian approach to induction and confirmation rose to dominance in philosophy of science only slowly in the course of the second half of the twentieth century,³ with its importance codified by such works as Howson and Urbach [2006].

Given this history of competing traditions in induction and confirmation rising and falling over the centuries, it seems only prudent to expect that the Bayesian approach will recede from its present prominence and once again be merely one of several useful instruments for assessing inductive inference relations. The goal of this chapter is to review the weaknesses recounted in the literature that may drive this decline. It will draw on two sources. First are those outside the Bayesian fold who present direct challenges; second are shortcomings identified by Bayesians, who in turn have sought modifications within the Bayesian system. These two

²Mill's methods remain to this day the single most important methodological idea for computer repair people and auto mechanics seeking to diagnose problems. Their reach persists in the methodological literature at least to Skyrms [1975], whose Ch. IV gives a detailed development. See the historical survey of Blake, Ducasse and Madden [1960] for a broader sense of the minor role probability theory has played in the longer term history of ideas of scientific methodology.

³Glymour [1980, p. 64] identifies the influence of Carnap [1950] as decisive.

literatures overlap sufficiently for it to be impractical for me to disentangle them.⁴

The survey will proceed from larger to smaller issues. Section 2 will review global challenges to Bayesian confirmation theory. These challenges derive from differences in the basic conception of inductive inference and the principles that govern it. They include the view that notions of inductive inference are merely artifacts subservient to whatever methods may reliably get us to the truth; and the proposal that there is no universally valid logic of induction, but only localized logics adapted to individual domains. Sections 3, 4 and 5 are based on a decomposition of Bayesian confirmation theory into independent components, which are presented as intuitive notions pertaining to confirmation. Additivity amounts to a decision that degrees of probability span belief and disbelief, not belief and ignorance; the essence of Bayesian dynamics is glossed as the incorporation of new evidence by a simple rule of “refute and rescale” prior belief. This decomposition supports a catalog of challenges to the specific Bayesian commitment. Among them is “the problems of the priors,” which proves to be multiple problems aggregated under one label. One — here called first problem — is that additivity precludes prior probabilities that represent a state of complete ignorance. The second problem derives from the simplicity of refute and rescale dynamics. It precludes specification of a neutral prior probability distribution that exerts little influence on the future course of conditionalization. Section 6 reports some further challenges and Section 7 is a brief conclusion.

Some topics lie largely outside the compass of this chapter. In statistical practice, as opposed to philosophy of science, an epochal battle continues to be waged between proponents of traditional Neyman-Pearson hypothesis testing and Bayesian approaches. That debate has been of far less importance in philosophy of science and will be addressed by Deborah Mayo and Aris Spanos (this volume).⁵ Within the Bayesian literature itself, the dominant problem has been to find the correct interpretation of probability: subjective, objective or logical? To someone who is at a slight distance from Bayesianism, the debate seems misdirected in setting standards of precision for the interpretation of a central term of a theory that are rarely achieved elsewhere and are probably unachievable for probability. Its energy is reminiscent of an internecine feud among sect members who agree vastly more than they differ, but nonetheless tolerate no doctrinal deviations. This topic will be addressed only in so far as it arises in broader challenges to the Bayesian approach. These debates are addressed by Sandy Zabell (This Volume) and Philip Dawid (This Volume) as well as survey volumes [Galavotti, 2005; Gillies, 2000; Mellor, 2005].

⁴Hence I apologize in advance to Bayesians who feel that their work has been mischaracterized as a challenge to Bayesianism.

⁵For a brief account from the perspective of a philosopher sympathetic with the Bayesian approach, see [Howson, 1997]; and for a brief history from someone with opposing sympathies, see [Mayo, manuscript].

2 COMPETING ACCOUNTS OF THE NATURE OF INDUCTIVE INFERENCE

Any theory of inductive inference depends upon one or more principles or presumptions that distinguish the right inductive inference relations. These principles must be there, whether they are made explicit, or, as is the more usual case, left tacit. The most fundamental of the challenges to Bayesian confirmation theory come from differing views of these principles.

2.1 *Alternatives*

Given the large number of approaches to inductive inference, one might imagine that there is a plethora of competing principles. A recent survey [Norton, 2005], however, shows that most accounts of inductive inference can be grouped into one of three families, there called “inductive generalization,” “hypothetical induction” and “probabilistic induction.” Each family is based upon a distinct inductive principal and the different members emerge from efforts to remedy the deficiencies of the principle.

While it is impossible to give detailed descriptions of each family, it is useful here to indicate the breadth and longevity of the first two, which compete with the third probabilistic family. Accounts of induction belonging to inductive generalization depend on the principle that the instance of a hypothesis confirms the generalization. This ancient notion was implemented in syllogistic logic as enumerative induction: an A that is B confirms that all A s are B . The principal weakness of this archetype of the family is that it may allow rather little to be inferred. We cannot use it to pass from the evidence of a 3°K cosmic background radiation to the theory of big bang cosmology. The embellishments of the archetype are devoted to extending the reach of enumerative induction. Hempel’s satisfaction criterion reformulates the basic notion within the richer, first order predicate logic. Mill’s methods extend it by licensing the inference to a new notion, cause. In the Method of Agreement, if A is seen to follow a , then we infer this pattern will persist generally and label a as the cause of A . Glymour’s “bootstrap” account of confirmation uses instance confirmation to enable evidence to confirm hypotheses that employ theoretical notions than are not explicitly mentioned in the evidence. For our purposes, what matters is that all these approaches to induction depend upon the basic principle that an instance confirms the generalization.

Accounts of induction belonging to hypothetical induction depend on the principle that the ability of an hypothesis deductively to entail the evidence is a mark of its truth. The archetype of accounts in this family is the saving of the phenomena in astronomy. That some model of planetary motions correctly fits and predicts the observed positions of the planets is a mark of its truth. The principal weakness of this principle is that it assigns the mark too indiscriminately. There are many planetary systems — geocentric, heliocentric and more — that can save the phenomena of astronomical appearances. They are all equally assigned the mark of

truth, even though there must be some sense in which one saves the appearances better than another. Embellishments of this basic notion are intended to rein in the reach of hypothetical induction by adding further conditions that must be met to earn the mark. The hypothesis that saves the phenomena must be simple, or simpler than its competitors; or we must in addition be assured of an exclusionary clause: that were the hypothesis that saves the phenomena false, then it would be unlikely for the phenomena to obtain.⁶ Other versions require hypotheses not just deductively to entail the evidence, but also to explain it. Ptolemy's model of planetary motion entails that Venus and Mercury will always appear close to the sun. Copernicus' model explains why: these planets orbit the sun. Finally, another sort of embellishment enjoins us to take into account the historical process through which the hypothesis was generated. It must be generated by a method known to be reliable. Generating hypotheses *ad hoc* is not judged reliable. Parapsychologists hypothesize retrospectively that their experiments failed in the presence of skeptics because of the unintended obstructive influence emanating from the skeptics. We are free to discount this as an *ad hoc* hypothesis.

2.2 Bayesian Annexation

All these accounts represent challenges to Bayesian confirmation theory in that they are grounded in and can be developed in a framework that does not need the familiar structures of Bayesian confirmation theory. The natural Bayesian rejoinder is that the two basic principles and their embellishments can be absorbed and vindicated by the Bayesian system.

Whether one finds this possibility compelling depends on whether one finds this gain worth the cost of adopting the extra structures and commitments that Bayesian confirmation theory requires (such as will be discussed in Sections 3-5 below). The absorption and vindication can sometimes be quite successful. The clearest case is the success of the Bayesian analysis of when the success of an hypothesis H entailing true evidence E should count as inductive support for H [Earman and Salmon, 1992, §2.7]. A familiar application of Bayes' theorem⁷

$$\frac{P(H|E)}{P(\sim H|E)} = \frac{P(E|H)}{P(E|\sim H)} \frac{P(H)}{P(\sim H)}$$

shows that the degree of support is controlled essentially by the likelihood ratio $P(E|H)/P(E|\sim H)$. If it is high, so that the evidence is much less likely to come

⁶One important form of this is Mayo's [1996] error statistical approach to inductive inference in science. It requires evidence to provide a "severe test" of an hypothesis. The above exclusionary clause is realized in Mayo's [1996, p. 180] severity requirement: "There is a very low probability that test procedure T would yield such a passing result, if [the hypothesis] H is false." It of special interest here since its debate with Bayesian confirmation theory replicates in philosophy of inductive logic the tensions between Neyman-Pearson hypothesis testing and Bayesian approaches in statistics. For a Bayesian rejoinder, see [Howson, 1997, §4].

⁷The terms have their obvious meanings. For example, $P(H|E)$ is the (posterior) probability of H conditioned on E ; $P(H)$ is the (prior) probability of H against the tacit background. See Section 5 below.

about if H is false than if it is true, then the posterior probability $P(H|E)$ increases correspondingly with respect to the prior probability $P(H)$. If the likelihood ratio is close to one so that E is pretty much as likely to come about whether or not H is true, then $P(H|E)$ enjoys no such increase. The analysis is appealing in that it vindicates the basic principle of hypothetical induction, as well as going beyond it in indicating when it will not work.

Other efforts at absorption have been less successful. In Section 5 below, we shall the difficulties facing Bayesian efforts to explicate notions like predictive power and simplicity. The problem is that the simple structures of the Bayesian theory do not have enough resources to distinguish the case of an hypothesis H merely entailing evidence E from it entailing it in a virtuous way that merits confirmatory rewards.

There also seems to be some troubling elasticity in the way that a Bayesian vindication of independent inductive norms is achieved. Sometimes it seems that almost any norm could be justified. The problem is illustrated by difficulties faced in a simple example due to Carnap, as summarized by Earman and Salmon [1992, pp. 85-89]. The traditional form of inductive inference called “example” is a weakening of enumerative induction that merely infers from instances to instances: these cases of x are X ; so that new case of x will also be X . The illustration deals with finitely many individuals $a, b, c, \dots$ that may carry a property F (written “ Fa ” etc.) or fail to carry it (written “ $\sim Fa$ ” etc.) The individual possibilities are described by state descriptions, such as $Fa \& Fb \& \sim Fc \& \dots$. Carnap initially assigned equal probability to each state description. The immediate outcome was that the argument form “example” failed. Learning that Fa and Fb are true leaves the probabilities of Fc and $\sim Fc$ unaffected. That could be offered as a proof of the failure of “example,” if one was inclined against the argument form. Carnap, however, was not so inclined. He chose to readjust the probabilities on the state descriptions in a way that favored correlations between the properties, so that “example” ends up being vindicated.

Closer reflection on Carnap’s illustration shows that the vindication (or otherwise) of the inductive argument form “example” does not depend on anything inherent in the probability calculus. Rather it depends upon additional assumptions we make. In this case it is our determination of whether the individuals under consideration are correlated in the loose sense that one individual carrying F tends to go with another also doing so. If we decide they are so correlated, then “example” is vindicated, whether we express the vindication formally within a probabilistic analysis or merely as a restatement of the informal notion of correlation.

This suggests that the Bayesian vindication of inductive norms is less a matter of extracting them from the probability calculus and more one of our introducing them as independent assumptions that can be expressed in probabilistic language.⁸

⁸Other examples are easy to find. The Bayesian analysis of hypothetical induction above depended essentially on our specifying as an independent assumption that the ratio $P(E|H)/P(E|\sim H)$ is large. That amounts to the external assumption that E would much

That makes them far less impressive than they initially seemed. One virtue does persist: since we are likely to want to combine many possibly competing inductive principles, if we do it within the framework of the probability calculus, we have some assurance that they will be combined consistently.

2.3 *Bayesian Foundational Principles*

The strongest response to these challenges of other principled accounts would be for Bayesian confirmation theory to display its own principled foundation, justify it and argue for its superiority. However prospects here are not strong. Bayesian confirmation theory is distinctive among accounts of confirmation theory in that a great deal of effort has been expended in seeking to justify it. These efforts are usually associated with different interpretations of probability and vary markedly in character according to the interpretation. While the literature on this problem is enormous, there seems to be no vindication that is widely agreed to be successful. Rather the agreement lies only with the conclusion that has to be reached — that the probability calculus is the logic of induction. Each Bayesian seems to find his or her own path to this conclusion, often from widely diverging starting points and with critical dismissal of other starting points.

A few examples will illustrate the divergences in the starting points and their difficulties. The most promising approach is associated with a relative frequency interpretation of probability, such as sought in [Salmon, 1966, pp. 83-96]. In it, the probability of a scientific hypothesis would simply be the relative frequency of true hypotheses in a reference class of hypotheses of the appropriate type. Since relative frequencies behave mostly like probabilities, it is easy to conclude that the probability calculus is the natural logic of induction, which is largely reduced to the simple task keeping count of truth in relevant reference classes. The proposal fails because of the insurmountable difficulty of defining which are the appropriate reference classes, let alone assuring that the appropriate relative frequencies are unique and well defined in the resulting infinite sets.

Another popular approach seeks to identify necessary conditions that any inductive logic must meet. A version has recently been developed by Jaynes [2003] following Cox [1961]. The attendant interpretation of probability is objective, in the sense that, in any one circumstance, there is one correct probability assignment, even though we may not know it; and is “logical” in the sense that the probabilities are treated as degrees of a logical relationship of support. The approach is strong in that, if there is a unique logic of induction to be found, circumscribing it by an ever-narrowing fence of necessary conditions will isolate it. Its difficulty is that it needs to convince the reader of the necessity of a list of conditions: that belief comes in numerical degrees that are always comparable; that the degree of belief assigned some proposition must be a function of those assigned to its disjunctive parts; and further, more specialized suppositions. We

more likely likely come about if H is the case than if it is not, which already pretty much gives us the principle wanted.

shall dissect these sorts of conditions in Sections 3, 4 and 5 below and find that they seem necessary only if one believes from the start that degrees of belief are probabilities.⁹

Finally, in association with a subjective approach to probability, there are arguments that arrive at the probability calculus by urging that any distribution of belief not in accord with the probability calculus exposes us to the danger of a combination of wagers that assure a loss. These are the Dutch book arguments, best known from the work of de Finetti [1937]. In a related approach, such as Savage [1972], natural conditions on our actions are translated into conditions on beliefs. The first and most evident weakness of these proposals is that they require a quite substantial body of presumptions on our actions and preferences and the rules that should be used to relate them to our beliefs. The arguments can be defeated by denying a major presumption. I may simply be unwilling to bet; or I may never find a bet such that I am indifferent to which side I take; or I may harbor intransitive preferences.¹⁰

However there is a deeper problem. These approaches fundamentally change the question asked. It was: when is this observation good evidence for that hypothesis? The presumption of the question is that evidential support obtains independently of our beliefs and actions. These arguments then change the question in two steps that introduce dependencies on our beliefs and actions. First, the question of evidential support (Does the fossil record support evolutionary theory?) is replaced by the question of distributions of belief (How do we distribute our belief over evolutionary theory and its competitors?). At this stage, prior to consideration of

⁹For example, the belief assigned to $A \& B$ is assumed to be a function of the belief assigned to A and to B given A . It is natural to assume this if one imagines all along that “belief” is just a code word for probabilities or frequencies. However, if one does not prejudge what “belief” must be, the assumption of this specific functional dependency is alien and arbitrary. For a natural counterexample that violates the rule, see [Norton, 2007, p. 144, fn. 6]. Similar difficulties arise for attempts to use “loss functions” or “scoring rules” as a means of vindicating the Bayesian system, such as in [Joyce, 1998]. They depend on using a scoring rule that picks out the subspace of additive measures as local optima in the larger space of all belief functions. Any such scoring rule must be selected carefully so as to enforce a coupling between belief in an outcome A and its negation $\sim A$; for an arbitrary change in the belief assigned to A must force a suitable change in the belief in $\sim A$ to ensure that optimal scoring occurs only in the subspace of additive measures. That coupling is just the sort of dependency noted in Section 4.1 and in (A') that is tantamount to choosing additivity. However natural they may seem, the specific conditions chosen to endow scoring rules with this particular property amount to a prior decision to interpret low belief values as disbelief rather than ignorance or some mix of disbelief and ignorance. In the latter case, belief in an outcome is no longer closely coupled to the belief in its negation.

¹⁰Hájek [2008] argues that four of the principal means of vindicating Bayesianism fail, each in the same way. The means are the Dutch book arguments, representation theorem arguments, calibration arguments and gradational accuracy arguments. Each depends on displaying a theorem that is interpreted to show that rationality forces Bayesianism. Each vindication, Hájek urges, overlooks a “mirror-image” theorem that shows the reverse, that rationality requires degrees of belief *not* to be probabilities. For example, the Dutch book theorem assures the existence of a combination of fair bets that forces a sure loss for you if your numerical beliefs are not probabilities. Its mirror-image is the “Czech book theorem.” It assures that a benevolent Czech bookie can find a combination of fair bets that assures a certain gain for you just if your numerical degrees of belief are not probabilities.

actions, we are to accept that there is no one right distribution of belief. Yours is presumed as good as mine. Second, rules are provided for translating beliefs into actions, so that the consequences of our actions can be assessed and used to restrict how we distribute our beliefs, while in general never eliminating the possibility of different belief distributions. These rules make most sense if we are at a racetrack or engaged in an engineering project where beliefs must be translated into design decisions. They are strained if our concern is the question originally asked: whether, for example, the observed motion of the galaxies provides good evidence for the hypothesized heat death of the universe, billions of years in the future — a circumstance independent of our beliefs and actions. The two-step process has produced an answer to a question, but it is not the question with which we started.¹¹

2.4 *Is There a Unique Logic of Inductive Inference?*

So far, all the approaches considered — Bayesian and otherwise — have presumed that there is a unique logic of induction to be found. Another type of challenge to Bayesian confirmation theory denies this. Rather these challenges portray the conditions surrounding individual problems as determining which are the appropriate inductive strategies and, from this perspective, it turns out that there is no single logic of induction to be identified. Rather than having principled grounds for identifying the right inductive logic, these approaches offer principled grounds for denying that such a thing is possible.

A well articulated version of this approach is the learning theoretic paradigm, as developed in [Kelly, 1996]. It presumes that our goal is getting to the truth through some reliable method of discovery. At its simplest, an agent receives a potentially unlimited stream of data and must formulate hypotheses expressing the pertinent truths of the world, as the data accumulates. The central problem of the literature is to discern the sorts of strategies that will lead to convergence to the truth in the hypotheses formulated successively, with more rapid convergence prized. In certain simple worlds, these methods can be governed by familiar inductive notions. A data stream “day, night, day, night, day, . . .” rapidly succumbs to the hypothesis that day follows night and night follows day. In this case, enumerative induction is all that an agent needs to formulate the true hypothesis quite early. Matters get more complicated as the problems get harder. Looking for regularities in what follows what, leads to limited statistical success if the problem is weather prediction at one location and the data stream is “rain, shine, shine, shine, rain, . . .”; or stock market prediction with a data stream “up 3 points, up 2 points, down 20 points, up 3 points, . . .” There is a small amount of inertia in both the weather and the stock market, so that conditions one day tend somewhat to be

¹¹For a spirited assault on Dutch book arguments, see [Bacchus *et al.*, 1990]. They urge these arguments fail since an agent should simply refuse to accept a combination of bets that assures a loss; it is (pp. 504-505) “a matter of deductive logic and not of propriety of belief.” Schick’s [1986] view is similar.

replicated the next. These methods fail completely in inhospitable worlds. If the data stream is derived from a well-designed roulette wheel, “red, black, black, red, red, ...,” no hypotheses on when red will follow black (other than chance) can succeed. And matters can be worse in a demonic universe, which delivers data maliciously tailored to deceive the particular agent in question.

In this reliabilist paradigm, inductive norms are parasitic on convergence to the truth. It is simply a matter of ingenuity to conceive universes in which a given inductive norm will be fruitful or harmful. For example, a standard supposition of scientific methodology is that we should only advance hypotheses that are consistent with the data at hand. However it turns out that, once we allow for the computational limitations of agents, a demand for consistency can conflict with reliability. So it may be efficient for us to offer an hypothesis inconsistent with the evidence, even though most accounts of induction, including the Bayesian, insists the hypothesis has been refuted. For a critique of Bayesian confirmation theory from the reliabilist perspective, see [Kelly and Glymour, 2004].

The idea that inductive strategies will vary according to the domain at hand is the central idea of the “material theory of induction” [Norton, 2003]. Its name derives from the idea that the license of an inductive inference does not come from the *form* of the sentences involved, as it does in deductive logic. Rather the license comes from the subject *matter* of the inference. Therefore the inductive logic applicable will vary from domain to domain as the licensing facts change; there is no universally applicable logic of induction.

In one restricted area, this notion of material facts fixing the inductive logic is already familiar to subjective Bayesians through Lewis’ [1980] “principal principle.” For processes for which objective chances are available, the principle enjoins subjective Bayesians to match their subjective degrees of belief to the physical chances. Quantum mechanical laws deliver chances for the timing of the decay of a radioactive atom. If we match our subjective degrees of belief concerning those timings to the quantum mechanical chances, then we are assured high belief in outcomes that are most likely to happen and low belief in those unlikely to happen. A by-product is that our degrees of belief automatically conform to the probability calculus.¹²

The principal principle naturally generalizes to the central claim of the material theory of induction: that the factual properties of the system under consideration will determine the appropriate inductive logic. It follows that there may be systems whose facts dictate that the applicable logic of induction is *not* probabilistic. Norton [2007, §8.3; forthcoming] describes certain physical systems whose governing laws are indeterministic in the sense that the complete specification of

¹²The facts that license probabilistic inferences need not be the sorts of physical chances recovered from stochastic theories like quantum mechanics. For example, imagine that one is a bookmaker accepting wagers on a definite matter of presently unknown fact: say, whether some recently discovered infectious illness is viral or bacterial in origin. If the factual conditions surrounding the bookmaking conform to all those assumed in Dutch book arguments, then those facts determine that one’s degree of belief ought to conform to the probability calculus, on pain of exposure to a sure loss.

the present state of the system does not fix the future state; yet these governing laws provide no probabilities for the various possible futures. They simply tell us that, given such-and-such a present state, this or that future state is possible. In the corresponding inference problem, we are required to distribute beliefs over these possible futures, knowing the state of the present. If we distributed beliefs as probabilities rather than in some weaker form, we end up with beliefs that outstrip the full physical specification as given by the initial conditions and the applicable physical laws. We must assert that one possible outcome is twice as probable as another; or as probable as another; or half as probable. That is, we must pretend to know more than the physical laws, which are only able to assert that each outcome is possible, without any sense of “twice as possible” or “half as possible.”

3 CHALLENGES TO FRAMEWORK ASSUMPTIONS

The challenges considered so far in Section 2 above have been global in the sense that they arise from tensions between Bayesian confirmation theory as a whole and other approaches to inductive inference. If we recall that Bayesian confirmation theory is a composite of many assumptions, then another class of challenges to Bayesian confirmation theory is identifiable. These are challenges to specific assumptions of Bayesian confirmation theory. In order to structure the inventory of these challenges, the assumptions comprising Bayesian confirmation theory will be divided into three parts, called: “Framework,” to be discussed in this section; “Additivity,” to be discussed in Section 4; and “Bayes’ dynamics: to be discussed in Section 5.¹³

3.1 *Framework Assumptions*

What is common to all versions of Bayesian confirmation theory is the assumption that there is a real-valued magnitude, $P(H|E)$, that is one’s subjective degree of belief or some objective measure of support for the hypothesis H from evidence E . This apparently simple assumption of a real valued magnitude can be decomposed further into a series of assumptions, each of which presumes the one before:

Precision. There is a magnitude $P(A|B)$ that represents the degree of belief or support accrued to A given B .

Universal comparability. It is always possible to compare two such degrees, finding one to be less than, equal to or greater than the other.

Partial order. The comparison relation “no less than” $\leq$ is a partial

¹³This disassembly is similar to the one effected in [Norton, 2007], which in turn draws on an extensive literature in qualitative probability. For surveys see [Fine, 1973; Fishburn, 1986].

order: it is reflexive¹⁴, antisymmetric¹⁵ and transitive¹⁶.

Real values. The degrees are real-valued.

Different challenges would require us to halt at different stages as we pass up this hierarchy. Lest that seem odd, the notion that a cautious ascent is needed may be more palatable if we consider an analogous problem of assigning degrees of simplicity to hypotheses. The same hierarchy can be formed, but few are likely to ascend all the way to the idea of single real value that measures the degree of simplicity of all hypotheses we encounter. We may stall at the start, judging the very idea of a degree of simplicity inadmissible. Or we may allow comparison among closely related hypotheses only. Jeffreys [1961, p.47], for example, wished to assign greater prior probability to simpler hypotheses. So he defined the complexity m of certain differential equations as the sum of their order, degree and the absolute values of their coefficients. Unit prior probability was then distributed over the resulting complexity classes as a decreasing function of m , such as 2^{-m} or $6/\pi^2 m^2$. Jeffreys' degree is defined only for a specific class of equations. It would not sustain universal comparability in that we could not compare the degree of simplicity of the hypothesis of a linear relationship between the velocity and distance to the galaxies with that of the hypothesis of the germ theory of disease.

3.2 *The Very Idea.*

The most fundamental objection is that there is something mistaken about the very idea of a degree of belief or support that evidence lends to a hypothesis. This objection can rest on the visceral sense that such assignments are simply spurious precision: that we can conjure up a number does not mean that we are measuring something. A colorful statement of the mismatch of Bayesian analysis with diagnostic clinical practice is Feinstein's [1977] "The haze of Bayes, the aerial palaces of decision analysis, and the computerized Ouija board."

In a different direction, Glymour's [1980, Ch. III] "problem of old evidence" speaks against the idea that the fundamental quantity $P(A|B)$ presumed in the framework is an adequate basis for an account of inductive inferences. The import of evidence E on hypothesis H is gauged by comparing the prior probability $P(H|B)$ with the posterior $P(H|E \& B)$, where B is the totality of background knowledge. It frequently happens that the evidence E is already part of our background knowledge. The celebrated case concerns the motion of Mercury, which, at the start of the 20th century, was known to deviate slightly from the predictions of Newtonian gravitation theory. In November 1915, Einstein found that his new-born general theory of relativity " H " predicted exactly the observed deviations

¹⁴For all admissible $A, B, P(A|B) \leq P(A|B)$.

¹⁵For all admissible A, B, C and D , if $P(A|B) \leq P(C|D)$ and $P(C|D) \leq P(A|B)$, then $P(A|B) = P(C|D)$.

¹⁶For all admissible A, B, C, D, E and F , if $P(A|B) \leq P(C|D)$ and $P(C|D) \leq P(E|F)$, then $P(A|B) \leq P(E|F)$.

“ E ”. The universal agreement is that E provided strong support for H . However, since E is part of the background then known to Einstein, we have

$$E \& B = B$$

so that

$$P(H|E \& B) = P(H|B)$$

which shows that E is evidentially inert.

If this objection is sustained, it means that a confirmation theory employing only the quantities $P(H|E)$ and $P(H|E \& B)$ cannot support the judgment that E is good evidence for H .¹⁷ The obvious escape is to replace the prior $P(H|B)$ with an adjusted prior $P(H|B')$, where B' is the background B with E somehow excised. Glymour’s original presentation sought to block this escape by urging that it is quite unclear how this excision could be effected. Other, more elaborate escapes are reviewed in [Earman, 1992, Ch. 5], including the notion that we assign evidential value to learning the logical relation between some hypothesis H and the evidence E that is already, unbeknown to us, part of our background B .

3.3 Universal Comparability

Once we allow that the degree $P(A|B)$ is always defined, it may still not follow that all degrees are comparable. Keynes [1921, pp. 37-40] allows that all degrees will lie between those assigned to impossibility and certainty, but urges that not all intermediate degrees are mutually comparable.

The clearest way that comparability can fail is when we conditionalize on very different sorts of backgrounds so that we are distributing belief over incommensurable possibilities. For example, Norton [2007, pp. 147-48] expands a concern expressed by Humphreys [1985] in the context of propensity interpretations to suggest that the direct degrees $P(E|H)$ and the inverse $P(H|E)$ may measure very different things. $P(E|H)$ may be derived from the computing of a quantity in a stochastic theory: the chance of E , the decay of some particular radioactive atom in some time as deduced from H , the laws of quantum physics. The inverse $P(H|E)$ assigns degrees to different physical laws on the basis of the evidence of radioactive decay. The first deals with possibilities licensed by a known theory and is grounded in the science; the second deals with speculation on how the world might admit different laws whose range is delimited by our imagination. Even if we can find protocols that allow us to assign numbers in each case, it is not clear that there is anything meaningful in the comparison of those numbers — just as it is meaningless to compare degree of temperature and degrees Baumé of specific gravity, even though both are real numbers and arithmetic allows the

¹⁷Glymour’s original presentation used Bayes’ theorem to deduce $P(H|E \& B) = P(H|B)$, although the problem is independent of Bayes’ theorem.

comparison.¹⁸

That there are different sorts of uncertainty even in decision theoretic contexts is illustrated by Ellsberg's [1961] urns. Consider two urns, each containing 100 balls. The first has some combination of red and black balls. Any ratio from 0-100% red is possible and we are completely uncertain over which. The second urn has exactly 50 red and 50 black balls. A ball will be drawn from each urn. What is our belief that it will be red? For both urns, the symmetry of the arrangements requires that we be equally certain of a red or a black, so, in drawing from both urns we should be indifferent to making bets on red or on black. Any protocol for converting these inclinations to bet into degrees of belief must respect that symmetry. Therefore, if the protocol generates probabilities solely on the basis of these inclinations, in each case it will return a probability of 0.5. However Ellsberg suggests that most of us prefer the bet on the urn of known composition, so that the degrees of belief assigned to red ought to differ in the two cases. For our purposes, this illustrates the possibility of different senses of uncertainty that may not be readily comparable if we try to characterize them in a single magnitude. These different senses reflect a distinction to be made in Section 4 below between disbelief and ignorance.

Jaynes [2003, pp. 658-59] regards such quibbles over different senses of uncertainty with characteristic, entertaining disdain. He makes an analogy to a mineralogist who may classify rocks with different parameters but can still find a way to trade off changes in one parameter against the other to recover a single scale of comparison.

The idea that universal comparability must be discarded is a consequence of a significant development of the Bayesian approach that has adopted several forms. What they have in common is the notion that a single number may not be rich enough to capture the extent of belief. That concern is readily motivated whenever we are pressed to assign a definite probability to some outcome — say, whether it will rain tomorrow. Assigning a probability 0.6 is too precise. Our belief is better represented by something vaguer, a probability somewhere around 0.6. In one approach, we might replace definite values by intervals of values; the probability of rain is 0.5 to 0.7. In another we might take set of probability measures as representing our belief states — say the set of all probability measures that assign a probability of 0.5 to 0.7 to rain tomorrow. Developing these proposals into well functioning theories is not straightforward. For developments of these notions, see [Kaplan, 1998; Kyburg, 1959; Kyburg and Teng, 2001; Levi, 1974; 1980, Ch. 9; Walley, 1991].

For our purposes, what matters is that these proposals break with the framework sketched in that they deny universal comparability. Take the simplest case of interval-valued degrees of belief. The belief in some outcome represented by the

¹⁸On the basis of their being "sufficiently disparate," Fishburn [1986, p. 339] offers the following pair as candidates for failure of comparability: "A=Mexico's City's population will exceed 20,000,000 by 1994; B=The first card drawn from this old and probably incomplete bridge deck will be a heart."

interval $[0.8, 0.9]$ is greater (in the sense of being closer to certainty) than the interval $[0.1, 0.2]$. But the intervals $[0.3, 0.7]$ and $[0.4, 0.6]$ are none of greater than, less than or equal to each other. Analogous effects arise for richer structures that dispense with single valued degrees of belief.

3.4 *Transitivity*

If we allow that the degrees are always comparable, it does not yet follow that they form a partial order, which is reflexive, antisymmetric and transitive. Of these three properties, transitivity is of the greatest concern. Norton [2007, §3.2.2] has suggested that transitivity may fail in certain complicated cases if common wisdoms on inductive support obtain. That is, if we have a case of evidence entailed by several different hypotheses, it is routine to say that the evidence supports one of these hypotheses more than another if the first hypothesis displays more of some relevant virtue, such as greater simplicity, greater explanatory power or greater fecundity. Using three hypotheses and three virtues, it is easy to envisage structures in which the first hypothesis is better confirmed than the second, the second better than the third and the third better than the first. This violates transitivity.¹⁹

3.5 *Real Values*

Once it is allowed that the degrees may be partially ordered, it is still not assured that they must be real valued, or isomorphic to the reals or some interval of real numbers. A great deal of effort has been expended in determining what additional assumptions are needed to assure that the degrees are real valued. These additional assumptions have been given many forms and are generally known as “Archimedean axioms.” [Fishburn, 1986, pp. 341-42]. Their function is to block the possibility of value sets that extend beyond real values in admitting infinitesimally small or infinitely large values.

The familiar illustration in this literature of how partially ordered degrees may fail to be real valued is generate by degrees that are ordered pairs $\langle x, y \rangle$ of reals x and y [Jeffreys, 1961, pp. 19-20]. They are partially ordered by the rule

$$\begin{aligned} \langle X, Y \rangle > \langle x, y \rangle & \text{ when } X > x, \text{ no matter the values of } Y, y \\ & \text{ or, if } X = x, \text{ when } Y > y \end{aligned}$$

There is no way to map these degrees $\langle x, y \rangle$ onto the reals so that the partial order is preserved.

We could imagine a two parameter family of hypotheses $H_{x,y}$ such that our degrees of belief in the various hypotheses may be ordered according to the rule

¹⁹For the example of a slightly bent coin, Fishburn [1986, p. 339] describes three propositions for which intransitivity “do[es] not seem unreasonable”: “ A =The next 101 flips will give at least 40 heads; B =The next 100 flips will give at least 40 heads; C =The next 1000 flips will give at least 460 heads.” The suggestion is that we could have the same belief in A and in C ; the same belief in C and in B , but that we must have more belief in A than B .

just sketched. One might doubt whether such a family could arise in realistic problems and take that doubt as a reason to discount the challenge. Or one might be concerned that adopting a Bayesian confirmation theory presumes in advance that such a distribution of belief is inadmissible. Such a presumption is not compatible with the universal applicability supposed for the theory.

4 ADDITIVITY

4.1 *The Property*

Probability measures are additive measures. That means that they conform to the condition:

$$\begin{aligned} &\text{If } A \text{ and } B \text{ are mutually exclusive then, for any } C, \\ &P(A \vee B|C) = P(A|C) + P(B|C). \end{aligned} \tag{A}$$

Additivity is the most distinctive feature of the probability calculus. In the context of confirmation theory, it amounts to assigning a particular character to the degrees of belief of the theory. The degrees of this additive measure will span from a maximum value for an assured outcome (conventionally chosen as one) to a minimum value for an impossible outcome (which must be zero²⁰). If the high values are interpreted as belief with unit probability certainty, then it follows that the low values must represent disbelief, with zero probability complete disbelief. To see this, recall that near certain or certain belief in A corresponds to near complete or complete disbelief in its negation, $\sim A$.²¹ If we assign high probability 0.99 or unit probability to some outcome A , then by additivity the probability assigned to the negation $\sim A$ is 0.01 or zero, so that these low or zero values correspond to near complete disbelief or complete disbelief.

This interpretation of low probability as disbelief is already expressed in the functional dependency between the probabilities of outcomes and those of their negations. We have $P(\sim A|C) = 1 - P(A|C)$, which entails

$$P(\sim A|C) \text{ is a strictly decreasing function of } P(A|C) \tag{A'}$$

More generally, taking relative negations, $P(\sim A \& B|C) = P(B|C) - P(A \& B|C)$, we recover the more functional dependency

$$P(\sim A \& B|C) \text{ is a strictly decreasing function of } P(A \& B|C) \tag{A''}$$

and a strictly increasing function of $P(B|C)$.

These dependencies tell us that a high probability assigned to an outcome corresponds to a low probability assigned to its negation or relative negation, which is the characteristic property of a scale of degrees that spans from belief to disbelief.

²⁰If Imp is an impossible outcome, we have $\text{Imp} = \text{Imp} \vee \text{Imp}$. Since Imp and Imp are mutually exclusive in the sense that $\text{Imp} \& \text{Imp} = \text{Imp}$, additivity applies so that $P(\text{Imp}) = P(\text{Imp}) + P(\text{Imp})$, from which we have $P(\text{Imp})=0$.

²¹I take it as a definition that *disbelief* in A is the same thing as belief in *not-A*.

While additivity (A) entails weaker functional dependencies (A') and (A''), the gap between them and (A) is not great. If the functional dependencies (A') and (A'') are embedded in a natural context, it is a standard result in the literature that the resulting degrees can be rescaled to an additive measure satisfying (A). (See [Aczel, 1966, pp. 319-24; Norton, 2007, §7].)

4.2 Non-Additive Measures: Disbelief versus Ignorance

Once it is recognized that the additivity (A) of the probability calculus amounts to selecting a particular interpretation of the degrees, then the ensuing challenge becomes inevitable. In many epistemic situations, we may want low degrees to represent ignorance or some mix of ignorance and disbelief, where ignorance amounts to a failure to commit to belief or, more simply, an absence of belief or disbelief. Shafer [1976, pp. 22-25] considers how we might assign degrees of belief to the proposition that there are living beings in orbit around Sirius, an issue about which we should suppose we know nothing. No additive measure is appropriate. If we assign low probability to “life,” additivity then requires us to assign high probability to “no-life.” That asserts high certainty in there being no life, something we do know. The natural intermediate of probability 1/2 for each of the two outcomes “life” and “no-life” fails to be a usable ignorance value since it cannot be used if we are ignorant over more than two mutually exclusive outcomes.

This example makes clear that introducing an element of ignorance requires a relaxation of the functional dependencies (A') and (A''). It must be possible to assign a low degree of belief to some outcome without being thereby forced to assign a high degree to its negation. Measures that allow this sort of violation of additivity are “superadditive”: if A and B are mutually exclusive then, for any C , $P(A \vee B|C) \geq P(A|C) + P(B|C)$. The extent to which the equality is replaced by an inequality is the extent to which the measure allows representation of ignorance.

The best-known superadditive calculus is the Shafer-Dempster calculus. In it, an additive measure m , called a “basic probability assignment,” is defined over the power set of the “frame of discernment” Θ , so that $\sum_{A \subseteq \Theta} m(A) = 1$, where the summation extends overall all subsets of Θ , and $m(\{\}) = 0$. This basic measure probability assignment is used to generate the quantity of interest, the “belief function” Bel , which is defined as

$$\text{Bel}(A) = \sum_{B \subseteq A} \text{Bel}(B)$$

for any A in Θ , where the summation is taken over all subsets B of A . These belief functions allow blending of disbelief and ignorance. For example, that we are largely ignorant over the truth of “life” can be represented by

$$m(\text{life}) = 0.1 \quad m(\sim \text{life}) = 0.1 \quad m(\text{life} \vee \sim \text{life}) = 0.8$$

which induces the belief function

$$\text{Bel}(\text{life}) = 0.1 \quad \text{Bel}(\sim \text{life}) = 0.1 \quad \text{Bel}(\text{life} \vee \sim \text{life}) = 1$$

4.3 Complete Ignorance: The First Problem of the Priors

In Bayesian confirmation theory, prior probability distributions are adjusted by Bayes' theorem to posterior probability distributions that incorporate new evidence learned. As we trace this chain back, the prior probability distributions represent states of greater ignorance. Bayesian confirmation theory can only give a complete account of this learning process if it admits an initial state representing complete ignorance. However representing ignorance is a long-standing difficulty for Bayesian confirmation theory and the case of complete ignorance has been especially recalcitrant. It is designated here as the first problem of the priors, to distinguish it from another problem with prior probability delineated below in Section 5.6 below.

There are two instruments already in the probability literature that are able to delimit the representation of this extreme case, the epistemic state of complete ignorance. Norton [2008] has given an extended analysis of how the two may be used to do this. The first instrument is the principle of indifference. It asserts that if we have no grounds for preferring one outcome to a second, then we should assign equal belief to both. This platitude of evidence is routinely used to ground the classical interpretation of probability and famously runs into trouble when we redescribe the outcome space. Complete ignorance in one description is equivalent to complete ignorance in another. That fact allows one to infer quite rapidly that the degree of belief assigned to some compound proposition $A \vee B$ should be the same as the degree of belief assigned to each of the disjunctive parts A and B , even though A and B may be mutually exclusively.²² Since no probability distribution can have this property, it is generally concluded that there is something wrong with the principle of indifference.

The difficulty is that the principle of indifference is not so easily discarded. It is a platitude of evidence. If beliefs are grounded in reasons and we have no reasons to distinguish two outcomes, then we should have the same belief in each. The alternative is to retain the principle of indifference and discard the notion that a probability distribution can adequately represent complete ignorance. Instead we are led to a representation of complete ignorance by a non-probabilistic distribution with three values: *Max* and *Min* for the extreme values of certainty and complete disbelief and *Ig* ("ignorance") for everything in between

$$\begin{aligned} \text{Degree}(A) &= \text{Max}, \text{ for } A \text{ an assuredly true outcome} \\ &= \text{Min}, \text{ for } A \text{ an assuredly false outcome} \\ &= \text{Ig}, \text{ for } A \text{ any contingent}^{23} \text{ outcome} \end{aligned} \tag{I}$$

²²An example is von Mises' famous wine-water problem. We have a glass with a mixture of wine and water, knowing only that the ratio of wine to water lies in 1/2 to 2. So we are indifferent to each of the intervals wine to water: (a) 1/2 to 1, (b) 1 to 1 1/2, (c) 1 1/2 to 2; and assign equal probability of 1/3 to each. However if we redescribe the problem in terms of the ratio of water to wine, we end up assigning equal probability of 1/3 to the intervals water to wine (a') 1/2 to 1, (b') 1 to 1 1/2, (c') 1 1/2 to 2. Now the interval (a) describes the same outcome as the disjunction (b') $\vee$ (c'). So we assign probability 1/3 to the disjunction (b') $\vee$ (c') and also to each of its parts.

The essential property here is that we can assign the ignorance degree I_g to some contingent outcome $A \vee B$ and that same ignorance degree to each of its mutually exclusive, disjunctive parts, A and B . This is the only distribution for which this is true over all contingent outcomes.

The second instrument used to delineate the epistemic state of complete ignorance is the notion of invariance, used so effectively by objective Bayesians, but here used in a way that objective Bayesians may not endorse. The notion ultimately produces serious problems for Bayesian confirmation theory. The greater our ignorance, the more symmetries we have under which the epistemic state should be invariant. It is quite easy to accumulate so many symmetries that the epistemic state cannot be a probability distribution. For example, let us say we know only

x is a real number the interval $(0,1)$. (DATUM)

This information (DATUM) remains true if x is replaced by $x' = 1 - x$, where the function $x'(x)$ is self-inverting. It also remains true if x is replaced by $x'' = 1 - (1 - (1 - x)^2)^{1/2}$, where once again the function $x''(x)$ is self-inverting. So our epistemic state must remain unchanged under each of these transformations. It is easy to show that no probability distribution can be invariant under both. (See [Norton, 2008, §3.2].)

Invariance requirements can be used to pick out the unique state of complete ignorance, which turns out to be (I) above. To see the relevant invariance, consider some proposition A over whose truth we may be completely ignorant. Our belief would be unchanged were A replaced by $\sim A$.²⁴ That is, the state of complete ignorance remains unchanged under a transformation that replaces every contingent proposition with its negation. It can readily be seen that the ignorance distribution (I) satisfies this invariance requirement. Each contingent proposition and its negation are assigned the same degree I_g .²⁵ That this ignorance distribution (I) is the only distribution satisfying this invariance that we are likely to encounter is made more precise by the demonstration that it is the only monotonic²⁶ distribution of belief with the requisite invariance [Norton, 2008, §6].

²³Contingent propositions are defined here as propositions that may be either true or false.

²⁴Describing this transformation more figuratively makes the invariance intuitive. Imagine that the content of proposition A has been written in a normal English sentence by a scribe on a slip of paper, folded before us on the table. We form our belief over the truth of the sentence on the paper: it is complete ignorance because we have no idea what the sentence says. We are now told that the scribe erred and mistakenly wrote the content of $\sim A$ on the paper instead. That new information would not change our belief state at all.

²⁵It is assumed that the ignorance does not extend to logical truths, so that we know which propositions are assuredly true and false and we assign Max and Min to them. We could define a broader ignorance state in which logical truths are presumed unknown as well by assigning the same value I_g to all propositions.

²⁶A distribution of belief “degree $(.)$ ” is monotonic if, whenever A logically entails B , $\text{degree}(A) \leq \text{degree}(B)$.

The negation map — the transformation that replaces each contingent proposition with its negation — is a little more complicated than it may initially seem. To see the difficulty, imagine that the outcome space is exhausted by n mutually exclusive atomic propositions, $A_1, A_2, \dots, A_n$. The transformation replaces atomic propositions, such as A_1 by compound propositions, such as $A_2 \vee A_3 \vee \dots \vee A_n$, so it may not be evident that the transformation is a symmetry of the outcome space. It is, in the sense that it maps the outcome space back to itself and is self-inverting; $A_2 \vee A_3 \vee \dots \vee A_n$ is mapped to A_1 . However, it does not preserve additive measures. The transformation takes an additive measure m to what Norton [2007a] describes as a “dual additive measure” M . These dual additive measures have properties that are, on first acquaintance, odd looking. They are additive, but their additivity is attached to conjunctions. If we have propositions A and B such that $A \vee B$ is always true, then we can add their measures as $M(A \& B) = M(A) + M(B)$. The notion of a dual measure allows a simple characterization of the ignorance distribution (I): it is the unique, monotonic measure that is self-dual.

If one does not see that an epistemic state of complete ignorance is represented by the non-probabilistic (I), one is susceptible to the “inductive disjunctive fallacy” [Norton, forthcoming a]. Let $a_1, a_2, a_3, \dots$ be a large number of mutually exclusive outcomes over which we are in complete ignorance. According to (I), we remain in that state of complete ignorance for any contingent disjunction of these outcomes, $a_1 \vee a_2 \vee a_3 \vee \dots$. If one applies probabilities thoughtlessly, one might try to represent the state of complete ignorance by a broadly spread probability distribution over the outcomes. Then the probability of the disjunction can be brought close to unity merely by adding more outcomes. Hence one would infer fallaciously to near certainty for a sufficiently large contingent disjunction of outcomes over which we are individually in complete ignorance. The fallacy is surprisingly widespread. A striking example is supplied by van Inwagen [1996] in answer to the cosmic question “Why is there anything at all?” There is, he asserts, one way for no thing to be, but infinitely many ways for different things to be. Distributing probabilities over these outcomes fairly uniformly, we infer that the disjunction representing the infinitely many ways things can be must attract all the probability mass so that we assign probability one to it.

4.4 *Bayesian Responses*

The literature in Bayesian confirmation theory has long grappled with this problem of representing ignorance. That is especially so for prior probability distributions, where the presumption that ignorance must be representable is most pressing. Perhaps the most satisfactory response comes through the basic supposition of subjective Bayesians that the probabilities are subjective and may vary from person to person as long as the axioms of the probability calculus are respected. So the necessary deviation from the non-probabilistic ignorance distribution (I) in some agent’s prior probability distribution is discounted as an individual aberration not reflecting the true evidential situation. The price paid in adopting this

response, the injection of subjectivity and necessity of every prior being aberrant, is too high a price for objective Bayesians, who are committed to there being one probability distribution appropriate for each circumstance.

However objective Bayesian methods have not proven able to deliver true “ignorance priors,” even though the term does appear over-optimistically in the objective Bayesian literature [Jaynes, 2003, Ch.12]. One approach is to identify the ignorance priors by invariance properties. That meets only with limited success, since greater ignorance generates more invariances and, as we saw in Section 4.3 above, eventually there are so many invariances that no probability measure is admissible. An alternative approach is to seek ignorance priors in distributions of maximum entropy.²⁷ Maximum entropy distributions do supply what are, in an intuitive sense, the most uniform distributions admissible. If, for example, we have an outcome space comprising n atomic propositions, without further constraints, the maximum entropy distribution is the one uniform distribution that assigns probability $1/n$ to each atomic proposition. However, if there is sufficient ignorance, there will be invariances under which the property of having attained maximum entropy will not be preserved. In the end, it is inevitable that these methods cannot deliver the ignorance distribution (I), for (I) is not a probability distribution. So the best that can be expected is that they will deliver a distribution that captures ignorance over one aspect of the problem, but not all. The tendency in the literature now is to replace the misleading terminology of “ignorance prior” by more neutral terms such as “noninformative priors,” “reference priors” or, most clearly “priors constructed by some formal rule” [Kass and Wasserman, 1996].

Another popular approach to representing ignorance with Bayesian confirmation theory is to allow that an agent’s epistemic state is not given by any one probability measure, but by a set of them, possibly convex. (See for example [Levi, 1980, Ch. 9].) The deepest concern with this strategy is that it amounts to an attempt to simulate the failure of additivity that is associated with the representation of ignorance. Something like the complete ignorance state (I) can be simulated, for example, by taking the set of all probability measures over the same outcome space. The resulting structure is vastly more complicated than (I), the state it tries to simulate. It has become non-local in the sense that a single ignorance value is no longer attributed to an outcome. Each of the many probabilities assigned to some outcome must be interpreted in the context of the other values assigned to other propositions and in cognizance that there are many other distributions in the set.²⁸

Finally, as pointed out in Norton [2007a; 2008], a set of probability measures necessarily falls short of simulating (I). For no set of additive measures can have the requisite invariance property of a complete ignorance state — invariance under the negation map. For a set of additive measures is transformed by the negation

²⁷For further objections to maximum entropy methods, see [Seidenfeld, 1979].

²⁸For a discussion of convex sets of probability measures and the how they contain the Shafer Dempster belief functions as a special case, see [Kyburg, 1987].

map into a set of dual additive measures. In informal terms, any set of additive measures on an outcome space preserves a directedness. For each measure in the set, as one proceeds from the assuredly false proposition to the assuredly true by taking disjunctions, the measures assigned are non-decreasing and must, at some point, increase strictly. Invariance under the negation map precludes such directedness.

4.5 Ignorance over a Countable Infinity of Outcomes

The difficulties of representing ignorance have been explored in the literature in some detail in the particular problem of identifying a state of ignorance over a countable infinity of outcomes. It has driven Bayesians to some extreme proposals none of which appear able to handle the problem in its most intractable form.²⁹

The traditional starting place — already “well-known” when de Finetti [1972, p.86] outlined it— is to determine our beliefs concerning a natural number “chosen at random.” Its more figurative version is “de Finetti’s Lottery” [Bartha, 2004] in which a lottery ticket is picked at random from a countable infinity of tickets. If we write the prior probability for numbers $1, 2, 3, \dots$ as $p_1, p_2, p_3, \dots$, we cannot reconcile two conditions. First, since we have no preference for any number over any other, we assign the same probability to each number

$$p_i = p_k \text{ all } i, k$$

Second, the sum of all probabilities must be unity

$$p_1 + p_2 + p_3 + \dots = 1 \tag{CA}$$

No set of values for p_i can satisfy both conditions.³⁰

De Finetti’s own solution was to note that the condition (CA), “countable additivity” (as applied to this example), is logically stronger than finite additivity (A). The latter applies only to a finite set of outcomes — say, that the number chosen is a number in $\{1, 2, \dots, n\}$. It asserts that the probability of this finite set is the finite sum $p_1 + p_2 + p_3 + \dots + p_n$. Condition (CA) adds the requirement that this relation continues to hold in the limit of $n \rightarrow \infty$. De Finetti asserted that the

²⁹The problem of ignorance over a countable infinity of outcomes is actually no worse than the corresponding problem with a continuum outcome space. The latter problem contains the former in that a continuum outcome space can be partitioned into a countable infinity of subsets. Perhaps the countable case is deemed more problematic since outcomes can still be counted so it appears (incorrectly) that there will be a unique, natural probability measure recoverable from ratios of counts. In the continuum case, no such illusions are possible. The continuum case is more problematic in the sense that in it non-zero probabilities cannot be assigned to individual outcomes. Instead a probability density is used to assign probabilities to certain sets of outcomes. That presumes the outcome space has a topology that admits an additive measure. This device of a probability density becomes harder to use as the outcome space becomes larger. If the space consists of all real numbers, then the only uniform probability density is an improper density that cannot be normalized to unity.

³⁰Let $p = p_i = p_k$, then $p_1 + p_2 + p_3 + \dots = 0$ or ∞ , according to whether p is zero or non-zero, both of which contradict (CA).

probability distribution representing our epistemic state in this problem need only be finitely additive, not countably additive. That allows us to set $p_i = 0$ for all i , without forcing the probability of infinite sets of outcomes to be zero. So if $odd = \{1, 3, 5, \dots\}$ and $even = \{2, 4, 6, \dots\}$ we can still set $P(odd) = P(even) = 0.5$.

Solving the problem by dropping countable additivity has proven to be a popular and well-understood solution. While proponents of the restriction to finite additivity are typically not frequentists (who identify probabilities with relative frequencies), the connection to frequentism is natural. The frequency of any particular natural number among all is zero and the frequency of even numbers is $1/2$, in a naturally defined limit. Kadane and O'Hagan [1995] have mapped out which uniform, finitely additive probability distributions are possible over the natural numbers, noting how these are delimited by natural conditions such as agreement with limiting frequencies and invariance of probability under translation of a set of numbers. There are also variant forms of the proposal, such as the use of Popper functions and the notion of "relative probability" [Bartha, and Johns, 2001, Bartha, 2004]. However dropping countable additivity is not without disadvantages and enthusiasm for it is not universal. There are Dutch book arguments that favor countable additivity; and important limit theorems, including Bayesian convergence of opinion theorems, depend upon countable additivity. See [Williamson, 1999; Howson and Urbach, 2006, pp. 26-29; Kelly, 1996, Ch. 13].³¹ Other approaches explore the possibility of assigning infinitesimally small probabilities to what would otherwise be zero probability outcomes [McGee, 1994].

While all these solutions come at some cost, they are eventually unavailing. For they have not addressed the problem in its most acute form. They deal only with the case of ignorance over natural numbers, where this set of a countable infinity of outcomes has a natural order. If we presume that we know of no such natural order, then all these solutions fail, as has been shown in a paradox reported by Bartha [2004, §5.1] and in the work of Arntzenius [manuscript].

Imagine that we have a countable infinity of outcomes with no way to order them. If we have some labeling of the outcomes by natural numbers, that numbering is completely arbitrary.³² Let us pick some arbitrary labeling of the outcomes, $1, 2, 3, \dots$, and seek an ignorance distribution over these labels. That ignorance distribution should be unaffected by any one-to-one relabeling of the outcomes; that is, the ignorance distribution is invariant under a permutation of the labels, for permutations are a symmetry of this system. Consider the outcomes $odd = \{1, 3, 5, \dots\}$ and $even = \{2, 4, 6, \dots\}$. There is a permutation that simply

³¹To someone with an interest in physics, where probabilities are routinely summed over infinitely many outcomes, the restriction to finite additivity appears ruinous to ordinary physical theorizing. Countable additivity is hidden in many places. It is presumed whenever we normalize a probability distribution $p(x)$ over some real-valued parameter x . For example $1 = \int_0^1 p(x)dx = \int_0^{1/2} p(x)dx + \int_{1/2}^{3/4} p(x)dx + \int_{3/4}^{7/8} p(x)dx + \dots$

³²For a concrete example, imagine that an infinite space is partitioned into a countable infinity of geometrically identical cubes and that our cosmology tells us that a single hydrogen atom will appear in one of them without favoring any one of them. We arbitrarily labeling the cubes as $1, 2, 3, \dots$

switches the labels of the two sets ($1 \leftrightarrow 2, 3 \leftrightarrow 4, 5 \leftrightarrow 6, \dots$), so that outcomes in each set are exchanged. Since our belief distribution is invariant under such a permutation, it follows that the permutation does not alter our belief and we must have the same belief in each outcome set *odd* and *even*. Now consider the four outcome sets

$$\begin{aligned} one &= \{1, 5, 9, 13, \dots\} = \{4i + 1 : i = 0, 1, 2, 3, \dots\} \\ two &= \{2, 6, 10, 14, \dots\} = \{4i + 2 : i = 0, 1, 2, 3, \dots\} \\ three &= \{3, 7, 11, 15, \dots\} = \{4i + 3 : i = 0, 1, 2, 3, \dots\} \\ four &= \{4, 8, 12, 16, \dots\} = \{4i + 4 : i = 0, 1, 2, 3, \dots\} \end{aligned}$$

There is a permutation that switches labels of *one* and *two*; so we have equal belief in *one* and *two*. Proceeding pairwise through the sets, we find we must have equal belief in each of *one*, *two*, *three* and *four*. Now there is also a pairwise permutation that switches the labels of *one* with those of $two \cup three \cup four$, where

$$two \cup three \cup four = \{2, 3, 4, 6, 7, 8, 10, 11, 12, 14, 15, 16, \dots\}$$

It is just the obvious permutation read off the above sets

$$1 \leftrightarrow 2, 5 \leftrightarrow 3, 9 \leftrightarrow 4, 13 \leftrightarrow 6, 17 \leftrightarrow 7, 21 \leftrightarrow 8, \dots$$

So we now infer that we must have the same belief in *one* as in $two \cup three \cup four$. Combining we find: we must have the same belief in $two \cup three \cup four$ and in each of its disjunctive parts *two*, *three* and *four*. This requirement cannot be met by a probability distribution for it contradicts (finite) additivity.³³ It is however compatible with the ignorance distribution (I).

Finally, it is sometimes remarked that the very idea of uniform ignorance over a countable set is somehow illicit for there is no mechanical contrivance that could select outcomes so that they have equal chances. No lottery commission could build a device that would implement the de Finetti lottery. For references to these concerns, see [Bartha, 2004, pp. 304-305], who correctly objects that the inevitable non-uniformity of probabilities of the lottery machine does not force non-uniformity of beliefs. What needs to be added is that the entire objection is based on a circularity. In it, the notion of a mechanical contrivance tacitly supposes a contrivance whose outcomes are governed by a probability distribution. So it amounts to saying that no machine whose outcomes are governed by a probability distribution can generate outcomes governed by a non-probabilistic distribution. Recent work in philosophy of physics has identified idealized physical mechanisms that produce indeterministic outcomes that are *not* governed by a probability distribution. An example is the “dome,” described in [Norton, 2007, §8.3; forthcoming], where it is shown that the dome’s outcomes are governed by a non-probabilistic distribution

³³Another possibility is the improper probability distribution that assigns some small probability ε to each, individual outcome and infinite probability to the total outcome space. This improper distribution is badly behaved under conditionalization. For example $P(2|even) = 0$, not $\varepsilon/2$; and $P(two|even) = \infty/\infty = \text{undefined}$, not $1/2$.

with the same structure as (I). There are many more examples of these sorts of indeterministic systems in the “supertask” literature. (For a survey, see [Laraudogoitia, 2004].) Many of these mechanisms conform with Newtonian mechanics, but depend on idealizations some find “unphysical” for their distinctive behavior.³⁴ These processes could be the physical basis of an idealized contrivance that implements something like the de Finetti lottery.

5 BAYESIAN DYNAMICS

5.1 *The Property: Refute and Rescale*

The properties investigated so far — framework and additivity — are essential to Bayesian confirmation theory. However we are still missing the essential part of the theory from which its name is derived. For so far, we have no means to relate probability measures conditioned on different propositions, so that we cannot yet relate the posterior probability $P(H|E)$ of some hypothesis H conditioned on evidence E with its prior probability $P(H) = P(H|B)$, for some background B . (Here and henceforth, for notational convenience, “ $P(A)$ ” will be written as shorthand for “ $P(A|B)$ ”, for some assumed background B .)

These means for relating prior and posterior probabilities are supplied by two properties. The first property is “narrowness” and asserts for any A and C that

$$P(A \& C|C) = P(A|C) \quad (N)$$

The second property is “multiplication,” which asserts for any A and C that

$$P(A \& C) = P(A \& C|C) \cdot P(C) = P(C \& A|A) \cdot P(C) \quad (M)$$

These properties combined entail³⁵

$$P(A|C) = P(A \& C)/P(C)$$

when $P(C)$ is not zero. That this formula arises through compounding the two properties (N) and (M) is not generally reported. That compounding is important in the discussion that follows, since these two properties make distinct assertions in the context of confirmation theory and require separate analysis.

These two properties (N) and (M) can also be combined to yield Bayes’ theorem. For hypothesis H and evidence E :

$$\begin{aligned} P(H \& E) &= P(H \& E|E) \cdot P(E) = P(H|E) \cdot P(E) \\ &= P(E \& H) = P(E \& H|H) \cdot P(H) = P(E|H) \cdot P(H) \end{aligned}$$

³⁴For an analysis of the notion of “physical” in this context and the suggestion that we discount these concerns, see [Norton, 2008a].

³⁵The combined formula is sometimes regarded as the definition of conditional probability in terms of unconditional probability. That view is not taken here since $P(A)$ is not unconditional, but shorthand for $P(A|B)$. Hájek [2003] has objected that this candidate definition fails to yield an everywhere serviceable notion of conditional probability.

so that when $P(E)$ is not zero we have Bayes' theorem

$$P(H|E) = \frac{P(E|H)}{P(E)}P(H) \quad (B)$$

It tells us how learning evidence E should lead us to update our beliefs, as expressed in the transition from the prior $P(H)$ to the posterior probability $P(H|E)$.

Bayes' theorem embodies a very simple model of belief dynamics whose two steps essentially correspond to the two properties (N) and (M) above.

Refute. When evidence E is learned, those parts of the hypothesis H logically incompatible with the evidence are discarded as irrelevant to the bearing of evidence. That is, the hypothesis H has two parts in relation to evidence E : $H = (H \& E) \vee (H \& \sim E)$. The second part $(H \& \sim E)$ is that part which is refuted by E . The first part $(H \& E)$ is the part of H that entails E . H can only accrue support through this part, since we have from (N) that $P(H|E) = P(H \& E|E)$.

Rescale. The import of evidence E on competing hypotheses $H, H', \dots$ is then simply the prior probability $P(H \& E), P(H' \& E), \dots$ of those parts of the hypotheses logically compatible with the evidence, linearly rescaled so as to preserve normalization to unity; for, we have from (B) and (N) that

$$\frac{P(H|E)}{P(H'|E)} = \frac{P(H \& E|E)}{P(H' \& E|E)} = \frac{P(E|H \& E) \cdot P(H \& E)}{P(E|H' \& E) \cdot P(H' \& E)} = \frac{P(H \& E)}{P(H' \& E)}$$

since $P(E|H \& E) = 1 = P(E|H' \& E)$.

A figurative model of this dynamics underscore its very great simplicity. In his "muddy Venn diagram," Van Fraassen's [1990. p. 161-62] pictures the total outcome space as a surface, such as a table top, and the degrees of belief assigned to outcome sets are represented by volumes of mud piled over the areas corresponding to each outcome. Conditionalization on evidence E occurs in two steps, as shown in Figure 1. First ("*refute*"), all the mud piled outside E is carefully swept away. Second ("*rescale*"), that swept-away mud is redeposited on E in such a way that the proportions in the contours over E are retained.³⁶

It is informally evident that this qualitative notion of refute and rescale dynamics captures the essence of Bayesian dynamics. That informal idea has been made more precise in [Norton, 2007, §4], where the dependencies of refute and rescale dynamics are used to generate an axiom system for degrees, which turn out to be ordinary probabilities up to monotonic rescalings.

³⁶For example, if the mud is twice as high at one spot in relation to another prior to the redepositing of the mud, it will be piled twice as high after, as well.

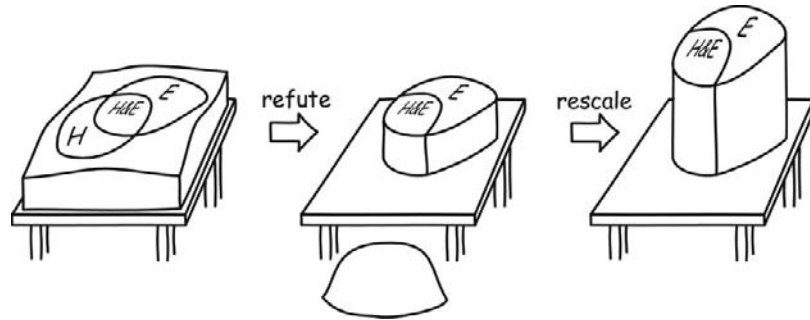


Figure 1. Muddy Venn diagram illustrates rescale and refute dynamics

5.2 Non- Bayesian Shifts

Under refute and rescale dynamics, the import of evidence is always to modify prior beliefs. Earman [1992, pp. 195-198] and Brown [1996] state the obvious difficulty. The dynamics cannot apply if there is no prior belief since the hypothesis or outcome in question is not in the outcome space. Just such a circumstance occurs at decisive moments in the history of science. At moments of rapid change — best known through Thomas Kuhn's notion of a scientific revolution — the inconceivable becomes conceivable and, we are told, that evidence compels us to believe it. When Einstein proposed his special theory of relativity in 1905, he urged that there is no absolute fact as to whether spatially separated events are simultaneous; the simultaneity of events depends upon the inertial reference frame from which they are assessed. That notion was regarded as bizarre by Einstein's critics, many of whom continued to refuse to take it seriously long after 1905. Another example is Bohr's theory of the atom of 1913. That theory seemed to require commitment to an inconsistency: in different processes, accelerated electrons bound in an atom would be supposed to radiate, as classical electrodynamics required, or would be supposed not to radiate, in contradiction with classical electrodynamics, according to which supposition was expedient for the recovery of the results Bohr needed.

Closely connected to the problem of outcomes not in the space are those to which it is natural to assign zero probability. A logical inconsistency, such as Bohr's theory, is a natural candidate. It follows immediately from Bayes' theorem that, once we have assigned a zero prior probability to any hypothesis, conditionalization cannot change its probability from zero. The only recovery is to reassign a non-zero probability to the hypothesis by a process that contradicts Bayes' theorem.

We might discount as rarities these inversions of what is conceivable and the oddity of accepting inconsistencies as established law. However even if they are rarities, they are important rarities that reappear throughout the history of our science and cannot be neglected by a comprehensive account of scientific rationality. The natural Bayesian rejoinder is to protest that too much is being asked of it. The formation of new outcome spaces belongs to Reichenbach's context of discov-

ery. Bayesian confirmation theory, indeed any logic at all, cannot be responsible both for logical relations among propositions and also for the discovery methods used to create new outcome spaces.

5.3 *Is Bayes' Theorem Sensitive Enough to the Connections between Evidence and Theory?*

The simplicity of refute and rescale dynamics invites the next challenge: the dynamics is too simple and too insensitive to the ways that evidence and theory can relate.

5.3.1 *Problems Arising from "Refute..."*

The first "refute" step of Bayesian conditionalization depends essentially on the idea that the disjunctive part of $H = (H \& E) \vee (H \& \sim E)$ that extends beyond the total evidence E , that is $H \& \sim E$, does not affect the bearing of evidence E on H . That is, "narrowness" directly asserts that $P(H|E) = P(H \& E|E)$. While this idea is so familiar as generally to pass without comment, it does represent a blindness in Bayesian confirmation theory. As an illustration of it, imagine that we seek to identify some animal. Our evidence is that it is a bird. What support does that evidence lend to the possibility that the animal is a canary or, alternatively, a canary or a whale? Narrowness asserts that it lends the same support:

$$P(\text{canary or whale} \mid \text{bird}) = P(\text{canary} \mid \text{bird})$$

In some scenarios, this can make sense. As we check instances of birds, we will find the frequency of "canary or whale" to arise exactly as often as "canary." If these frequencies are all that matters, then we would say the two outcomes are equally supported. However few would halt there. We would discount the "whale" disjunct as a nuisance embellishment that should be dismissed as a distraction precisely because it is deductively refuted by the evidence "bird." The frequency count just mentioned is blind to its nuisance character in so far as it assigns the same frequency of success to "canary" as to "canary or whale."

It is possible to devise a logic that is sensitive to this problem and punishes an hypothesis in the degree that it extends beyond the evidence. An example³⁷ is the "specific conditioning" logic defined in [Norton, forthcoming b, Section 10.2]. In a formulation compatible with developments here, the logic amounts to a new definition for the conditional probability $P(H|E)$. It starts with an additive measure

³⁷Specific conditioning resembles some of the measures of incremental support that have been investigated in the Bayesian literature, such as the ratio measure $r(H, E) = P(H|E)/P(H) = P(H \& E)/(P(E).P(H))$. (See [Eels and Fitelson, 2002].) The two differ essentially. The incremental measures are functions of three arguments, H, E and B , that represent the incremental support accorded H by evidence E alone with respect to a tacit background B . $P_{SC}(H|E)$ is a function of two arguments, H and $E \& B$, that measures the total support accorded to H by evidence E conjoined with the background B , that is, by the total evidence $E \& B$.

$P(\cdot)$ on the background and defines the alternative

$$P_{SC}(H|E) = P(H \& E)^2 / (P(E) \cdot P(H)) = P(H \& E) / (P(E) \cdot P(H \& E) / (P(H))) \quad (SC)$$

This alternative rule of conditionalization rewards hypotheses H with more support the closer they come to the total evidence $E = (E \text{ and background})$ — hence the term “specific conditioning.” $P_{SC}(H|E) = 1$ only when $H = E$ excepting measure zero differences. H can differ from E in two ways and the logic of (SC) penalizes it for both. In the first way, H can fail to exhaust E in measure. Then the first factor of (SC), that is, $P(H \& E) / (P(E))$, is less than one (which alone would correspond to the usual case of Bayesian conditionalization). In the second way, H can extend beyond E in measure. Then the second factor of (SC), that is $P(H \& E) / (P(H))$, is less than one. Deviations in both ways are penalized equally and it turns out that this is expressed in a striking symmetry in P_{SC} . That is $P_{SC}(H|E) = P_{SC}(E|H)$.

Returning to the bird example, with plausible values for $P_{SC}(\cdot)$, we will have

$$P_{SC}(\text{canary or whale} \mid \text{bird}) < P_{SC}(\text{canary} \mid \text{bird})$$

Indeed we will have

$$P_{SC}(\text{canary or whale} \mid \text{canary}) < P_{SC}(\text{canary} \mid \text{canary}) = 1$$

That is, even though “canary” deductively entails “canary or whale,” the latter is not accorded full support from “canary.” The logic of (SC) judges “canary or whale” to be supported less specifically by the evidence “canary.” In the case of Bayesian confirmation theory, evidence accords unit support to any of its deductive consequences, no matter how much they are weakened logically. P_{SC} reduces the support accorded to these deductive consequences according to how much they are weakened logically.

5.3.2 Problems Arising from “... Rescale”

In the special case in which the hypothesis H deductively entails the evidence E , the first “refute” step of Bayesian conditionalization is not invoked and the process reduces to rescaling only by means of the property “multiplication” (M). This special case provides a means of isolating the second step and plumbing its problems.

Consider two hypotheses H and H' both of which entail the evidence E . In that case, it follows from Bayes’ theorem that the ratio of the posterior probabilities $P(H|E)/P(H'|E)$ is the same as the ratio of the priors. It now follows that the incremental confirmation, as measured by the ratio of posterior and prior $P(H|E)/P(H)$, is the same in both cases. That is, when two hypotheses both entail the same true evidence, they get the same confirmatory boost. In particular, if H is more probable than, less probable than or just as probable as H' prior to conditionalization, it will remain so afterwards.

This failure to separate H and H' evidentially, according to critics, is to overlook that their entailments of the evidence E can differ in more subtle qualities that have epistemic import. For example, recall efforts in the late 19th century to detect the earth's motion through the luminiferous ether. These experiments yielded a null result (E). We may entertain two hypotheses:

H : There is no ether state of rest against which the earth can move.

H' : There is an ether state of rest but we happen to be motionless with respect to it each time the experiments are performed.

Each of H and H' entail the evidence E , but it seems wrong that both should receive the same support from E . H succeeds, as it were, by honest toil; but H' by theft. There are many ways that this metaphor of honest toil and theft is explicated.

One way notes that H succeeds because it explains why the experiments yielded null results, where as H' does not. Considerations like this lead Achinstein [2001, Ch. 7] to argue that it is not sufficient to count as evidence that some datum incrementally increases the probability of an hypothesis or that the probability of the hypothesis conditioned on the datum is high.³⁸ In addition there must be an explanatory connection.

Other approaches are variants of this idea that some successful entailments are virtuous and others not. The hypothesis H makes the prediction that all ether current experiments will fail and that prediction is verified by the evidence. Since H' supposes that it is by happenstance that we were at rest in the ether just at the moment of the experiments, H' simply accommodates the evidence in the sense that it has been adjusted to accommodate experimental reports at hand. Predictions, for example, are virtuous and their success merits an increase in belief; accommodations are not virtuous and are epistemically inert. See [Horwich, 1982, Ch. 5] for discussion and an argument that we should not differentially reward prediction and accommodation. Another way of expressing these concerns is to note that H' (but not H) was cooked up “ad hoc” specifically to match the evidence. A classic example is the creationist hypothesis that the world was created in 4004 BC, complete with a fossil record that perfectly simulates a more ancient, evolutionary past. It is an evidential truism that such ad hoc hypotheses gain no support from the observations they accommodate. For an account of how ad hoc hypotheses can be treated in a Bayesian context, see [Howson and Urbach, 2006, pp. 121-126], who also suggest that sometimes there are “good” ad hoc hypotheses that deserve support.

The literature on these evidential virtues in Bayesian confirmation theory is large and convoluted. In general, however, there are two broad strategies that Bayesian confirmation theory can employ when faced with a virtuous and non-virtuous pair of hypotheses. The first is to expand the outcome space so that there

³⁸See [Achinstein, 2001, Ch.7] and *passim*) for numerous examples. For example, learning that most lottery tickets were unsold may increase greatly my belief that mine is the winning ticket, even though the absolute probability may remain small. See also [Laudan, 1997].

are more resources available to pick the hypotheses apart.³⁹ That risks merely postponing the problem since it may now return for another pair of hypotheses in the larger space. The second is to use prior probabilities to reward virtuous hypotheses and punish non-virtuous ones. That is, since the ratio of the posterior probabilities $P(H|E)/P(H'|E)$ equals the ratio of the priors $P(H)/P(H')$, we may reward the explanatory, predictive or non-ad hoc H by making the ratio $P(H)/P(H')$ very large, so that there is a corresponding advantage for H in the posterior probabilities. As observed in [Norton, 2007, §5.3.3], this use of prior probabilities requires a curious prescience: we are to penalize the prior probability of an ad hoc hypothesis in advance in just the right degree so that the punishment perfectly cancels the as yet unknown confirmatory reward that will be accorded to the hypotheses by Bayes' theorem. Further, prior probabilities for hypotheses or theories are assigned once, globally, whereas virtues are manifested locally and may differ from domain to domain. So we may find a theory explanatory in one domain and want to reward it with a high prior probability; but may find it explanatorily deficient in another, and may want to punish it with a low prior probability.

5.4 The “Likelihood Theory of Evidence” and “Likelihoodism”

The above discussion has dealt with the case of hypotheses that entail the evidence. Matters are slightly more complicated when the hypothesis only makes the evidence more or less probable. In that case, Bayes' theorem allows us to compare the import of evidence E for two hypotheses H and H' as

$$\frac{P(H|E)}{P(H'|E)} = \frac{P(E|H)}{P(E|H')} \frac{P(H)}{P(H')} \quad (B')$$

It follows that the relative import of the evidence E is fully determined by the two likelihoods, $P(E|H)$ and $P(E|H')$. For the ratio of the prior probabilities $P(H)/P(H')$ reflects our comparative belief in H and H' prior to any consideration of evidence E ; the ratio of the posteriors $P(H|E)/P(H'|E)$ reflects our comparative belief in H and H' after the import of evidence E has been accommodated. The ratio of the likelihoods $P(E|H)/P(E|H')$ is what takes us from the ratio of the prior to the ratio of the posteriors. Therefore, it expresses the relative import of evidence E on these two hypotheses.

That the likelihoods capture all needed information for evaluating the bearing of evidence is the core notion of a “likelihood theory of evidence.” (In the special case in which the hypotheses entail the evidence, $P(E|H) = P(E|H') = 1$. Then the ratio of the posteriors equals the ratio of the priors and E cannot discriminate evidentially between H and H' .)

Since so many of the problems of Bayesian confirmation theory focus on prior probabilities, an attractive modification to the theory is a view that excises prior

³⁹For example Maher treats prediction and accommodation by supposing that hypotheses are generated by methods that may or may not have certain items of evidence available to them when they generate the hypotheses. He then shifts assessment to the credibility of the methods. See [Achinstein, 2001, pp. 215–221] for discussion.

probabilities but leaves the rest intact. The view, popularly known as “likelihoodism,” has been advocated by Edwards [1972] and Royall [1997]. As we have just seen, in Bayes’ theorem (B'), the evidence E favors the hypothesis H over H' just in the ratio of their likelihoods, $P(E|H)/P(E|H')$. The proposal is that this consequence of Bayes’ theorem be extracted and elevated to an independent principle, the “law of likelihood.” The judgment of evidential import is always comparative — evidence favors this hypothesis more or less than that — and the troublesome prior probabilities never need enter.

The principal difficulty for likelihoodism is that it faces all the problems of the Bayesian refute and rescale dynamics recounted here, but without the resources of prior probabilities to ameliorate them. In particular, likelihoodists must say that evidence entailed by two hypotheses is equally favorable to both hypotheses no matter what their virtues or vices. For likelihoodists have lost the Bayesian mechanism of using prior probabilities to reward simpler hypotheses or ones with anticipated predictive power and to penalize ad hoc or contrived hypotheses. Royall [1997, §1.7] considers and offers likelihoodist rejoinders to such concerns.

5.5 *Model Selection, Prediction and Simplicity*

While the notion that the likelihoods capture the incremental import of evidence has enjoyed notable successes, its overall history has been troubled. There are recalcitrant cases in which the likelihoods alone are clearly too coarse a measure of the import of evidence. Likelihood rewards only accuracy in fitting data at hand. It has difficulty accommodating our common preference for simpler hypotheses that may be less accurate. In rewarding fit to the data at hand, likelihood may be a weaker guide to the greater truth sought that extends beyond these data; that is, likelihood tends to favor accommodation to present data as opposed to predictive success with future data.

The familiar and long-standing illustration of these difficulties comes in the problem of curve fitting. Let us say that we wish to find the relationship between variables x and y :⁴⁰

$$\langle x_1, y_1 \rangle = \langle 0.7, 1.0 \rangle, \langle x_2, y_2 \rangle = \langle 1.5, 1.8 \rangle, \langle x_3, y_3 \rangle = \langle 2.1, 2.0 \rangle, \dots, \quad (DATA)$$

which are shown on the graph of Figure 2 below. Our presumption is that these data were generated by the relation

$$y_i = f(x_i) + \text{error}_i$$

where error_i is a random error term affecting the i -th pair of $x - y$ values. (It is common to assume that these error terms are normally distributed and independent of one another.) The goal of our analysis is identification of the function

⁴⁰The full data set is $\langle 0.7, 1.0 \rangle, \langle 1.5, 1.8 \rangle, \langle 2.1, 2.0 \rangle, \langle 2.3, 0.6 \rangle, \langle 2.6, 1.0 \rangle, \langle 3.8, 2.1 \rangle, \langle 4.5, 3.1 \rangle, \langle 4.7, 6.0 \rangle, \langle 5.6, 6.9 \rangle, \langle 5.6, 7.7 \rangle, \langle 5.8, 4.9 \rangle, \langle 6.2, 4.4 \rangle, \langle 7.1, 7.7 \rangle, \langle 7.6, 6.7 \rangle, \langle 8.8, 10.1 \rangle, \langle 8.9, 8.2 \rangle, \langle 9.1, 8.1 \rangle, \langle 9.3, 7.4 \rangle$.

$f(x)$. The standard method is to find that function that generates a curve of best fit to the data. For data such as in Figure 2, we would normally seek a straight line. That is, we would assume that the unknown function is linear, so that the data was generated by

$$y_i = A + Bx_i + \text{error}_i \quad (LIN)$$

That straight line turns out to be

$$y_i = -0.332 + 0.997x_i + \text{error}_i \quad (LIN_{best})$$

The straight line LIN_{best} is the best fit to the data in the sense that it is the straight line drawn from LIN that makes the DATA most probable. That is, its values of A and B maximize the likelihood $P(\text{DATA} \mid LIN_{best})$; and $A = -0.332$ and $B = 0.997$ are the “maximum likelihood estimators” of A and B .

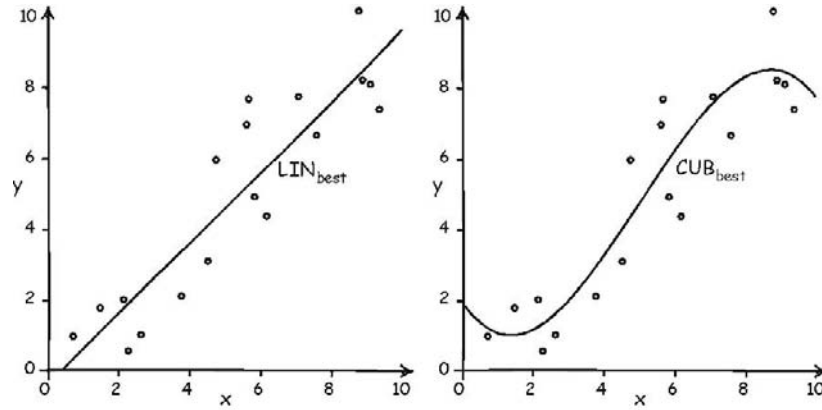


Figure 2. Linear and cubic curves of best fit

The complication is that we can find other functions $f(x)$ that fit the data more closely and make it even more probable. Consider, for example, cubic functions

$$y_i = A + Bx_i + Cx_i^2 + Dx_i^3 + \text{error}_i \quad (CUB)$$

The particular cubic plotted in Figure 2 against the background of the same data,

$$y_i = 1.952 - 1.377x_i + 0.581x_i^2 - 0.0389x_i^3 + \text{error}_i \quad (CUB_{best})$$

maximizes the likelihood $P(\text{DATA} \mid CUB_{best})$.

As one can see from comparing the two curves visually, the cubic CUB_{best} fits the data more closely than the linear LIN_{best} and it turns out that

$$P(\text{DATA} \mid CUB_{best}) > P(\text{DATA} \mid LIN_{best})$$

Therefore, according to the likelihood theory of evidence, we should conclude that the data better supports the cubic CUB_{best} over the linear LIN_{best} .

It is evident from a cursory scan of the data that this is most likely a mistake. Whatever trend $f(x)$ may be embodied within the data, it is heavily confounded by noise from the error term. While the data definitely gives us warrant to believe that $f(x)$ is an increasing function of x , we have no warrant for the specifics of the cubic CUB_{best} .⁴¹ We should be much more modest in what we infer from data this noisy. The better fit of the cubic CUB_{best} has been achieved by the greater ability of a cubic to conform to random fluctuations from the trend induced by noise; that is, the curve is “overfitted.” This overfitting is most evident around $x = 1$, where the curve reverses direction; and again around $x = 9$, where the curve also reverses direction. Given the sparsity and scattering of the data, cubic CUB_{best} is merely tracking noise at these places, whereas the linear LIN_{best} is responding to a real increase with x in the trend.

The informal discussion of the last paragraph captures the practical thinking of curve fitting. It is grounded tacitly in the considerations of simplicity and prediction at issue here. We know that we can always find a curve that fits the data better by looking to more complicated functional forms, such as higher order polynomials. However, at some stage, we judge that we should prefer the simpler curve over the more complicated one that fits the data better. The reason is that our interests extend beyond the narrow problem of finding a curve that merely fits this particular data set well. Our data is only a sample of many more values of x and y ; and we are interested in finding a function $f(x)$ that will also fit these other as yet unseen values. That is, our goal is to find a function $f(x)$ that will support prediction. We expect that we can best achieve that goal by forgoing some accuracy of fit with the present data of an overfitted curve in favor of a simpler functional form for $f(x)$ that will enable future data to be fitted better. For an overfitted curve responds to random noise whose specific patterns of deviations will probably not reappear in future data; however a curve drawn from the simpler model responds more to the real relation that drives present and, we suppose, future data as well.

The challenge to a Bayesian analysis is to find a way of capturing these last informal thoughts in a more precise analysis. The natural way to do this is to break up the inference problem into two parts.⁴² The first part requires us to consider only models. These are sets of hypotheses indexed by parameters. For example, LIN above is the set of linear functions with parameters A and B taking all real values. CUB above is the set of all cubic functions with the parameters A, B, C and D taking all real values. In the first part we decide which model is

⁴¹Perhaps it is unfair to introduce a more elevated perspective that affirms this judgment against the cubic CUB_{best} . The data plotted was generated artificially from the model $y_i = x_i + \text{error}_i$, so the linear LIN_{best} has actually done a very good job of identifying the trend $y = x$.

⁴²Traditional non-Bayesian statistical methodology implements this two part procedure by testing whether deviations in the coefficients B and C of CUB from zero are statistically significant. If they are not, the null hypothesis of $B=C=0$ is accepted and estimation proceeds with the simple model LIN .

appropriate to the data. For DATA above, we presume that would turn out to be LIN and not CUB. In the second part of the inference problem, we then find the curve that fits the data best just within that selected model.

Implementing this program within Bayesianism runs into several difficulties. The functions of LIN are a subset of those of CUB, for CUB reverts to LIN if we set $C = D = 0$. So if we compute the posterior probabilities of the models on the DATA, we will always find

$$P(\text{LIN} \mid \text{DATA}) \leq P(\text{CUB} \mid \text{DATA})$$

from which it follows that the model LIN can never be more probable on the DATA than the model CUB. So it seems that we cannot have a straightforward Bayesian justification for preferring LIN over CUB.

Slightly less than straightforward grounds can be found. We compare not the posterior probabilities of models, but the boosts in probability each model sustains upon conditionalizing on the data. That is, in this modified approach, we would prefer LIN over CUB if the ratio $P(\text{LIN} \mid \text{DATA}) / P(\text{LIN})$ exceeds $P(\text{CUB} \mid \text{DATA}) / P(\text{CUB})$. The ratio of these two ratios is itself called the “Bayes’ factor.”

This modified criterion of the Bayes’ factor has some plausibility since it conforms to the spirit of the likelihood theory in that we consider changes in probability under conditionalization, not absolute values of the posterior probability. However the modified criterion rapidly runs into difficulties with prior probabilities. To compute $P(\text{LIN} \mid \text{DATA})$, for example, we would use Bayes’ theorem, in which the likelihood $P(\text{DATA} \mid \text{LIN})$ appears. This likelihood is really a compound quantity. It is a summation over the infinitely many curves that appear in the model LIN, corresponding to all possible values of A and B . That is,

$$P(\text{DATA} \mid \text{LIN}) = \int_{\text{all } A, B} p(\text{DATA} \mid A, B) p(A, B) dA dB$$

The problematic term is $p(A, B)$, which is the prior probability density for the parameters A and B given that LIN is the correct model. Since it expresses our distribution of belief over different values of A and B prior to evidence, $p(A, B)$ should favor no particular values of A or B . However it is familiar problem that no probability distribution can do this. If $p(A, B)$ is any non-zero constant, then it cannot be normalized to unity when we integrate over all possible values of A and B .

While no unqualified solution to these problems has emerged, it has proven possible to show that, for large data sets, the Bayes factor depends only weakly on the choice of the prior probability density. In that case, the model favored by the Bayes factor is the one that maximizes what is known as the Bayes Information Criterion (BIC):

$$\text{BIC} = \log L_{\text{best}} - (k/2) \log n$$

where the maximum likelihood L_{best} is the likelihood of the data when conditionalized on the best fitting curve in the model, k is the number of parameters in the model and n is the size of the data set. (For LIN, k is 2; for CUB, k is 4.) For further discussion, see [Wasserman, 2000].

In maximizing BIC, we tradeoff accuracy of fit (expressed in the maximizing of likelihood) against the simplicity of the model (expressed in the number of parameters k). This was not the goal in constructing BIC; it was generated from an analysis that seeks the most probable model on the evidence. An analogous criterion can be generated by directly seeking the model with the greatest predictive power, as opposed to the one that is most probable. This criterion, the Akaike Information Criterion (AIC), advocated by Forster and Sober [1994], is given by

$$AIC = \log L_{best} - k$$

The rationale is that overtly seeking predictive power is a surer way to get to a truth that extends beyond the particular data at hand. Seeking the most probable model risks favoring the vagaries of the particular data at hand.

The connection to prediction is achieved through the notion of “cross-validation.” While we cannot now test the predictive powers of a model against presently unknown data, we can simulate such a test by dividing the data at hand into two parts. We use the first part to generate a best fitting hypothesis in the model and then we check how well the resulting hypothesis fits the remaining data. In a large set of N data, we leave out one datum and use the remaining $N - 1$ to generate hypotheses from the model. We then average the errors in fitting the original datum. If we allow each datum successively to be left out in the procedure and average the resulting scores, we recover an estimate of the predictive powers of the model. That estimate turns out to approach the AIC statistic for large N . [Forster, 2002, p. S128; Browne, 2000]. The presumption is that predictive powers manifested within a known data set in this computation will persist beyond the confines of that data set.

It remains an open question whether likelihoods may be used as the basis of assessment of models when we need to find an appropriate accommodation of accuracy, simplicity and predictive power. That all the proposals above fail has been suggested recently by Forster [2006; 2007]. He has devised examples of pairs of models with the same number of parameters k that also deliver the same likelihood and maximum likelihood for a specified set of data. As a result, considerations of likelihood alone or the more refined BIC or AIC are unable to discriminate between the models. However it is evident from informal considerations that one model is predictively superior to the other.

Forster’s simplest example⁴³ draws on a set of data with three points $\langle x_i, y_i \rangle$:

$$\langle 1, 1 \rangle, \langle 2, 2 \rangle, \langle 3, 3 \rangle \quad (\text{DATA}')$$

⁴³This example was presented in a talk by Malcolm Forster, “Is Scientific Reasoning Really that Simple?” in “Confirmation, Induction and Science” London School of Economics, March 8 – March 10, 2007, on March 8.

The first model, $H_a : y_i = ax_i$, comprises four functions

$$H_{1/2} : y_i = (1/2)x_i \quad H_1 : y_i = x_i \quad H_2 : y_i = 2x_i \quad H_3 : y_i = 3x_i$$

corresponding to the parameter values $a = 1/2, 1, 2, 3$. The relation of this model to DATA' is shown in Figure 3.

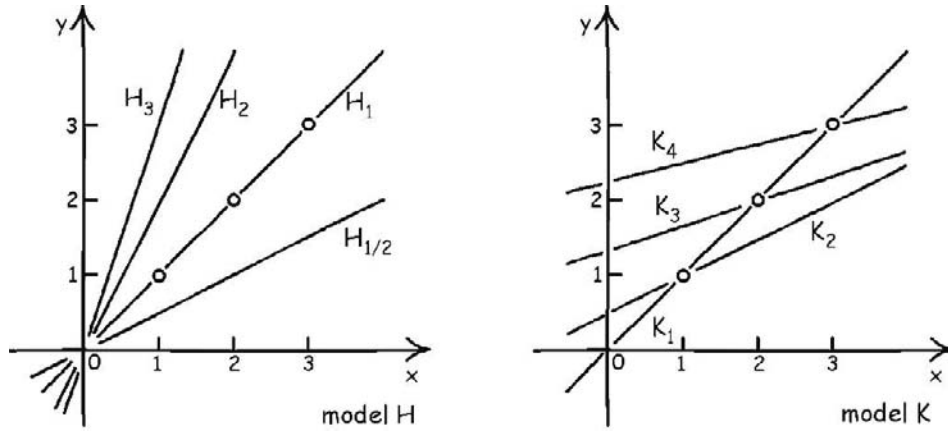


Figure 3. Models H and K

The second model, $K_a : y_i = x_i/a + (a-1)^2/a$, comprises four functions

$$K_1 : y_i = x_i \quad K_2 : y_i = x_i/2 + 1/2 \quad K_3 : y_i = x_i/3 + 4/3 \quad K_4 : y_i = x_i/4 + 9/4$$

corresponding to parameter values $a = 1, 2, 3, 4$. The relation of this model to DATA' is shown in Figure 3.

Which model does DATA' favor? The computation of the likelihoods of the individual functions is easy since they are all zero or one. The only non-zero likelihoods are

$$P(\text{DATA}' | H_1) = 1 \quad P(\text{DATA}' | K_1) = 1$$

We assume that the prior probabilities of functions within each model is the same. That is⁴⁴

$$P(H_a|H) = 1/4 \text{ for } a = 1/2, 1, 2, 3 \\ P(K_a|K) = 1/4 \text{ for } a = 1, 2, 3, 4$$

Thus the likelihood associated with each model is the same

$$P(\text{DATA}'|H) = P(\text{DATA}'|H_1) \times P(H_1|H) = 1/4 \\ P(\text{DATA}'|K) = P(\text{DATA}'|K_1) \times P(K_1|K) = 1/4$$

⁴⁴In these formulae, read $H = H_{1/2} \vee H_1 \vee H_2 \vee H_3$ and $K = K_1 \vee K_2 \vee K_3 \vee K_4$.

Hence, we cannot use DATA' to discriminate between the two models H and K by means of their maximum likelihoods or the likelihoods of the models. Moreover, we cannot discriminate between them by means of BIC or AIC; for both models agree on the maximum likelihood and the number of parameters, $k = 1$.

While all the likelihood based instruments fail to discriminate between the models, informal considerations favor the model H . For it requires only one datum from DATA', such as $\langle 1, 1 \rangle$, to fix the function $y = x$ in H that fits the remaining data. The remaining two data points act as tests of the model; or, we may view them as successful predictions. However it takes two data points from DATA' to fix the function in K that fits the remaining data. The datum $\langle 1, 1 \rangle$, for example, is compatible with both K_1 and K_2 . A second datum is needed to decide between them. Hence only one datum remains to test the function fitted; or the model needs more data before it can make successful predictions. As a result we informally judge that the model H is better tested and predictively stronger and thus better warranted by the evidence. (For discussion of this aspect of curve fitting, see [Glymour, 1980, Ch. VIII.]⁴⁵)

For further discussion of Bayesian and Likelihood based analyses of this problem, see [Howson and Urbach, 2006, 288-96] and [Forster and Sober, 2004] (which includes commentary by Michael Kruse and Robert J. Boik).

5.6 *A Neutral Starting Point: The Second Problem of the Priors*

Bayesian confirmation theory depicts learning from evidence as the successive conditionalization of a prior probability distribution on evidence so that the resulting posterior probability distributions increasingly reflect just the evidence learned. As we trace back along this chain of conditionalization, we expect to come to epistemically more neutral probability distributions, reflecting our lesser knowledge earlier in the process. The problem of the priors in both forms developed here is that the structure of Bayesian confirmation theory essentially precludes neutral probability distributions. In the first problem of the priors (Section 4.3 above), we saw that the additivity of probability measures precludes their representing epistemic states of complete ignorance. Here we shall see that the Bayesian dynamics of refute and rescale also precludes a different sort of neutrality of the probability distribution. Bayesian dynamics makes it impossible to select a prior probability distribution that does not exercise a controlling influence on subsequent course of conditionalization.⁴⁶

The essential problem is suggested by the notion of refute and rescale dynamics. The results of conditionalization must already be present in the prior probability; all that conditionalization does is to remove those parts of the prior probability

⁴⁵The informal considerations invoked here may appear more familiar if we vary the example slightly. Imagine that we are to choose between the model $H : y = ax$ and $L : y = ax + b$ for DATA'. Then we would allow that H is better warranted since a single datum is needed to fix the parameter a in H ; whereas two data are needed to fix the two parameters a and b in L .

⁴⁶The phrase "problem of the priors" seems to mean somewhat different to things to different authors. For further discussion, see [Earman, 1992, pp. 57-59; Mellor, 2005, pp. 94-95].

distribution attached to outcomes refuted by the evidence to reveal the pre-existing distribution of belief between the unrefuted outcomes. This same point is evident if we recall the formula governing conditional probabilities

$$P(H|E) = P(H \& E) / P(H)$$

This formula tells us that $P(H|E)$, the degree of belief we should have in H given that we have learned E , is fixed by the two prior probabilities $P(H \& E)$ and $P(H)$. That means that if we fully specify our prior probabilities over some outcome space, we have also delivered an exhaustive catalog of what our belief in each outcome would be, given that we have learned some other outcome. For from being neutral, the prior probability distribution anticipates how our beliefs will evolve as we learn any sequence of admissible outcomes and different prior probability distributions can yield courses that differ greatly. This consideration gives further grounds for the abandoning of such terms as “ignorance priors” or “informationless priors” in favor of “priors constructed by some formal rule” as reported in Section 4.4 above.

As with the first problem, objectivists have the most acute difficulty with this second problem of the priors. For they are committed to there being one, appropriate prior probability distribution in any given epistemic situation. Yet none can supply the neutral starting point appropriate in the absence of all evidence. So the objectivist must either accept the unpalatable conclusion that the initial prior probability contains information without evidential warrant; or that Bayesian dynamics can only be used once we have learned enough by other means to make a properly grounded selection of the prior probability.

Subjectivists at first seem to have a better response. For them, a prior probability is simply a statement of personal opinion with no pretensions of evidential warrant. As evidence accumulates and is incorporated into the probability distributions by conditionalization, limit theorems assure us of a merging of opinion onto a unique distribution that properly represents the bearing of evidence. (For discussion of these theorems, see [Earman, 1992, Ch. 6].) However this solution comes at some cost. At any finite stage of the process, the posterior probability distribution is an unknown mix of unwarranted opinion and warranted support. While the limit theorems may assure us that, in particular circumstances, the mix will eventually converge onto warranted support, at any definite state we may be arbitrarily far from it.⁴⁷ Imagine that we have collected some very large set of evidence E_{large} and we are interested in its bearing on some hypothesis H . It will always be possible for a determined mischief maker to identify some prior probabilities for $P(H \& E_{large})$ and $P(E_{large})$ so that $P(H|E_{large})$ is as close to one as you like or as close to zero as you like. The two differ essentially. The incremental measures are functions of three arguments, H , E and B , that represent the incremental support accorded H by evidence E alone with respect to a tacit background B . $P_{SC}(H|E)$ is a function of two arguments, H and $E \& B$, that measures

⁴⁷Here is the obligatory reporting of Keynes’ quip: “In the long run we are all dead.”

the total support accorded to H by evidence E conjoined with the background B , that is, by the total evidence $E \& B$.

5.7 The Prior Probability of Universal Generalizations

Neither objectivist nor subjectivist fares well if zero or unit prior probability has been assigned injudiciously to some outcome. Then Bayesian dynamics becomes completely dogmatic. For once they have been assigned zero or unit probability, it follows from Bayes' theorem that these assignments are unrevisable by conditionalization.⁴⁸ The complete disbelief of a zero prior probability and the complete belief of a unit probability will persist indefinitely. In concrete terms, this sort of dogmatism is a familiar if unwelcome phenomenon. Consider the dogmatic conspiracy theorist who discounts the benign explanation of a catastrophe in favor of an elaborate conspiracy committed by some secret agency. Once a zero prior probability has been assigned to the benign explanation, no failure to reveal the secret agency's intervention or even existence can restore belief in the benign explanation. Rather Bayes' theorem will provide mounting assurance that each failure is further evidence of the perfection of the secret agency's cover-up.

While the best Bayesian advice would seem to be very cautious before assigning a zero prior probability, there are suggestions, reviewed in more detail in [Earman, 1992, §4.2, 4.3], that hypotheses with the power of universal generalizations ought to be assigned a zero prior probability. These concerns arise for any hypotheses H that entails a countable infinity of consequences $E_1, E_2, \dots$ ⁴⁹ Considering only the first n consequences we have

$$P(H) = P(H|E_1 \& \dots \& E_n) \cdot P(E_n|E_1 \& \dots \& E_{n-1}) \cdot P(E_{n-1}|E_1 \& \dots \& E_{n-2}) \cdot \dots \cdot P(E_2|E_1) \cdot P(E_1)$$

Popper [1959, Appendix VII] urged that we should expect the instances $E_1, E_2, \dots$ of a universal generalization H to be equiprobable and independent of one another. That is,

$$P(E_n|E_1 \& \dots \& E_{n-1}) \cdot P(E_{n-1}|E_1 \& \dots \& E_{n-2}) \cdot \dots \cdot P(E_2|E_1) \cdot P(E_1) \\ = P(E_n) \cdot P(E_{n-1}) \cdot \dots \cdot P(E_2) \cdot P(E_1) = P(E_1)^n$$

But since this must hold for arbitrarily large n and with $P(E_1) < 1$, it follows that the prior $P(H) = 0$. The obvious weakness of Popper's proposal is that the very fact that $E_1, E_2, \dots$ are instances of a universal generalization H suggests that they are *not* independent. However, just a little dependence between $E_1, E_2, \dots$ is not enough to allow a non-zero prior for H . Inspection of the above expression for $P(H)$ shows that $P(H) > 0$ entails

⁴⁸Once $P(H) = 0$, Bayes' theorem (B) requires $P(H|E) = 0$ for all admissible E . (Conditionalizing on evidence E for which $P(E) = 0$ leads to an undefined posterior $P(H|E)$.) It follows that for any $H' = \sim H$ for which $P(H') = 1$, $P(H'|E) = 1$ as well.

⁴⁹The simplest example is a universal generalization that asserts, "For all $x, Q(x)$," where x ranges over a countably infinite set of individuals $a, b, c, \dots$. Its consequences are $Q(a), Q(b), \dots$

$$\lim_{n \rightarrow \infty} P(E_n | E_1 \& \dots \& E_{n-1}) = 1$$

That is, finitely many favorable instances of H eventually make us arbitrarily sure that the next instance will be favorable. That was an outcome that Jeffrey [1983, p. 194] felt sufficiently akin to “jumping to conclusions” to want to renounce non-zero prior probabilities on universal hypotheses.

Lest this commitment to the projectability of the hypothesis not seem immodest, the analysis needs only a slight modification to make it more striking. Group the instances $E_1, E_2, \dots$ into sets that grow very rapidly in size. For example

$$F_1 = E_1 \quad F_2 = E_2 \& \dots \& E_{10} \quad F_3 = E_{11} \& \dots \& E_{100} \dots$$

so that

$$F_1 \& \dots \& F_n = E_1 \& \dots \& E_{(10^{n-1})}$$

Each of the F_i is a consequence of H , so the above analysis can be repeated with F_i substituted for E_i to infer that $p(H) > 0$ entails

$$\lim_{n \rightarrow \infty} P(E_1 \& \dots \& E_{(10^{n+1})} | E_1 \& \dots \& E_{(10^n)}) = 1$$

That is, if we assign non-zero prior probability to an hypotheses with the power of a universal generalization, eventually we are willing to project from some finite set of its positive instances to arbitrarily many more at arbitrarily high probability.

The point is *not* that there is something intrinsically wrong with such immodesty. Indeed for the right hypothesis, this may be just the appropriate epistemic attitude. The point is that assigning a zero or a non-zero prior probability to a hypothesis with the power of a universal generalization is never neutral. The former commits us dogmatically never to learning the hypothesis; the second, no matter how small the non-zero prior, commits us to its arbitrary projectability on finite favorable evidence. There is no intermediate value to assign that leaves our options open. Our inductive course is set once the prior is assigned.

Finally, one sees that some significant prior knowledge is needed if we are to assign the non-zero prior probabilities prudently. For a countably infinite set of outcomes $E_1, E_2, \dots$ corresponds to an uncountable infinity of universal hypotheses: each hypothesis corresponds to one of the uncountably many combinations of the E_i and their negations $\sim E_i$. However we can assign non-zero prior probability to at most countably many of these hypotheses. That is, loosely speaking, when we assign our priors, we are selecting in advance the measure zero subset that we will be prepared to learn and dogmatically banishing the rest to disbelief no matter what evidence may accrue.

5.8 Uncertain Evidence

The versions of Bayesian confirmation theory considered so far depict learning as an absolute: if we are given evidence E , it cannot be doubted. That is quite unrealistic. No datum is ever certain and it is not uncommon in science that the accumulation of later evidence leads one to doubt the correctness of evidence collected

earlier. Something like this could be incorporated into traditional Bayesianism if we separate, say, the inerrant sense datum D : “my retina senses a plesiosaur like shape on the surface of Loch Ness” from the fallible evidence E : “I saw a plesiosaur on Loch Ness.” That fallibility could then enter the analysis through the conditional probability $P(E|D)$. This solution complicates the analysis without ultimately allowing that all evidence is fallible, for it merely pushes the inerrancy deeper into an elaborated outcome space.

A direct solution has been proposed by Jeffrey [1983, Ch. 11]. Standard Bayesianism introduced evidence E by conditionalizing, which forces unit probability onto E . Jeffrey instead proposed that we merely shift a lesser amount of probability to E , commensurate with our confidence in it. This non-Bayesian shift from initial probability P_i to final probability P_f is then propagated through the outcome space by Jeffrey’s rule. For this simplest case of evidence that partitions the outcome space into $\{E, \sim E\}$, the rule asserts that, for each proposition A ,

$$P_f(H) = P_i(H|E).P_f(E) + P_i(H|\sim E).P_f(\sim E)$$

The distinctive property of this rule is that it leaves unaffected our judgments what the probability of H would be were E definitely true or definitely false. That is, for all H , it satisfies $P_i(H|E) = P_f(H|E)$ and $P_i(H|\sim E) = P_f(H|\sim E)$. The uniqueness of Jeffrey’s rule follows in that it proves to be the only rule that satisfies these two conditions [Jeffrey, 1983, p. 169; Howson and Urbach, 2006, p. 85]. Jeffrey’s rule reduces to ordinary conditionalization if the evidence E is certain, $P_f(E) = 1$.

One unappealing aspect of Jeffrey’s rule is that the order in which it is applied in successive applications matters. That is, if we apply it first on a partition $\{E, \sim E\}$ and then $\{F, \sim F\}$, in general we do not arrive at the same final probability distribution as when the order is reversed. (See [Diaconis and Zabell, 1982].)

In the context of the Shafer-Dempster theory (Section 4.2 above), Dempster’s rule of combination allows uncertain evidence to be combined with other belief distributions in an analogous way, but one that is independent of the order of combining. With appropriate restrictions, Dempster’s rule of combination can yield ordinary Bayesian conditionalization as a special case. See [Diaconis and Zabell, 1986] for discussion of the relationship between Jeffrey’s rule and Dempster’s rules.

6 FURTHER CHALLENGES

This concluding section collects mention of some of the many further challenges to Bayesian confirmation theory that can be found in the literature.

In a letter to *Nature* and a subsequent article, Popper and Miller [1983, 1987] urged that the probabilistic support evidence E provides an hypothesis H in Bayesian confirmation theory is not inductive support. Their argument depended upon decomposing H into conjunctive parts by the logical identity

$$H = (H \leftarrow E) \& (H \vee E)$$

where the first conjunct, “ H , if E ,” is $(H \leftarrow E) = (H \vee \sim E)$. Since the second conjunct $(H \vee E)$ is deductively entailed by E , they identify the first, $(H \leftarrow E)$, as “containing all of $[H]$ that goes beyond $[E]$ ” [1983, p. 687] and require that E supplies inductive support for H only in so far as it supports this part that is not deductively entailed by E . However it is easy to show that $P(H \leftarrow E|E) \leq P(H \leftarrow E)$ so that $(H \leftarrow E)$ never accrues probabilistic support from E .

While Popper and Miller’s argument drew much discussion (for a survey see [Earman, 1992, pp. 95-98]), the decisive flaw was already identified immediately by Jeffrey [1984], who objected that $(H \leftarrow E)$ is not the part of H that goes beyond E . It rests on a confusion of “part” with “deductive consequence.” While we might think of theorems as logical parts of the axioms that entail them, that sense of part no longer works when inductive relations are the concern. For example, let H be that my lost keys are in my office (“office”) or in my bedroom (“bedroom”) and E is the evidence that they are in my house. Then $(H \leftarrow E)$ is the enormous disjunction

$$\text{office} \vee \text{bedroom} \vee \text{Mars} \vee \text{Jupiter} \vee \text{Saturn} \vee \dots$$

where the ellipses include all places not in my house. It is as peculiar to call this disjunction a part of the original hypothesis as it is unsurprising that it is not probabilistically confirmed by the evidence.

Another challenge to Bayesian confirmation theory comes from what first appears as an unlikely source. Over recent decades, there has been a mounting body of evidence that people do not assess uncertainty in accord with the strictures of probability theory. The most famous example comes from Kahneman and Tversky’s [1982] experiments that showed that people are quite ready to contradict the monotonicity of a probability measure — the property that an outcome is never less probable than each of its logical consequences. In the case of “Linda,” they found that people would affirm that a suitably described character Linda is more likely to be a bank teller and feminist than a consequence of this possibility, that she is a bank teller. One response to this body of work is to lament the unreliability of people’s intuitions about uncertainty and the need for them to receive training in the probability calculus. Another response notes that we humans have been dealing more or less successfully with uncertainty for a long time; this success extends to much of science, where uncertainties are rarely treated by systematic Bayesian methods. So perhaps we should seek another approach based a little more closely on what people naturally do. Cases like “Linda” suggest that we naturally try to select the one conclusion that best balances the risk of error against informativeness and detach it from the evidence. That is quite compatible with preferring to detach less probable outcomes. When told a coin was tossed, most people would infer to the outcome “heads or tails” as opposed to “heads or tails or on edge” even though the latter is more probable and a certainty.

In philosophy of science, if a numerical calculus is used to assess the bearing of evidence, it is most often the probability calculus. In other domains, that is less so. In the artificial intelligence literature, alternative calculi proliferate. They include the Shafer-Dempster theory, possibility theory [Zadeh, 1978] as well as numerical systems generated for particular contexts, such as the MYCIN system applied in medical diagnostics [Shortliffe and Buchanan, 1985]. One motivating factor has been pragmatic. A probabilistic analysis requires a prohibitively large number of quantities to be fixed for the general case. In an outcome space with n atomic propositions, there are of the order of $n!$ conditional probabilities. Seen from that perspective, possibility theory's simple rule for disjunction is appealing: the possibility of $A \vee B$ is just the larger of the possibilities of A or B individually. That same rule would seem disastrous to the systematic philosopher of science schooled with Bayesian intuitions, who will wonder why the rule does not distinguish the cases of A and B coincident or mutually exclusive and all those in between. The debate between proponents of the different calculi, including probabilistic systems, has been fierce. The principle goal seems to be to establish that each proponent's system has the same or greater expressive power than another's. (See for example [Cheeseman, 1986; Zadeh, 1986].)

The issue of "stopping rules" has entered the philosophy of science literature from the debate between Bayesian and Neyman-Pearson statisticians. It pertains to a deceptively simple plan for assuring belief in a false hypothesis. Assume we have a fair coin. On average, on repeated tosses, it will show heads with a frequency of about 0.5. The actual frequency will fluctuate from this expected 0.5 and there will always be some definite but small probability that the observed frequency is arbitrarily far removed from the expected. In order to advance the misperception that the coin is not fair, we toss it repeatedly, waiting mischievously for chance to realize one of these rare removals that is possible but very unlikely for a fair coin. That is our "stopping rule." When the rare removal happens we report the resulting count of heads and tails as evidence that the coin is not fair. The plan can be generalized to seek disconfirmation of any statistical hypothesis. We accumulate observations, waiting for chance fluctuations to deliver results that can only arise with low probability if the hypothesis is true.

The result is troubling for Bayesians. For them, the import of evidence is fully determined by the likelihood ratio and that likelihood ratio has been contrived to favor arbitrarily strongly the falsity of the hypothesis. That the experimenter intended to generate a deceptive result is irrelevant; the intentions of the experimenter collecting the data do not appear in Bayes' theorem. Neyman-Pearson statisticians have a ready escape that is not available to the Bayesian: the strategy of "try and try again until you get a statistically significant result" is not a good test of the hypothesis. Correspondingly, the procedure that uses this stopping rule does not count as a severe test of the hypothesis in the sense of Mayo [1996, §11.3, 11.4].

The Bayesian response has been to reaffirm that the likelihoods are all that is relevant to discerning the bearing of evidence and also to point out that the

stopping rule may never be satisfied, so that the planned disconfirmation can fail. That is, assume that a Bayesian agent assigns a prior probability p to some hypothesis and decides to accumulate evidence concerning the hypothesis until the posterior probability has risen to $q > p$. Kadane, Schervish and Seidenfeld [1996] show that, if the hypothesis is false, the agent's probability that the stopping rule is eventually satisfied is given by

$$\text{Probability of stopping} \leq \frac{p(1-q)}{q(1-p)}$$

Therefore, the closer the desired posterior probability q is set to one, the closer the probability of stopping comes to zero; that is, the probability as assessed by the agent approaches one that the experiment will continue indefinitely without returning the result sought in the stopping rule.⁵⁰ For more discussion see [Savage *et al.*, 1962, pp. 17-20 and subsequent discussion; Mayo, 1996, §11.3, 11.4; and Kadane *et al.*, 1996].

7 CONCLUSION

This chapter has reviewed many challenges to Bayesian confirmation theory. Some are global, reflecting differences in the very conception of inductive inference and the inductive enterprise in science. Others are more narrowly targeted to specific commitments of Bayesian confirmation theory. That there are so many challenges is a reflection of the dominance that Bayesian confirmation theory now enjoys in philosophy of science. It offers a well-articulated account of inductive inference in science that is of unrivalled precision, scope and power. In this regard it greatly outshines its competitors. Abductive accounts — inference to the best explanation — are still stalled by the demand for a precise account of their central notion, explanation, whereas Bayesian confirmation theory suffers from the reverse: excessive elucidation of its central notion, probability. This precision and depth of articulation is part of what enables the challenges to Bayesian confirmation theory to be mounted. For it is only after a position is clearly articulated that its weaknesses may also be discerned. When a new Goliath enters the arena, many would be Davids come with their pebbles.

ACKNOWLEDGEMENTS

I am grateful to Malcolm Forster for helpful discussion on an earlier draft.

⁵⁰If a tenfold boost in probability is sought, so $q/p = 10$, we must have $p \leq 0.1$ to start with and the formula assures us that the probability of stopping is less than $(p/q - p)/(1 - p) = (0.1 - p)/(1 - p) < 0.1$.

BIBLIOGRAPHY

- [Achinstein, 2001] P. Achinstein. *The Book of Evidence*. New York: Oxford University Press, 2001.
- [Aczel, 1966] J. Aczel. *Lectures on Functional Equations and their Applications*. New York: Academic Press, 1966.
- [Artzenius, manuscript] F. Artzenius. Goodman's and Dorr's Riddles of Induction.
- [Bacchus *et al.*, 1990] F. Bacchus, H. E. Kyburg, Jr., and M. Thalos. Against Conditionalization, *Synthese*, **85**, pp. 475-506, 1990.
- [Bartha, 2004] P. Bartha. Countable Additivity and the de Finetti Lottery, *British Journal for the Philosophy of Science*, **55**, pp. 301-321, 2004.
- [Bartha and Johns, 2001] P. Bartha and R. Johns. Probability and Symmetry, *Philosophy of Science*, **68** (Proceedings), pp. S109-S122, 2001.
- [Blake *et al.*, 1960] R. M. Blake, C. J. Ducasse, and E. H. Madden. *Theories of Scientific Method: The Renaissance Through the Nineteenth Century*. E. H. Madden, ed. Seattle: University of Washington Press, 1960.
- [Brown, 1994] H. Brown. Reason, Judgment and Bayes's Law, *Philosophy of Science*, **61**, pp. 351-69, 1994.
- [Browne, 2000] M. W. Browne. Cross-Validation Methods, *Journal of Mathematical Psychology* **44**, pp. 108-132, 2000.
- [Carnap, 1950] R. Carnap. *Logical Foundations of Probability*. Chicago: University of Chicago Press, 1950.
- [Cheesman, 1986] P. Cheesman. Probabilistic versus Fuzzy Reasoning, pp. 85-102 in L. N. Kanal and J. F. Lemmer, eds., *Uncertainty in Artificial Intelligence: Machine Intelligence and Pattern Recognition. Volume 4*. Amsterdam: North Holland, 1986.
- [Cox, 1961] R. T. Cox. *The Algebra of Probable Inference*, Baltimore: Johns Hopkins Press, 1961.
- [Diaconis and Zabell, 1982] P. Diaconis and S. Zabell. Updating Subjective Probability *Journal of the American Statistical Association*. **77**, pp. 822-30, 1982.
- [Diaconis and Zabell, 1986] P. Diaconis and S. Zabell. Some Alternatives to Bayes's Rule, pp. 25-38 in *Information and Group Decision Making: Proceedings of the Second University of California, Irvine, Conference on Political Economy*. B. Grofman and G. Owen, eds. Greenwich, Conn.: JAI Press, 1986.
- [Duhem, 1969] P. Duhem. *To save the phenomena, an essay on the idea of physical theory from Plato to Galileo*. Chicago, University of Chicago Press, 1969.
- [Earman, 1992] J. Earman. *Bayes or Bust*. Cambridge, MA: Bradford-MIT, 1992.
- [Earman and Salmon, 1992] J. Earman and W. Salmon. The Confirmation of Scientific Hypotheses, ch. 2 in *Introduction to the Philosophy of Science*. M. H. Salmon, et al. Prentice-Hall, 1992; repr. Indianapolis: Hackett, 1999.
- [Ells and Fitelson, 2002] E. Eells and B. Fitelson. Symmetries and Asymmetries in Evidential Support, *Philosophical Studies*, **107**, pp. 129-142, 2002.
- [Edwards, 1972] A. W. F. Edwards. *Likelihood*. London: Cambridge University Press, 1972.
- [Ellsberg, 1961] D. Ellsberg. Risk, Ambiguity, and the Savage Axioms, *Quarterly Journal of Economics*, **75**, pp. 643-69, 1961.
- [Feinstein, 1977] A. R. Feinstein. Clinical Biostatistics XXXIX. The Haze of Bayes, the Aerial Palaces of Decision Analysis, and the Computerized Ouija Board, *Clinical Pharmacology and Therapeutics* **21**, No. 4, pp. 482-496, 1977.
- [de Finetti, 1937] B. de Finetti. Foresight: Its Logical Laws, Its Subjective Sources, trans. H. E. Kyburg, in H. E. Kyburg and H. Smokler (eds), *Studies in Subjective Probability*, New York: John Wiley & Sons, 1964, pp. 93-158.
- [Fine, 1973] T. L. Fine. *Theories of Probability*. New York: Academic Press, 1973.
- [Fishburn, 1986] P. C. Fishburn. The Axioms of Subjective Probability, *Statistical Science*, **1**, pp. 335-58, 1986.
- [Forster, 2002] M. Forster. Predictive Accuracy as Achievable Goal of Science, *Philosophy of Science*, **69**, pp. S124-34, 2002.
- [Forster, 2006] M. Forster. Counterexamples to a Likelihood Theory of Evidence, *Minds and Machines*, **16**, 319-338, 2006.

- [Forster, 2007] M. Forster. A Philosopher's Guide to Empirical Success, *Philosophy of Science*, 74, pp. 588–600, 2007.
- [Forster and Sober, 1994] M. Forster and E. Sober. How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide, *British Journal for the Philosophy of Science*, 45, pp. 1–35, 1994.
- [Forster and Sober, 2004] M. Forster and E. Sober. Why Likelihood? pp. 153–190 in M. L. Taper and S. R. Lee, *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. Chicago and London: University of Chicago Press, 2004.
- [Galavotti, 2005] M. C. Galavotti. *Philosophical Introduction to Probability*. Stanford: CSLI Publications, 2005.
- [Gillies, 2000] D. Gillies. *Philosophical Theories of Probability*. London: Routledge, 2000.
- [Glymour, 1980] C. Glymour. *Theory and Evidence*. Princeton: Princeton University Press, 1980.
- [Hájek, 2003] A. Hájek. What Conditional Probability Could Not Be, *Synthese*, 137, pp. 273–323, 2003.
- [Hájek, 2008] A. Hájek. Arguments for — or against — Probabilism? *British Journal for the Philosophy of Science*, 59, pp. 793–819, 2008.
- [Hempel, 1945] C. G. Hempel. Studies in the Logic of Confirmation, *Mind*, 54, pp. 1–26, 97–121, 1945; reprinted with changes, comments and Postscript (1964) in Carl G. Hempel, *Aspects of Scientific Explanation*. New York: Free Press, 1965, Ch.1.
- [Horwich, 1982] P. Horwich. *Probability and Evidence*. Cambridge: Cambridge University Press, 1982.
- [Howson, 1997] C. Howson. A Logic of Induction, *Philosophy of Science*, 64, pp.268–90, 1997.
- [Howson and Urbach, 2006] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*. 3rd ed. La Salle, Illinois: Open Court, 2006.
- [Humphreys, 1985] P. Humphreys. Why Propensities Cannot be Probabilities, *Philosophical Review*, 94, pp. 557–570, 1985.
- [Jaynes, 2003] E. T. Jaynes. *Probability Theory: The Logic of Science*, Cambridge: Cambridge University Press, 2003.
- [Jeffrey, 1983] R. Jeffrey. *The Logic of Decision*. 2nd ed. Chicago: University of Chicago Press, 1983.
- [Jeffreys, 1961] H. Jeffreys. *Theory of Probability*. 3rd ed. Oxford: Clarendon Press, 1961.
- [Joyce, 1998] J. Joyce. A Nonpragmatic Vindication of Probabilism, *Philosophy of Science*, 65, pp. 575–603, 1998.
- [Kadane and O'Hagan, 1995] J. B. Kadane and A. O'Hagan. Using Finitely Additive Probability: Uniform Distributions on Natural Numbers, *Journal of the American Statistical Association*, 90, pp. 626–31, 1995.
- [Kadane et al., 1996] J. B. Kadane, M. J. Schervish, and T. Seidenfeld. Reasoning to a Foregone Conclusion, *Journal of the American Statistical Association*, 91, pp. 1228–36, 1996.
- [Kass and Wasserman, 1996] R. E. Kass and L. Wasserman. The Selection of Prior Distributions by Formal Rules, *Journal of the American Statistical Association*, 91, pp. 1343–70, 1996.
- [Kaplan, 1998] M. Kaplan. *Decision Theory as Philosophy*. Cambridge: Cambridge University Press, 1998.
- [Kelly, 1996] K. T. Kelly. *The Logic of Reliable Inquiry*. New York: Oxford, 1996.
- [Kelly and Glymour, 2004] K. T. Kelly and C. Glymour. Why Probability Does not Capture the Logic of Justification, in C. Hitchcock, ed., *Contemporary Debates in Philosophy of Science*. Malden, MA: Blackwell, 2004.
- [Keynes, 1921] J. M. Keynes. *A Treatise of Probability*. London: Macmillan, 1921; repr. New York, AMS Press, 1979.
- [Kyburg, 1959] H. E. Kyburg. Probability and Randomness I and II, *Journal of Symbolic Logic*, 24, pp. 316–18, 1949.
- [Kyburg, 1987] H. E. Kyburg. Bayesian and non-Bayesian Evidential Updating, *Artificial Intelligence*, 31, pp. 271–93, 1987.
- [Kyburg and Teng, 2001] H. E. Kyburg and C. M. Teng. *Uncertain Inference*. New York: Cambridge University Press, 2001.
- [Laraudogoitia, 2004] J. P. Laraudogoitia. Supertasks, *The Stanford Encyclopedia of Philosophy* (Winter 2004 Edition), Edward N. Zalta (ed.), 2004. <http://plato.stanford.edu/archives/win2004/entries/spacetime-supertasks/>.

- [Laudan, 1997] L. Laudan. How About Bust? Factoring Explanatory Power Back into Theory Evaluation, *Philosophy of Science*, **64**, pp. 306-16, 1997.
- [Levi, 1974] I. Levi. On Indeterminate Probabilities, *Journal of Philosophy*, **71**, pp. 391-418, 1974.
- [Levi, 1980] I. Levi. *The Enterprise of Knowledge*. Cambridge, MA: MIT Press, 1980.
- [Lewis, 1980] D. Lewis. A Subjectivist's Guide to Objective Chance, in R.C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*. Berkeley: University of California Press, pp. 263-93, 1980.
- [Mayo, 1996] D. Mayo. *Error and the Growth of Knowledge*. Chicago: University of Chicago Press, 1996.
- [Mayo, manuscript] D. Mayo. Three Battle Waves in this Philosophy of Statistics.
- [McGee, 1994] V. Mc Gee. Learning the Impossible, pp. 179-199 in E. Eels and B. Skyrms, eds., *Probability and Conditionals: Belief Revision and Probability Conditions*. Cambridge: Cambridge University Press, 1994.
- [Mellor, 2005] D. H. Mellor. *Probability: A Philosophical Introduction*. Abingdon: Routledge, 2005.
- [Mill, 1891] J. S. Mill. *A System of Logic: Ratiocinative and Inductive*. 8th ed. Honolulu, Hawaii: University Press of the Pacific, 1892/2002.
- [Norton, 2003] J. D. Norton. A Material Theory of Induction *Philosophy of Science*, **70**, pp. 647-70, 2003.
- [Norton, 2005] J. D. Norton. A Little Survey of Induction, in P. Achinstein, ed., *Scientific Evidence: Philosophical Theories and Applications*. Johns Hopkins University Press. pp. 9-34, 2005.
- [Norton, 2007] J. D. Norton. Probability Disassembled, *British Journal for the Philosophy of Science*, **58**, pp. 141-171, 2007.
- [Norton, 2007 a] J. D. Norton. Disbelief as the Dual of Belief, *International Studies in the Philosophy of Science*, **21**, pp. 231-252, 2007.
- [Norton, 2008] J. D. Norton. Ignorance and Indifference. *Philosophy of Science*, **75**, pp. 45-68, 2008.
- [Norton, 2008 a] J. D. Norton. The Dome: An Unexpectedly Simple Failure of Determinism, *Philosophy of Science*, **74**, pp. 786-98, 2008.
- [Norton, forthcoming] J. D. Norton. There are No Universal Rules for Induction, *Philosophy of Science*, forthcoming.
- [Norton, forthcoming a] J. D. Norton. Cosmic Confusions: Not Supporting versus Supporting Not-, *Philosophy of Science*, forthcoming.
- [Norton, forthcoming b] J. D. Norton. Deductively Definable Logics of Induction, *Journal of Philosophical Logic*, forthcoming.
- [Popper, 1959] K. R. Popper. *Logic of Scientific Discovery*. London: Hutchinson, 1959.
- [Popper and Miller, 1983] K. R. Popper and D. Miller, David (1983) A Proof of the Impossibility of Inductive Logic, *Nature*, **302**, pp. 687-88, 1983.
- [Popper and Miller, 1987] K. R. Popper and D. Miller. Why Probabilistic Support is not Inductive, *Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences*. **321**, pp. 569-91, 1987.
- [Royall, 1997] R. M. Royall. *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall, 1997.
- [Salmon, 1966] W. Salmon. *The Foundation of Scientific Inference*. Pittsburgh: University of Pittsburgh Press, 1966.
- [Savage, 1972] L. J. Savage. *The Foundations of Statistics*, Second revised ed. New York: Dover, 1972.
- [Savage et al., 1962] L. J. Savage et al. *The Foundations of Statistical Inference*. London: Methuen, 1962.
- [Schick, 1986] F. Schick. Dutch Bookies and Money Pumps,' *Journal of Philosophy*, **83**, pp. 112-119, 1986.
- [Seidenfeld, 1979] T. Seidenfeld. Why I am not an Objective Bayesian: Some Reflections Prompted by Rosenkrantz, *Theory and Decision*, **11**, pp.413-440, 1979.
- [Shortliffe and Buchanan, 1985] E. H. Shortliffe and B. G. Buchanan. A Model of Inexact Reasoning in Medicine, Ch. 11 in Buchanan, B. G., and Shortliffe, E. H., eds., *Rule-Based Expert Systems: The MYCIN Experiments and the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley, 1985.

- [Skyrms, 1975] B. Skyrms. *Choice and Chance: An Introduction to Inductive Logic*. 2nd ed. Encino, CA: Dickenson, 1975.
- [Tversky and Kahneman, 1982] A. Tversky and D. Kahneman. Judgments of and by representativeness, pp. 84-98 in D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press, 1982.
- [van Fraassen, 1990] B. Van Fraassen. *Laws and Symmetry*. Oxford: Clarendon, 1990.
- [van Inwagen, 1996] P. van Inwagen. Why is There Anything at All? *Proceedings of the Aristotelian Society*, Supplementary Volumes, 70, pp. 95-120, 1996.
- [Walley, 1991] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.
- [Wasserman, 2000] L. Wasserman. Bayesian Model Selection and Model Averaging, *Journal of Mathematical Psychology* **44**, pp. 92-107, 2000.
- [Williamson, 1999] J. Williamson. Countable Additivity and Subjective Probability, *British Journal for the Philosophy of Science*, **50**, pp. 401-16, 1999.
- [Zadeh, 1978] L. Zadeh. Fuzzy Sets as the Basis for a Theory of Possibility, *Fuzzy Sets and Systems*, **1**, pp. 3-28, 1978.
- [Zadeh, 1986] L. Zadeh. Is Probability Theory Sufficient for Dealing with Uncertainty in AI: A Negative View, pp. 103-116 in L. N. Kanal and J. F. Lemmer, eds., *Uncertainty in Artificial Intelligence: Machine Intelligence and Pattern Recognition. Volume 4*. Amsterdam: North Holland, 1986.

This page intentionally left blank

BAYESIANISM AS A PURE LOGIC OF INFERENCE

Colin Howson

The calculus of probability can say absolutely nothing about reality ... As with the logic of certainty, the logic of the probable adds nothing of its own: it merely helps one to see the implications contained in what has gone before (Bruno de Finetti, *Theory of Probability*, vol.1, p.215).

1 INTRODUCTION

The seventeenth century saw the beginning of that great scientific revolution from which mathematics emerged as the language of a unified physics. This was the continuous mathematics of the differential and integral calculus and (eventually) the very rich theory of complex numbers and analytic functions. But the late seventeenth century also saw another seminal scientific development: a connection forged between the idea of a graded probability and another new branch of mathematics, the discrete mathematics of combinations and permutations. Starting from mundane beginnings,¹ it developed into a completely novel science of mathematical probability and statistics.

What was not realised for a long time after the first seminal treatises of Huygens, Montmort and James Bernoulli was that two quite distinct notions seemed to be subsumed under the common title ‘probability’, notions which today we now variously distinguish by the names of ‘epistemic probability’ and ‘physical probability’, ‘Bayesian probability’ and ‘chance’, and by Carnap ‘probability₁’ and ‘probability₂’.² The focus of this discussion will be on the former, and in particular on the question of whether its laws should be classified as laws of *logic*. The idea that there might be an intimate relationship between logic and probability, at any rate epistemic probability, has been the subject of exploration and controversy for over three centuries. Both disciplines specify rules of valid non-domain-specific reasoning, and it would seem a reasonable question why one should be distinguished as logic and the other not. The first great post-Renaissance treatise on

¹‘A problem in games of chance, proposed to an austere Jansenist [Pascal] by a man of the world [the Chevalier de Méré, an assiduous gambler] was the origin of the calculus of probabilities.’ ([Poisson 1837, p.1]; my translation)

²Poisson, who was the first to explicitly note the duality, used the words ‘probability’ and ‘chance’ [1837, p.80].

logic, the so-called Port Royal Logic (*La logique, ou l'art de penser*), subsumed both the logic of certainty (deductive logic) and that of uncertainty (probability) under the common heading of logic, and the idea that they are complementary subdisciplines runs like a thread, sometimes visible, sometimes not, through the subsequent history.

Be that as it may, most contemporary Bayesians see their discipline less as a part of logic, which in common with most contemporary deductive logicians they regard as comprising just deductive logic, than of a general theory of *rational belief and decision*. Savage, for example, tells us in his classic Bayesian text [1954] that he is about to develop a theory of 'the behaviour of a "rational" person with respect to decisions' [1954, p.7]. In locating Bayesian probability within the theoretical milieu of utility and rational decision he was, of course, following one of the two great pioneers of modern Bayesianism, Frank Ramsey, who was the first to develop the theory of probability within an axiomatic theory of preference. The other, Bruno de Finetti, is best known for pointing out that if you use your probability-evaluations as the basis for your assessment of fair odds (i.e. odds at which you reckon neither side of the bet has an advantage: this relation between probability-evaluations and fair odds goes back to the beginnings of the theory of probability) then the constraints imposed by the finitely additive probability calculus protect you from making bets which you could be made to lose in any event, and more generally from making inadmissible decisions (an admissible decision is one which cannot be dominated — i.e. there is no alternative decision which will produce a better outcome, in terms of gain or loss, regardless of what the true state of affairs is).

Probability-evaluations possessing these features de Finetti termed *coherent*.³ But, paradoxically, it is in his work that another, apparently not at all decision-theoretically oriented view of coherence emerges, as a species of *intrinsic consistency* which according to de Finetti the evaluations possess - or not - independently of who makes them or why:

[I]t is better to speak of coherence (consistency) of probability evaluations rather than of individuals ... because the notion belongs strictly to the evaluations and only indirectly to the individuals. [1937, p.103, footnote (b)]

The parenthetical gloss of 'consistency' for 'coherence' is de Finetti's own, which he reinforced with the observation that incoherent probability assignments actually 'contain an intrinsic contradiction' (*ibid.*). This is not the only reference to a specifically logical character of coherent evaluations in that paper. Its title, in the original French, is 'La prévision: ses lois logiques, ses sources subjectives' ('Foresight: its logical laws, its subjective sources' in the Kyburg translation,⁴)

³There are several qualifications that need to be added to this brief summary of de Finetti's theory of coherence. I shall do so later.

⁴'Foresight' is not the best of translations of what was for de Finetti a technical term meaning roughly 'expectation'.

and ‘logical laws’ was by no means an idle phrase: in another paper published a year earlier, de Finetti wrote that

it is beyond doubt that probability theory can be considered as a multi-valued logic ... and that this point of view is the most suitable to clarify the foundational aspects of the notion and the logic of probability’ [1936, p.183]; quoted in [Coletti and Scozzafava 2002, p.61].

Yet despite such advocacy from probably the most influential of all modern Bayesians, a view of the rules of probability as furnishing an authentic logic, though with values in $[0,1]$ rather than $\{0,1\}$ (deductive logic), not only did not go on to command widespread acceptance, but even de Finetti himself seemed in his later work to have relinquished it in favour of the view now almost universally associated with his work, that the rules are merely prudential safeguards, protecting the agent from ‘decisions whose consequences are manifestly undesirable (leading to certain loss)’ [1974, vol. 1, p.72].

But there is more to the general failure to see in Bayesian probability an authentic logic than de Finetti’s own personal evolution from apparent logicist to determined decision-theorist. There are also some important and seemingly recalcitrant facts: modern deductive logic deals in discrete (two) truth-values, and its central notions of consistency and consequence as properties of, and relations between, sentences seem to have no analogues in the Bayesian formalism. One can certainly *say* that a set of probability-evaluations is consistent if it obeys the probability axioms, but this is to give the word ‘consistent’ an altogether different meaning, or so it seems, than the one it has in deductive logic. It certainly seemed so to Henry Kyburg, who in his translator’s preface to the English version of de Finetti’s 1937 paper tells us that it was for this reason that he translated de Finetti’s original ‘cohérence’ as ‘coherence’:

“Consistency” is used by some English and American authors, and is perfectly acceptable to de Finetti, but it is ambiguous (from the logician’s point of view) because, as applied to beliefs, it has another very precise and explicit meaning in formal logic. As the words are used in this translation, to say that a body of beliefs is “consistent” is to say (as in logic) that it contains no two beliefs that are contradictory. To say that in addition the body of beliefs is “coherent” is to say that the *degrees* of belief satisfy certain further conditions ([1964, p.95]; parenthesis in original)⁵

I shall show later that the apparently very strong disanalogy between deductive consistency and probabilistic ‘coherence’ that Kyburg thought he had discerned vanishes on closer examination. This is remarkable enough, but even more remarkable, as we shall see, is the very close formal kinship that emerges between the two

⁵Kyburg should strictly have used the word ‘implies’ rather than ‘contains’. David Makinson pointed out that it is not necessary for inconsistency for a body of beliefs to *contain* two contradictory beliefs.

notions, which at a certain level of description are actually identical. This in my opinion fully justifies the use of the one word ‘consistency’ to apply to both. We shall also see that although de Finetti did not describe his work in the language of modern logic, the concepts he forged and the results he proved are fundamentally logical in character.

This is not to say that in the preceding three centuries people did not work hard to elicit some fruitful relationship between logic and probability. They did, but despite their efforts very little of a positive nature emerged, and when it did it was the result of taking a quite different approach. The two fundamental notions of modern deductive logic are (semantic) *consistency* and (semantic) *logical consequence*, which in classical logic at least are interdefinable. Practically from the start it had been agreed that in some sense probability generalised deductive logic (‘the logic of certainty’ as it was often described), but it was also agreed that it was the deductive consequence relation that was generalised. For reasons which will become apparent in the next few sections this was the wrong choice, but it is instructive to see why it does not work, particularly since there is still a large number of people who think it does.

2 BOLZANO AND PARTIAL ENTAILMENT

The first significant marker in the probability-as-generalised-deduction programme was put down in the mid-nineteenth century by the mathematician and philosopher Bernard Bolzano. In fact, Bolzano made a very interesting discovery. Recorded in his *Theory of Science* 1837, section 161), his discovery was this. Suppose we partition the components of declarative sentences into fixed and variable terms (or fixed and variable ‘ideas’, as he called them), rather as we do today when we parse sentences into *logical* and *extralogical* components. Bolzano considered a partition according to such a logical/extralogical criterion, calling those sentences which remain uniformly true under variation of their extralogical components ‘analytic’, but in general he treats the partition merely as an exogenously given parameter. His criterion for the valid deduction of a conclusion C from finitely many premises $A_1, \dots, A_n$, relative to any such partition, is that there is no way of interpreting the variable terms which results in all the A_i being jointly true and C false. We might paraphrase this in our current terminology as: *C is validly inferred from $A_1, \dots, A_n$, given the partition, if every model of the set $\{A_1, \dots, A_n\}$ is a model of C* (where a model is just an interpretation making the sentences in question jointly true).

If Bolzano had stopped there he might have already done enough to merit being regarded as the progenitor of modern model theory. But he went further, pointing out that, where the sentences in question admit only finite numbers of distinct models, we can consider not just the two limit cases (i) every model of $\{A_1, \dots, A_n\}$ is a model of C , and (ii) no model of $\{A_1, \dots, A_n\}$ is a model of C (in the second case C is of course *inconsistent* with those premises), but also the case where an intermediate *proportion* of C ’s models are also models of $\{A_1, \dots, A_n\}$; i.e., setting

$\Pi = \{A_1, \dots, A_n\}$ we can consider a spectrum of different possible values of the quotient $|\text{Mod}(C) \cap \text{Mod}(\Pi)| / |\text{Mod}(C)|$, where $|\cdot|$ around a set signifies the number of its members, i.e. its cardinality, $\text{Mod}(\Sigma)$ is the set of all models of a set Σ of sentences,⁶ and $\cap$ is set-intersection. Bolzano terms this, in remarkably modern language, the relation of *relative satisfiability*. This proportion will of course be a number between 0 and 1 including the two bounds, and it is not difficult to see that the proportion of models of the disjunction ‘ C or D ’ which are also models of any set Σ of sentences is the *sum* of the proportion of models of C which are models of Σ and the proportion of models of D which are models of Σ . Bolzano proves this and many of the other results we now call theorems of the probability calculus.

We can illustrate Bolzano’s idea with a simple example. Suppose a die is to be thrown once, and let A be the sentence ‘The die comes to rest with a numeral between 1 and 6 uppermost’ (again we shall avoid tiresome pedantry by just writing n instead of the strictly correct ‘ n ’). Let C be ‘the outcome is of type Φ ’ where Φ is some canonical way of describing the outcome of the throw. Bolzano did not have the flexible symbolic notation of modern formal logic with which to illustrate his ideas adequately, but we do and at the cost of being charged with ahistoricity I shall use them now. Accordingly, in the die example we will employ a vocabulary whose ‘fixed terms’ are just the usual logical constants with their customary notation $\vee$ (‘or’), $\neg$ (‘not’), $\wedge$ (‘and’), $\exists$ (‘some’), $\forall$ (‘all’), and whose ‘variable terms’ are a constant a (to be read ‘the throw of the die’), and six monadic predicates B_1 (‘comes to rest 1 uppermost’), B_2 (‘comes to rest 2 uppermost’), $\dots$, and B_6 (‘comes to rest 6 uppermost’). One of the few serious symbolic defects of modern classical logic is that the domain of the variables is never signalled by an explicit symbol, though it is just as ‘variable’ as the rest of the extralogical material. In this case the intended domain, or ‘universe of discourse’ as it is sometimes called, is the set consisting just of the single throw of the die. We can write A in this notation, together with multiple conjunction and disjunction operators, as

$$\forall x(x = a) \wedge \forall x \left[\bigvee_{0 < i < 7} B_i(x) \wedge \bigwedge_{0 < i < j < 7} (B_i(x) \rightarrow \neg B_j(x)) \right]$$

Strictly speaking a single universal quantifier $\forall$ would have sufficed, but I have written A as the conjunction of two universally quantified sentences to make it clear that A implies that the intended domain consists of just the individual denoted by a (that is what the first conjunct states).

Varying the ‘variable terms’ (including the suppressed domain-denoter) gives us the class of interpretative structures of this simple formal language. Any one of these which makes a sentence or set of sentences jointly true is now usually called a *model* of that sentence or set of sentences. It is not difficult to see that there are six models of A distinct up to isomorphism, each being characterised by a being the extension of exactly one B_i (the others are all empty). If C is ‘the outcome

⁶Strictly speaking we should write $\text{Mod}(\{C\})$, since Mod is applied to *sets* of sentences, but to avoid the resulting nasty notation I shall omit the set brackets.

is an even number', i.e. C is the sentence $B_2(a) \vee B_4(a) \vee B_6(a)$, or logically equivalent to it, then of these six models three satisfy C , so the ratio in question is $1/2$. Following Bolzano, we can call this the *probability* of C relative to A . If C is 'the outcome is a 3', i.e. C is $B_3(a)$, the probability of C relative to A is $1/6$, and so on.

The knowledgeable reader will immediately recognise that Bolzano has in effect re-stated Laplace's famous definition of probability as the ratio of favourable to possible cases (the 'definition' in fact was not original to Laplace but goes back to the earliest work in the mathematical theory of probability). Laplace had advanced this definition subject to the proviso that the cases in question were what he called 'equally possible'. Bolzano, like many subsequent commentators, believed that since possibility does not admit of degrees but probability does, the only intelligible meaning Laplace's proviso could have is 'equally probable', and hence that his definition was circular. But Bolzano also believed that the assignment of the same probability of $1/6$ a priori to each of the propositions $B_1(a), B_2(a), \dots, B_6(a)$, from which by the addition principle one obtains the Laplace definition, does require some justification, and in providing one he employs another seminal idea: an appeal to a criterion of *symmetry*.

Bolzano's own illustration of this idea was given in an example structurally very similar to the one above. In this one Caius is drawing a ball from an urn containing numbered balls. Let A simply state that this is what is going on (it is straightforward to write this in the same notation as the earlier example, but I shall leave this to any reader who wants to) Consider the two statements 'Caius has drawn ball no. 1' and 'Caius has drawn ball no. 2'. According to Bolzano, since these differ only in that one of them has 'no. 1' whereas the other has 'no. 2', while A makes no mention of numbers at all, *each stands in exactly the same logical relation to A* [1837, p.241]. Thus, given only the information in A , we are according to Bolzano justified in putting each of the propositions 'Caius has drawn ball no. 1' and 'Caius has drawn ball no. 2' on exactly the same probabilistic footing; to put it another way, the probability should be *permutation-invariant* relative to the indices 1 and 2. Although Bolzano does not say this, his own discussion might be seen as a sort of vindication of Laplace, since one can easily envisage Laplace claiming that in saying that cases are 'equally possible' he meant that the propositions in question were logically symmetrical with respect to the background information.

Symmetry considerations of one sort or another will turn out to be central to all the 'logical' approaches to epistemic probability which see the latter as generalising the deductive consequence relation. Symmetry is nowadays identified mathematically with invariance under the action of a group, and we have already seen an example in Bolzano's example where the symmetry group is that of permutations of two and more generally n objects. In the twentieth century the criterion of symmetry with respect to the action of a more extensive class of transformation groups became a powerful and flexible tool in the hands of the logical-Bayesian physicist E.T. Jaynes.

3 SYMMETRY AND ITS PROBLEMS

It probably goes without saying, but is nevertheless worth reminding ourselves, that the theory (or maybe theories) of epistemic probability, whether thought of as logic or not, was not intended to be a piece of pure mathematics. It was an object of increasingly great interest because it promised to furnish a mathematically rigorous and *applicable* science of uncertainty or, as James Bernoulli put it, of

the art of measuring as exactly as possible the probabilities of things with this end in mind: that in our decisions or actions we may be able always to choose or to follow what has been perceived as being superior, more advantageous, safer, or better considered; in this alone lies all the wisdom of the philosopher and all the discretion of the statesman. [1715, p.13]

The goal of science may be truth, but truth is unfortunately seldom if ever manifest, and certainly not scientific truth. We are forced to conjecture, and we need to know how credible our conjectures are.

The problem these otherwise rather commonplace observations pose for the theory Bolzano was trying to develop is that practical inferential and decision problems are typically not of the simple ‘We know that Caius is going to pick a ball from an urn containing n balls’ type. They involve complex background information which is typically far from symmetrical between the various logical possibilities it leaves open. It is, for example, not strictly ruled out by anything we currently reckon ourselves reasonably certain of that there exist angels; while some people may profess themselves certain of just that, most would not gamble very much on it. One possible way out, canvassed at various times in the nineteenth and twentieth centuries, was this: assume a *tabula rasa* condition in which at the outset we have no information, and relative to which we can invoke the symmetry condition, and then update the resulting distribution by conditionalising on successive pieces of experiential information. If we are currently rational in the way we informally assess probabilities, so the story goes, we should be able to retrieve these assessments as the terminal stage of the process. All learning, on this view, is Bayesian learning, via conditionalisation from a symmetrically distributed prior ignorance.

It is difficult to see, in retrospect, why such a view should ever have seemed plausible. There are compelling reasons for concluding that it is impossible that the current collective epistemic state of mankind could ever have been arrived at in this way: the dynamics of theory change, which Kuhn in his seminal study [1962] aptly summarised as periods of normal science punctuated by revolutionary upheavals, rather obviously do not fit such a cumulative model. Two of the most powerful objections concern the supposed initial *tabula-rasa* state, relating to (i) infinity, and (ii) the notion of symmetry itself.

(i) *Infinity*

Suppose you are asked to consider all the logical possibilities that there are, subject

to no constraints at all. There is a broad agreement that this possibility-space is too vast even to be well-defined: everyone aware of the paradoxes of so-called naive set theory would, I think rightly, suspect that there can be no such space, on pain of contradiction. Even restricted in some appropriate way it is certainly not finite; on the contrary, whatever bounds one might want to put on it, it must be vastly infinite. It is certainly too vast and amorphous to have any familiar metric structure that would permit the application of a symmetry principle like Bolzano's to it. Even familiar and well-behaved sets that do have an intrinsic metric structure based on their usual topology pose serious problems. Consider the smallest infinite sets, the countably infinite ones (i.e. the sets whose members can be indexed exhaustively by the positive integers, with cardinal number $\aleph_0$). If one applies Bolzano's symmetry criterion to these then each singleton member must presumably be assigned the same probability of 0, which is problematic for a number of reasons principal among which is that the distribution does not sum to 1. The uniform zero distribution, in other words, is finitely but not countably additive, while a countable additive distribution will necessarily be heavily skewed, with a finite set carrying all but an epsilon's worth of the probability, where epsilon tends to zero as the size increases. I said earlier that the finite/countable additivity issue is controversial and that we would return to it later. So we will, but for the time being it is enough to note that the appeal to the sorts of logical symmetry Bolzano thought relevant seem to decide the question, for better or worse, against countable additivity.

Bounded intervals in Euclidean n -space seem on the face of it more straightforward, even though they are much larger in terms of cardinality than countable sets (though exactly how much larger in terms of where that cardinality, $2^{\aleph_0}$, occurs on the $\aleph_\alpha$ scale remains one of the great open problems of mathematics), because their usual metric structure looks amenable to a straightforward application of a symmetry principle. Take any bounded, open or closed, interval with end-points $a, b, a < b$, and subdivide it into m equal subintervals $I_1, \dots, I_m$. If the only information is that the value of the quantity in question lies in that interval then the symmetry between the possibilities implicit in that information seems to dictate that each of the subintervals should receive the same probability $1/m$. If we now let m go to infinity the distribution above tends to a limiting distribution, the so-called uniform *probability density* distribution, taking the value $1/(b-a)$ at each point. It seems to have been just such an appeal to symmetry that led Bayes in his celebrated posthumously published *Memoir* (1763) to advocate the uniform prior density $p(x) = 1$ over the closed unit interval (though there is some controversy about whether this really was his reasoning; see [Dale, 1982]).

Of course, the same strategy won't work if the interval is unbounded; for example, if it is the entire real line or half-line. There is no uniform density over either of these that satisfies the condition that it integrates to 1 (or any finite value). Some Bayesians have argued nevertheless for using a so-called 'improper' uniform prior density over the real line, which in certain cases can still result in a 'proper' posterior density. The argument justifying this procedure, where it works, is that

the uniform prior is an approximation to a truncated ‘proper’ prior density, i.e. one that is integrable. As should not be altogether surprising, however, paradoxes can result from such highly non-standard devices (see [Dawid *et al.*, 1973; Seidenfeld, 1979]). Unfortunately, paradoxes are not associated only with improper distributions: they afflict even the apparently straightforward case of possibility spaces represented by bounded intervals. Suppose that A is the proposition that t is a quantity whose value is in the interval $[a, b]$ above. Dividing $[a, b]$ into n equal subintervals gives us the set of ‘equal’ possibilities $I_1, \dots, I_n$ to which we can apply Bolzano’s symmetry principle. But suppose instead of t we consider the new quantity y where y is a continuous invertible function $y(t)$ of t . We can now divide the interval $[y(a), y(b)]$ into n equal subintervals $I'_1, \dots, I'_n$, and so by the same symmetry principle these should also each receive probability $1/n$. But it is not difficult to see that these probability distributions will be consistent with each other just in case y is a linear function of x . If that is not the case we will have different probabilities assigned some interval $u < t < v$ and $y(u) < y < y(v)$, even though the two conditions are equivalent and hence, by the basic rules of probability, must receive the same probability.

A classic example was produced in the late nineteenth century by the French mathematician Bertrand. Bertrand considered the problem of determining the probability that an otherwise unspecified chord to a circle will have a length less than the length l of the side of the inscribed equilateral triangle (there is of course an uncountable infinity of inscribed triangles, but they are all congruent). However, there are various equivalent ways of specifying a chord of a given length. Three that Bertrand considered are where (the exposition follows [Gillies, 2000, pp. 37–39]; the reader might like to write out the parametric transformations by which these are generated from each other): (1) the midpoint of the chord intersects a radius; (2) the angle between an end-point of the chord and the tangent at that point; and (3) the position, in polar coordinates, of the midpoint in the circle. (1), (2) and (3) determine different possibility-spaces: (S_1) all the points on the radius; (S_2) the angles from 0 to π , and (S_3) the set $\{(r, \theta) : 0 \leq r \leq R, 0 \leq \theta \leq 2\pi\}$, where r is the length of the radius. Subsets T_1, T_2 and T_3 respectively correspond to the event, call it C , that the chord-length is less than l . T_1 is the set of points on a radius between its perpendicular intersection with a side of the inscribed triangle and its intersection with the circumference of the circle; T_2 is the 60° angle between the tangent and the nearest side of the inscribed triangle whose vertex is at the tangent; and T_3 is the set of points in the interior circle of radius $r/2$. The pairs (S_i, T_i) have a ‘natural’ measure m_i induced by their geometrical structure: in the first, length, in the second, angular measure, and in the third, area measure. The same symmetry considerations we have mentioned seem to suggest that relative to each such pair the probability of C is equal to $m_i(T_i)/m_i(S_i)$. However, the probability of C relative to the first pair (S_1, T_1) is $1/2$, relative to the second is $1/3$, and relative to the third is $1/4$. So we have a contradiction, or, as inconsistencies were often termed in the nineteenth century, a paradox. In honour of Bertrand, this one is known as Bertrand’s Paradox.

Symmetry, an intuitively appealing idea, unfortunately seems capable of doing too much: in this case, overdetermining the probability. The verdict might well seem to be that there are simply too many symmetries which cannot be consistently combined. It was a verdict famously disputed by Jaynes, according to whom there is only an *appearance* of too many symmetries: in the Bertrand example there are, he claimed, just enough to determine $1/2$ as the *unique* solution: symmetry with respect to the orientation of the circle (rotational), with respect to the position of its centre (translational), and with respect to its size (scale). It follows that $f(x, y)$, the probability-density that the midpoint of the chord lies at the point (x, y) in or on the circle (assumed to be centred at the origin) should be rotation, translation and scale-invariant. Rotational invariance implies that f depends on (x, y) through $\sqrt{(x^2 + y^2)}$ only. Transforming to polar coordinates, the joint probability density $p(r, \theta)$ is $rf(r)$, $0 \leq r \leq R$, $0 \leq \theta \leq 2\pi$. Now consider a scale transformation $r' = ar$, $a \geq 0$, which we can think of as blowing up or shrinking the original circle. Suppose it is shrunk, i.e. $a \leq 1$. The probability densities $p(r)$ and $p(r')$ are related in the usual way by

$$p'(r') \frac{dr'}{dr} = p(r)$$

We cannot immediately invoke scale invariance to identify the two density functions p and p' , since they have different domains, $[0, aR]$ and $[0, R]$ respectively. What we can do, though, is to restrict r to $[0, aR]$ by conditioning p on the information that r lies in $[0, aR]$, and then equate the resulting *conditional* density with p' . This is what Jaynes does, and differentiates the resulting equation with respect to a to obtain a differential equation whose solution is of the form

$$f(r) = qr^{q-2}/2\pi R^q$$

where $0 \leq q$. Finally, Jaynes shows that translation-invariance uniquely determines $q = 1$, giving $f(r) = 1/2\pi Rr$ (however, as he pointed out, translation-invariance is by itself so strong a condition that it already determines the result). Thus the required probability density is uniform, and the probability that the chord is less than the length of the inscribed triangle is easily seen to be $(1/2\pi R)$

$$\int_0^{R/2} \int_0^{2\pi}$$

$drd\theta = 1/2 = 1/2$, i.e. the first of the solutions above.

Jaynes went on to apply a similar analysis to other examples with what seemed to be very fruitful results. For example, if scale-invariance is implicit in a problem about a quantity whose possible values are the non-negative real axis $[0, \infty)$, then it must be distributed in this interval with an improper density proportional to $1/x$ [Howson and Urbach, 2006, p.281-2]. However, it must be regarded as highly doubtful that Jaynes has really solved the transformational problems afflicting appeals to symmetrical states of information. In the Bertrand example Jaynes picked

three groups of transformations whose action appears to leave the information in the problem unaltered. These, happily, determined a solution. But it is only Jaynes's reading of that information that declares the alternatives mentioned by Bertrand illegitimate, a reading which Jaynes himself confessed was due only to a reformulation of the problem into that of a person of ordinary height and eyesight throwing long straws into a circle inscribed on a flat piece of ground [1973, p.478]. It is a logical truism that if one loads more detail into a set of assumption one will diminish the space of solutions. Bertrand's original problem, as he proved, admits too many solutions, solutions inconsistent with each other.

(ii) *Logical Symmetry*

It is often stated, erroneously, that the problem of inconsistency only genuinely arises in the context of continuous spaces, of which the Bertrand example is an example. In fact, it appears just as destructively in the finite case. Consider the following simple example of the type Bolzano himself discussed. An urn contains just two balls, and we are told that they may be both black, both white, or one may be black and the other white. What is the probability that both are black? We have enumerated three possible cases: both black, both white, and one black and one white, so the answer would appear to be $1/3$. However, it is also the case that there are two balls, in principle distinguishable. If we include this in the reckoning, the space of distinct possibilities is now fourfold, namely: both black; both white; one ball, call it ball 1, black and the other, ball 2, white; ball 1 white and ball 2 black, and the answer is now $1/4$. The answers are inconsistent with each other, though both model the initial information.

It might be objected that the first answer, $1/3$, is wrong because it does not take into account the fact that the case 'one ball black and one white' is in reality composite, decomposing into two further exclusive possibilities. This was Keynes's view. Though Keynes himself seemed to be unaware of Bolzano's work, his ideas are, modulo minor differences of style and expression, very similar to Bolzano's. Keynes argued that probability must be calculated relative to the finest partition of logical space [1921, p.60], pointing out that failure to observe this rule will in suitable circumstances lead to inconsistency, just as it did above.⁷

This may sound plausible, but consider again the balls in the urn. The premissory information is that each of the two balls may or may not be coloured black (or white). Both this piece of information, call it A , and the statement that both balls are black, call it C , can be expressed in a language, call it $L^1(=)$, containing just the one monadic predicate $B(x)$ (' x is black'), identity ($=$) and the connectives and first order quantifiers, as follows

$$\begin{array}{ll} A & \exists x \exists y [\neg(x = y) \wedge \forall z (z = x \vee z = y)] \\ B & \forall x B(x) \end{array}$$

⁷He produced his own famous example. Given that the cover of a book is coloured, two exclusive and exhaustive possibilities are that it is red or not red. But if we reckon each to have a probability $1/2$ we need only repeat the calculation with respect to the partitions {blue, not blue} and {green, not green} to obtain probabilities greater than 1 [1921, pp.46-47].

A and B are also statements in the language, call it $L_2^1(=)$, which is just like $L^1(=)$ above but augmented by two individual names (constants) a and b . But now observe that there are just three models of A (up to isomorphism) in $L^1(=)$, two of which are also models of B , giving a probability of B given A of $1/3$, while in $L_2^1(=)$ there are four distinct models of A (assuming we stipulate that a and b have distinct designata), three of which are models of B , giving a probability of B given A equal to $1/4$. But now Keynes's rule does not pronounce definitely in favour of the second, as it seemed to do in our original discussion, because relative to the languages $L^1(=)$ and $L_2^1(=)$ each answer, $1/3$ and $1/4$ respectively, is determined by the finest partition of logical space. The fact is, and it is one that Keynes failed to appreciate, that what is to count as the finest partition of logical space is language-dependent because language itself, i.e. the set of one's descriptive categories, *determines* logical space. And that being the case, there is simply no determinate answer to the question of what the 'correct' probabilities are.

Even so a latter-day Keynes might still object that since $L^1(=)$ is a sublanguage of $L_2^1(=)$ its logical space is a refinement of the latter's and that we should always refer any problem to that language which permits the finest partition. But the objection is to no avail since it begs the question. To be precise, it begs two questions. Firstly, scientific languages and their associated logical spaces evolve over time, so there is no single finest partition of logical space. Secondly, it is simply not true that the language with individual names necessarily makes finer distinctions than the one without. Paired bosons, for example, are indistinguishable particles according to quantum mechanics and have the same statistics as $L^1(=)$.⁸ Less exotically, dollars in bank accounts are also indistinguishable: it makes no sense to say that there are four distinct ways in which two dollars can be distributed between two accounts. In these cases there *is* no finer-grained reality than that described by $L^1(=)$. Even if there were an unequivocally finer partition, logic by itself does not dictate that one should choose it, or indeed that one *should* do anything at all.

The phenomenon of language-dependence not only presents a strong disanalogy with the logic of full entailment, or deductive logic (it is a theorem of classical, first or higher order, logic that if Π entails C then it does so independently of the language of which Π and C are part). But in view of the intended application of this extension of deductive logic — providing a rigorous theory of the degree of confidence one should entertain in a hypothesis on the basis of the data one currently has — the lack of univocality raises a serious problem. Keynes, introducing a terminology that was to become fairly entrenched, described the presumptive degree of partial entailment, where it exists,⁹ as the measure of the *degree of rational*

⁸Though not all quantum-mechanically indistinguishable particles have these statistics: fermions, which can only be in antisymmetric coupled states, obey the very different Fermi-Dirac statistics.

⁹He did not believe that it did exist for all pairs of propositions, or even that the pairs were totally ordered.

belief warranted by the data for that hypothesis. We thus have an intended interpretation that is supposed to be determined by the underlying logical structure, but manifestly is not.

4 CARNAP'S LOGICAL PROBABILITY

The problems of language-dependence, and of the possibility of the formal and interpretative aspects of this theory coming apart, reach a sort of crescendo in Carnap's work on what he called logical probability from 1950 onwards. Although Bolzano receives no mention in Carnap's monumental study [1950], Carnap appeared to be working broadly within the tradition he initiated, of regarding formal probability as a relation of weak logical consequence: he interpreted a suitable conditional probability function, defined on pairs (h, e) of sentences of a specified formal language,¹⁰ as a measure both of the partial entailment of h by e and also of the degree of rational belief warranted in h by e . Symbolised $c(h, e)$,¹¹ this function is defined in the manner of Bolzano as a quotient $m(h \& e)/m(e)$, where m is a normalised measure of the logical ranges of propositions, i.e. of the sets of distinct ways in which the corresponding sentences can be true. Unlike Keynes, Carnap certainly appreciated the fact of language-dependence implicit in this account, and in his 1950 book he explicitly relativised his measures to specified languages, only in his later work attempting to create a more language-invariant theory.

Carnap's most significant innovation was to define logical probabilities, at any rate as he understood them, directly on the sentences of completely specified formal languages, a move which in its turn resulted in the term 'logical probability' being understood subsequently as almost any function so defined. Carnap's languages were simple monadic predicate languages *without* identity, each containing a finite number of individual constants. Let L_N^k be the language with N constants and k predicate symbols $P_1, \dots, P_k$. A quasi-logical axiom constraining the admissible interpretations of L_N^k which Carnap adopted was that the domain should contain exactly n individuals each named by a distinct a_i (this constraint allows the m -measure to be normalised). Since the identity relation is needed to express this, it perforce remains a tacit constraint. Given it, however, it is straightforward to show that the class of distinct interpretations of L_N^k is finite, and that each is specified by a single sentence of L_N^k , called by Carnap a *state-description*, of L_N^k , of the form $Q_{i1}(a_1) \& \dots \& Q_{iN}(a_N)$ where each $Q_{ij}(x)$ is one of the 2^k distinct, mutually incompatible formulas $\pm P_1(x) \& \dots \& \pm P_k(x)$ ($+P = P, -P = \neg P$). These

¹⁰'Suitable' means among other things 'regular', in Carnap's terminology, i.e. it takes the value 0 only when h is inconsistent with e ; e is assumed always to be consistent.

¹¹The ' c ' stands for 'confirmation': where e represents evidential information and h a hypothesis, Carnap thought that $c(h, e)$ also measures the degree to which e confirms h . Whether degree of confirmation really is formally a probability, and if it is not what formal properties does it have, are philosophical questions that have agitated many people with little sight of a resolution, but they are fortunately not ones on which I feel obliged to pronounce here. In the discussion of Carnap's theory I shall usually follow his notation and refer to sentences by lower-case Roman letters.

predicates, which partition the domain into disjoint cells, were called by Carnap *Q-predicates*. Note that, given the constraint mentioned, the disjunction of all the 2^{kN} state-descriptions is a logical truth.

Their simple structure means that these languages are nearer to propositional languages generated by finitely many propositional variables, with atoms corresponding to the state-descriptions, than more ‘typical’ first order languages with equality and a non-empty set of relation symbols of arity greater than one. This makes for a particularly straightforward probabilistic treatment, for every consistent sentence s of L_N^k is equivalent to a finite disjunction of state-descriptions (it is a type of disjunctive normal form of s), and hence an assignment to the state-descriptions fully determines the probabilities on all sentences. Since each state-description is in effect a distinct model of s up to isomorphism one might have thought that Carnap would follow the Bolzano-Keynes prescription and identify the basic measure $m(s)$ as the normalised number of state-descriptions in the normal form of A . Carnap did consider this policy, calling the resulting measure function $m^\dagger$. But, significantly, he rejected it because it is a product-measure: it makes the conjuncts $Q_{ij}(a_j)$ probabilistically independent with the same probability.¹² And this mattered very much for Carnap, since it implied that, gauged in terms of $c^\dagger$, *there could be no learning from experience*. For suppose that e consists of the data that n individuals $a_1, \dots, a_n, n < N$, have manifested the same Q -predicate Q , and consider $f(n+1) = c^\dagger(Q(a_{n+1}), e)$. Since the $Q_{ij}(a_j)$ are independent relative to $m^\dagger$ with constant probability 2^{-kN} , $f(n+1) = 2^{-kN}$, which is of course independent of n . It is not difficult to see that the independence of r remains a feature, whatever complex predicate in L_N^k one substitutes for Q . Thus no amount of data can affect the probability that the next observation will manifest the same property as the previous ones. Given that Carnap had set up this formal machinery to provide an authentic logic of induction (‘inductive logic’ was his own coinage), this feature of $c^\dagger$ is certainly anomalous.

Nor is it the only one. Consider the case where h is a universal hypothesis over an infinite domain. Here of course we run up against the problem of infinity discussed earlier, but Carnap had a way of dealing with this: to let the parameter N tend to infinity, with the limiting values of $c^\dagger$, where they exist, now regarded as those for a presumptive limit language L_∞^k . It follows from a result of Fagin [1976] that the limits exist for all the sentences in every $L_M^k, M > N$, and they determine a well-defined m -function. Since $f(n) = c^\dagger(Q(a_{n+1}), Q(a_1) \& \dots \& Q(a_n))$ is independent of n its value will of course remain unchanged in this passage to the limit, but we also have the following zero-one law: the limiting value of $c^\dagger(h, e) = 0$ for every $N > 0$ if h is a contingent pure universally quantified sentence and e is a contingent sentence without quantifiers entailed by h (so e is ‘predicted’ by h); and the limiting value of $c^\dagger(h, e)$ is 1 if h is a contingent pure existentially quantified sentence.

¹²This is easy to see. The function $c^\dagger(., t)$ based on $m^\dagger$, where t is a logical truth, is a finitely additive probability function $P(.)$ on the sentences of L_N^k . $P(Q_{ij}(a_j)) = 2^{k(N-1)}/2^{kN} = 2^{-k}$, and $P(Q_{i1}(a_1) \& \dots \& Q_{iN}(a_N)) = 2^{-kN} = \prod_j P(Q_{ij}(a_j))$.

Thus $c^\dagger$ is very strongly inimical to ‘learning from experience’: the predictive probability $f(n)$ tells us that no finite number of instances of a property can raise the probability that the next individual to be observed will have the property over its initial probability, while no finite number of successful predictions will ever raise the probability of a universal hypothesis above that of a logical falsehood. Nor, perhaps surprisingly, is that situation in any way altered by considering more sophisticated first order languages than Carnap’s monadic languages without equality: in particular, as Fagin shows [1976], it remains unchanged for arbitrary first order languages even with equality.

Carnap’s response was to declare $c^\dagger$ inadmissible as a foundation for his inductive logic, and to contrast it with one he found more congenial, which he called c^* . Defined as before on a language L_N^k , the corresponding m -function, m^* , assigns equal absolute probabilities not to all state-descriptions but to those disjunctions of state-descriptions, called by Carnap *structure-descriptions*, containing exactly the same numbers of each Q -predicate (with the probabilities of the constituent state-descriptions distributed equally within each state-description). Where, to simplify the discussion, $k = 1$, consider the predictive probability $c^*(P(a_{n+1}), e_n)$, where e_n is the conjunction $\pm P(a_1) \& \dots \& \pm P(a_n)$, $n < N$, with r unnegated P s. A little combinatorial arithmetic quickly elicits a historically famous formula

$$c^*(P(a_{n+1}), e_n) = (r + 1)/(n + 2) \quad (1)$$

where for $r = n$ the right hand side clearly tends to 1 as n tends to infinity (notice that the result is independent of N).

Known as the *Rule of Succession*, (1) was first derived by Laplace; it is in fact a straightforward consequence of the posterior distribution for a binomial parameter obtained by Bayes in his celebrated *Memoir* (1763), using the uniform prior over $[0,1]$ (there is a deep mathematical connection between Bayes’s distribution and Carnap’s logical probability functions which I shall come to later). Although Carnap was not overly enthusiastic about c^* ,¹³ it had (for him) the virtue of endorsing predictive inferences (in the general case one if one replaces P in (1) by any L_N^k -definable predicate $A_i(x)$, where the $A_i(x)$ determine a partition, the resulting right hand side for $j = n$ will remain a strictly increasing function of n which tends to 1).

The relation between $c^\dagger$ and c^* reconnects with the question of language-dependence. Again to simplify, let us focus the discussion on L_2^1 . There are now just two Q -predicates, $P(x)$ and $\neg P(x)$, and three structure-descriptions: $P(a_1) \& P(a_2)$, $(P(a_1) \& \neg P(a_2)) \vee (\neg P(a_1) \& P(a_2))$, and $\neg P(a_1) \& \neg P(a_2)$. So $m^*(\exists x P(x)) = 1/3$, where t is any logical truth in L_2^1 , whereas $m^\dagger(\exists x P(x)) = 1/4$. This is very significant and suggestive. Indeed, let $L_0^1(=)$ be like L_0^1 except that $L_0^1(=)$ has equality. Where t_2 is the $L_0^1(=)$ sentence $\exists x \exists y (\neg(x = y) \& \forall z (z = x \vee z = y))$ it is easy to see that the value assigned to a sentence s containing no constants in L_2^1 by c^* is equal to $m(s \& t_2)/m(t_2)$ where for any sentence w , $m(w)$ is the number

¹³Of the two m -functions $[m^\dagger$ and $m^*]$, m^* is the only one which is not entirely inadequate’ [1950, p.565].

of distinct $L_0^1(=)$ -models of w up to isomorphism. Call m the Bolzano measure (as we know it will be a different function depending on whether the language in question has individual names or not). But where s is a sentence in L_2^1 containing at least one constant and m is the corresponding Bolzano measure, $m(s \& t_2)$ is in general *not* equal to $m^*(s)$. For example, if s is $\neg P(a_1) \& P(a_2)$, then $m^*(s) = 1/6$, while $m(s \& t_2) = 1/4 = m^\dagger(s)$.

That c^* in effect pastes together Bolzano measures for two distinct languages becomes even more apparent when we compute $m^*(s(n, r))$ where $s(n, r)$ is the disjunction of all sentences $\pm P(a_{i1}) \& \dots \& \pm P(a_{in})$, $n < N$, where as before $+P = P$ and $-P = \neg P$, in which there are r occurrences of P . Let $\text{Str}(R)$ be the structure-description in which there are R unnegated P s. Then

$$m^*(s(n, r)) = \sum_R c^*(s(n, r), \text{Str}(R)) m^*(\text{Str}(R))$$

where $m^*(\text{Str}(R)) = (N+1)^{-1}$ and $c^*(s(n, r), \text{Str}(R)) = {}^R C_r^{N-R} C_{n-r}^N / {}^N C_n$. For large N this is approximately equal to ${}^n C_r p^r (1-p)^{n-r}$ where $p = R/N$. As N tends to infinity $g(R/N) = m^*(\text{Str}(R))$ for $R = 0, 1, \dots, N$ approximates the uniform distribution over $[0, 1]$, and consequently $m^*(s(n, r))$ tends to ${}^n C_r \int_0^1 p^r (1-p)^{n-r} dp$. A completely rigorous proof of this result, a special case of his representation theorem for exchangeable events, was first provided by de Finetti [de Finetti, 1937]. Now let m be the Bolzano measure relative to $L_N^1(=)$, and let t_N be a sentence of $L_N^1(=)$ asserting that there are exactly N individuals, each named by $a_1, \dots, a_N$.¹⁴ $m(s(n, r), t_N \& \text{Str}(R)) = {}^R C_r^{N-R} C_{n-r}^N / {}^N C_n = c^*(s(n, r), \text{Str}(R))$. But $m(\text{Str}(R), t_N) = m^*(\text{Str}(R))$ only relative to the language $L_0^1(=)$, not $L_2^1(=)$.

Carnap nevertheless insisted that despite its mongrel character c^* is an authentic *logical probability function*. What might be thought to justify the terminology is his claim that the disjoined state-descriptions of any structure-description are *isomorphic*, as if m^* is just identifying isomorphic models. But this is not the case. According to the definition of isomorphism for relational structures, the models of $\neg P(a_1) \& P(a_2) \& t_2$ and $P(a_1) \& \neg P(a_2) \& t_2$ are *not* isomorphic, since the constants of a language L play an essential role in determining isomorphism-classes of L -structures (to see this intuitively, recall that every constant a in L can be defined by adding a corresponding unary predicate P_a and stipulating that x satisfies P_a just in case $x = a$). Formally, the state-descriptions in a given structure-description are simply an orbit of the group S_N of permutations of $\{1, \dots, N\}$ acting on the set of all state-descriptions.

Carnap's clearly rather tolerant criteria of what it means to be logical, at any rate as far as probability was concerned, became even more inclusive after 1950, progressing far beyond c^* as his emphasis shifted from identifying measures of the proportional overlap of logical ranges to providing a general quantitative theory of rational belief and decision.¹⁵ His retention of the descriptor 'logical' during this

¹⁴We know from the earlier discussion of Bolzano how to write t_N (though there is of course no unique representation).

¹⁵'Far beyond' if anything understates the case: c^* is just a single point in his so-called

evolution was justified on the very moot ground that the constraints he imposed were describable in purely mathematico-logical terms: for example, the invariance principles imposed on the c -functions, like invariance under permutations of the individuals and predicates in h and e , and finally — in an attempt to free the system from any language-dependence — under additions of arbitrarily many new extralogical items to the language.¹⁶ Even the earlier idea that the functions should be defined on the sentences of a suitable formal language was relaxed in his final publications, where the descendants of the earlier c -functions now become measures on explicitly set-theoretical probability spaces of the type introduced by Kolmogorov, i.e. fields of subsets of a basic set, with the only link to the earlier account provided by the stipulation that sets of model-structures are relative to certain language systems [Carnap, 1970]. The link is a weak one, however, since we are told that these spaces can and usually will be σ -fields, and hence can contain elements represented by no sentences of any classical (first order) language.

Judged by anything other than an excessively tolerant criterion of what is logical, these constraints are almost certainly *not* a part of logic. In c^* we something like the culmination of a process in which logic is being systematically confused with normative methodological criteria: that a certain formally specified system can learn from experience in some way¹⁷ might well be seen as a worthy AI goal, but it hardly seems a dictate of logic. Nor are the invariance postulates themselves obviously logical, or even always sensible. The postulate of symmetry in the sense of functional symmetry, i.e. permutation-invariance for individuals, for example, is equivalent to a universal assumption of the exchangeability of the propositions $Q(a_{ij})$. Although some people working in the Carnap programme saw this as showing that Carnapian inductive logic had come of age, from its humble beginnings of probabilities defined on simple finite sampling languages (the L_N^k) to the point where it could now connect significantly with modern Bayesian statistics, the fact is that as far as most people are concerned, and this includes de Finetti himself, exchangeability was not by any means a universal property: the example de Finetti gives in his 1937 paper is tossing a coin n times, where he says

In general, different probabilities will be assigned [the 2^n possible outcomes], depending on the order, whether it is supposed that one toss has an influence on the other one which follows it immediately, or whether the exterior circumstances are supposed to vary, etc.; nevertheless it is particularly interesting to study the case where the probability does not depend on the order of the trials [i.e. where exchangeability can be assumed]. [1937, p.121]

continuum of inductive methods [1952], a family of c -functions parametrised by a real number λ lying in the half-open interval $[0, \infty)$; for $L_2^1 c^*$ corresponds to $\lambda = 2$.

¹⁶Carnap [1950; 1971]. It turns out that these demands are not all simultaneously satisfiable; see [Fine, 1973 p.193].

¹⁷As noted earlier, c^* obeys the Rule of Succession, and also the further 'inductive' condition called the Principle of Positive Instantial Relevance by Carnap: $c(Q(a_{n+1}), Q(a_n) \& e) > c(Q(a_{n+1}), e)$.

In other words, exchangeability is a *modelling assumption*, appropriate for some situations, i.e. where one's background knowledge entitles one to regard the position of individuals in an ordered sample as irrelevant, but not for others where there is believed to be dependency from trial to trial, as in Markov chains for example.

Saying that some constraint is logical *because* it can be described in purely logico-mathematical terms therefore begs the question. So too, though possibly less obviously, do the range-measure probability functions defined in the way Bolzano prescribed. It is not just that the phenomenon of language-dependence does not arise in deductive logic, nor even that there are serious problems with infinity: there is also the fundamental question of why the different models of a set of sentences should be counted equally. No appeal to informational symmetry can answer this question, because that too begs the same question. Implicit in such measures is a substantive principle, the Principle of Indifference, no less substantive — and arbitrary — because it is often concealed in the formalism. It might be (and sometimes is) objected that in assigning different weights to the members of the possibility-space one would be equally, if not more, begging the question. What could justify such an a priori discrimination? The objection fails to take into account the fact that there will *always* be a priori discrimination because of the inevitable language-dependence associated with this measure. As we have seen, in the logical space determined by two individuals and one predicate P , but without individuating names, ' P is instantiated twice' has the same probability as ' P is instantiated once' ($1/3$), whereas if one adds the individuating names the sentences are ranked as differently in probability ($1/4$ and $1/3$ respectively).

Objections like these have convinced many if not most commentators that the idea of an epistemically neutral logical probability *metric* which tells us how, for given data, we *should* evaluate the probabilities of contingent propositions is untenable: the constraints imposed in order that various consequences regarded as epistemic goods can be delivered beg the question. Without such constraints, however, we are left with the probability axioms and nothing more, and if this is logic then, so it would seem, it is too weak to be of any use whatever. Indeed, it would seem to be inevitably a subjective matter how, within those very liberal constraints, one chooses probability-values.

One of the surprising, and at first sight paradoxical, findings to emerge in the aftermath of Carnap's work was a recognition that indeterminacy of this type need not preclude a purely subjective theory of probability from being seen as an authentically logical probability. Carnap's 'logical probability' has speciated into a variety of what have become called probability logics, mostly retaining the original Carnapian feature that probabilities are assigned to the sentences (and sometimes even open formulas/formulas also) of an appropriate formal language, usually some extension of a first order language with equality. These developments fall into two broad classes, which in their different ways mirror developments in deductive logic itself, in particular its treatment of the modalities 'necessarily' and its dual 'possibly'. One aspect of necessary truth is automatically provided for in the model

theory of first and higher order logic, where it is identified with the category of logical truth. But in addition there are the modal logics which generalise first order logic by adding a new sentential operator to the first order vocabulary,¹⁸ where the usual formula-closure conditions mean that one can have arbitrary degrees of embedding of modal operators. The post-Carnapian treatments of logical probability reflect this bifurcation, with one generalising the model theory of classical logic and the other adding a probability-operator to the object language.¹⁹ The first of these ways of extending classical logic seems to connect more naturally with the developments in Bayesian probability, due to Ramsey and in particular de Finetti, whose central concern was the *consistency* of sets of probability assignments, and it is on that that I shall largely focus on in the remaining sections.

5 FROM LOGICAL PROBABILITY TO PROBABILISTIC LOGIC

I observed earlier that one can take either deductive consequence or deductive consistency as the primary deductive concept, since these are interdefinable. Bolzano, Keynes and Carnap saw in the formal properties of a conditional probability function a natural generalisation of the former. It was a seductive analogy, but in retrospect the wrong one. But a more promising way had already been pointed, if implicitly, when in a memorable phrase Frege called the rules of classical deductive logic ‘laws of truth’ [1956]. After Tarski’s seminal work Frege’s characterisation could be refined to ‘the laws of truth-valuations’. Truth-valuations (with respect to any given structure) can be represented by functions from sentences in a formalised language into the set $\{0,1\}$, and a fundamental if logically trivial property of truth-valuations is that semantically equivalent sentences receive the same value. This means that factoring the class of sentences of the appropriate language by equivalence yields a Boolean algebra, and the valuations can therefore be viewed as defined on the algebra itself. We have the elements of a familiar structure for probabilists: a Boolean algebra on which a finitely additive measure is defined. It remains to take the decisive step and allow the measure to take arbitrary values in the entire unit interval.

In addition, an almost trivial rewriting of classical first order logic shows that deductive consistency can be seen equivalently as a property of sets of *truth-value assignments*. Because in classical logic there are only two truth-values they don’t have to be, and usually are not, explicitly represented in the primitive vocabulary (negation ‘defines’ falsity). But in his book [1968] developing a very user-friendly version of semantic tableaux for sentences in ordinary first order languages, Raymond Smullyan also presented an elegant account of consistency and consequence for what he called ‘signed’ sentences, i.e. first order sentences sentences to which can be adjoined the letters *T* and *F*. If *A* is such a sentence, *AT* and *AF* can

¹⁸In a celebrated paper Montague showed that if the usual modal axioms are preserved together with the assumption that necessary sentences are true, then necessity cannot be an object-language *predicate* under any suitable codings, e.g. Gödel numbering, of sentences.

¹⁹As for example does Halpern [1990].

of course be equivalently represented by $A = 1$ and $A = 0$.²⁰ A set of signed propositional sentences is true in a model, i.e. a full Boolean valuation, just in case the model assigns the sentences those values, and the set is consistent just in case it has a model.

This is of course highly suggestive: once consistency is seen to be a property of value-assignments to sentences, and hence to members the sentence Boolean algebra of the language obtained by factoring by equivalence, the generalisation to a corresponding account for value-assignments in more extensive number-domains, like $[0,1]$, is of probabilistic consistency and consequence in the probabilistic case is obvious. Deductive consistency and probabilistic consistency can then be seen to be subspecies of a single species of consistency, the familiar notion of consistency as the solvability of equations subject to constraints, with the constraints being the differentiae: in the deductive case the clauses of a classical truth-definition, and in the probabilistic case the laws of probability.²¹

That said, signing propositions with arbitrary real numbers in $[0,1]$ is clearly a more complex affair than signing them with values in $\{0,1\}$, a problem further complicated by the fact that one might want not just to specify values for selected sentences but assert that the values satisfy various algebraic relations specifiable within some suitable mathematical language/theory T . Another problem is that the typical structure investigated in mathematical probability is a σ -algebra with a countably additive probability function defined on it. A first order language L determines only an ordinary Boolean algebra (via the map from sentences A to either $\text{Mod}(A)$, the class of models of A , or to the corresponding member $|A|$ of the Lindenbaum sentence algebra of L), not closed under countable sup and inf.

These issues were addressed in a paper by Dana Scott and Peter Krauss, published in 1966 and building on earlier work by Haim Gaifman [1964]. Firstly, they took as the language to which probability assignments are to be made an *infinitary language* of type $L_{\omega_1, \omega}$ in the $L_{\kappa, \lambda}$ family, where κ and λ are infinite cardinal numbers and $L_{\kappa, \lambda}$ is a first order language closed under the formation of fewer than κ conjunctions and disjunctions, and strings of quantifiers of length less than λ . Their particular choice of first order basis language L contained, apart from identity, just one binary relation $\mathbf{R}$ (since all of mathematics can be expressed in first order set theory, and first order set theory has just one binary relation symbol for membership as its only extralogical vocabulary, this allows a degree of universality to the range of propositions that can be expressed).

²⁰For fairly obvious reasons no embedding is permitted: despite what looks like a blurring of the object-language/metalanguage distinction, this is a calculus of truth-value assignments to object-language sentences in a semantically open language.

²¹Cf de Finetti:

Just as the ordinary logic of two values is the necessary instrument of all reasoning where only the fact that an event happens or does not happen enters in, so the logic of the probable, the logic of a continuous scale of values, is the necessary instrument of all reasoning into which enters, visibly or concealed, a degree of doubt' [1937, p.155]

$L_{\omega_1, \omega}$ languages have of course much greater expressive power than first order languages.²² For example, the standard model of arithmetic can be defined by a single sentence. More relevant to the present discussion is the fact that the Lindenbaum algebra is isomorphic to a σ -field of sets, of the form $\text{Mod}(\varphi)$ for sentences φ in $L_{\omega_1, \omega}$, and as a consequence many of the characteristic structures of mathematical probability theory are definable. Suppose for example that the sample space, call it C , is the set, 2^ω , the set of all functions from ω (i.e. the natural numbers) into $\{0,1\}$, which in descriptive set theory is often identified with the real numbers. If the base first order language for $L_{\omega_1, \omega}$ contain a denumerably infinite set of constants a_i and a one-place predicate symbol $B(x)$, the members of C can be identified with the structures (up to isomorphism) of the appropriate type in $\text{Mod}(\sigma)$ where $\sigma = \bigwedge_{j < \omega} \bigwedge_{i < j} \neg(a_i = a_j) \wedge \forall x [\bigvee_{i < \omega} (x = a_i)]$. The basic neighbourhoods are definable by ordinary first order sentences conjoined with σ , and hence the Borel sets in C are definable in $L_{\omega_1, \omega}$:

... L_{ω_1, ω_0} is the proper generalisation of L_{ω_0, ω_0} to denumerable formulas. In fact we have seen several reasons for claiming that L_{ω_1, ω_0} plays the same role for L_{ω_0, ω_0} that the theory of Borel sets and σ -fields plays for the ordinary field of sets'. [Scott, 1965, p.341]

Probability-assignments to sentences of $L_{\omega_1, \omega}$, which Scott and Krauss call 'probability assertions', are defined by them as sequences $(\Phi, \mathbf{A})$, where Φ is a formula of $L(T)$ with n free variables and $\mathbf{A} = (A_1, \dots, A_n)$ is a finite sequence of sentences of $L_{\omega_1, \omega}$. Informally, $(\Phi, \mathbf{A})$ 'says' that the probabilities of $A_1, \dots, A_n$ satisfy Φ . The class of interpretative structures, called 'probability-systems' by Scott and Krauss, for these probability-assertions are ordered quintuples (D, R, Id, B, m) where D is a non-empty set (the domain), B is a Boolean σ -algebra, R is a mapping from $D \times D$ into B , Id is also a mapping from $D \times D$ into B (having the substitution property and with $Id(a, a) = 1$, for all a in B), and m is a strictly positive countably additive probability function on B . A valuation function homomorphically associates the members of the Lindenbaum sentence algebra of the language with corresponding members of B and induces a distribution of probability values from m over all the sentences. A probability-assertion $(\Phi, \mathbf{A})$ is true in such a generalised relational structure just in case $\Phi[m(A_1), \dots, m(A_n)]$ holds in the intended model of T . 'Valid probability assertion' and 'valid consequence' are defined analogously to the deductive case: where α is a probability assertion and Σ a set of assertions ' α is valid' = ' α is true in all probability models' and ' α is a consequence of Σ ' = 'every probability model of Σ is a model of α '. For the choice of T Scott and Krauss adopted the first order theory M of algebraic real number theory, in which among other things inequalities and equalities between polynomials with real coefficients can be expressed and evaluated.

The strict positivity of m ensures that if it is two-valued then B is the Boolean

²²Despite this there is something like an effectively-specifiable proof system, with the property that every semantic consequence of a countable set of sentences is a provable consequence (though this does not extend to consequences of uncountable sets).

algebra **2** and the probability system is a classical relational structure. So we have an authentic probabilistic generalisation of classical model theory. What of the proof theory? A famous result of Tarski is that M admits quantifier-elimination and so is decidable (and complete). The existence of a proof procedure for $L_{\omega_1, \omega}$ -logic, complete for countable sets of assumption-sentences, and the decidability of M , were exploited by Scott and Krauss in demonstrating that the set of valid probability assertions is effectively enumerable [1966, Theorem 6.7]: this is not proved in the usual way via some axiomatisation; instead Scott and Krauss use the familiar fact that any probability statement can be put in an equivalent normal form involving only propositional atoms, and then show that there is an algorithm for enumerating the valid ones). But the natural coding structure for the syntax and proof theory of $L_{\omega_1, \omega}$ is the set HC of hereditarily countable sets (sets of countable sets whose members are countable, etc.). But HC is a very much larger set than the coding structure for countable first order languages, which is N if you use ordinary Gödel numbering or the class HF of hereditarily finite sets (i.e. the set of sets of finite rank in the cumulative hierarchy). The number-codings associated with HC , by contrast, are the real numbers. Hence ‘effectively enumerable’ in the context of $L_{\omega_1, \omega}$ -validity does not mean recursively enumerable.²³ ‘Effectively enumerable’ in the discussion above just means ‘definable in HC by a Σ_1 formula’, by analogy with ‘recursively enumerable’ in the sense of ordinary recursion theory, viz. ‘definable in N by a Σ_1^0 relation’).

Recursion theory tells us that a Σ_1 -definable subset of HC translates into a Σ_2^1 subset of 2^ω , i.e. the real numbers. This is certainly not the meaning of ‘effectively enumerable’ in the sense of being in the output of a Turing machine (i.e. a Σ_1^0 subset of N). In seminal work in the nineteen seventies Barwise showed that by restricting the set of HC -codes of the uncountably many $L_{\omega_1, \omega}$ formulas to a certain type of *countable* set A (technically, a *countable admissible set*), and hence permitting a coding in N , several of the distinctive features of first order logic can be reproduced — notably, compactness. In addition, such a restriction provides a generalisation of recursion theory on N known as A -recursion. But the corresponding notions of A -recursiveness (Δ_0 -definable in A) and A -recursive enumerability (Σ_1 -definable in A), though still characterising sets of natural numbers, still remain far stronger than ordinary recursiveness and recursive enumerability for any admissible set: even for the smallest countable admissible set.²⁴ which does not collapse $L_{\omega_1, \omega}$ into $L_{\omega, \omega}$, the corresponding natural-number classes all lie in the analytic hierarchy²⁵

So for anyone accustomed to a notion of logical reasoning as generating a recur-

²³A formula in the language of first order set theory is said to be $\Sigma_1(\Pi_1)$ if it is built up from atomic formulas $x \in y$ using only the connectives, bounded quantifiers $\exists x \in w (\forall x \in w)$, and unbounded $\exists x (\forall x)$. The set of valid probability assertions of $L_{\omega, \omega}$, as might be expected, is recursively enumerable [Scott and Krauss, 1966, Theorem 7.6].

²⁴i.e. $L(\omega_1^{CK})$, i.e. the set of all sets constructible in Gödel’s sense before the first nonrecursive ordinal (CK stands for ‘Church-Kleene’).

²⁵The $L(\omega_1^{CK})$ -recursively enumerable sets are Π_1^1 subsets of N , and the $L(\omega_1^{CK})$ -recursive sets are hyperarithmetical (Δ_1^1).

sively enumerable set of theorems, the proof theory for probability assertions based on $L_{\omega_1, \omega}$ goes a long way beyond that. On the other hand, as far as capturing ordinary probability theory is concerned, Scott and Krauss's logic of probability assertions is still relatively weak: M is only capable of describing (and deciding) algebraic relations among probabilities, while there are well-known limitations on the expressive power of even the $L_{\omega_1, \omega}$ languages. But there is of course a very familiar ready-made language *and* effective proof theory already at hand for real-valued probability functions defined on arbitrary σ -fields of sets: first order set theory, in which probability-systems à la Kolmogorov can be defined and investigated. A predictable reply to this proposal is that it takes us out of the domain of pure logic and puts us firmly back in that of special mathematical theories. One might therefore think that Scott and Krauss have gone as far as it is possible to go in providing a probabilistic semantics and proof theory which generalise those of classical deductive logic, with simultaneous barriers of expressiveness and effectiveness preventing an adequate representation of the usual subject-matter and proof-techniques of informal mathematical probability. The latter, it seems, is irreducibly mathematics, and not logic.

But that does not follow, any more than it follows that the metatheory of ordinary first order logic, whose semantic part also classifies assignments of values to assertions with an underlying Boolean structure, is not logic. Indeed, I have already pointed out that at a certain level of description the semantics of probability and deductive logic are identical. The emphasis on formalised languages, somewhat uncritically inherited from Carnap, has been partly to blame for obscuring this fact. Formalisation is a red herring, but if one wants to find a formal analogue it is not predicate logic that one should turn to but propositional — where the Boolean connectives ‘and’, ‘or’ and ‘not’ can just as well operate on interpreted sentences — and in particular the version of propositional logic due to Smullyan [1968], mentioned earlier, which deals with ‘signed’ sentences (we shall make the ‘signings’ 0 and 1 rather than Smullyan's F and T) in some propositional language L . The signed sentences are simply *truth-value assignments* to the unsigned sentences, assignments which may or may not be inconsistent (Smullyan's refinement of the tableau system provides a decision-procedure). Following Smullyan further, we can define a *model* of a set Q of such assignments as a single-valued extension of Q to all the sentences of L satisfying the classical truth-table constraints [1968, p.10], and say that Q is *consistent* just in case it has a model. We can further define an assignment Q' to other sentences of L to be a *consequence* of Q just in case every model of Q is a model of Q' .

The kinship with probability becomes even more apparent if we factor the unsigned sentences by logical equivalence, yielding a Boolean algebra $L/\Leftrightarrow$ on which corresponding definitions of model, consistency etc. are induced. It is an important and well-known fact that, whether defined on the sentences of the language or on its Lindenbaum algebra, consistency and consequence are *absolute*, or in de Finetti's terminology ‘intrinsic’: they do not depend on the ambient language L , or on any other sentences except those mentioned in the assignments. Also, the

logic is *compact*: an infinite assignment is consistent just in case every finite subset is though there is now no relation to an underlying topology.

All that changes in the generalisation to probabilistic model, probabilistic consistency and probabilistic consequence is that the value-set is $[0,1]$ instead of $\{0,1\}$ and the constraints are those of the probability axioms rather than the truth-table rules. If only finite additivity is imposed the properties of absoluteness (extendability to any including algebra) and compactness are also maintained. Extendability follows from a very deep result of de Finetti which exhibits his concept of coherence as a fundamental *theoretical* tool. The result is an extension theorem somewhat analogous to the well-known extension theorem for countably additive measures, except that in this case only finite additivity is assumed. Let us call it *de Finetti's Extension Theorem* (he himself called it the 'Fundamental Theorem of Probability': [1974, vol. 1]. It states that an assignment $Q(E)$ of probability-evaluations to any set E of propositions is coherent if, by definition, for every finite sum of bets at the odds given by Q there is no set of stakes which will generate a loss or gain for every distribution of truth-values over E .²⁶ De Finetti's theorem states that if E contains the certain proposition then Q is a coherent assignment of probabilities to a set of propositions which includes the certain proposition just in case Q can be extended to any including algebra in a way that satisfies the finite probability axioms [1972, p.78]. So suppose Q is consistent in the sense given earlier. It follows from de Finetti's theorem that Q is coherent, from which it follows that it can be extended consistently to any including algebra. QED. Compactness is another corollary of the same theorem. Suppose Q is inconsistent. Then it cannot be extended to the least including algebra in a way that satisfies the finitely additive probability axioms. Hence by de Finetti's theorem Q is incoherent. Hence a finite subset Q_0 is incoherent and hence, again by de Finetti's theorem, is inconsistent.

6 CONDITIONAL PROBABILITY

A natural question to ask at this point is where conditional probabilities fit into this picture. An initially seductive idea, which despite severe problems continues to seduce some, is to adopt the principle that a conditional probability is the probability of a conditional *proposition*, with the attractive corollary that a separate treatment of conditional probability is unnecessary. De Finetti held this view, as have others, most notably Adams [1975], after whom it is often called Adams's Principle. It appeared to most people to have foundered on the celebrated Triviality Theorem of Lewis [1976], which effectively presents a dilemma neither horn of

²⁶De Finetti gave *two* criteria of coherence corresponding to different operational definitions of an agent's subjective probability: one as their fair betting quotient in a forced-betting scenario, and the other as their estimate of a random quantity X subject to a penalty proportional to the square of the differences between the evaluation and the actual value of X . The criterion of coherence stated above is that appropriate to the first definition. For the second, a set of evaluations is coherent just in case for every finite subset of them no other choices uniformly reduce the penalty. A proof that the two criteria are equivalent is in [de Finetti, 1974, vol. 1].

which is very appealing: if a conditional probability is the probability of a conditional proposition and the rules of the finitely additive probability calculus hold, together with the condition

$$P(A|B) = P(A \& B) / P(B)$$

where $P(B) > 0$, then either the conditional propositions occupy some non-Boolean extension of classical propositional logic, or the probability function is absurdly limited in the values it can take. The second possibility is held to be self-dismissing, and most people reject the first too, and with it Adams's Principle.

De Finetti was an exception, who in the nineteen thirties, long before Adams or Lewis wrote, actually developed a non-classical propositional logic which supports Adams's Principle as well as the rule

$$P(A|B)P(B) = P(A \& B). \quad (1)$$

On the other hand, the additivity rule fails for conditional propositions, a fact which allows de Finetti's theory to evade the thrust of Lewis's theorem but also destroys the simplicity possessed by the classical probability calculus.²⁷ An understandable reluctance of probabilists to embrace a non-classical logic and the ensuing complications of the probability calculus resulted in this theory failing to find a ready acceptance. Nor, probably for that reason, did de Finetti himself ever make that account part of his 'official' corpus, where he took the more traditional route of treating $P(A|B)$ as a function of *two* classical (two-valued) propositional variables A and B , representing an agent's fair betting quotient in a bet on A which is called off if B is discovered to be false (in de Finetti's three-valued logic $A|B$ would then be void and a bet on $A|B$ would of course itself be void). A simple coherence argument then justifies (1) [1937, p.109], as well as the additivity principle for unconditional probabilities.

Notice that (1) does not constrain $P(A|B)$ at all if $P(B) = 0$, for then (1) simply becomes the equation $0 = 0$, permitting a variety of 'natural' evaluations of $P(A|B)$: e.g., if only finite additivity is imposed one might set $P(X = 3 | X \in \{1, 2, 3, 4\}) = 1/4$ where there is a uniform unconditional distribution (identically 0) over the positive integers; we shall see in the next section that de Finetti himself was happy to admit countable-additivity-violating distributions, among them this one. On the other hand, as far as (1) is concerned $P(A|B)$ can be consistently be greater than 1 or even negative. This seems unsatisfactory, particularly in view of the fact that if B is consistent it might be true, in which case one should allow for the possibility of updating $P(\cdot)$ to $P(\cdot|B)$ on learning B (and nothing stronger; $P_B(\cdot) = P(\cdot|B)$ is then said to be obtained from P by *conditionalising* on B). Thus it seems natural to require that $P(\cdot|B)$ should be a coherent unconditional probability for fixed consistent B , with $P(B|B) = 1$ (this is de Finetti's Axiom 3, [1974, vol. 2, p. 339]). But coherence-plus-conditionalisation also demands that

²⁷Mura [2009] contains a thorough discussion of de Finetti's system.

conditional probabilities relative to any function P_B obtained by conditionalisation should also satisfy (1), i.e. we should have

$$P_B(A|E)P_B(E) = P_B(A \& E) \quad (2)$$

Given de Finetti's operational definition of conditional probabilities as conditional betting quotients, $P_B(A|E)$ is a fair betting quotient in a doubly-conditional bet which is called off iff B is false or if B is true and E is false, i.e. iff $B \& E$ is false. Hence $P_B(A|E) = P(A|E \& B)$ and (2) becomes

$$P(A|E \& B)P(E|B) = P(A \& E|B), \quad (3)$$

often called the *multiplication rule of probabilities*. Setting $B = T$, necessary truth, (3) subsumes (1) as a special case.

Putting these observations together we obtain the following coherence-backed rules for conditional probabilities:

- i. $P(.|B)$ is a finitely additive unconditional probability for consistent B .
- ii. $P(B|B) = 1$ for consistent B .
- iii. $P(A \& E|B) = P(A|E \& B)P(E|B)$ for consistent $E \& B$.

i–iii. are in fact a standard axiomatisation of conditional probability: they are the axioms of Rényi [1956], minus the axiom of countable additivity and also follow from the scale-free axioms of Cox [1961] by regenerating any of his admissible measures using a suitable increasing continuous function. The definition of probabilistic consistency can now be extended to a set Q of conditional probability assignments $P(X|Y) : Q$ is consistent just in case it has an extension to $F \times F \setminus \{\perp\}$ satisfying i.–iii., where F is any Boolean algebra [Coletti and Scozzafava, 2002, p.76]. As in the unconditional case, it can be shown that an infinite assignment is consistent just in case every finite subset is.

7 COUNTABLE VERSUS FINITE ADDITIVITY

De Finetti's theory of coherence is a matter of controversy, generating an enormous literature that cannot be adequately evaluated here.²⁸ The fact is that almost everyone agrees on the finitely additive probability axioms, and on the axioms i.–iii. of conditional probability,²⁹ however they may differ on how exactly those rules are to be justified. The question of whether countable additivity should be added is even more controversial, with de Finetti being one of the leading

²⁸The objections are virtually all directed against coherence as a theory of rational behaviour, but as I noted in the previous section coherence also plays a fundamental theoretical role in de Finetti's theory, having nothing at all to do with rationality considerations.

²⁹Some prefer to take i.–iii. as the sole axioms, given that a finitely additive unconditional probability is obtained from i.–iii. by setting $B = T$.

figures to repudiate it. His opposition has been regarded as paradoxical by many commentators, since it would appear that the same rationale, coherence, that he invoked to justify the finitely additive axioms and the rule (1) extends also to mandating countable additivity. Indeed, it is straightforwardly proved by a direct argument that violating countable additivity leaves an (infinite) assignment vulnerable to a certain loss or gain. A simple example will suffice here, one actually provided by de Finetti himself: a countably infinite lottery where the probability of each ticket is stipulated to be zero [1972, p.91]. A unit stake on each outcome will generate a certain gain of 1; hence, it would seem, the lottery-assignment is incoherent.

There is no paradox, however. Recall that coherence stipulates that an assignment is coherent just in case for any *finite* set of bets at the corresponding odds there is no set of stakes which generates a certain loss or gain for every distribution of truth-values over the constituent propositions. The infinite lottery generates a sure loss only if all the individual bets are summed: for no finite subset is the loss surely zero, and hence it is coherent. Underwriting the definition of coherence is an assumption, which de Finetti sometimes calls ‘rigidity’ [1974, vol. 1], that the class of fair bets, in the sense of bets at the agent’s subjective probability evaluations, is closed under finite sums only.³⁰ It entails the conclusion that a finite sum of fair bets which generates a certain gain or loss is fair, which is *prima facie* inconsistent and thereby explains de Finetti’s claim (see section 1) that the incoherence of a set of probability-evaluations implies that they are ‘intrinsically contradictory’. At any rate, the only conclusion that one is entitled to draw from the example above is that a countable infinity of fair bets need not sum to a fair bet. Indeed, the principle of countable additivity is easily shown to be equivalent to the assumption that it should which, as de Finetti points out, begs the question (*ibid*).³¹ In fact, it was precisely to avoid the commitment to countable additivity that de Finetti restricted rigidity to finite sums of bets,

Exactly why he wanted to avoid that commitment I shall come to shortly: his reasons are, I believe, very compelling if one takes the logical view of the rules of probability that I believe de Finetti himself endorsed. His repudiation of countable additivity as a general rule has been widely criticised, however, largely because of an entrenched view of mathematical probability as the theory of countably additive normalised measures with the powerful and unifying methods that point of view affords. It guarantees the ability to commute probabilities with limits and supports a powerful theory of integration based on the Lebesgue integral. The effects of abandoning countable additivity are felt at a relatively low level: for example, where $\{B_i : i = 1, 2, 3, \dots\}$ is a countable partition, the marginal decomposition

$$P(A) = \sum P(A|B_i)P(B_i)$$

is not universally valid. It follows that what de Finetti calls the *conglomerative*

³⁰The assumption is presented explicitly as a *definition* of fairness for bets in [de Finetti, 1972, p.77].

³¹This is a subtle point, discussed in more detail in [Howson, 2008].

property also fails: this is the property, possessed by all countably additive functions P , that $P(A)$ lies within any common set of bounds for all the $P(A|B_i)$.

How genuinely ‘negative’ these features are is debatable, however, and de Finetti showed that in many cases they actually accord with intuitive judgments (see, for example, [1972, p.102; 1974, vol. 1, p.178]). Moreover, there are other consequences of maintaining only finite additivity that have an arguably very positive character. Finitely additive probabilities are extendable to arbitrary algebras; countable ones are not. Finitely additive distributions are a closed set; countable ones are not. There are finitely additive uniform distributions over countably infinite partitions; there are none under countable additivity. In any event, de Finetti’s response to the supposedly deleterious mathematical consequences was always that mathematics be the servant, not the master, of correct principles,³² and that correct principles in fact pronounce against, and not for, countable additivity. One in particular was, for him, fundamental. It is, in effect, that the laws of probability are laws of logic:

The calculus of probability can say absolutely nothing about reality ...
As with the logic of certainty [deductive logic], the logic of the probable adds nothing of its own: it merely helps one to see the implications contained in what has gone before. [1974, vol. 1, p.215], emphasis in the original.

There is probably no clearer statement in the literature on epistemic probability of the claim that probability and deductive logic are united under the rubric of universally valid rules empty of substantive content, and it is precisely this principle that de Finetti claims countable additivity violates. As we saw earlier, countable additivity entails that any distribution over a countably infinite partition places nearly all the probability on a finite subset. More precisely, for any $\varepsilon > 0$, however small, if $\{B_i : i = 1, 2, 3, \dots\}$ is a disjoint family then for some n , $P(B_1 \vee \dots \vee B_n) > 1 - \varepsilon$. In particular, the distribution cannot be uniform. This is of course not true for either a finite set or an uncountably infinite interval of real numbers, as de Finetti observes. Indeed, suppose you have a uniform distribution over $[0, 1]$, and then learn that the random quantity in question is a rational number in that interval. If countable additivity is assumed you are compelled to adopt the skewed distribution over the rationals under some enumeration. As de Finetti points out, no metrical consideration can justify this bias. The probability formalism *itself* produces it, in effect manufacturing information where there was none before [1974, vol. 1, pp.121-123] — contrary to the fundamental principle above.

It might be objected that deductive logic itself supplies a counterexample to de Finetti’s principle, since $\exists x(A(x) \vee \neg A(x))$, i.e. *something exists*, is a theorem of classical first order logic. That is true, and it is because it seems to contravene the principle that logic should be empty of content, and in particular not endorse unconditional existential assertions, that several people have proposed a variant,

³²[1972, p.89].

called *free logic*, which is the closest one can get to first order logic without entailing the thesis that something necessarily exists. This is not the place to go further into that issue, but it is for much the same reason that most probabilists endorse the axiom of countable additivity — computational simplicity - that most logicians are reluctant to go over to free logic. But I think that de Finetti’s response would be that they are all merely sacrificing principle for ease, a judgment that in my opinion it is difficult to fault.

8 CONCLUSION

I have ended this discussion as I began it, with de Finetti. I believe he rightly deserves to be given such prominence because, despite the strongly decision-theoretic focus he gave his later work, he is so far the only person to provide, if at times implicitly, a defensible account of Bayesian probability as an authentic *logic*. His account was never systematically developed in one place or at one time — one has to pick up what appear to be sporadic *obiter dicta* scattered across his works and across decades. Nor was it couched in the terminology of modern logic. But if the account given in the preceding sections is correct then that is not very difficult to do, and in a way that exhibits deductive and probabilistic logic as parallel disciplines distinguished only in the nature of the constraints they impose on consistent evaluations. What more than anything else has militated against a recognition of the close kinship of the two disciplines (compare the quote from Kyburg in section 1.) has been that they seem to be ‘about’ such very different things: internal relations between sentences versus numerical assignments to propositions, ‘laws of truth’ versus subjective beliefs, and so on. De Finetti, unlike nearly everyone else, did not fail to see the wood for the trees, with a clarity of vision evidenced early on in the title of his most famous paper: ‘Prevision: its *logical laws*, its *subjective sources*’ (my emphasis).

ACKNOWLEDGEMENTS

I am grateful for very helpful comments by an anonymous referee.

BIBLIOGRAPHY

- [Adams, 1975] E. W. Adams. *The Logic of Conditionals*, Dordrecht: Reidel, 1975.
- [Bayes, 1763] T. Bayes. An Essay Towards Solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society of London*, vol. 53, 370-418, 1763.
- [Bolzano, 1837] B. Bolzano. *Wissenschaftslehre. Versuch einer ausführlichen und grösstentheils neuen Darstellung der Logik mit steter Rücksicht auf deren bisherige Bearbeiter*, 1837. Page references to the English translation, *Theory of Science. Attempt at a Detailed and in the main Novel Exposition of Logic with Constant Attention to Earlier Authors*, by Rolf George (Oxford: Blackwell, 1972).
- [Carnap, 1950] R. Carnap. *Logical Foundations of Probability*, Chicago: University of Chicago Press, 1950.

- [Carnap, 1952] R. Carnap. *The Continuum of Inductive Methods*, Chicago: University of Chicago Press, 1952.
- [Carnap and Jeffery, 1971] R. Carnap and R. C. Jeffery, eds. *Studies in Inductive Logic and Probability*, volume 1, Berkeley: University of California Press, 1971.
- [Coletti and Scozzafava, 2002] G. Coletti and R. Scozzafava. *Probabilistic Logic in a Coherent Setting*, Dordrecht: Kluwer, 2002.
- [Cox, 1961] R. T. Cox. *The Algebra of Probable Inference*, Baltimore: The Johns Hopkins Press, 1961.
- [Dale, 1982] A. I. Dale. Bayes or Laplace? An Examination of the Origin and Early Application of Bayes's Theorem, *Archive for History of Exact Sciences*, vol. 27, 33-47, 1982.
- [Dawid et al., 1973] A. P. Dawid, M. Stone, and J. V. Zidek. Marginalisation Paradoxes in Bayesian and Structural Inference, *Journal of the Royal Statistical Society, B*, 189-223, 1973.
- [Fagin, 1976] R. Fagin. Probabilities on Finite Models, *Journal of Symbolic Logic*, vol. 41, 50-58, 1976.
- [de Finetti, 1936] B. de Finetti. La logique de la probabilité', *Actes du Congrès International de Philosophie Scientifique*, vol. IV, 1-9 1936.
- [de Finetti, 1937] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives, Paris, Institut Henri Poincaré, 1937. Page references to the English translation 'Foresight: its Logical Laws, its Subjective Sources', by H. E. Kyburg, in *Studies in Subjective Probability*, Second edition (1964), eds. H.E. Kyburg and H. Smokler, New York: Wiley,
- [de Finetti, 1972] B. de Finetti. *Probability, Induction and Statistics*, New York: Wiley, 1972.
- [de Finetti, 1974] B. de Finetti. *Theory of Probability*, vols. 1 and 2, 1974. (English translation of *Teorie delle probabilità*, Einaudi, 1970)
- [Fine, 1973] T. L. Fine. *Theories of Probability*, New York: Academic Press, 1973.
- [Frege, 1956] G. Frege. The Thought: A Logical Enquiry, *Mind*, vol. 65, 289-311, 1956.
- [Gaifman, 1964] H. Gaifman. Concerning Measures in First Order Calculi, *Israel Journal of Mathematics*, vol. 1, 1-18, 1964.
- [Gaifman and Snir, 1982] H. Gaifman and M. Snir. Probabilities Over Rich Languages, Testing and Randomness, *The Journal of Symbolic Logic*, vol. 47, 495-548, 1982.
- [Gillies, 2000] D. A. Gillies. *Philosophical Theories of Probability*, London: Routledge, 2000.
- [Halpern, 1990] J. Y. Halpern. An Analysis of First-Order Logics of Probability, *Artificial Intelligence*, vol. 46, 311-350, 1990.
- [Howson, 2000] C. Howson. *Hume's Problem: Induction and the Justification of Belief*, Oxford: The Clarendon Press, 2000.
- [Howson, 2001] C. Howson. The Logic of Bayesian Probability, *Foundations of Bayesianism*, eds. D. Corfield and J. Williamson, Dordrecht: Kluwer, 137-161, 2001.
- [Howson, 2008] C. Howson. De Finetti, Countable Additivity, Coherence and Consistency, *British Journal for the Philosophy of Science*, vol. 59, 1-23, 2008.
- [Howson and Urbach, 2006] C. Howson and P. Urbach. *Scientific Inference: the Bayesian Approach* (3rd edition), Chicago: Open Court, 2006.
- [Jaynes, 1973] E. T. Jaynes. The Well-Posed Problem, *Foundations of Physics*, vol. 3, 413-500, 1973.
- [Jech, 1978] T. Jech. *Set Theory*, New York: Springer, 1978.
- [Keynes, 1921] J. M. Keynes. *A Treatise on Probability*, London: Macmillan, 1921.
- [Kolmogorov, 1933] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, 1933. (Page References are to the English translation *Foundations of the Theory of Probability*, New York: Chelsea, 1951.)
- [Kuhn, 1962] T. S. Kuhn. *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press, 1962.
- [Lewis, 1976] D. Lewis. Probabilities of Conditionals and Conditional Probabilities, *Philosophical Review*, LXXXV, 581-589, 1976.
- [Mura, 2009] A. Mura. Probability and the Logic of de Finetti's Tries, in *Bruno de Finetti: Radical Probabilist*, ed. M.C. Galavotti, 201-242, 2009.
- [Popper, 1959] K. R. Popper. *The Logic of Scientific Discovery*, New York: Harper & Row, 1959.
- [Ramsey, 1926] F. P. Ramsey. Truth and Probability, *The Foundations of Mathematics*, ed. R.B. Braithwaite, London: Kegan Paul, Trench, Trubner, 156-199, 1926.
- [Rényi, 1955] A. Rényi. On a New Axiomatic Theory of Probability, *Acta Mathematica Academiae Scientiarum Hungaricae*, vol. VI, 285-335, 1955.

- [Savage, 1954] L. J. Savage. *The Foundations of Statistics*, New York: Dover, 1954.
- [Scott and Krauss, 1966] D. Scott and P. Krauss. Assigning Probabilities to Logical Formulas, *Aspects of Inductive Logic*, eds. J. Hintikka and P. Suppes, Amsterdam: North Holland, 219-264, 1966.
- [Seidenfeld, 1979] T. Seidenfeld. Why I am not an Objective Bayesian: Some Reflections Prompted by Rosenkrantz, *Theory and Decision*, vol. 11, 413-440, 1979.
- [Smullyan, 1968] R. Smullyan. *First Order Logic*, New York: Dover, 1968.
- [Tarski, 1951] A. Tarski. *A Decision Method for Elementary Algebra and Geometry*, Berkeley: University of California Press, 1951.

This page intentionally left blank

BAYESIAN INDUCTIVE LOGIC, VERISIMILITUDE, AND STATISTICS

Roberto Festa

Below we will consider the relations between inductive logic and statistics. More specifically, we will show that some concepts and methods of inductive logic may be applied in the rational reconstruction of several statistical notions and procedures and that, in addition, inductive logic suggests some new methods which can be used for different kinds of statistical inference. Although there are several approaches to inductive logic and statistics, here we will focus on some versions of the Bayesian approach and, thereby, on the relations between Bayesian inductive logic and Bayesian statistics. The paper is organized as follows. The subjects of inductive logic and statistics will be shortly illustrated in Section 1, where it will be suggested that statistics can be seen as a special field of inductive logic. Two important theories developed within Bayesian inductive logic are the theory of inductive probabilities, started by Rudolf Carnap in the forties of the past century, and the theory of confirmation: the conceptual relations between such theories and statistics will be considered in Section 2. A recent version of Bayesian inductive logic, proposed by Ilkka Niiniluoto and others, has been developed by using the notion of verisimilitude, introduced in philosophy of science by Karl Popper; the key ideas of the verisimilitudinarian version of Bayesian inductive logic will be illustrated in Section 3, where it will be argued that it provides useful conceptual tools for the analysis of some important kinds of statistical inference.

1 BAYESIAN INDUCTIVE LOGIC AND BAYESIAN STATISTICS

1.1 Inductive logic and statistics

Empirical sciences make use of different kinds of inferences leading from a set of one or more statements called premises to another statement, called conclusion. According to a traditional distinction inferences are divided into deductive and inductive. While in deductive inferences the information conveyed by the conclusion is already included in the premises, in inductive inferences the conclusion says something extra with respect to the premises. Hence, even though the truth of the premises is taken as guaranteed, there is the inevitable risk that the extra information conveyed by the conclusion of an inductive inference may be false. This means that the conclusion of an inductive inference is inevitably uncertain. The systematic analysis of inductive inferences is the subject matter of inductive

logic. Suppose that, in an inquiry on the hypothesis H , evidence is expressed by a sentence E , where H cannot be deduced from E . In this case we might ask what are the inductive relations between E and H . A good inductive logic should provide, among other things, a satisfying analysis of the inductive relations between evidence and hypotheses. In particular, it should answer the following questions about the plausibility, confirmation, and acceptability of H in the light of E : (1) What is the degree of *plausibility* of H in the light of E ? (2) What is the degree of *confirmation* conveyed by E to H , i.e., what is the increase of the initial plausibility of H determined by E ? (3) Given evidence E , should we *accept* H , in some suitably specified sense of acceptance? The inductive inferences involved in the questions (1)–(3) may be labelled as plausibilistic, confirmational, and acceptance-based inferences, respectively.

A statistical inference is a particular type of inductive inference where evidence is (the description of) a sample taken from a given population and the hypothesis concerns the whole population or certain samples to be drawn from it. More generally, a statistical inference is an inductive inference where evidence is (the description of) a sequence of trials of a given experimental process and the hypothesis concerns certain parameters of the process or certain future sequences of trials. The main goal of statistics is the formulation of a good theory of statistical inferences, i.e., of a systematic corpus of procedures applicable to any kind of statistical inference.

The above characterization of inductive logic and statistics suggests that statistics is a province of the wide continent of inductive logic. Indeed, if inductive logic is construed as the study of inductive reasoning, wherever and however it occurs, then an ideal inductive logician looks for stable patterns underlying the various forms and areas of inductive reasoning. Hence an ideal inductive logician looks, among other things, for stable patterns underlying the statistical reasoning in empirical sciences. This means that the subject matter of statistics belongs to inductive logic, i.e., that the basic relation between the two disciplines is the very close relation occurring between a whole and a part of it. However, in spite of this intimate relation, starting at least from the beginning of the past century, inductive logic and statistics have undergone a separate development. While inductive logic has been developed within the ground of philosophy, especially within epistemology and philosophy of science, statistics has impetuously grown as a highly specialized discipline, with its own departments and faculties, in a constant interaction with scientific practice, including the more sophisticated outcomes of probability theory and, more generally, of mathematical research. This separate growth of inductive logic and statistics has led, among other things, to the development of different languages, conceptual systems, and formal tools, so that it is difficult to recognize that in many cases the two disciplines deal with the same problems, and sometimes propose essentially identical solutions. Fortunately, in the last few decades there have been some signals of an opposite tendency, i.e., of an increasing awareness of the close conceptual relations and the potential fruitful interactions between inductive logic and statistics.

1.2 The Bayesian approach to inductive logic and statistics

The most popular approach to inductive logic is the Bayesian approach. The basic idea of *Bayesian inductive logic* (for short: *BIL*) is that any kind of inductive inference should be analysed in terms of inductive — or, equivalently, epistemic — probabilities. This means that any inductive inference from evidence E to the hypothesis H amounts to — or, at least, is based on — the attribution of the epistemic probability $p(H|E)$ to H , where $p(H|E)$ expresses the degree of belief that a rational agent X should attribute to H in the light of E . Within the different versions of BIL this basic idea has been applied in the analysis of plausibilistic, confirmational, and acceptance-based inferences. For instance, the plausibility of H in the light of E is usually identified with $p(H|E)$ and the confirmation conveyed by E to H is usually identified with some measure of the probability increase in the shift from the initial probability $p(H)$ of H to its final probability $p(H|E)$. Different versions of BIL may be distinguished on the basis of the methodological importance attributed to the different kinds of inductive inference. For instance, most supporters of BIL advocate a non-acceptance-based approach, according to which acceptance-based inferences are dispensable both in science and everyday life, where we need only plausibilistic and confirmational inferences. In conflict with this view, some Bayesian inductive logicians defend an acceptance-based approach which emphasizes the methodological indispensability of acceptance-based inferences. Within this approach two basic kinds of acceptance rules have been proposed: (i) purely probabilistic acceptance rules, where the acceptance of a hypothesis H depends *only* on its probability; (ii) decision-based acceptance rules, where the acceptance of H is seen as a cognitive decision depending not only on the probability of H but also on the (suitably defined) cognitive utilities associated with the acceptance of H in the possible states of nature: for instance, the acceptance rules stated within the so-called cognitive decision theories prescribe to accept the hypothesis which maximizes the expected cognitive utility [Levi, 1967; 1980; Niiniluoto, 1987; Festa, 1999b; 2003].

While in the last century Bayesianism has been the dominant view among inductive logicians, the orthodoxy among statisticians has been anti-Bayesian. Indeed, the so-called orthodox, or frequentist, statistics developed in the twentieth century — on the track of Jerzy Neyman, Egon S. Pearson, and Ronald A. Fisher [1925; 1935; 1950; 1956] — is based on the explicit rebuttal of the Bayesian idea that statistical inferences can be made by attributing epistemic probabilities to statistical hypotheses. In conflict with this idea, orthodox statisticians maintain that statistical frequencies are the only probabilities to be used in statistical inferences. However, starting from the second half of the twentieth century, the Bayesian approach to statistics has known — thank to the work by Bruno de Finetti [1937/1964; 1974], Harold Jeffreys [1931; 1937], Leonard J. Savage [1954], and many others — an impressive development, so that nowadays *Bayesian statistics* (for short: *BS*) represents a serious alternative to frequentist statistics.

2 THEORY OF INDUCTIVE PROBABILITIES, CONFIRMATION, AND STATISTICS

2.1 *Inductive methods for multinomial inferences*

Most of the work on BIL made by Rudolf Carnap between the beginning of the forties and his death, in 1970, is directed to the development of a theory of epistemic probabilities for the analysis of a well known kind of statistical inference, i.e., multinomial inferences. His theory is usually known under the label of *theory of inductive probabilities* (for short: TIP). Below we will shortly illustrate the notion of inductive method for multinomial inferences — for short: *multinomial method* — and the different strategies used within BS and TIP for the construction of such methods. Afterwards, we will consider some (non-analogical and analogical) multinomial methods introduced within TIP and BS and we will show that, in spite of their different conceptual basis, the methods worked out within TIP are essentially identical to those used within BS.

2.1.1 *Multinomial methods within TIP and BS*

Consider an experimental process Ex whose trials are described by using a qualitative character, or property, $\mathbf{Q} = \{Q_1, \dots, Q_k\}$, where the k (> 2) categories $Q_1, \dots, Q_k$ are mutually exclusive and jointly exhaustive. We say that Ex is a *multinomial process* in the case where, for any $Q_i \in \mathbf{Q}$, there is a physical probability $\mathbf{q}_i$ that the outcome of an arbitrary trial of Ex is Q_i , where $\mathbf{q}_i$ is constant across trials — neither influenced by the trial number nor by the outcomes of previous trials. The background knowledge available in many empirical inquiries includes the assumption that the investigated process Ex is a multinomial process. However, typically a researcher does not know the true value of the parameter vector $\mathbf{q} = (\mathbf{q}_1, \dots, \mathbf{q}_k)$ of Ex , but can only formulate different kinds of hypotheses about $\mathbf{q}$ and the results of future trials of Ex . The inductive inferences concerning such hypotheses will be called *multinomial inferences*. More precisely, a multinomial inference about Ex amounts to the determination of the epistemic probability $p(H|E)$, where each of the sentences H and E may concern either the value of $\mathbf{q}$ or the result of some sequence of trials of Ex . A *multinomial method* is an inductive procedure allowing to make any kind of multinomial inference about any multinomial process Ex .

Within BS the starting point for the definition of a multinomial method is given by a suitable representation of the researcher's initial opinions about the value of the parameter vector $\mathbf{q}$ of the investigated multinomial process Ex , i.e., by an appropriate *prior distribution* $F(\mathbf{q})$ of epistemic probabilities on the possible values of $\mathbf{q}$. A different strategy is adopted within TIP, where the starting point for the construction of a multinomial method is given by the definition of an *inductive rule* R specifying, for any possible sequence e_n of outcomes of n trials of Ex and any category $Q_i \in \mathbf{Q}$, the *predictive probability* $p(Q_i|e_n)$ of the hypothesis that the result of the next trial of Ex will belong to Q_i . Then the predictive

probabilities $p(Q_i|e_n)$ may be used to determine the prior probability $p(e_n)$ of any sequence e_n of outcomes of n future trials of Ex , the posterior probability $p(e_n|e_m)$, where e_m describes a sequence of outcomes of m past trials of Ex , and any other epistemic probability $p(H|E)$ associated to Ex . In Carnap's terminology predictive probabilities $p(Q_i|e_n)$ are referred to as the *special values* of a multinomial method, where "special" refers to their special importance in the definition of a multinomial method.

The conceptual differences between the strategies for the construction of multinomial methods adopted within BS and TIP can be construed as the differences between a *globalistic* and a *predictivistic* approach to multinomial inferences. Indeed, the strategy adopted within BS is based on the attribution of appropriate epistemic probabilities to the possible values of the parameter vector $\mathbf{q}$ of the multinomial process Ex which is, as it were, a *global* feature of Ex . On the contrary, the strategy adopted within TIP is based on the attribution of appropriate epistemic probabilities to any possible *prediction* concerning a future trial of Ex in the light of any possible sequence of past trials of Ex .

Although this conceptual difference between TIP and BS cannot be neglected, it should be recalled that there is also an important conceptual equivalency between TIP and BS. We refer to the circumstance that an important class of multinomial methods investigated within TIP, i.e., the exchangeable multinomial methods, is essentially identical to the multinomial methods considered within BS. Let n_i denote the frequency of Q_i in sequence e_n of trials of Ex , i.e., the number of outcomes in e_n belonging to the category Q_i . Then a multinomial method for Ex is called *exchangeable* in the case where, for any n and any e_n , the prior probability $p(e_n)$ depends *only* on $n_1, \dots, n_k$ — and not on the order in which the outcomes $Q_1, \dots, Q_k$ occur in e_n . De Finetti's representation theorem implies that any exchangeable multinomial method I for Ex is essentially identical to a corresponding (unique) prior distribution $F_I(q)$ defined on the parameter vector $\mathbf{q}$ of Ex [de Finetti, 1937/1964; 1974]. This means that the exchangeable multinomial methods worked out within TIP are essentially identical to the multinomial methods considered within BS.

2.1.2 The equivalence between GC-methods and Dirichlet distributions.

A well known family of exchangeable multinomial methods is given by *generalized Carnapian methods* (for short: *GC-methods*), which are characterized by the following inductive rule [Carnap and Stegmüller, 1959]:

$$(\text{GC-rule}) \quad p(Q_i|e_n) = \frac{n_i + \lambda\gamma_i}{n + \lambda}$$

where $\gamma_i > 0$, $\sum \gamma_i = 1$, and $0 \leq \lambda \leq \infty$.

The multinomial methods used within BS are typically based on the so-called Dirichlet distributions, which can be defined on the parameter vector $\mathbf{q}$ of any multinomial process Ex . It can be easily proved that — as a consequence of de

Finetti's theorem — any GC-method is equivalent to a corresponding Dirichlet distribution. It may be interesting to point out that, in spite of its relative triviality, the equivalence between GC-methods and Dirichlet distributions is not mentioned in Carnap's papers and, more generally, that it passed unnoticed until 1965 [Good, 1965; Festa, 1993]. This circumstance clearly reveals the almost surprising lack of interactions between TIP and BS in the forties and fifties of the past century. However, starting from the seventies, the conceptual relations between TIP and BS have been increasingly understood.

2.1.3 The axiomatization of GC-methods and Dirichlet distributions

Between the fifties and the seventies of the past century, inductive logicians and Bayesian statisticians succeeded in axiomatizing GC-methods and Dirichlet distributions.

Axiomatizing a class of inductive methods amounts to proving that the class includes all and only those methods which satisfy a given set of axioms. In particular, GC-methods have been axiomatized by several authors [Carnap and Stegmüller 1959; Carnap 1980, § 19; Kuipers 1978, Ch. 5]. Their results may be summarized as follows:

(T.1) An exchangeable multinomial method I for a family $\mathbf{Q} = \{Q_1, \dots, Q_k\}$ of k categories, where $k > 2$, is a GC-method in the case where the predictive probabilities $p(Q_i|e_n)$ determined on the basis of I satisfy the following principles:

(IP) *Principle of initial possibility.* $p(Q_i) \equiv \gamma_i > 0$

(RR) *Principle of restricted relevance.* If e_n and e'_n are such that $n_i = n'_i$, then $p(Q_i|e_n) = p(Q_i|e'_n)$.

Recalling that Dirichlet distributions are prior distributions defined on the parameter vector $\mathbf{q}$ of a multinomial process Ex , one may ask whether Dirichlet distributions can be axiomatized on the basis of appropriate principles expressing plausible constraints on the way in which we assign our prior probabilities to the possible values of $\mathbf{q}$. We will see below that a positive answer to this question can be given.

Axiomatizing a class of prior distributions amounts to proving that the class includes all and only those distributions which satisfy a given set of axioms. In particular, Dirichlet distributions have been axiomatized by Fabius [1973]. His axiomatization is based on the concept of *neutrality*. Put simply, given the proportions $\mathbf{q}_1, \dots, \mathbf{q}_k$ — which are defined as non-negative parameters satisfying the constraint $\sum \mathbf{q}_i = 1$ — the proportion $\mathbf{q}_1$ is said to be *neutral* w.r.t. $\mathbf{q} = (\mathbf{q}_1, \dots, \mathbf{q}_k)$ when no information on the value of $\mathbf{q}_1$ can influence our beliefs concerning “the manner in which the remaining proportions $\mathbf{q}_2, \dots, \mathbf{q}_k$ *proportionally* divide the remainder of the unit interval” [Connor and Mosimann, 1969, p. 196]. Fabius's axiomatization of Dirichlet distributions (*ibid.*) can be stated as follows:

(T.2) A prior distribution $F(q)$ on the parameter vector $\mathbf{q} = (\mathbf{q}_1, \dots, \mathbf{q}_k)$ of a multinomial process Ex is a Dirichlet distribution in the case where $F(q)$ satisfies the following principle:

(N) *Principle of neutrality.* $F(q)$ is such that any $\mathbf{q}_i$ is *neutral* w.r.t. $\mathbf{q}$.

An important difference between the axiomatizations of GC-methods and Dirichlet distributions should be noticed. The principles (IP) and (RR), used in the axiomatization of GC-methods, are *predictive* principles concerning the behaviour of the inductive rule used to determine the predictive probabilities $p(Q_i e_n)$; on the contrary, the principle (N), used in the axiomatization of Dirichlet distributions, is a *global* principle concerning our opinions about the global feature $\mathbf{q}$ of Ex , i.e., the probabilities attributed to the possible values of $\mathbf{q}$.

Given the essential identity between GC-methods and Dirichlet distributions, we might say that each of the theorems (T.1) and (T.2) provides an axiomatic foundations both of GC-methods and Dirichlet distributions. However, while the axiomatization stated in (T.1) provides a *predictivistic justification* of GC-methods and Dirichlet distributions, the axiomatization stated in (T.2) provides a *globalistic justification* of such methods.¹

2.1.4 The non-analogical nature of GC-methods and Dirichlet distributions

The principles (RR) and (N), used in the axiomatizations of GC-methods and Dirichlet distributions immediately reveal the non-analogical nature of such multinomial methods. Indeed, the principle (RR), used in the axiomatization of GC-methods, implies that the predictive probabilities $p(Q_i | e_n)$ of a GC-method depend on the empirical evidence e_n only via n and n_i . In other words, (RR) implies that $p(Q_i | e_n)$ should in no way be affected by the empirical frequencies n_j — with $j \neq i$ — in e_n ; this independently of the possible, more or less strong, similarity occurring between categories Q_j and Q_i . This means that GC-methods cannot take into account any analogy by similarity, i.e., that GC-methods are not analogical methods. Similar anti-analogical intuitions seem to underlie the principle (N), used in the axiomatization of Dirichlet distributions. Indeed (N) excludes any analogy by similarity by requiring that our beliefs about $\mathbf{q}_i$ should in no way be affected by information concerning the value of the fraction $\mathbf{q}_j / (1 - \mathbf{q}_i)$ — with $j \neq i$ — in the remainder $(1 - \mathbf{q}_i)$ of the unit interval; this independently of the possible, more or less strong, similarity occurring between categories Q_j and Q_i .

2.1.5 Analogical methods for one property

Although most of Carnap's work on TIP was devoted to the investigation of GC-methods and other non-analogical multinomial methods, the construction of ana-

¹On the axiomatic foundation of GC-methods and Dirichlet distribution — and the justification of the use of such multinomial methods in empirical inquiries — see [Festa, 1993, Ch. 6 and 7.6].

logical inductive methods applicable to *multicategorical experiments* — i.e., to experiments whose outcomes are described by one or more categories — has always been on Carnap's agenda.

Let us consider, first of all, the problem of defining appropriate *analogical methods for one property*. Such methods may be introduced w.r.t. to any kind of experimental process described by a property $\mathbf{Q} = \{Q_1, \dots, Q_k\}$ with $k > 2$. This means that Ex might be a multinomial process, but it might also be a different kind of physical process, or a process whose nature is not exactly known.

Given three categories Q_i, Q_j , and Q_k belonging to $\mathbf{Q}$, suppose that the similarity between Q_j and Q_i is higher than that between Q_k and Q_i . If our predictive probabilities are determined by using a GC-method, then equality $p(Q_i|Q_j) = p(Q_i|Q_k)$ holds; more generally, for any sequence e_n of trials of Ex , equality $p(Q_i|e_n Q_j) = p(Q_i|e_n Q_k)$ holds. This means that, if a GC-method is used, then the predictive probabilities are not affected by the similarity relations occurring among Q_i, Q_j , and Q_k . However, in certain cases we might require that our inductive method is sensible to such similarity relations; for instance, we might require that it satisfies the following *weak principle of analogy by similarity*:

(WAS) For any triple Q_i, Q_j , and Q_k such that Q_j is more similar than Q_k to Q_i , $p(Q_i|Q_j) > p(Q_i|Q_k)$.

In some cases we might also require that our inductive method satisfies the following *strong principle of analogy by similarity*:

(SAS) For any sequence e_n of trials of Ex and any triple Q_i, Q_j , and Q_k such that Q_j is more similar than Q_k to Q_i , $p(Q_i|e_n Q_j) > p(Q_i|e_n Q_k)$.

We will say that an inductive method is a *weakly analogical method* in the case where it satisfies (WAS) or akin weak principles of analogy by similarity, while it is a *strongly analogical method* in the case where it satisfies (SAS) or akin strong principles of analogy by similarity. Within TIP the problem of defining appropriate analogical methods for one property has been investigated by Carnap [1963; 1980], Niiniluoto [1981], Costantini [1983], Kuipers [1984a; 1984b; 1988], Skyrms [1993a], Festa [1996], and others. The methods proposed by the just mentioned authors can be classified as follows:

1. **STRONGLY ANALOGICAL NON-EXCHANGEABLE METHODS.** The analogical methods proposed by Carnap [1980], Niiniluoto [1981], Costantini [1983], and Kuipers [1984a; 1984b; 1988] are of this kind. Their methods are introduced in the TIP-style, i.e., on the basis of simple and intuitively transparent inductive rules specifying the predictive probabilities $p(Q_i|e_n)$. Such inductive rules are explicitly designed to guarantee the satisfaction of some strong principles of analogy by similarity, such as (SAS) (Carnap), or akin principles (Niiniluoto and Kuipers). Unfortunately, the above methods exhibit also three related and not very attractive features: (i) they are not exchangeable; (ii) thereby, they are not multinomial methods; indeed, the adoption of

a non-exchangeable method implies that the investigated process Ex is not multinomial, since any prior distribution $F(q)$ on the parameter vector $\mathbf{q}$ of a multinomial process is essentially identical to an exchangeable method; (iii) finally, it is not clear whether the adoption of such methods imposes some constraint on the (non-multinomial) physical structure of Ex .

2. **WEAKLY ANALOGICAL EXCHANGEABLE METHODS.** A method of this kind has been proposed by Skyrms [1993a]. Such method — which will be called Sk — is introduced in the BS-style, i.e., as a prior distribution $F_{Sk}(q)$ on the parameter vector $\mathbf{q}$ of a particular multinomial process Ex , given by a wheel of fortune whose possible outcomes are described by the property $\mathbf{Q} = \{\text{North, East, South, West}\}$. More precisely, $F_{Sk}(q)$ is given by a mixture of four Dirichlet distributions defined on $\mathbf{q}$. This implies that Sk is exchangeable, i.e., that Sk is a multinomial method. One can easily check that Sk is weakly analogical since it satisfies (WAS). However, none of the strong principles of analogy so far proposed in the literature is satisfied by Sk [Festa, 1996]. Finally, it is not clear whether Sk can be generalized so to obtain a class of strongly analogical methods applicable to different kinds of multinomial process.
3. **STRONGLY ANALOGICAL EXCHANGEABLE METHODS.** A method of this kind has been proposed by Festa [1996]. Also Festa's method — which will be called Fe — is introduced in the BS-style, i.e., as a prior distribution $F_{Fe}(q)$ on the parameter vector $\mathbf{q}$ of a particular multinomial process Ex , described by the property $\mathbf{Q} = \{Q_1, Q_2, Q_3\}$ with three categories. More precisely, $F_{Fe}(q)$ is given by a mixture of two Dirichlet distributions defined on $\mathbf{q}$. This implies that Fe is exchangeable, i.e., that it is a multinomial method. Moreover, it can be proved that Fe is strongly analogical since it satisfies (SAS). However, together with such attractive features, Fe shares with Sk an unpleasant feature; indeed, it is not clear whether and how Fe can be generalized to different kinds of multinomial process.

The research on analogical methods for one property seems to be an important field for the cooperative development of TIP and BS: indeed, some of the above mentioned analogical methods have been developed by using the conceptual tools both of TIP and BS. However, one has to recognize that so far the research on analogical methods for one property has not led to outcomes comparable with those of the research on non-analogical methods and, in particular, on GC-methods and Dirichlet distributions. Among other things, it should be recalled that, so far, (i) no general class of exchangeable analogical methods has been proposed, and (ii) no axiomatic foundation of exchangeable analogical methods has been provided.

2.1.6 *Analogical methods for many properties*

Another interesting problem considered within TIP is the definition of appropriate *analogical methods for many properties*. Such methods have been investigated with

reference to multinomial processes whose possible outcomes are cross classified on the basis of two or more properties [Carnap, 1963; Hesse, 1964; Maher, 2000; 2001; Romeyn, 2005; 2006]. In the most simple case, the outcomes are described by using two properties $\mathbf{F} = \{F_1, F_2\}$ and $\mathbf{G} = \{G_1, G_2\}$ where $F_2 \equiv \text{non-}F_1$ and $G_2 \equiv \text{non-}G_1$. Given two trials a and b of a multicategorical experiment Ex , we might require that our predictive probabilities satisfy the following analogical principle:

$$(A) \quad p(G_1b|F_2G_1a \& F_1b) > p(G_1b|F_2G_2a \& F_1b)$$

Suppose, for instance, that Ex is given by the random drawing of an individual from the population of swans and that the members of such population are described by the properties $\mathbf{F}$ and $\mathbf{G}$, where F_1 stands for “Australian” and G_1 for “white”. Then a and b will refer to individuals randomly drawn from the population of swans. With reference to this example, the intuitive meaning of (A) can be expressed as follows: the probability that the next observed Australian swan is white, given that a previously observed non-Australian swan is white, is greater than the probability of such hypothesis given that a previously observed non-Australian swan is non-white.

If an inductive method satisfies (A) or other akin analogical principles, we will say that it is an analogical method for many properties. Several analogical methods for many properties have been introduced in the literature. While the investigation of analogical methods for many properties was initially carried out using only the conceptual tools of TIP [Carnap, 1963; Hesse, 1964; Niiniluoto, 1980; 1988], the research on this subject made in the last decade is characterized by the extensive use also of the conceptual tools of BS [Maher, 2000; 2001; Romeyn, 2005; 2006]. In particular, such research has shown that analogical methods for many properties have close — and still not fully explored — relations with the statistical research on contingency tables [Albert and Gupta, 1983; Epstein and Feinberg, 1992; Good, 1965; 1983; Lindley, 1964].²

2.2 Inductive methods for non-multinomial statistical inferences

While the research on analogical methods has not led to outcomes comparable with those of the research on GC-methods and Dirichlet distributions, a much more successful story can be told about the research made on other two issues which were on Carnap’s agenda. We are referring to the construction of inductive methods for a value continuum and of inductive methods sensitive to analogy by proximity. Although here there is no room for this successful story, the reader can find an lucid survey of the research in this area in Skyrms [1996].³ Among other things, Skyrms clearly shows that such research is characterized by a deep integration of TIP and BS methods. For instance, within three years of Carnap’s

²For an explicit reference to the relation between Carnap’s analogical methods for two properties and the problem of estimating cell probabilities in the case of a two-dimensional contingency table, see [Costantini and Garibaldi, 1996].

³See also [Skyrms, 1991; 1993b].

death, the problem of the construction of inductive methods for a value continuum was “completely solved in a way quite consonant with Carnapian techniques” [Skyrms *ib.*, p. 321], by three Bayesian statisticians, i.e., Blackwell and MacQueen [1973] and Ferguson [1973].

2.3 Confirmation and statistics

As said in Section 1.2, the confirmation conveyed by evidence E to a hypothesis H is usually identified with some measure of the probability increase in the shift from the initial probability $p(H)$ of H to its final probability $p(H|E)$. Two popular incremental measures of confirmation are the probability difference $c_d(H, E) \equiv p(H|E) - p(H)$ and the probability ratio $c_r(H, E) \equiv p(H|E)/p(H)$. Another popular incremental measure is defined in terms of the initial and final odds of H , i.e., in terms of $o(H) \equiv p(H)/(1 - p(H))$ and $o(H|E) \equiv p(H|E)/(1 - p(H|E))$. We are referring to the so-called *odds ratio* $c_{or}(H, E) \equiv o(H|E)/o(H)$. An attractive feature of $c_{or}(H, E)$ is given by the easily proved equality $c_{or}(H, E) = p(E|H)/p(E|\neg H)$. The quantity $p(E|H)/p(E|\neg H)$ is commonly known as *Bayes factor* (in favour of H). It should be noticed that, while $c_{or}(H, E)$ is well defined only in the case where $p(H)$ is positive, in principle the Bayes factor $p(E|H)/p(E|\neg H)$ may be well defined also in the case where $p(H)$ is zero. More generally, one immediately sees that the initial probability of H is not necessary for the determination of the Bayes factor in favour of H . These feature of Bayes factor explains why it has been considered as an attractive measure of confirmation for statistical hypotheses, also by a number of non-Bayesian statisticians, who would be scarcely inclined to adopt confirmation measures stated in terms of the initial or final probabilities of statistical hypotheses.

Although the research on the confirmation of scientific hypotheses has been carried out mainly within the general framework of inductive logic [Festa, 1999a; 2009; Festa *et al.*, 2010; Fitelson, 1999], in the last few decades the concept of confirmation — or, equivalently, empirical support — has attracted an increasing attentions among statisticians. For instance, several frequentist statisticians have suggested that the so-called p -values of statistical hypotheses can be construed as an appropriate measure of their degree of empirical support. The first Bayesian statistician who has devoted a lot of attention to the confirmation measures suggested within inductive logic — and to the possibility to apply such measures to statistical hypotheses — is I. J. Good. Among other things, Good provides a thorough analysis of Bayes factor and suggests a Bayesian rational reconstruction of the measures of corroboration proposed by Karl Popper as an alternative to Bayesian measures of confirmation.⁴

⁴See especially [Good, 1950; 1960; 1960/1968; 1961; 1968; 1975; 1980; 1981a; 1981b; 1982; 1983; 1984; 1985a; 1985b; 1989a; 1989b; 1989c; 1991; 1997] and [Good and McMichael, 1984].

3 VERISIMILITUDE AND STATISTICS

3.1 *Verisimilitude and inductive logic: towards a verisimilitudinarian Bayesian inductive logic*

A recent version of BIL, which may be called *verisimilitudinarian BIL*, is based on the integration, within the conceptual framework of BIL, of the notion of verisimilitude. The first formal notion of verisimilitude — or approximation to the truth — was worked out by Karl Popper within his falsificationist and anti-inductivist philosophy of science [Popper, 1963]. Afterward such notion, and its methodological applications, have been extensively investigated within the post-Popperian theories of verisimilitude, emerged since 1975.⁵ Two basic issues dealt with by such theories are the following:

1. The analysis of the logical problem of verisimilitude, in order to provide appropriate notions of verisimilitude and distance from the truth. Such notions should allow us to compare any two theories w.r.t. their closeness to the truth, even in the case where they are false. Post-Popperian theories of verisimilitude generally admit that adequate notions of verisimilitude and distance from the truth should leave room for the possibility that, under certain conditions, a false theory is closer to the truth than a true one.
2. The analysis of the epistemic problem of verisimilitude, in order to define the notions of estimated verisimilitude and estimated distance from the truth. Such notions should allow to compare, on the basis of the available data, any two theories w.r.t. their estimated closeness to the truth, even in the case where they have been falsified by such data. It is widely agreed that adequate notions of estimated verisimilitude and estimated distance from the truth should allow that, under certain conditions, the estimated verisimilitude of a falsified theory is higher than that of a non-falsified one.

The notion of verisimilitude may be used in the analysis of all the three kinds of inductive inference defined in Section 1.2, i.e., plausibilistic, confirmational, and acceptance-based inferences. In fact one may define the plausibility of H in the light of E as the estimated verisimilitude of H in the light of E and may identify the degree of confirmation conveyed by E to H is with an appropriate measure of the increase of the initial estimated verisimilitude of H determined by E . Finally, one may state verisimilitudinarian rules of inductive acceptance based on the idea that verisimilitude is the proper cognitive utility of science, so that, within a set of rival hypotheses, one should accept the hypothesis which maximizes the expected verisimilitude.

⁵An excellent survey of the post-Popperian theories of verisimilitude is provided by Niiniluoto [1998].

3.2 *Verisimilitudinarian Bayesian inductive logic and statistics*

In the last twenty years the notions and principles of verisimilitudinarian BIL have been applied in the rational reconstruction of a number of widely used statistical procedures. Some of these applications are shortly illustrated below.

1. THE CHOICE OF THE OPTIMUM MULTINOMIAL METHODS. Multinomial inferences might be defined as statistical inferences aiming at establishing the true value of the parameter vector $\mathbf{q} = (\mathbf{q}_1, \dots, \mathbf{q}_k)$ of a multinomial process *Ex.* A major problem for the Bayesian analysis of multinomial inferences — which has been thoroughly investigated both in TIP and BS — is the choice of an optimum prior distribution $F^*(q)$ on $\mathbf{q}$. Festa [1993; 1995] considers this problem with reference to Dirichlet distributions (or, equivalently, GC-methods) and suggests a verisimilitude solution, inspired by the intuitive idea that the optimum Dirichlet distribution $F^*(q)$ should be chosen so that the estimates of $\mathbf{q}_1, \dots, \mathbf{q}_k$ based on $F^*(q)$ approach the truth most effectively. This means that a good reason for selecting $F^*(q)$, within the class of Dirichlet distributions, is that there are grounds to believe that $F^*(q)$ is the optimum tool for achieving a high degree of verisimilitude. The main points of Festa's verisimilitude solution can be informally stated as follows: (i) the efficiency of a Dirichlet distribution (GC-method) in approaching the truth depends on the degree of order in the sector of the universe under investigation; (ii) hence, the adoption of a given Dirichlet distribution (GC-method) as the optimum multinomial method — i.e., as the presumably most efficient tool to approach the truth — should be based on appropriate presuppositions about the degree of order in the investigated sector of the universe; (iii) such presuppositions are given by the scientists' initial estimates — grounded on their background knowledge — of the degree of order of the sector of the universe under examination within a given empirical inquiry. indexepistemic probability
2. BAYESIAN POINT AND INTERVAL ESTIMATES. Given a distribution of epistemic probabilities on the relevant set of statements, one can calculate, on the basis of available evidence E , the expected verisimilitude and the expected distance from the truth of any hypothesis H under inquiry. One can construct a theory of inductive acceptance based on the idea that the distance from the truth is the cognitive loss whose expected value should be minimized when a hypothesis is selected for inductive acceptance. This idea has been applied to the rational reconstruction of certain standard results about point estimation obtained in Bayesian statistics. More specifically, Niiniluoto [1982a, 1987, pp. 27-29] shows that, given a suitable loss function, accepting a Bayesian point estimate is equivalent to minimizing the expected distance from the truth. Furthermore, Niiniluoto [1982b; 1986; 1987, pp. 430-41] and others [Festa, 1986; Maher, 1993, pp. 143-47] extend this perspective to Bayesian interval estimation by approaching interval es-

timization in decision-theoretic terms, where the loss of accepting an interval is defined by the distance of this interval from the truth.

3. **BAYESIAN ANALYSIS OF OBSERVATIONAL ERRORS.** The expected verisimilitude of a hypothesis H can be seen as a reasonable estimate of the degree of verisimilitude of H , while the probable verisimilitude of H expresses our degree of belief in the possibility that H is highly verisimilar, i.e., that the verisimilitude of H exceeds a fixed threshold [Niiniluoto, 1987, Ch. 7.3]. The notions of expected and probable verisimilitude have been applied to the analysis of significant statistical problems such as the problem of observational errors. This problem typically arises in the cases where the evidence is obtained by measuring a given quantity: in fact, in such cases there is a positive probability that, due to errors in measurement, the result of measurement deviates from the true value of the quantity. After discussing in general terms the possibility of appraising verisimilitude on the basis of false evidence, Niiniluoto [1987, Ch. 7.4] suggests that the classical Bayesian treatment of observational errors can be understood as an attempt to evaluate the expected and probable verisimilitude of hypotheses on the basis of evidence which is known (with probability one) to be erroneous.
4. **PROBLEMS OF CURVE-FITTING AND REGRESSION ANALYSIS.** In many cognitive problems — for instance, whenever the inquiry is based upon counterfactual idealizing conditions — all the relevant hypotheses under inquiry are known to be false. A well known example of this sort of cognitive problems is given by the statistical problems of curve-fitting, where all the relevant quantitative hypotheses, expressed by sufficiently simple linear functions, are known to be false. A verisimilitude interpretation of the typical statistical methods based on the Least Square Difference between a curve and a finite set of data points is provided by Niiniluoto [1987, Ch. 7.5]. He suggests also that verisimilitude could be used for the rational reconstruction of the statistical methods of regression analysis which are applied to problems in which all the relevant hypotheses are false (idealizations) and the evidence is known to be erroneous [Niiniluoto, 1987, p. 289].
5. **APPROACHING THE STATISTICAL TRUTH BY QUALITATIVE THEORIES.** In the last three decades several measures of verisimilitude have been defined for different kinds of qualitative and quantitative theories, and their methodological applications have been thoroughly explored. Above we have seen, with reference to quantitative hypotheses, that the notion of verisimilitude can be fruitfully applied in the analysis of some important kinds of statistical inferences. However, recently it has been argued that also qualitative theories may have interesting statistical applications [Festa, 2007a; 2007b; 2008]. More precisely, it has been argued that qualitative theories stated in qualitative languages with two or more properties can be used in the description of the statistical structure of cross classified populations and that

their adequacy in this task can be evaluated by appropriate measures of their *statistical verisimilitude*. Among other things, Festa's notion of statistical verisimilitude has been used for the rational reconstruction of some aspects of the so-called prediction logic, a relatively recent approach to the statistical analysis of cross classified populations developed by the statisticians and social scientists David K. Hildebrand, James D. Laing, and Howard Rosenthal in a series of papers published between 1974 and 1977, which culminated in their not yet adequately appreciated book *Prediction Analysis of Cross Classification* (1976).

The above, necessarily incomplete, survey of some research areas on the border between inductive logic and statistics shows the abundance of conceptual relations between the two disciplines. It also suggests that inductive logicians and statisticians should engage in a serious effort to reduce the inessential differences in their languages and conceptual systems and to explore the numerous possibilities of interaction between — and cooperative development of — the two disciplines.⁶

ACKNOWLEDGEMENTS

I would like to thank an anonymous referee for helpful comments on the first draft of this paper.

BIBLIOGRAPHY

- [Albert and Gupta, 1983] J. H. Albert and A. K. Gupta. Bayesian Estimation Models for 2×2 Contingency Tables Using Mixtures of Dirichlet Distributions. *Journal of the American Statistical Association* 78, 708-717, 1983.
- [Blackwell and MacQueen, 1973] D. Blackwell and J. B. MacQueen. Ferguson Distributions via Polya Urn Schemes. *Annals of Statistics* 1, 353-355, 1973.
- [Carnap, 1963] R. Carnap. Variety, Analogy and Periodicity in Inductive Logic. *Philosophy of Science* 30, 222-227 1963.
- [Carnap, 1980] R. Carnap. A Basic System of Inductive Logic, Part II. In R. Jeffrey (ed.), *Studies in Inductive Logic and Probability, vol. II*. Berkeley: University of California Press, pp. 7-155, 1980.
- [Carnap and Stegmüller, 1959] R. Carnap and W. Stegmüller. *Induktive Logik und Wahrscheinlichkeit*. Wien: Springer-Verlag, 1959.
- [Connor and Mosimann, 1969] R. J. Connor and J. E. Mosimann. Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. *JASA* 64, 194-206, 1969.
- [Costantini, 1983] D. Costantini. Analogy by Similarity. *Erkenntnis* 20, 103-114, 1983.
- [Costantini and Garibaldi, 1996] D. Costantini and U. Garibaldi. Predictive Laws of Association in Statistics and Physics. *Erkenntnis* 45, 259-261, 1996.
- [de Finetti, 1937/1964] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales Institute Henri Poincaré* 7, 1-68, 1937. English translation in Kyburg H. E. Jr. and Smokler, H. (eds.), *Studies in Subjective Probability*. New York: Wiley, pp. 93-158, 1964.
- [de Finetti, 1974] B. de Finetti. *Probability, Induction and Statistics*. New York: Wiley, 1974.

⁶For an excellent survey of the conceptual relations between statistics and inductive logic — and, more generally, between statistics and philosophy of science — see [Good, 1988].

- [Epstein and Feinberg, 1992] L. D. Epstein and S. E. Feinberg. Bayesian Estimation in Multi-dimensional Contingency Tables. In P. K. Goel and N. S. Iyengar (eds.), *Bayesian Analysis in Statistics and Econometrics*. New York: Springer, pp. 27-41, 1992.
- [Fabius, 1973] J. Fabius. Two Characterizations of the Dirichlet distributions. *Annals of Mathematical Statistics* 1, 583-587, 1973.
- [Ferguson, 1973] T. Ferguson. A Bayesian Analysis of Some Non-Parametric Problems. *Annals of Statistics* 1, 209-230, 1973.
- [Festa, 1986] R. Festa. A Measure for the Distance Between an Interval Hypothesis and the Truth. *Synthese* 67, 273-320, 1986.
- [Festa, 1993] R. Festa. *Optimum Inductive Methods. A Study in Inductive Probability, Bayesian Statistics, and Verisimilitude*. Dordrecht: Kluwer, 1993.
- [Festa, 1995] R. Festa. Verisimilitude, Disorder, and Optimum Prior Probabilities. In T. K. Kuipers and A. R. Mackor (eds.), *Cognitive Patterns in Science and Common Sense*. Rodopi: Amsterdam, pp. 299-320, 1995.
- [Festa, 1996] R. Festa. Analogy and Exchangeability in Predictive Inferences. *Erkenntnis* 45, 229-252, 1996.
- [Festa, 1999a] R. Festa. Bayesian Confirmation. In M. C. Galavotti e A. Pagnini (eds.), *Experience, Reality, and Scientific Explanation*. Dordrecht: Kluwer, pp. 55-87, 1999.
- [Festa, 1999b] R. Festa. Scientific Values, Probability, and Acceptance. In R. Rossini Favretti, G. Sandri e R. Scazzieri (eds.), *Incommensurability and Translation*. Cheltenham: Edward Elgar, pp. 323-338, 1999.
- [Festa, 2003] R. Festa. Induction, Probability, and Bayesian Epistemology. In L. Haaparanta and I. Niiniluoto (eds.), *Analytic Philosophy in Finland*. Amsterdam: Rodopi, pp. 251-284, 2003.
- [Festa, 2007a] R. Festa. Verisimilitude, Cross Classification, and Prediction Logic. Approaching the Statistical Truth by Falsified Qualitative Theories. *Mind and Society* 6, 37-62, 2007.
- [Festa, 2007b] R. Festa. The Qualitative and Statistical Verisimilitude of Qualitative Theories. *La Nuova Critica* 47-48, 91-114, 2007.
- [Festa, 2008] R. Festa. Verisimilitude, Qualitative Theories, and Statistical Inferences. In M. Sintonen, S. Pihlström and P. Raatikainen (eds.), *Approaching Truth: Essays in Honour of Ilkka Niiniluoto*. London: College Publications, pp. 143-177, 2008.
- [Festa, 2009] R. Festa 'For Unto Every One That Hath Shall Be Given': Matthew Properties for Incremental Confirmation, *Synthese*. DOI 10.1007/s11229-009-9695-5. Forthcoming
- [Festa et al., 2010] R. Festa, V. Crupi, and C. Buttasi. Towards a Grammar of Bayesian Confirmation. In M. Suárez, M. Dorato, and M. Rédei (eds.), *EPSA Epistemology and Methodology of Science, vol 1*. Dordrecht: Springer, pp. 73-93, 2010.
- [Fisher, 1925] R. A. Fisher. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd, 1925.
- [Fisher, 1935] R. A. Fisher. *The Design of Experiments*. Edinburgh: Oliver & Boyd. 1935.
- [Fisher, 1950] R. A. Fisher. *Contributions to Mathematical Statistics*. New York: Wiley, 1950.
- [Fisher, 1956] R. A. Fisher. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver & Boyd, 1956.
- [Fitelson, 1999] B. Fitelson. The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity. *Philosophy of Science* 66, S362-S378, 1999.
- [Good, 1950] I. J. Good. *Probability and the Weighing of Evidence*. London: Charles Griffin and Co., 1950.
- [Good, 1960] I. J. Good. The Paradox of Confirmation. *BJPS* 11, 145-149, 1960.
- [Good, 1960/1968] I. J. Good. Weight Of Evidence, Corroboration, Explanatory Power, Information, and the Utility Of Experiments. *JRSS B* 22, 319-331, 1960; Corrigenda 30, 203, 1968.
- [Good, 1961] I. J. Good. The Paradox of Confirmation (II). *BJPS* 12, 63-64, 1961.
- [Good, 1965] I. J. Good. *The Estimation of Probabilities*. Cambridge, Mass.: The MIT Press, 1965.
- [Good, 1968] I. J. Good. Corroboration, Explanation, Evolving Probability, Simplicity, and A Sharpened Razor. *BJPS* 19, 123-143, 1968.
- [Good, 1975] I. J. Good. Explicativity, Corroboration, and the Relative Odds of Hypotheses. *Synthese* 30, 39-73 and 83-93, 1975.
- [Good, 1980] I. J. Good. Another Relationship Between Weight of Evidence and Errors of the First and Second Kinds. *JSCS* 10, 315-316, 1980.

- [Good, 1981a] I. J. Good. The Weight of Evidence Provided by Uncertain Testimony or from an Uncertain Event. *JSCS* 13, 56-60, 1981.
- [Good, 1981b] I. J. Good. An Error by Peirce Concerning Weight of Evidence. *JSCS* 13, 155-157, 1981.
- [Good, 1982] I. J. Good. A Good Explanation of an Event Is Not Necessarily Corroborated by the Event. *Philosophy of Science* 49, 251-253, 1982.
- [Good, 1983] I. J. Good. *Good Thinking*. Minneapolis: The University of Minnesota Press, 1983.
- [Good, 1984] I. J. Good. The Best Explication for Weight of Evidence. *JSCS* 19, 294-299, 1984.
- [Good, 1985a] I. J. Good. Weight of Evidence: A Brief Survey. In J. M. Bernardo, M. H. DeGroot, D.V. Lindley and A. F. M. Smith (eds.), *Bayesian Statistic 2*. New York: North Holland, 249-269, 1985.
- [Good, 1985b] I. J. Good. Sufficiency and Weight of Evidence. *JSCS* 21, 334-336, 1985.
- [Good, 1988] I. J. Good. The Interface Between Statistics and Philosophy of Science. *Statistical Science* 3, 386-397, 1988.
- [Good, 1989a] I. J. Good. Yet Another Argument for the Explication of Weight of Evidence. *JSCS* 31, 58-59, 1989.
- [Good, 1989b] I. J. Good. Weight of Evidence and a Compelling Metaprinciple. *JSCS* 31, 121-123, 1989.
- [Good, 1989c] I. J. Good. The Theorem of Corroboration and Undermining, and Popper's Demarcation Rule. *JSCS* 31, 119-120, 1989.
- [Good, 1991] I. J. Good. Weight of Evidence and the Bayesian Likelihood Ratio. In C. G. G. Aitken and D. A. Stoney (eds.), *The Use of Statistics in Forensic Science*. Boca Raton, Florida: CRC Press, 84-106, 1991.
- [Good, 1997] I. J. Good. Bayes Factors, Weights of Evidence and the Law. Review of J. B. Kadane and D. A. Schum, *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*, New York: Wiley, 1966. *JSPI* 64, 171-191, 1997.
- [Good and McMichael, 1984] I. J. Good and A. F. McMichael. A Pragmatic Modification of Explicativity for The Acceptance of Hypotheses. *Philosophy of Science* 51, 120-127, 1984.
- [Hesse, 1964] M. Hesse. Analogy and Confirmation Theory. *Philosophy of Science* 31, 319-327, 1964.
- [Hildebrand et al., 1976] D. K. Hildebrand, J. D. Laing, and H. Rosenthal. *Prediction Analysis of Cross Classification*. New York: Wiley, 1976.
- [Jeffreys, 1931] H. Jeffreys. *Scientific Inference*. Cambridge: Cambridge University Press, 1931.
- [Jeffreys, 1939] H. Jeffreys. *Theory of Probability*. Oxford: University Press, 1939.
- [Kuipers, 1978] T. A. F. Kuipers. *Studies in Inductive Probability and Rational Expectation*. Reidel: Dordrecht, 1978.
- [Kuipers, 1984a] T. A. F. Kuipers. Two Types of Inductive Analogy by Similarity', *Erkenntnis* 21, 63-87, 1984.
- [Kuipers, 1984b] T. A. F. Kuipers. Inductive Analogy in Carnapian Spirit. In P. Asquith and P. Kitcher (eds.), *PSA 1984, Vol. I*. East Lansing, Michigan.: Philosophy of Science Association, pp. 157-167, 1984.
- [Kuipers, 1988] T. A. F. Kuipers. Inductive Analogy by Similarity and Proximity. In D. H. Helman (ed.), *Analogical Reasoning*. Dordrecht: Kluwer, 299-313, 1988.
- [Levi, 1967] I. Levi. *Gambling with Truth*. New York: Alfred A. Knopf, 1967.
- [Levi, 1980] I. Levi. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*. Cambridge, Mass.: The MIT Press, 1980.
- [Lindley, 1964] D. V. Lindley. The Bayesian Analysis of Contingency Tables. *Annals of Mathematical Statistics* 34, 1622-1643, 1964.
- [Maher, 1993] P. Maher. *Betting on Theories*. Cambridge: Cambridge U. P., 1993.
- [Maher, 2000] P. Maher. Probabilities for Two Properties. *Erkenntnis* 52, 6-91, 2000.
- [Maher, 2001] P. Maher. Probabilities for Multiple Properties: The Models of Hesse and Carnap and Kemeny. *Erkenntnis* 55, 183-216, 2001.
- [Niiniluoto, 1980] I. Niiniluoto. Analogy, Transitivity, and the Confirmation of Theories. In L. J. Cohen and M. Hesse (eds.), *Applications of Inductive Logic*. Oxford: Oxford University Press, pp. 218-234, 1980.
- [Niiniluoto, 1981] I. Niiniluoto. Analogy and Inductive Logic. *Erkenntnis* 16, 1-34, 1981.
- [Niiniluoto, 1988] I. Niiniluoto. Analogy and Similarity in Scientific Reasoning. In D. H. Helman (ed.), *Analogical Reasoning*. Dordrecht: Kluwer, pp. 271-298, 1988.

- [Niiniluoto, 1982a] I. Niiniluoto. What Shall We Do with Verisimilitude?. *Philosophy of Science* 49, 181-197, 1982.
- [Niiniluoto, 1982b] I. Niiniluoto. Truthlikeness for Quantitative Statements. In P.D. Asquith and T. Nickles (eds.), *PSA 1982, Vol. 1*. East Lansing, Michigan.: Philosophy of Science Association, pp. 208-216, 1982.
- [Niiniluoto, 1986] I. Niiniluoto. Truthlikeness and Bayesian Estimation. *Synthese* 67, 321-346, 1986.
- [Niiniluoto, 1987] I. Niiniluoto. *Truthlikeness*. Dordrecht: Reidel, 1987.
- [Niiniluoto, 1998] I. Niiniluoto. Verisimilitude: The Third Period. *BJPS* 49, 11-29, 1998.
- [Popper, 1963] K. R. Popper. *Conjectures and Refutations*. London: Routledge and Kegan Paul, 1963.
- [Romeyn, 2005] J. W. Romeyn. *Bayesian Inductive Logic*. PhD dissertation, University of Groningen, 2005.
- [Romeyn, 2006] J. W. Romeyn. Analogical Predictions For Explicit Similarity. *Erkenntnis* 64, 253-280, 2006.
- [Savage, 1954] L. J. Savage. *The Foundations of Statistics*. New York: Wiley, 1954.
- [Skyrms, 1991] B. Skyrms. Carnapian Inductive Logic for Markov Chains. *Erkenntnis* 35, 439-460, 1991.
- [Skyrms, 1993a] B. Skyrms. Analogy by Similarity in HyperCarnapian Inductive Logic. In J. Earman (ed.), *Philosophical Problems of the Internal and External Worlds. Essays in the Philosophy of Adolf Grünbaum*. Pittsburgh: University of Pittsburgh Press, pp. 273-282, 1993.
- [Skyrms, 1993b] B. Skyrms. Carnapian Inductive Logic for a Value Continuum. In H. Wettstein (ed.), *The Philosophy of Science. Midwest Studies in Philosophy vol.18*, South Bend, Indiana: University of Notre Dame Press, pp. 78-89, 1993.
- [Skyrms, 1996] B. Skyrms. Carnapian Inductive Logic and Bayesian Statistics. *Statistics, Probability and Game, Theory IMS Lecture Notes - Monograph Series Volume 30*, 321-336, 1996.

Likelihood Paradigm

This page intentionally left blank

LIKELIHOOD AND ITS EVIDENTIAL FRAMEWORK

Jeffrey D. Blume

1 INTRODUCTION

Statistics is the discipline responsible for the interpretation of data as scientific evidence. Not surprisingly, there is a broad statistical literature dealing with the interpretation of data as statistical evidence, the foundations of statistical inference, and the various statistical paradigms for measuring statistical inference. However, this literature is surprisingly diverse; the range of viewpoints, opinions and recommendations regarding methods for measuring statistical evidence is as varied as it is vast. This diversity is due, in part, to the complex philosophical nature of the problem. But it is also due to the absence of a generally accepted framework for characterizing and evaluating paradigms that purport to measure statistical evidence.

This paper proffers a general framework that enables the comparison and the evaluation of statistical paradigms claiming to measure the strength of statistical evidence in data. The framework is simple and general, consisting of only three key components (i.e., three key definitions). These essential components may, at first, appear obvious. For example, the first component is nothing more than a definition of the mathematical quantity used to measure the strength of evidence in data.¹ Unfortunately, the first component is always obvious and its (critical) definition is sometimes missing.² Once defined, however, the behavior of competing measures can be detailed and contrasted. As we will see, more than just a measure of evidence is needed to understand and evaluate a paradigm for measuring statistical evidence. The absence of a well defined framework can lead to controversies like those surrounding ad-hoc adjustments to p-values for multiple looks.³ or for multiple comparisons.⁴

¹For example, in classical frequentist inference this measure would be the p-value.

²For an example consider Bayesian inference. Is it the Bayes factor or the posterior probability that measures the strength of evidence in data?

³Reexamining accumulating data during the course of a study

⁴Evaluating several different scientific endpoints (e.g., overall survival, cause-specific survival, safety) in a single study.

1.1 *The three evidential quantities*

Three essential quantities for assessing and interpreting the strength of statistical evidence in data are

1. the measure of the strength of evidence,
2. the probability that a particular study design will generate misleading evidence,⁵
3. the probability that *observed* evidence is misleading.

For brevity I will sometimes denote these evidential quantities as EQ1, EQ2, and EQ3.⁶ The first and the third quantities inform the statistical evaluation of data as scientific evidence. The second quantity informs the data collection process. All three quantities are essential to science and to statistics.⁷

Each evidential quantity represents the answer to a critical scientific question. The first provides an answer to the question, “How strong is the evidence in these data?” The second answers the question, “What is the chance that my study will yield data that are misleading?” The third answers the question, “What is the chance that these observed data are misleading?” The first and third quantities depend on, and pertain to, the observed data; they reflect characteristics of an existing set of data. The second quantity depends on, and pertains to, the study design; it does not inform the interpretation of data because it presupposes that data have not yet been collected.

Each quantity provides unique information regarding the interpretation (EQ1), collection (EQ2) and reliability (EQ3) of statistical evidence. In this paper, a statistical paradigm is said to have a well defined ‘evidential framework’ if that paradigm clearly defines, and distinguishes between, these three quantities. Once established, evidential frameworks are evaluated and contrasted by examining the statistical performance of these quantities in various scenarios.⁸

1.2 *An analogy for the second and third evidential quantities*

The second (EQ2) and third (EQ3) evidential quantities are routinely mistaken for one another. Once a set of data is actually collected, the probability of collecting

⁵Misleading as measured by EQ1. It helps to think of misleading evidence as strong evidence in support of a false hypothesis. This allows for the possibility of weak or inconclusive evidence in support of a false hypothesis, which is typically not considered misleading in scientific applications. I will be more precise in upcoming illustrations.

⁶I assert the importance of these quantities based on my experience and their immediate obviousness. Others have alluded to the same thing. For example, see Royall’s [1997] preface, [Blume, 2002; 2007], and [Strug and Hodge, 2006a; 2006b] as examples.

⁷The natural temporal order in which these quantities are considered during scientific research is: EQ2, EQ1, EQ3.

⁸For example, the rate at which an EQ1 converges to the identification of the correct hypothesis might be a comparison metric between frameworks. The framework with the quicker rate of convergence would probably require less data and therefore be preferable.

some other set of data that turns out to be misleading (i.e., EQ2) is irrelevant. The observed data are either misleading or not, and we know not which. What is of interest is the potential for the observed data to be misleading (i.e., EQ3). After observing data, it makes perfect sense to ask “What is the probability that the data I just collected are misleading?” EQ2 characterizes the chance that the study will yield a misleading result. EQ3 characterizes the chance that the observed result is misleading.

A simple example may help illuminate the distinction. Jena and Jamie each play the lottery.⁹ Jena buys one ticket. Jamie buys ten tickets, but one of the tickets is identical to Jena’s ticket. If Jena wins, so does Jamie. But Jamie can also win with one of her other nine tickets. Therefore, Jamie’s chance of winning is ten times that of Jena’s, but still very small at 10 out of 195,249,054. Here the probability of winning is analogous to EQ2 (e.g., “What is the chance of winning with these tickets?”).¹⁰ Both Jena and Jamie have different chances of winning due to their different game playing strategies.

The next morning Jena and Jamie look in the newspaper for the winning lottery numbers. Unfortunately, the one of the numbers is smudged and unreadable.¹¹ The remaining numbers, however, match Jena’s ticket. And because Jamie also has an identical ticket, both women may have a winning ticket. At this point, Jamie’s and Jena’s chance of winning is identical and is equal to 2.5%.¹² Here the (updated) probability of winning is analogous to EQ3 (e.g., “What is the chance of winning after matching all but one number?”). The fact that Jamie had a ten-fold chance to win the lottery yesterday is completely irrelevant. Her ticket is just as valid as Jena’s, she stands to win just as much money, and she has the exact same chance of doing so. The insight is this: Once data are collected, EQ2 becomes irrelevant. What ‘might have been’ becomes irrelevant. It is EQ3, what ‘might be’, that is the relevant quantity once data are collected.

1.3 *Absence of an evidential framework*

The absence of a well defined evidential framework can lead to irresolvable controversies, such as those surrounding the proper use and interpretation of p-values or those concerning adjustments to p-values for multiple comparisons and multiple looks at data [Blume 2003, Goodman and Royall 1988, Goodman 1998, and Royall 1986, 1997]. I will revisit this later, but it suffices to note that the lack of clarity on which evidential quantity the p-value represents causes considerable confusion.¹³

⁹By lottery I mean Powerball. To win the jackpot, players must correctly match 5 white balls (numbered 1 through 59) and one red ball (numbered 1 to 39) that are drawn randomly without replacement. The order in which the white balls are drawn does not matter. The odds of winning the jackpot are 1 in 195,249,054.

¹⁰Here winning the lottery is akin to collecting misleading evidence and the tickets are akin to the study design.

¹¹I assume that the red ball number is unreadable, but the details are ancillary to this analogy.

¹²Because only the red ball is in question, the chance of winning is now 1 in 39 or approximately 2.5%.

¹³The p-value is often interpreted as each of the three quantities, sometimes at the same time.

Bayesians are not immune to this criticism either, as it remains unclear which of the evidential quantities a posterior probability¹⁴ is intended to represent.

The advantages of a well defined evidential framework will be illustrated by examining the Likelihood paradigm in the context of multiple examinations of data and multiple comparisons. Among the three¹⁵ prominent statistical paradigms for measuring statistical evidence, only the Likelihood paradigm has a well developed evidential framework. In fact, the decoupling of the three evidential quantities allows Likelihood to obey the Likelihood principle¹⁶ and retain good frequentist properties without having to use ad-hoc adjustments or prior probabilities. This is because its EQ1 may adhere to the Likelihood principle without forcing an adjustment to EQ2, which naturally depends on the study design.

Within a well defined evidential framework it is not a contradiction for two studies to yield identical data, and therefore equivalent statistical evidence, when each study initially had a different chance of generating misleading evidence (due to differing study designs). Moreover, we will see that neither set of *observed* data is more likely to be misleading than the other.¹⁷ In terms of evidential quantities, EQ2 may differ between studies, but those studies may generate identical data such that EQ1 and EQ3 are themselves identical (e.g., see the above lottery example). The often noted controversial case of this is when two different studies, each with a different stopping rule, yield the exact same data.¹⁸

2 THE LIKELIHOOD PARADIGM

The Likelihood paradigm is based on the Law of Likelihood, which explains when the data represent statistical evidence for one hypothesis over another. Simply put, the data better support the hypothesis that does a better job of predicting the observed events and the likelihood ratio measures the degree to which one hypothesis is better supported over the other [Hacking, 1965; Edwards, 1971; Royall, 1997]. Likelihood ratios are non-negative. Suggested benchmarks of 8 and 32 signify a transition from a weak to moderate level of evidence and from

¹⁴A posterior probability is the probability of the hypothesis given the observed data. This uses the prior probability of the hypothesis which is set before the data were observed.

¹⁵The three paradigms are Bayesian, Likelihood and frequentist.

¹⁶The likelihood principle states that two studies yielding the same data and using the same probabilistic model must have equivalent measurements of the strength of the evidence in those data. P-values violate the likelihood principle.

¹⁷It may seem obvious to assert that identical sets of data have exactly the same propensity to be misleading. The lottery example would certainly support this point of view. However just the opposite is the current dogma in statistics. It is commonly believed that two identical sets of data generally have different propensities to be misleading when the study designs from which they came differ. This dogma is a consequence of the failure to distinguish between EQ2 and EQ3.

¹⁸The controversy arises because the classical approach does not yield the same strength of evidence in each study (despite the fact that they generated identical datasets). This is because the p-value is adjusted differently based on the study design (i.e. stopping rule). The validity of this approach has been debated for decades.

a moderate to strong level evidence, respectively. Introductory material on this approach and extensions are readily available (e.g., [Blume, 2002; 2005; 2007; 2008; Goodman and Royall, 1988; Strug and Hodge, 2006a; 2006b; Royall and Tsou, 2003; Tsou and Royall, 1995; Van der Tweel, 2005]).

The Law of Likelihood provides a measure of the strength of evidence between two hypotheses.¹⁹ In contrast, the Likelihood principle sets forth the conditions under which two experiments yield equivalent statistical evidence for two hypotheses of interest. This condition is met when their likelihood functions are proportional [Birnbaum, 1962; Barnard, 1949]. Acceptance of the law implies acceptance of the principle: if all the likelihood ratios between the two experiments are identical, then their likelihood functions must be proportional [Royall, 2000].

2.1 Illustration of the three evidential quantities

The most accessible illustration comes from Royall's [1997] diagnostic test example. Suppose we use a diagnostic test to generate evidence about the disease state of an individual. The properties of this test are given in Table 1. Its sensitivity is $0.94 = P(T+|D+)$ and specificity is $0.98 = P(T-|D-)$.²⁰ According to the law of likelihood a positive test result would be evidence supporting H_+ (disease is present) over H_- (disease is absent) because the likelihood ratio (LR) would be $47 [= 0.94/0.02 = P(T+|D+)/P(T+|D-)]$. Likewise, a negative result would represent evidence supporting H_- over H_+ because the likelihood ratio would be $16.3 [= 0.98/0.06 = P(T-|D-)/P(T-|D+)]$. The likelihood ratio is EQ1; it measures the degree to which the data support one hypothesis over another.

Table 1. Properties of the diagnostic test

		Test Result (T)	
		Positive (+)	Negative (-)
Disease Status (D)	Yes (+)	0.94	0.06
	No (-)	0.02	0.98

However, this test may generate evidence that is misleading. A positive result is correctly interpreted as evidence for H_+ over H_- , but positive results can occur when the disease is absent. This happens only 2% of the time in people without the disease, but when it does happen the test is said to have generated misleading evidence.²¹ Of course, the test can also generate misleading negative

¹⁹This representation of evidence is essentially comparative between simple hypotheses. Simple hypothesis specify a single probability distribution. Composite hypothesis specify a family of distributions.

²⁰The vertical bar ' $|$ ' is read as 'given'. This is a conditioning argument, as events after the bar are considered fixed.

²¹Statistical tests (i.e., tests that are not deterministic) must be able to generate misleading

results and this happens 6% of the time in people who have the disease. These two probabilities are second evidential quantities (EQ2); they are the probabilities of observing misleading evidence under this study design. They are analogous to the error rates of hypothesis testing and they play an important role in defining the quality of the diagnostic test and the data collection process. A good diagnostic test maximizes sensitivity and specificity, which here is the same as minimizing the second evidential quantities (i.e., minimizing potential to observe misleading positive and negative tests).

Upon observing a certain test result, the strength of the evidence will be clear from the likelihood ratio (e.g., a positive tests yields a LR of 47). We will never know if the observed test result is misleading or not. However, it is sometimes possible to determine the propensity for the observed test result to be misleading.²² For example, an observed positive result is misleading *if and only if* the subject does not have the disease. The probability that the subject does not have the disease is $P(D-|T+)$. By Bayes theorem, this is $P(D-|T+) = 1/[1 + 47\pi_+/\pi_-]$, where $\pi_+ = P(H_+)$ and $\pi_- = P(H_-) = 1 - \pi_+$ are the prior probabilities of the individual's disease state. By identical reasoning, an observed negative result is misleading *if and only if* the subject has the disease. Here that probability is $P(D+|T-) = 1/[1 + 16.3\pi_-/\pi_+]$. So we see that the probability that the observed evidence is misleading, EQ3, is nothing more than the posterior probability.

The obvious (non-computational) difficulty with EQ3 is the specification of the prior probabilities, which are inherently subjective.²³ However diagnostic tests are a rare example in which there exists a broad consensus regarding the prior probabilities. Typically, the disease prevalence is used to set the prior when it is reasonable to assume that the subject was selected at random from the population. Suppose the disease prevalence is $\pi_+ = 0.015$. Then $P(D-|T+) = 0.583$ and $P(D+|T-) = 0.0009$. This means that positive results are not as reliable as negative results. In fact, observed positive results are misleading more than half the time *in this population*. This does not mean we are wrong when we interpret a positive result as evidence that the disease is present.²⁴ It just means that *in this population* the strength of evidence provided by a positive result is not strong enough to outweigh our (very strong) prior knowledge about the presence or absence of disease. Before the test is given, the probability that the individual is diseased is only 1.5%. This probability increases to 41.7% after observing a positive test result.²⁵ However, it remains more likely (58.3%) that the individual does not have the disease.

Because EQ3 depends on prior probabilities, it remains context based. This is

evidence. If not, then only a single observation would be needed to correctly identify the true hypothesis.

²²We need only be willing to make certain assumption about the prior probability of the hypotheses or disease state.

²³Admittedly there are varying degrees of subjectivity, as priors come from many different sources.

²⁴Surely a positive result from this test is not evidence the disease is absent!

²⁵This large increase is due to having a large likelihood ratio — strong evidence — of 47.

why strong evidence in one study may not be strong enough evidence to provide the same sense of reliability in another study.²⁶ From the expressions for EQ3, we see that larger likelihood ratios are more reliable in the sense that they are less likely to be misleading.²⁷ EQ1 and EQ3 have an inverse relationship; as the evidence gets stronger, its potential to be misleading decreases. Lastly, it is worth reiterating that the likelihood principle indicates that EQ2 is irrelevant once the data have been collected. At that point, EQ1 and EQ3 are the only quantities of interest.

2.2 Hypothesis testing and significance testing

The cornerstone of mainstream statistical methods is an ad-hoc union of hypothesis testing and significance testing [Blume 2003, Goodman 1998]. Experiments are designed under the hypothesis testing framework and analyzed under a significance testing one. Because this union was unplanned, the tail area probability²⁸ and what it represents is a source of continuing confusion. In the hypothesis testing framework the tail area probability represents EQ2 (i.e., the Type I error), but in the significance testing framework it represents EQ1 (i.e., the p-value). Moreover, there is no EQ1 in the hypothesis testing framework and there is no EQ2 in the significance testing framework. So it seems reasonable (and even quite natural) to many non-statisticians to merge the two approaches.²⁹ Scientists will always look for each of the three evidential quantities; they represent important concepts and they have distinct roles in the scientific process.

To impress the point, let's consider a hypothesis test and a significance test in the diagnostic testing example from the previous section. A hypothesis test sets the null hypothesis as H_0 : disease absent and the alternative as H_1 : disease present. We 'reject the null hypothesis' when a positive result is observed and 'fail to reject H_0 ' when a negative result is observed. The test would have a Type I error rate of 2% and a Type II error rate of 6%.³⁰ According to various decision theoretic standards this is a good test.

But problems arise if we try to interpret the results of the test as statistical evidence. First, failure to reject the null hypothesis cannot be taken as evidence for the null hypothesis. In fact, a negative result (i.e., failure to reject the null hypothesis) can never be interpreted as evidence that the disease is absent³¹; instead it is always interpreted as statistically inconclusive. Second, there is no

²⁶Context affects the prior probabilities that are used in the calculation of EQ3.

²⁷Here the prior probabilities are considered fixed.

²⁸Tail area probability is a probabilistic term representing the core calculation in a p-value and in a Type I error.

²⁹Fisher was well aware of this and warned: "In fact, as a matter of principle, the infrequency with which, in particular circumstances, decisive evidence is obtained, should not be confused with the force, or cogency, of such evidence" [Fisher, 1959, p93].

³⁰A Type I error rate is the probability of rejecting a true null hypothesis and a Type II error rate is the probability of failing to reject a false null hypothesis. The Type II error rate is calculated under a pre-specified simple alternative hypothesis.

³¹Absence of evidence is not evidence of absence.

strength of evidence to report (i.e., no EQ1). The best we can do is report our decision to ‘reject’ or ‘fail to reject’, along with the error rates of our decision rule.

But this is largely unsatisfactory from a scientific viewpoint, especially if the desire is to report the strength of evidence for the hypotheses of interest (e.g., weak, moderate or strong). This is why significance testing is employed at the end of the study. Significance testing involves calculating and reporting the p -value, as a measure of the strength of evidence against the null hypothesis. In this case, the p -value is 2% which would be considered strong evidence against the null because it is less than the common benchmark of 5%. Here too, it is not possible to generate evidence in favor of the null hypothesis; large p -values are interpreted as being inconclusive. For example, a negative test result, which yields a p -value of one³², cannot be interpreted as evidence that the disease is absent. So we see from this example that while significance testing provides an EQ1 (i.e., the p -value), it does not provide an EQ2 or EQ3. EQ2, which would be calculated before the study is conducted, might be the probability of observing a small p -value when the null hypothesis is true (but EQ2 is not identically the p -value itself).³³

Consider also a Bayesian approach that focuses on the posterior probabilities. In this case, after observing a positive result, the posterior probability of disease, $P(D+|T+)$, is only 0.417. It remains unclear how that positive result ought to be interpreted if the posterior probability is used as the EQ1. This is because even after a positive test result the subject is more likely to be disease free (i.e., 41.7% < 50%). So should the positive result be considered evidence that the disease is absent? If yes, then this test can never generate statistical evidence that the disease is present (because the posterior probability of disease after a negative test result, $P(D+|T-) = 0.0009$ is very small). If no, then by what scale and context are we to interpret the posterior probability?³⁴ EQ1 needs to be explicitly defined and we need to be told how to use it. Also, does it make sense to define EQ2 in a way that is dependent on the prior probabilities?³⁵ So here too, the lack of an evidential framework leaves the approach so opened ended that its utility is not clear.

2.3 Background and notation

In order to consider more complex situations it is helpful to establish some notation, if only to help the reader distinguish between the three evidential quantities.

³²The p -value is the probability of observing the data or data more extreme. Thus, the p -value is the probability of observing a negative or positive test result when the null hypothesis is true. This probability is one.

³³The corresponding EQ3 would be the probability that the disease is absent given $p = 0.02$.

³⁴If the answer is “They should be interpreted in relation to how much they have changed from the prior probability” then this is nothing more than the likelihood paradigm. See [Royall, 1997].

³⁵One example is “What is the probability that the posterior will be as large as 0.95 under various hypotheses?” The subtle caveat here is that we are providing long run frequency properties that are dependent on the priors and this may or may not rely on the ‘validity’ of the specification of the priors.

Every effort is made to skip unnecessary mathematical complexities in the examples that follow.

Suppose observations $X_1, \dots, X_n$ are independent and identically distributed according to some density $f(X_i; \theta)$. For a fixed sample size of n , let the likelihood function be $L_n(\theta) \propto \prod_i f(X_i; \theta)$. There are two hypotheses of interest: the null, $H_0 : \theta = \theta_0$, and the alternative, $H_1 : \theta = \theta_1$. According to the Law of Likelihood, the strength of the evidence for H_1 over H_0 is measured by $LR = L_n(\theta_1)/L_n(\theta_0)$. An observed likelihood ratio will fall into one of three regions: $LR \in [0, 1/k]$ indicating strong evidence for H_0 over H_1 , $LR \in (1/k, k)$ indicating weak evidence in either direction, and $LR \in [k, \infty)$ indicating strong evidence for H_1 over H_0 for some $k \geq 1$. By convention, we use $k = 8$ or 32 to denote the transition from weak to moderate and from moderate to strong evidence.³⁶

2.4 Misleading evidence

Misleading evidence, by definition, is strong evidence in favor of the incorrect hypothesis (e.g., observing $LR = 47$ when H_0 is true or $LR = 1/47$ when H_1 is true).³⁷ We never know if observed evidence is misleading or not, but we can know the probabilities that a particular study will generate misleading evidence. They are represented, in general, by $\text{mis}_0 = P(LR \geq k|H_0)$ and $\text{mis}_1 = P(LR \leq 1/k|H_1)$ for some $k > 1$, typically 8 or 32. Here mis_0 is the probability of observing misleading evidence under the null hypothesis and mis_1 is the probability of observing misleading evidence under the alternative hypothesis. Both are EQ2s. Misleading evidence is seldom observed in the likelihood paradigm. Under quite general conditions, mis_0 and mis_1 are each bounded by $1/k$, the so-called universal bound, and their average is bounded by $1/(k+1)$ [Royall, 1997; 2000].

The universal bound is a crude device for controlling the probability of observing misleading evidence. The actual probability is often much less. As the sample size grows, both probabilities of observing misleading evidence (i.e., mis_0 and mis_1) converge to zero.³⁸ Under some mild parametric conditions and in moderate to large samples, the probability of observing misleading evidence is approximately $\Phi[-\ln k/c - c/2](\approx \text{mis}_0)$, where $c = |\theta_1 - \theta_0|/(nI(\theta_0))^{-1/2}$, $\Phi[\cdot]$ is the standard normal cumulative distribution function, and $I(\theta_0)$ is the Fisher information³⁹ [Royall, 2000]. From this we get an asymptotic maximum⁴⁰ probability of observ-

³⁶A likelihood ratio of 4 would be weak evidence in favor of the H_1 and a likelihood ratio (LR) of $1/47$ would be strong evidence in favor of H_0 . The weak evidence region — LRs between $1/8$ and 8 — is important to future discussions.

³⁷Weak evidence — LRs between $1/k$ and k or typically $1/8$ and 8 — are not considered to be misleading by definition. This is because weak evidence is inconclusive and has little force. In practice, Likelihoodists tend to ignore the directionality of weak evidence, characterizing it as inconclusive. See Royall [1997] for more on this topic.

³⁸This is an interesting and useful property for EQ2. Contrast this with hypothesis testing in which one of its EQ2s (the Type I error) remains constant regardless of the sample size. Thus errors are made even with an infinite sample size.

³⁹Under a normal model with known variance this approximation is exact.

⁴⁰The maximum is over all alternatives

ing misleading evidence of only $\Phi[-\sqrt{2 \ln k}] (\leq 1/[2k\sqrt{(\pi \ln k)}])$, which is typically much less than the universal bound, $1/k$. Understanding the behavior of EQ2 is important in evaluating study designs. Technical extensions are recently available [Royall, 2000; 2003; Blume, 2002; 2007].

EQ2 and EQ3 have different mathematical representations. In general, the third evidential quantities are $P(H_0|\text{data}) = [1 + k\pi_1/\pi_0]^{-1}$ and $P(H_1|\text{data}) = [1 + \pi_0/k\pi_1]^{-1}$ where $\pi_0 = P(H_0)$, $\pi_1 = 1 - \pi_0$, and $k > 1$ is the observed likelihood ratio in support of H_1 over H_0 . So, as the sample size increases, these probabilities are driven to zero or one by the likelihood ratio, which itself converges to 0 or ∞ in support of the correct (true) hypothesis. Thus, for a fixed prior probability, larger observed likelihood ratios are more reliable and the degree of reliability can always be improved by increasing the sample size.

3 REEXAMINATION OF ACCUMULATING DATA ('MULTIPLE LOOKS')

Scenarios involving repeated examination of accumulating data (i.e., multiple looks) provide an excellent opportunity to illustrate the value of having an evidential framework. A common misconception is that there is no 'penalty' for multiple looks at data in the Likelihood paradigm.⁴¹ It is true that likelihood ratios (EQ1) are not modified by the number of looks at the data. But it is also true that there is a 'penalty'. The caveat is that the penalty applies to EQ2 and not to EQ1 as it would with p -values. The probability of observing misleading evidence, EQ2, is indeed affected by the number of examinations of the data. Every additional examination of the data increase EQ2. The stopping rule is indeed relevant for determining how often a study design will yield misleading evidence. But once data are collected, the stopping rule is irrelevant; it does not affect the measure of evidence nor the potential for observed evidence to be misleading (EQ3).

What many classical statisticians seem to find unsettling (or at least unfamiliar) about the likelihood paradigm is the separation of EQ1 from EQ2. EQ1 is the likelihood ratio; EQ2 is the probability that the likelihood ratio will be misleading. This is unfamiliar because the p -value is essentially used as both EQ1 and EQ2 indiscriminately. For example, consider two studies, one of which examines the data as it accumulates. Both studies happen to generate the same data set. Despite having identical data, they will report different p -values and claim to have different amounts of evidence. This is because the calculation of a p -value depends on the study design; it is calculated as if it were an EQ2, but interpreted as if it were an EQ1.

The underlying problem is that just one number, the p -value, has to function as both EQ1 and EQ2 (and maybe even the EQ3). This problem is avoided when a well defined evidential framework exists. For example, under a likelihood approach these two studies would report the same likelihood ratio and therefore the same

⁴¹ P -values are penalized (i.e., discounted/inflated) for each look at accumulating data and for each planned look at the data that has not yet been executed. Likelihood ratios are not.

observed strength of evidence in the data. What is different is their study design; one study had a much larger chance of generating misleading evidence before any data were collected. But this probability is irrelevant now that data have been observed. What might have been is irrelevant. All that is relevant is what was observed (EQ1) and the propensity for the observed data to be misleading (EQ3).

Let's consider a numerical example to fix ideas. Suppose I design a study to collect evidence about the average change in systolic blood pressure due to a new medication. Assume that these changes are normally distributed with a standard deviation of $\sigma = 16$ mm Hg. My null hypothesis is that there is no change in mean blood pressure ($H_0 : \mu = 0$) and my alternative hypothesis is that the mean change is 8 mm Hg or one-half of the population standard deviation ($H_1 : \mu = 8$ or 0.5σ).⁴² With this information we can calculate how often certain study designs will generate misleading evidence in favor of the alternative hypothesis.

First, consider a fixed sample size design. The study enrolls 25 people, collecting their before and after blood pressure measurements, and the data are examined after all observations have been collected. The probability of observing misleading evidence of strength 8 or more in this design is only 0.0187, which decreases to 0.0042 for (misleading) likelihood ratios of 32 or more.⁴³

In contrast, consider instead a 'truncated sequential' design, where the data are examined after every observation is collected and the study is stopped as soon as strong evidence for the alternative hypothesis is observed or when data on 25 people is collected, whichever comes first. When the null hypothesis is true, the probability that this design generates misleading evidence of strength 8 or more is 0.0717, which decreases to 0.0123 for (misleading) likelihood ratios of 32 or more.⁴⁴

Lastly, suppose it was possible to continue collecting data forever, so we could plan to collect observations until we observed evidence in support of the alternative hypothesis. This is a very biased design. Data are collected until we find strong evidence in support of the alternative hypothesis while strong evidence in favor of the null hypothesis is ignored. Thus, this design provides the best chance of collecting misleading evidence, because even if the null hypothesis is true we continue sampling until we find misleading evidence. But even in this biased design the probability of observing misleading evidence is only 0.0934 (for evidence of strength 8 or more) and 0.0233 (for evidence of strength 32 or more).⁴⁵ Notice that there is a very large probability that misleading evidence will never be generated. This is an important scientific safeguard; an investigator searching for evidence to support his pet hypothesis over the correct hypothesis is likely never to find such

⁴² σ is the standard deviation, μ is the mean, and H_0 and H_1 are the null and alternative hypotheses. For convenience I expressed the alternative hypothesis in units of standard deviations. This makes the upcoming probability calculations simpler and adds to the generality of this example in the sense that any dependence on the stated value of the standard deviation is removed.

⁴³This probability is simply $\Phi[-\ln k/c - c/2]$, noted in an earlier section, with $c = 0.5$ and $k = 8$ or 32.

⁴⁴See [Blume, 2008] for the formula for this calculation.

⁴⁵This probability is $\exp(-0.583 * c)/k$ with $c = 0.5$ and $k = 8$ or 32 [Blume, 2002; 2007].

evidence when using the likelihood paradigm.

It should be clear that even within the Likelihood paradigm a price is exacted for each examination of the data. That price is the inflation in the probability of observing misleading evidence, EQ2. One important characteristic of this probability is that the amount by which it increases converges to zero as the sample size grows [Blume, 2008]. Thus the probability of generating misleading evidence remains bounded even with an infinite number of examinations of the data. Simply put, the chance of observing misleading evidence at a single point in time is less than the chance of observing misleading evidence at any point in time, although the latter remains bounded. More precisely we can write [Robbins. 1970]:⁴⁶

$$P(LR_n \geq k | H_0) < P(LR_n \geq k \text{ for any } n = 1, 2, \dots | H_0) \leq 1/k$$

As we have just seen, the probability of observing misleading evidence increases as the number of examinations increases. The chance of observing misleading evidence after collecting 25 participants is 0.0187, which increased to 0.0717 when the data were continuously monitored up to the 25 subjects, which increased to 0.0934 when the data were continually monitored until misleading evidence was obtained (i.e., there was no limit on the sample size). While the behavior of EQ2 is interesting and clearly depends on the study design, this behavior is independent of the observed EQ1 which does not depend on the study design.

4 MEASURING EVIDENCE ABOUT SEVERAL ENDPOINTS SIMULTANEOUSLY ('MULTIPLE COMPARISONS')

An evidential framework is also helpful when dealing with multiple comparisons for reasons similar to that just discussed in the context of reexamination of accumulating data. Consider a study in which 4 different measurements or endpoints are collected from each subject before and after a certain medical intervention. The goal of the study is to see if any of these endpoints change after the medical intervention. Our study will collect data from 6 participants.⁴⁷ For simplicity and without loss of generality, we will also assume that the 4 endpoints are independent and normally distributed with a known variance.⁴⁸

The null hypothesis is that of no change, $H_0 : \mu = 0$. Because our sample size is small, we set our alternative to be a change of one standard deviation, $H_1 : \mu = \mu_1 = \sigma$ (i.e., $(\mu_1 - \mu_0)/\sigma = 1$).⁴⁹ The likelihood probabilities of observing misleading evidence are $\text{mis}_0 = \Phi[-\ln k / \sqrt{n} - \sqrt{n}/2]$ and $\text{mis}_1 = \Phi[-\ln k / \sqrt{n} -$

⁴⁶A general approximation is $P(LR_n = k \text{ for any } n = 1, 2, \dots | H_0) \equiv \exp(-0.583 * c)/k$ [Blume 2002, 2007].

⁴⁷I picked this for ease of calculation. Small sample sizes like this are, unfortunately, not unusual in genomic or in proteomic studies, although the number of genes or proteins being examined is often much larger than 4.

⁴⁸These assumptions allow us to focus on the conceptual underpinnings of the problem instead of its mathematics. No generality is lost.

⁴⁹In the example from the previous section the alternative was set to $1/2$.

$\sqrt{n/2}$ with $n = 6$. When $k = 8$, these probabilities are both only 2%. The family-wise⁵⁰ probability — the probability of observing misleading evidence on at least one endpoint — is 7.8% ($=1-(1-0.02)^4$) under both the null and alternative hypotheses. When $k = 20$, the probability of observing misleading evidence on a single endpoint is 0.72% and the family-wise probability is 2.85%. We can see from this example that the sample size plays an important role. With $n=12$ observations and $k=8$, the probabilities of observing misleading evidence are only 1%, and the family-wise error rate drops to 3.9% from 7.8%.

Note that in the Likelihood paradigm we do not choose a cutoff for ‘ k ’ to denote which likelihood ratios are significant or not. Likelihood ratios are descriptive. If we want to control the probability of observing misleading evidence, then this is done completely through the sample size. Also, it should be clear that the probability of observing misleading evidence on at least one endpoint will always increase as the number of endpoints increase. However, this probability is also controlled through the sample size and can be driven to zero with a large enough sample.⁵¹

Likelihood is different from hypothesis testing in that there are effectively three evidence regions: strong evidence for the null over the alternative, strong evidence for the alternative over the null, and weak evidence supporting either hypothesis over the other.⁵² In this example, there is a 33% chance of observing weak evidence.⁵³ Strong evidence, when it is observed, tends to be reliable but one-third of the time we will be left with weak inconclusive evidence. In contrast, hypothesis testing has only two zones: ‘reject H_0 ’ and ‘fail to reject H_0 ’.⁵⁴

For comparison purposes, it is instructive to consider what happens when we handicap the likelihood approach by removing the weak evidence zone. We do this by using $k = 1$, so that any amount of evidence (not just strong evidence) is potentially misleading. When $k = 1$, the probabilities are symmetrical and $\text{mis}_0 = \text{mis}_1 = \Phi[-\sqrt{n/2}]$. The two corresponding design probabilities (e.g., the probability of observing any evidence in support of the alternative when the null is true) increase from 2% to 11% and the family-wise rates increase from 7.8% to 37%. This happens because weak evidence is naturally less reliable.⁵⁵

Let’s consider how this situation would be handled with a hypothesis test. The Type I and II errors of a hypothesis test are α and $\beta = \Phi[Z_{1-\alpha} - \sqrt{n}]$. With a one-sided Type I error rate of 5% and 6 observations, we get a Type II error rate

⁵⁰The family-wise error rate is defined as the probability of observing misleading evidence on at least one endpoint.

⁵¹This is not so with hypothesis testing, as we will see later.

⁵²This inferential refinement is important in many respects and it should not be overlooked; it is the weak evidence that is most often misleading and should not be taken too seriously.

⁵³It is the same under either hypothesis.

⁵⁴Allowing for three zones is a major plus in my opinion. It completely resolves the asymmetry in hypothesis testing and significance testing that prevents them from generating evidence in support of the null hypothesis.

⁵⁵Judging from this, it seems best to call weak evidence inconclusive rather than over-interpreting it in practice. However, the upcoming comparison to hypothesis testing is insightful and worth pondering for academic purposes.

of 21%. But when there are 4 endpoints with this identical structure, a Bonferroni adjustment is required to control the overall Type I error.⁵⁶ This adjustment results in a Type I error rate of 1.25% and a Type II error rate of 42% for each endpoint. The family-wise Type I error rate (i.e., the probability of making at least one Type I error) is now controlled at 5% ($=1-0.9875^4$), but the family-wise Type II error rate balloons to 89%.

Remember that the comparative family-wise rates for likelihood were both 37%. Certainly, 37% is far from 5%, but it is even further from 89%. For a small increase in one rate there is a large drop in the other, and this occurs without the complexity of post-hoc adjustments to the p -value. For example, the adjusted error rates of hypothesis testing are now dependent on how many other endpoints are being considered in the study. Thus, changing the number of endpoints in the study affects the accuracy (i.e., EQ2) with which you can test other endpoints. In contrast, the likelihood paradigm leaves the chance of observing misleading evidence (EQ2) on a single endpoint the same regardless of how many other endpoints are being considered. A price is still paid for multiple comparisons (the family-wise probabilities increase), but each endpoint remains consistent within itself because there is a clear evidential framework.

One way to balance the tradeoff between the different error rates is to use a metric like the average rate. The average error rate even has a nice interpretation — as the probability of making either error — if we are willing to assume that the null and alternative hypotheses are equally likely.⁵⁷ With Likelihood, the average rate increases from 2% to 11% when we eliminate the weak evidence region. This yields a family-wise average rate of 37%, which is the probability of making at least one error, in any direction, over all the endpoints. In comparison, the hypothesis test has an average error rate of 13% for a single endpoint, which increases to 22% when that endpoint is properly adjusted for the multiple comparisons, and this yields an average family-wise error rate of 62% ($0.62=1-(1-0.22)^4$).

This simple example is a good illustration of the general state of nature; there is no reason to assume that a lack of adjustment necessitates an uncontrollable or outrageous increase in observing misleading evidence. In fact, the worst case average rate at which likelihood ratios are misleading is 40% less than the same rate of a hypothesis test that adjusts for the multiple comparisons. Perhaps more importantly, the likelihood rates can always be driven to zero by increasing the sample size (unlike the Type I error), so a large enough sample size virtually guarantees there will not be any misleading evidence, regardless of the number of comparisons.

Figure 1, displays the family-wise average error rate for 1 and 4 endpoints under both likelihood and hypothesis testing (with multiple endpoints and a Bonferroni

⁵⁶Adjustments other than Bonferroni are available, but this is the most common. The type of adjustment is irrelevant to the thrust of this example in any case. All adjustments simply trade Type II errors for Type I errors.

⁵⁷These two types of errors can be weighted unequally in the likelihood paradigm when it makes sense to do so. But this is beyond the scope of this paper.

correction). Notice that for every sample size, the average error rate is smaller under Likelihood (when $\alpha = \beta$ in hypothesis testing the average error rates are equal). Notice also that the Likelihood average error rate (solid lines) converges to zero, instead of some fixed threshold, and hence can be driven as low as desired by increasing the sample size.

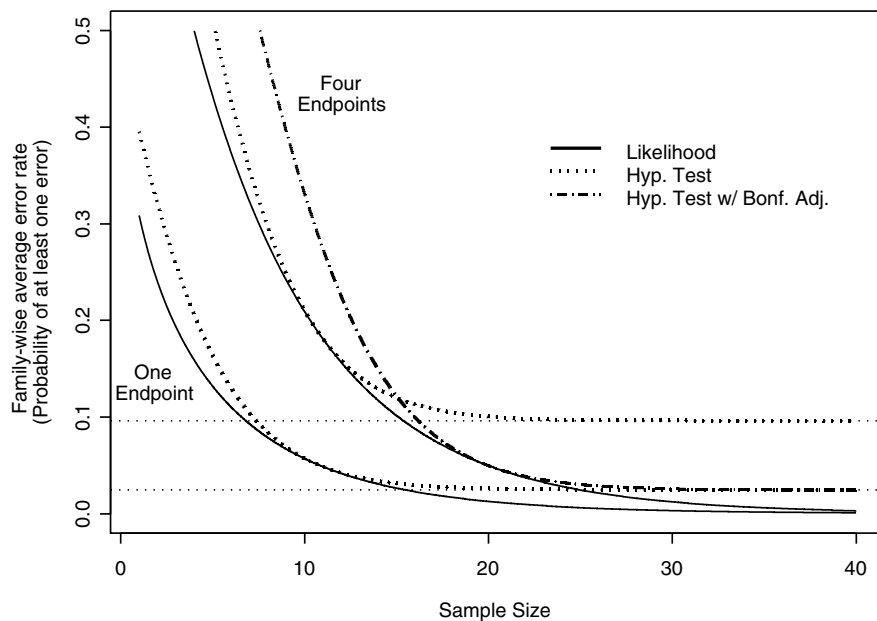


Figure 1. Average ‘error’ rates and family-wise average ‘error’ rates

Figure 2, displays the individual rates themselves. The single solid line represents both likelihood rates. Unlike hypothesis testing, how often misleading evidence is observed on a single endpoint does not depend on the total number of endpoints. For example, consider Fig. 2 at 11 observations, where, for a single endpoint, each of the design probabilities is 5% (all dashed and dotted lines cross at 5% with 11 observations). However, the hypothesis testing rates change when the three other endpoints are considered (solid and dashed line). What happened is this: the Type I error rate for each endpoint decreased to 1.25% and Type II error rate increased to 14%. As a result, the new family-wise average error rate is 27% (family-wise Type II rate is 45% and family-wise Type I rate is 5%) instead of 19% ($=1-0.95^4$) for likelihood, where the individual probabilities remain at 5% regardless of the number of endpoints.

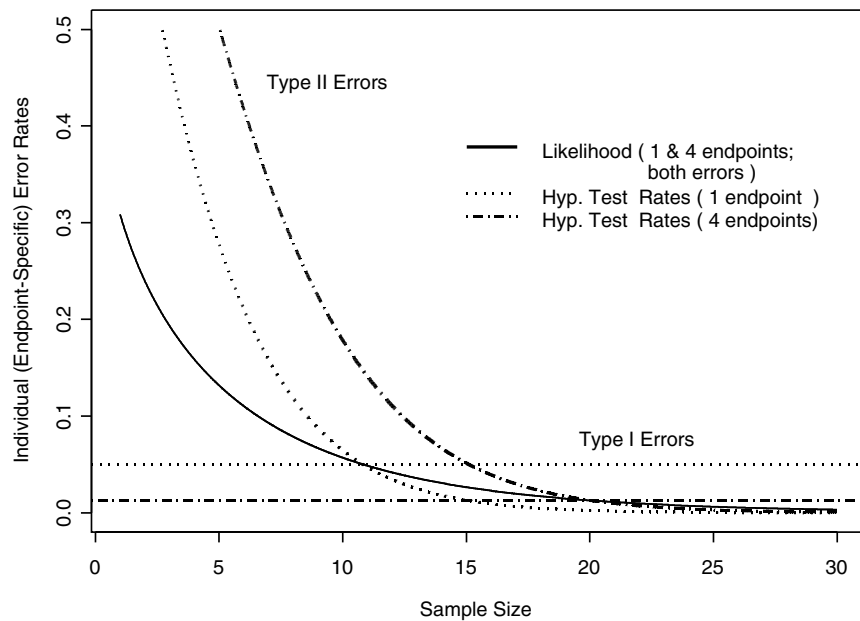


Figure 2. Type I and II error rates and probabilities of observing any evidence in the wrong direction (i.e., $k=1$) for Likelihood

Thus, even in the Likelihood paradigm there is a traditional frequentist price exacted for having multiple endpoints in a single study: The probability of observing misleading evidence on at least one endpoint increases with each additional endpoint. The important difference is that this price is not exacted on EQ1, so that the evaluation of the evidence from any individual endpoint does not depend on how many other endpoints are under consideration.⁵⁸ Moreover, the probability of observing misleading evidence on any single endpoint (EQ2) does not change with the number of endpoints (unlike hypothesis testing), thus conferring some internal consistency.

4.1 *The potential for observed data to be misleading*

The focus to this point has been on the probability of generating misleading evidence (EQ2) and how that varies with the addition of endpoints to the study

⁵⁸The only minor exception to this is when the underlying statistical model allows endpoints to be correlated, but this not the dependence that we are trying to avoid.

in question. While EQ2 is an important piece of the puzzle, it is irrelevant once observations are collected. At the end of a study, EQ2 is often confused with the probability that the *observed* evidence is misleading (EQ3). Once we collect data, the design probability is irrelevant — the data are either misleading or not and we do not know which. So what we really want to know is how likely it is that *the observed data are* misleading.

Unlike EQ2, the probability that the observed evidence is misleading, EQ3, does not depend on the study design. An example is helpful to fix ideas and I'll continue with the multiple comparisons example from the previous section. In that example, we planned to collect six observations under one of two different designs: Study A has just a single endpoint, while Study B has three additional endpoints (4 total). Table 2 displays their design probabilities which have already been discussed at length.

Table 2. Two study designs and their performance characteristics. The interior of table is expressed as percentages. *based on the average error rate.

		One endpoint (Study A)			Four endpoints (Study B)		
		$\alpha(\mathbf{k})$	$\beta(\mathbf{k})$	Overall*	$\alpha(\mathbf{k})$	$\beta(\mathbf{k})$	Overall*
Likelihood with $k = 8$	Primary endpoint	2	2	2	2	2	2
	Any endpoint	-	-	-	7.8	7.8	7.8
Likelihood with $k = 1$	Primary endpoint	11	11	11	11	11	11
	Any endpoint	-	-	-	37	37	37
Hypothesis Testing	Primary endpoint	5	21	13	1.25	42	22
	Any endpoint	-	-	-	5	89	62

Now suppose the observed data yield a mean change on the first endpoint that was 1.01 standard deviations from zero. This evidence favors the alternative hypothesis by a factor of 21.34 ($= L(\mu_1)/L(\mu_0)$) regardless of which study generated the data (EQ1 does not vary by study). But the correct p -value is different depending on the design: $p = 0.013$ under Study A and $p = 0.052$ ($=1-(1-0.013)^4$) under Study B.⁵⁹ This inevitably raises the following question: Do these data

⁵⁹Adjustment for p -values according to [Wright, 1992].

represent stronger evidence when they come from Study A? Or, when these data come from Study A, are they not “less likely to be misleading”, “more reliable” in some sense, or don’t they warrant more “confidence”?

The answer to both of these questions is a resounding no because EQ3 does not vary by study. These data are equivalent as evidence about the unknown mean regardless of the study design; they are no more likely to be misleading under one design or the other. Here is why: these data are misleading if and only if the null hypothesis is true. And here $(H_0|\text{data})$ is the same regardless of the design. An application of Bayes theorem yields $P(H_0|\text{data}) = 1/[1 + 21.34] = 0.045$, which does not differ by study design.⁶⁰ There is only a 4.5% chance that these data are misleading and this probability does not change if there were one, four or 2 million endpoints. Of course, changing the prior will change the degree to which we think these observed data are misleading. But EQ3 will not vary by study design unless the priors do. Thus we can use EQ3 to help assess the reliability of the observed evidence for a given prior, or as a sensitivity analysis across several priors.

5 COMMENTS

The Likelihood approach is used to measure the strength of evidence in observed data. It avoids the pitfalls of other statistical paradigms because it has a well defined evidential framework. The likelihood paradigm is often viewed as the common ground between the frequentists and Bayesians when it comes to measuring the strength of statistical evidence. This is because the likelihood function is often the only thing a frequentist and Bayesian can agree upon. Another reason is that likelihood ratios retain the desirable properties from both paradigms (irrelevance of sample space, good performance probabilities) while shedding the undesirable ones (dependence on prior distributions, ad-hoc adjustments to control error probabilities).

BIBLIOGRAPHY

- [Barnard, 1949] G. A. Barnard. Statistical Inference. Journal of the Royal Statistical Society, Series B, 11: 115-149, 1949.
- [Birnbaum, 1962] A. Birnbaum. On the foundations of statistical inference (with discussion). Journal of the American Statistical Association 53: 259-326, 1962.
- [Blume, 2002] J. D. Blume. Likelihood methods for measuring statistical evidence. Stat Med. Sep 15, 21(17): 2563-99, 2002.
- [Blume and Peipert, 2003] J. D. Blume and J. F. Peipert. What your statistician never told you about p -values. J AM Assoc Gynecol Laparosc 10(4):439-444, 2003.
- [Blume, 2005] J. D. Blume. How to choose a working model for measuring the statistical evidence about a regression parameter. International statistical review, 73(2): 351-363, 2005.
- [Blume *et al.*, 2007] J. D. Blume, L. Su, L. Acosta, R. M. Olveda, and S. T. McGarvey. Statistical evidence for GLM regression parameters: a robust likelihood approach Statistics in Medicine 26(15) 2919-36, 2007.

⁶⁰Assumes a non-informative prior, i.e., the hypotheses were equally likely before any data were collected.

- [Blume, 2008] J. D. Blume. How often likelihood ratio are misleading in sequential trials. *Communications on Statistics-Theory and Methods*, 38(8):1193-1206, 2008.
- [Cornfield, 1966] J. Cornfield. Sequential Trials, Sequential Analysis and the Likelihood Principle. *The American Statistician* 29(2): 18-23, 1966.
- [Edwards, 1971] A. W. F. Edwards. *Likelihood*. Cambridge University Press, London, 1971.
- [Fisher, 1959] R. A. Fisher. *Statistical Methods and Scientific Inference*, 2nd ed. New York, Hafner, 1959.
- [Goodman and Royall, 1988] S. N. Goodman and R. M. Royall. Evidence and scientific research. *American Journal of Public Health* 78(12):1568-1574, 1988.
- [Hacking, 1965] I. Hacking. *Logic of Statistical Evidence*. Cambridge University Press, New York, 1965.
- [Royall, 1997] R. M. Royall. *Statistical Evidence*. Chapman & Hall, London, 1997.
- [Royall, 2000] R. M. Royall. On the probability of observing misleading statistical evidence (with discussion). *Journal of the American Statistical Association* 95 (451), 760-767, 2000.
- [Royall and Tsou, 2003] R. M. Royall and T. S. Tsou. Interpreting statistical evidence using imperfect models: Robust adjusted likelihood functions. *Journal of the Royal statistical society, series B*. 65,part2,391-404, 2003.
- [Tsou and Royall, 1995] T. S. Tsou and R. M. Royall. Robust likelihoods. *Journal of the American Statistical Association* 90(419): 316-320, 1995.
- [Savage, 1964] L. J. Savage. The foundations of statistics reconsidered. In *Studies in Subjective probability* with Kyburg HE and Smokler HE (eds). New York: John Wiley and Sons, 1964.
- [Strug and Hodge, 2006a] L. Strug and S. E. Hodge. An alternative foundation for the planning and evaluation of linkage analysis: 1. Decoupling 'error probabilities' from 'measures of evidence'. *Human Heredity* 61:166-188, 2006.
- [Strug and Hodge, 2006b] L. Strug and S. E. Hodge. An alternative foundation for the planning and evaluation of linkage analysis: 1. Implications for multiple test adjustments. *Human Heredity* 61:200-209, 2006.
- [van der Tweel, 2005] I. Van der Tweel. Repeated looks at accumulating data: to correct or not to correct? *European journal of epidemiology*, 20: 205-211, 2005.
- [Wright, 1992] S. P. Wright. Adjusted *P*-Values for Simultaneous Inference. *Biometrics*, vol. 48 1005-1013, 1992.

This page intentionally left blank

EVIDENCE, EVIDENCE FUNCTIONS, AND ERROR PROBABILITIES

Mark L. Taper and Subhash R. Lele

1 INTRODUCTION

In a recent paper Malcolm Forster has stated a common understanding regarding modern statistics:

Contemporary statistics is divided into three camps; classical Neyman-Pearson statistics (see [Mayo, 1996] for a recent defense), Bayesianism (e.g., [Jefferys, 1961; Savage, 1976; Berger, 1985; Berger and Wolpert, 1988]), and third, but not last, Likelihoodism (e.g., [Hacking, 1965; Edwards, 1987; Royall, 1997]). [Forster, 2006]

We agree with this division of statistics into three camps, but feel that Likelihoodism is only an important special case of what we would like to call evidential statistics. In the sequel, we will try to justify our expansion of evidential statistics beyond the likelihood paradigm and to relate evidentialism to classical epistemology,¹ and to classical statistics.

For at least the last three quarters of a century a fierce battle has raged regarding foundations for statistical methods. Statistical methods are epistemological methods, that is, methods for gaining knowledge. What needs to be remembered is that epistemological methods are technological devices — tools. One does not ask if a tool is true or false, or right or wrong. One judges a tool as effective or ineffective for the task to which it will be applied. In this article, we are interested in statistics as a tool for the development of scientific knowledge. We develop our desiderata for knowledge developing tools in science and show how far the evidential statistical paradigm goes towards meeting these objectives. We also relate evidential statistics to the competing paradigms of Bayesian statistics and error statistics.

Richard Royall [1997; 2004] focuses attention on three kinds of questions: “What should I believe?”, “What should I do?”, and “How should I interpret this body of observations as evidence.” Royall says that these questions “define three distinct

¹We have tried to frame our discussion for a philosophical audience, but we are scientists and statisticians. Omission of any particular citation to the philosophical literature will most likely represent our ignorance not a judgment regarding the importance of the reference.

problem areas of statistics.” But, are they the right questions for science? Science and scientists do many things.

Individual scientists have personal beliefs regarding the theories and even the observations of science. And yes, these personal beliefs are critical for progress in science. Without as yet unjustified belief, what scientist would stick his or her neck out to drive a research program past the edge of the known? In a more applied context, scientists are often called to make or advise on decisions large and small. It is likely that this decision making function pays the bills for the majority of scientist. But, perhaps the most important activity that scientists aspire to is augmenting humanity’s accumulated store of scientific knowledge. It is in this activity that we believe Royall’s third question is critical.

Our thinking regarding the importance and nature of statistical evidence develops from our understanding (however crude) of a number of precepts drawn from the philosophy of science. We share the view, widely held since the eighteenth century, that science is a collective process carried out by vast numbers of researchers over long stretches of time [Nisbet, 1980].

Personally, we hold the view that models carry the meaning in science [Frigg, 2006; Giere, 2004; 2008]. This is, perhaps, a radical view, but an interest in statistical evidence can be motivated by more commonplace beliefs regarding models such as that they represent reality [Cartwright, 1999; Giere, 1988; 1999; 2004; Hughes, 1997; Morgan, 1999; Psillos, 1999; Suppe, 1989; van Fraassen, 1980; 2002] or serve as tools for learning about reality [Giere, 1999; Morgan, 1999].

We are strongly skeptical about the “truth” of any models or theories proposable by scientists [Miller, 2000]. We mean by this that although we believe there is a reality, which we refer to as “truth”, no humanly constructed model or theory completely captures it, and thus all models are necessarily false. Nevertheless, some models are better approximations of reality than other models [Lindsay, 2004], and some models are even useful [Box, 1979]. In the light of these basal concepts, we believe that growth in scientific knowledge can be seen as the continual replacement of current models with models that approximate reality more closely. Consequently, the question “what methods to use when selecting amongst models?” is perhaps the most critical one in developing a scientific method.

Undoubtedly the works that most strongly influenced 20th century scientists in their model choices were Karl Popper’s 1934 (German) and 1959 (English) versions of his book *Logic of Scientific Discovery*. Nobel Prize winning scientist Sir Peter Medawar called this book “one of the most important documents of the twentieth century.” Popper took the fallacy of affirming the consequent² seriously, stating that the fundamental principle of science is that hypotheses and theories can never be proved but only disproved. Hypotheses and theories are compared by comparing deductive consequences with empirical observations. This hypothetico-deductive framework for scientific investigation was popularized in the scientific community by Platt’s [1964] article on *Strong Inference*. Platt’s important contribution was

²The logical fallacy that is made according to the following faulty reasoning 1) If A then B, 2) B, 3) Therefore A.

his emphasis on multiple competing hypotheses.

Another difficulty with the falsificationist approach is the fact that not only can you not prove hypotheses, you can't disprove them. This was recognized by Quine [1951]; his discussion of the under-determination of theory by data concludes that a hypothesis³ is only testable as a bundle with all of the background statements on which it depends. Another block to disproving hypotheses is the modern realization that the world and our observation of it are awash with stochastic influences including process variation and measurement error. When random effects are taken into consideration, we frequently find that no data set is impossible under a model, only highly improbable.

Therefore, "truth" is inaccessible to scientist either because the models required to represent "truth" are complex beyond comprehension, or because so many elements are involved in a theory that might represent "truth" fully that an infinite number of experimental manipulations would be required to test such a theory. Finally, even if two full theories could be formulated and probed experimentally, it is not likely that either will be unequivocally excluded because in a stochastic world all outcomes are likely to be possible even if unlikely. What are we as scientists to do? We do not wish to lust after an unattainable goal; we are not so adolescent. Fortunately, there are several substitute goals that may be attainable. First, even if we can't make true statements about reality, it would be nice to be able to make true statements about the state of our knowledge of reality. Second, if our models are only approximations, it would be nice to be able to assess how close to truth they are [Forster, 2002].

Popper [1963] was the first to realize that although all theories are false, some might be more truthlike than others and proposed his concept of verisimilitude to measure this property. Popper's exact formulation was quickly discredited [Harris, 1974; Miller, 1974; Tichy, 1974], but the idea of verisimilitude continues to drive much thought in the philosophy of science (see [Niiniluoto, 1998; Zwart, 2001; Oddie 2007] for reviews). The results of this research have been mixed [Gemes, 2007]. The difficulty for the verisimilitude project is that, philosophically, theories are considered as sets of linguistic propositions. Ranking the overall truthlikeness of different theories on the basis of the truth values and content of their comprised propositions is quite arbitrary. Is theory A, with only one false logical consequence, truer than theory B, with several false consequences? Does it make a difference if the false proposition in A is really important, and the false propositions in B are trivial? Fortunately, as Popper noted [1976] verisimilitude is possible with numerical models where the distance of a model to truth can be represented by a single value.

We take evidence to be a three-place relation between data and two alternate models⁴. Evidence quantifies the relative support for one model over the other and

³A scientific hypothesis is a conjecture as to how the world is or operates.

⁴A reviewer has suggested that background information may be a necessary fourth part, but background information will enter the formalization either as part of the data, or as part of one or more of the models.

is a data based estimate of the relative distance from each of the models to reality. Under this conception, to speak of evidence for a model does not make sense. This then is what we call the evidential approach, to compare the truthlikeness of numerical models. The statistical evidence measures the differences of models from truth in a single dimension and consequently may flatten some of the richness of a linguistic theory. While statistical evidence is perhaps not as ambitious as Popper's verisimilitude, it is achievable and useful.

We term the quantitative measure of relative distance of models to truth an evidence function [Lele, 2004, Taper & Lele, 2004]. There will be no unique measure or the divergence between models and truth so a theory of evidence should guide the choice of measures in a useful fashion. To facilitate the use of statistical evidence functions as a tool for the accumulation of scientific knowledge we believe that a theory of evidence should have the following desiderata:

- D1) Evidence should be a data based estimate of the relative distance between two models and reality.
- D2) Evidence should be a continuous function of data. This means that there is no threshold that must be passed before something is counted as evidence.
- D3) The reliability of evidential statements should be quantifiable.
- D4) Evidence should be public not private or personal.
- D5) Evidence should be portable, that is it should be transferable from person to person.
- D6) Evidence should be accumulable: If two data sets relate the same pair of models, then the evidence should be combinable in some fashion, and any evidence collected should bear on any future inferences regarding the models in question.
- D7) Evidence should not depend on the personal idiosyncrasies of model formulation. By this we mean that evidence functions should be both scale and transformation invariant.⁵

We do not claim that inferential methods lacking some of these characteristics cannot be useful. Nor do we claim that evidential statistics is fully formulated. Much work needs to be done, but these are the characteristics that we hope a mature theory of evidence will contain.

Glossed over in Platt is the question of what to do if all of your hypotheses are refuted. Popper acknowledges that even if it is refuted, scientists need to keep their best hypothesis until a superior one is found [Popper, 1963]. Once we recognize

⁵An example of scale invariance is that whether one measures elevation in feet or meters should not influence the evidence that one mountain is higher than another. An example of transformation invariance is that it should not matter in conclusions regarding spread whether spread is measured as a standard deviation or as a variance.

that scientists are unwilling to discard all hypotheses [Thompson, 2007] then it is easy to recognize that the falsificationist paradigm is really a paradigm of relative confirmation — the hypothesis least refuted is most confirmed. Thus, the practice of science has been cryptically evidential for at least half a century. We believe that it is important to make this practice more explicit.

2 QUANTIFYING EVIDENCE, LIKELIHOOD RATIOS AND EVIDENCE FUNCTIONS:

The issue of quantifying evidence in the data has always vexed statisticians. The introduction of the concept of the likelihood function⁶ [Fisher, 1912; 1921; 1922] was a major advance in this direction. However, how should one use the likelihood function? The main uses of the likelihood function have been in terms of point estimation of the parameters of the statistical model and testing of statistical hypotheses⁷ [Neyman and Pearson, 1933]. Neyman and his associates couched statistical inference as a dichotomous decision making problem (violating D2) whereas Fisher seems to have been much more inclined to look at statistical inference in terms of quantification of evidence for competing models.⁸ The use of significance tests and the associated p-values⁹ as a measure of evidence is most popular in applied sciences. However, their use is not without controversy [Royall, 1986]. The main problem with the use of p-values as a measure of evidence is that they are not comparative measures (violating D1). There is no explicit alternative against which the hypothesis of interest is being compared = [Royall, 1992]. Similarly, the use of Bayesian posterior probabilities as a measure of evidence is problematic leading to a number of contradictions. Posterior probabilities are not invariant to parameterization making them an unsatisfactory measure of evidence (by violating D7). Many Bayesian formulations involve subjective prior probabilities there by violating D4. The likelihood ratio (LR) is a measure of evidence with a long history. LRs explicitly compare the relative support for two models given a set of observations (D1) and are invariant to parameter transformation and scale change (D7). Barnard [1949] is one of the earliest explicit expositions. Hacking [1965] made it even more explicit in his statement of the law of the likelihood.¹⁰ Ed-

⁶The likelihood is numerically the probability of the observations (data) under a model and is considered a function of the parameters, that is: $L(\theta; \mathbf{x}) = f(\mathbf{x}; \theta)$. Here L is the likelihood function, θ is the parameter or parameter vector, x is the datum or data vector, and f is a probability distribution function. The likelihood is not a probability, as it does not integrate to one over the parameter space.

⁷A statistical hypothesis is a conjecture that the data are drawn from a specified probability distribution. Operationally, one tests scientific hypotheses by translating them into statistical hypotheses and then testing the statistical hypotheses (see [Pickett *et al.*, 1994] for a discussion)

⁸"A likelihood based inference is used to analyze, summarize and communicate statistical evidence. . . ." [Fisher, 1973, page 75].

⁹the p-value is the probability (under a null hypothesis) of observing a result as or more extreme than the observed result.

¹⁰According to the law of likelihood model, 1 is supported over model 2 if based on the same data the likelihood of model 1 is greater than the likelihood of model 2.

wards [1992] is another exposition that promoted the law of the likelihood and the likelihood function. Royall [1997] perhaps makes the best pedagogic case for the law of the likelihood and expands its scope in many significant ways. In particular, his introduction of error probabilities and their use in designing experiments is extremely important. Further, his promotion and justification of profile likelihood and robust likelihood as a measure of evidence is a significant development.

Although, in the first part of his book Royall strongly emphasizes the likelihood principle¹¹ (which is different than the law of likelihood) his use of ad hoc adjustments to likelihoods such as the profile likelihood in the presence of nuisance parameters clearly violate the likelihood principle because consideration beside just the likelihoods influence the inference [Fraser, 1963; Berger and Wolpert, 1988]. Similarly, the consideration of error probabilities depends on the sample space and hence they violate the likelihood principle as well [Boik, 2004]. It is clear the error probabilities, if taken as part of the evidence evaluation, violate the likelihood principle. We hasten to point out that Royall does not suggest using error probabilities as part of evidence. We agree that error probabilities are not evidence, but feel that they can play an important role in inference. We expand on this discussion in the next section.

Royall and others take the law of likelihood as a given and then try to justify why it makes sense in terms of the adherence to the likelihood principle, universal bound for probability of misleading evidence and other intuitive criteria. On the other hand, when faced with the problem of nuisance parameters¹² or the desire for model robustness,¹³ they propose the use of other ad hoc methods but their justifications do not carry through. Nevertheless, in many practical situations, one may not want to specify the model completely and use methods based on mean and variance function specification alone such as Generalized Linear Models [McCulloch and Nelder, 1989]. A further problem is that the error probabilities are computed assuming one of the hypotheses is in fact true (violating D1). To circumvent these issues and to try to give a fundamental justification for the use of the likelihood ratio as a measure of evidence, Lele [2004] introduced the concept of evidence functions.

The first question that Lele [2004] poses is: what happens to the likelihood ratio when true distribution is different than either of the competing hypotheses? A simple application of the law of large numbers shows that as the sample size increases, the log-likelihood ratio converges to the difference between the Kullback-Leibler divergence¹⁴ between the true distribution and hypothesis A and Kullback-

¹¹The likelihood principle states that all evidential meaning in the data is contained in the likelihood function.

¹²Nuisance parameters are parameters that must be included in the model for scientific reality, but are not themselves the entities on which inference are desired. Nuisance parameters are discussed in more detail in the section on multiplicities.

¹³Model robust techniques are designed so that inference on the model elements of interest can be made even if nuisance portions of the model may be somewhat incorrect.

¹⁴The Kullback-Leibler divergence is one of the most commonly used measure of the difference of one distribution from another. If $f(x)$ and $g(x)$ are probability distributions, then $KL(f, g)$ is

Leibler divergence between the true distribution and hypothesis B . If hypothesis A is closer to the truth than hypothesis B is, the likelihood ratio leads us to hypothesis A . Thus, it follows that strength of evidence is a relative measure that compares distances between the true model and the competing models [Lele, 2004]. Immediate consequences of this observation are the questions: 1) Can we use divergence measures other than Kullback-Leibler to measure strength of evidence? 2) Is there anything special about Kullback-Leibler divergence? A study of these two questions led Lele [2004] to following conclusions: First, different divergence measure based quantification may be compared in terms of the rate at which the probability of strong evidence converges to one. And second, the Kullback-Leibler divergence measure has the best rate of convergence among all other measures of evidence.

This result holds provided full specification of the probabilistic model is available, there are no outliers in the data, and the true model is one of the competing hypotheses. However, one can make quantification of evidence robust against outliers, an important practical consideration, by using divergences other than Kullback-Leibler divergence.¹⁵ Further, one can quantify strength of evidence in situations where one may not want to specify the full probabilistic model but may be willing to specify only mean and variance functions by using Jeffrey's divergence measure. One may also use an empirical likelihood ratio or other divergence measures based on an estimating function. Thus, one can justify the use of a variety of modified forms for the likelihood ratio such as conditional likelihood ratios, profile likelihood ratios, and composite likelihood ratios as measures of evidence because they correspond to some form of relative divergence from "truth".

Other notable conclusions that follow from the generalization of the law of likelihood in terms of divergence measures are: 1) The design of experiment and stopping rules do matter in the quantification of evidence if divergences other than Kullback-Leibler divergence are used [Lele, 2004, discussion]. This goes against the pre-eminence of the likelihood principle in the development of Royall [1997] but criticized by Cox [2004]. And 2), the concept of error probabilities needs to be extended to allow for the fact that the class of hypothesized models seldom contains the true model. In the following, we suggest how the second issue could be addressed. It also leads to quantifying post-data reliability measures for the strength of evidence.

the average for observations, x , drawn from $f(x)$ of $\log(f(x)/g(x))$. $KL(f, g)$ is 0 when f and g are the same distribution and is always greater than 0 if the two distributions differ. Technically $KL(f, g)$ is a divergence not a distance because $KL(f, g)$ need not equal $KL(g, f)$.

¹⁵Other common divergences/distances between statistical distributions are the Pearson chi-squared distance, the Neyman chi-squared distance, the symmetric chi-squared distance, and the Hellinger distance [Linhart & Zucchini, 1986; Lindsay, 2004; Lindsay *et al.*, 2007].

3 THE PROBABILITY OF MISLEADING EVIDENCE AND INFERENCE RELIABILITY:

Richard Royall's introduction of the concepts of the probability of misleading evidence (M) and the probability of weak evidence (W) constituted a major advance in evidential thinking. Misleading evidence is defined as strong evidence for a hypothesis that is not true. The probability of misleading evidence is denoted by M or by $M(n, k)$ to emphasize that the probability of misleading evidence is a function of both sample size and the threshold, k , for considering evidence as strong. The probability of weak evidence is the probability that an experiment will not produce strong evidence for either hypothesis relative to the other. When one has weak evidence, one cannot say that the experiment distinguishes between the two alternative hypotheses in any meaningful way. These probabilities link evidential statistics to the error statistical thread in classical frequentist analysis. As experimental design criteria, M and W are superior to the type I (design based probability of rejecting a true null hypothesis = α) and type II (design based probability of failing to detect a true alternative hypothesis = β) error rates of classical frequentist statistics because both M and W can be simultaneously brought to zero by increasing sample size [Royall, 1997; 2004; Blume, 2002].

For Royall and his adherents there are three quantities of evidential interest: 1) the strength of evidence (likelihood ratio), 2) the probability of observing misleading evidence¹⁶ (M), and 3) the probability that observed evidence is misleading.¹⁷ This last is not the same as M and it requires prior probabilities for the two alternative hypotheses.¹⁸ Royall claims that M is irrelevant post data and that M is for design purposes only. In common scientific practice, all three measures have often been freighted on the p-value. There are a vast number of papers discussing common misconceptions on the interpretation of p-value (e.g. Blume & Peipert, 2003; Goodman 2008). The strength of Royall's approach is that these three quantities are split apart and can be thought about independently.

4 GLOBAL & LOCAL RELIABILITY

There is a deep reason why M and other flavors of error statistics are important in statistical approaches to scientific problems. We strongly believe that one of the foundations of effective epistemology is some form of reliabilism. Under reliabilism,

¹⁶Given two statistical hypotheses (H_1 and H_2) the probability of misleading evidence for H_2 over H_1 is $M = P_1([L(x; H_2)/L(x; H_1)] > k)$; where M is the probability of misleading evidence, $P_1(\cdot)$ is the probability of the argument under hypothesis 1, and k is an *a priori* boundary demarcating the lower limit of strong evidence.

¹⁷The probability that observed evidence for H_2 is misleading = $\pi(H_1)/[\pi(H_1) + \pi(H_2) \cdot LR_{ob}]$, where $\pi(H_1)$ is the prior probability of H_1 and LR_{ob} is the observed likelihood ratio (see [Blume, 2002]).

¹⁸We do not attach much importance to this third quantity, meaningful priors are rarely available, its primary purpose in its presentation is to clarify that it is indeed distinct from M .

a belief (or inference) is justified if it is formed from a reliable process [Goldman, 1986; 2008; Roush, 2006].

Reliability has two flavors, global reliability and local reliability. Global reliability describes the truth-tracking or error avoidance behavior of an inference procedure over all of its potential applications. Examples of global reliability measures in statistics are Neyman/Pearson test sizes (α and β) and confidence interval levels. These measures describe the reliability of the procedures not individual inferences. Individual inferences are deemed good if they are made with reliable procedures. For example if $\alpha=0.05$ the scientist can feel comfortable saying: “Only in 5% of the cases would this procedure reject a null hypothesis in error, so I can have confidence in the rejection that I have currently observed.” Royall’s probability of misleading evidence M is this global kind of a measure and hereafter we will refer to it as the global reliability of the design or M_G .

Local reliability on the other hand is the “truth-acquisition or error avoidance in scenarios linked to the actual scenario in question” [Goldman, 2008]. Fisherian p-values and Mayo’s test severity [Mayo, 2004; Mayo and Cox, 2006] are local reliability measures. It is easy to see that the p-value is a local reliability measure because the error probabilities are calculated relative to the specific observed results. Both local and global reliability are useful in generating scientific knowledge [Goldman, 1986]. A local reliability measure or measures would be useful within the context of the evidential paradigm.

5 LOCAL RELIABILITY AND THE EVIDENTIAL PARADIGM

5.1 Local reliability under the alternatives

We define the local reliability of the evidence, M_L , as the probability that evidence for one model is strong as or stronger than the evidence actually observed could have been generated under the alternative. As a post data measure, M_L is not the same as M_G conceptually or quantitatively. M_L is also distinct from the probability that the observed evidence is misleading in several aspects. First, M_L involves a tail sum and the probability that the evidence is misleading does not, and second, the probability that the evidence is misleading depends on the prior probability of the two models, while M_L does not.

Royall [1997] presents a surprising but powerful and simple result that sets bounds on the magnitude of M_G . He shows that

$$P_B \left(\frac{P_A(X)}{P_B(X)} \geq q \right) \leq \frac{1}{q}$$

where q is any constant.¹⁹ In particular if $q = k$, the threshold for strong evidence, we see that the probability of misleading evidence, M_G , must be less than $1/k$.

¹⁹One proof follows directly from substitution into a classic theorem in probability called Markov’s inequality which states that if Y is a nonnegative random variable then $P(Y \geq q) \leq E(Y)/q$ where $E(\cdot)$ denotes expectation. Substituting the likelihood ratio for Y we

Royall calls this the universal bound on the probability of misleading evidence. The actual probability of misleading evidence may often be much lower. Further, as the likelihood ratio of observed strong evidence (LR_{ob}) is by definition greater than or equal to k , the local probability of misleading evidence, M_L , must be less than or equal to the global probability of misleading evidence. That is:

$$M_L = P_B \left(\frac{P_A(X)}{P_B(X)} \geq LR_{ob} \right) \leq \frac{1}{LR_{ob}} \leq M_G \leq \frac{1}{k}.$$

One question that springs to mind is why was a post data reliability measure not included in Royall's original formulation of the evidential paradigm? While only Royall could really answer this question, should he choose to break his silence, it is easy to see that within Royall's context there is no need for an explicit measure of local reliability. Royall's work was focused on the comparison of simple or point models. In the comparison of simple models, the likelihood ratio and the p-value contain the same information allowing one to transform from one to the other [Sellke *et al.*, 2001] and M_L is redundant. However, when one begins to expand the evidential approach, as one must to develop a complete statistical toolkit, M_L does seem to become an interesting and useful evidential quantity.

5.2 Local reliability under the unknown truth

The likelihood ratio or any other measure of strength of evidence is a point estimate of the difference between divergences from truth to A and truth to B . The first issue we need to address is to quantify the distribution of the strength of evidence under hypothetical repetition of the experiment. In Royall's formulation, this is done under either hypothesis A or hypothesis B . But as we have noted, neither of these hypotheses need be the true distribution. Royall's formulation is useful for pre-data, sample size determination or optimal design issues. The local reliability M_L defined in the previous section is potentially a useful post data quantity, but it is still calculated under the explicit alternative hypotheses and not the underlying true distribution.

Once the experiment is conducted or the observations made, a non-parametric estimate of the true distribution is accessible. One can conduct a non-parametric bootstrap²⁰ to obtain an estimate of the distribution of the strength of the evidence under this true distribution. This can be used inferentially in several different ways:

First, one can obtain a bootstrap based confidence interval for the likelihood ratio: This tells us if the experiment is repeated infinitely often (under the true model), what would be the distribution of likelihood ratios? This information

have $P_B(P_A(x)/P_B(x) \geq q) \leq E_B(P_A(x)/P_B(x))/q$. By definition, $E_B(P_A(x)/P_B(x)) = \int P_B(x)(P_A(x)/P_B(x))dx$. This last integral simplifies to $\int P_A(x)dx$ which integrates to 1 because $P_A(x)$ is a probability distribution. Thus, $P_B(P_A(x)/P_B(x) \geq q) \leq 1/q$ as claimed.

²⁰The bootstrap is one of the profound statistical developments of the last quarter century. The unknown distribution of a statistic can be estimated non-parametrically by repeatedly re-sampling data sets from the observed data set and recalculating the statistic for each (see [Efron & Tibshirani, 1993]).

could be presented either as intervals, as a curve that represents a post data measure of the reliability of the estimated strength of evidence, or transformed to a support curve [Davison and Hinkley, 1992; Davison and Hinkley, 1997, Sellke *et al.*, 2001]. Both the upper and lower confidence limits are informative. The lower limit says “it is not likely that the true divergence is less than this”, while the upper limit says that “it is not likely that the true divergence is greater than this.” A second alternative measure that can be calculated using a bootstrap is the proportion of times hypothesis A will be chosen over hypothesis B (proportion of times $LR > 1$). This measure can be interpreted as a post data reliability of model choice. Although the bootstrap quantities defined above involve tail-sums, they are quite distinct from either the global or local probabilities of misleading evidence discussed in the previous section. Whereas M_G or M_L are counterfactual, answering the question if the correct model was the one not indicated by the evidence how probable is a mistaken evidential assessment as strong as the one observed, the bootstrap tail-sum is a direct assessment of the reliability of the observed evidence. Furthermore, this assessment is made under truth, and not under either hypothesized model.

6 EVIDENCE AND COMPOSITE HYPOTHESES

The evidential approach has been criticized (e.g. [Mayo and Spanos, 2006]) as a toy approach because the LR can’t compare composite hypotheses.²¹ This criticism is simultaneously true, a straw man, a good thing, and false. It is true because one can only strictly rank composite hypotheses if every member of one set is greater than every member of the other [Royall, 1997; Blume, 2002; Forster and Sober, 2004]. But, the statement is also a strawman because it implies that the evidential paradigm isn’t able to do the statistical and scientific work done using composite hypotheses, which is patently false. Classically, composite hypotheses are used to determine if a point null is statistically distinguishable from the best alternative, or to determine if the best supported alternative lies on a specified side of the point null. Royall [1997, chapter 6] gives a number of sophisticated examples of doing real scientific work using the tools of the support curve, the likelihood ratio, and the support interval. Further, the inability of the LR to compare composite hypotheses is a good thing because Royall is correct in that the composite H can lead to some ridiculous situations. Consider the comparison of hypotheses regarding the mean of a normal distribution with a known standard deviation of 2 as in [Mayo and Cox, 2006]. $H_0: \mu \leq 12$ vs: $H_1: \mu > 12$. A μ of 15 and a μ of 10,000 are both in H_1 . But, if 15 is the true mean, a model with $\mu = 0$ (an element of H_0) will describe data generated by the true model much better than will $\mu = 10,000$ (an element of H_1). This contradiction will require some awkward circumlocution by Neyman/Pearson adherents. Finally, the statement is false if under the evidence function concept discussed above we expand the evidential paradigm to include

²¹Composite hypotheses are hypotheses that subsume multiple hypotheses.

model selection using information criteria. Comparing composite hypotheses using information criteria is discussed in more detail in the next section.

7 SELECTING BETWEEN COMPOSITE HYPOTHESES

We suggest that, under the evidential paradigm, the composite hypothesis problem be recast as a model selection problem among models with different numbers of free parameters. In the simple example given above H_0 is a model with no free parameters while H_1 is a family of models indexed by the free parameter μ . Model selection using information criteria²² compares models by estimating from data their relative Kulback-Leibler distance to truth (Burnham and Anderson 2002). This is a reasonable evidence function. With multiple models, all models are compared to the model with the lowest estimated KL distance to truth. The model selection procedures are blind to whether the suite of candidate models is partitioned into composite hypotheses. One can consider that the hypothesis that contains the best supported model is the hypothesis best supported by the data. No longer comparing all points in one hypothesis to all points in another, but in effect, comparing the best to the best. Where best is defined as the model with the lowest information criterion value. This solution is neither ad hoc (to the particular case) nor post hoc (after the fact/data). The comparison of composite hypotheses using information criteria is not a toy procedure, and can do real scientific work. Taper and Gogan [2002] in their study of the population dynamics of the Yellowstone Park northern elk herd were interested in discerning whether population growth was density dependent or density independent. They fitted 15 population dynamic models to the data and selected amongst them using the Schwarz information criterion (SIC). The best model by this criterion was a density dependent population growth model and difference between the SIC value for this model and that of the best density independent model was more than 5, a very highly significant difference [Burnham and Anderson, 2002]. There were a number of statistically indistinguishable density dependent models that all fit the data well, making identifying the best model difficult. Nevertheless, it is clear that the best model is in the composite hypothesis of density dependence, not the composite hypothesis of density independence.

8 EVIDENCE AND THE CHALLENGES OF MULTIPLICITIES

As pointed out by Donald Berry [2007] multiplicities are the bane of all statistics. By multiplicities we mean the vague class of problem that are not simple, including

²²Information criteria are a class of measures for estimating the relative KL distance of models to “truth”. In general information criteria include both the number of parameters and the number of data points in their calculation. Information criteria attempt (with varying degrees of success) to overcome the problems of overfitting that would result if comparisons were made on the basis of likelihoods alone. Technically, only order consistent information criteria meet all the criteria of evidence functions.

multiple hypotheses, multiple comparisons, multiple parameters, multiple tests, and multiple looks at the data. Evidential statistics is not immune to the effects of multiplicities, but the evidential paradigm does have approaches to tackling these problems, which are in some cases superior to classical approaches.

8.1 *Nuisance parameters*

Nuisance parameters occur when reality and data are complex enough to require models with multiple parameters, but inferential interest is confined to a reduced set of parameters. Making inferences on the parameters of interest that isn't colored by the nuisance parameters is difficult. Marginal or conditional likelihoods can be used. These are proper likelihoods²³ so all the likelihood ratio based evidential techniques can be employed. Unfortunately, marginal and conditional likelihoods are not always obtainable.

Royall [2000] recommends the use of profile likelihood²⁴ ratio as a general solution. Royall feels that the profile likelihood ratio is an *ad hoc* solution because true likelihoods are not being compared. Nevertheless, he finds the performance of the profile likelihood ratio to be very satisfactory. In our expanded view of evidence, the profile likelihood ratio is not *ad hoc* because the profile likelihood ratio can be shown to be an evidence function. Royall [2000] shows that the probability of misleading evidence from a profile likelihood ratio is not constrained by the universal bound, and can exceed $1/k$. Thus, even in this first expansion of the concept of evidence from the likelihood ratio of two simple hypotheses we see that M_L is decoupled from the likelihood ratio and contains distinct information.

8.2 *Sequential analyses — multiple tests of the same hypothesis*

Another multiplicity that the evidential approach handles nicely is sequential analysis. Multiple looks at data while it is accumulating does not diminish the strength of evidence of the ultimate likelihood ratio, unlike p-value based approaches, which must carefully control the spending of test size in multiple analyses [Demets and Lan, 1994]. Further, the universal bound that $M_G \leq 1/k$ is still maintained. This subject is developed in detail by Blume [2008]. Under a sequential sampling design, observations will be terminated as soon as the likelihood ratio passes the threshold k . Consequently, the local probability of misleading evidence will only be slightly lower than the global probability of misleading evidence.

²³A proper likelihood is directly associated with and numerically equal to some probability distribution function for the observations. Proper likelihoods are often referred to in the literature as “true likelihoods.”

²⁴Profile likelihoods are functions of the data and the parameter or parameters of interest (i.e. not the nuisance parameters). The value of the profile likelihood is the maximum value the full likelihood could take under all possible values of the nuisance parameters.

8.3 Multiple comparisons: Many tests of different hypotheses

Multiple comparisons place a heavy burden on scientists. Scientists are bursting with questions, and design experiments and surveys to answer many questions simultaneously. As a classical error statistical analysis sets an *a priori* level of type I error on each test, increasing the number of tests increases the probability that at least one of them will be significant by chance alone. To control the family wide error rate, scientists have been forced to decrease the size of individual tests using lower type I error rates. The price of this move is that the power of the individual tests to detect small but real differences is diminished. The scientist makes fewer errors, but gets fewer things right as well.

An evidential analysis is not immune to the increase in the family wide probability of error with an increasing number of tests. If we define $M_G(\mathbf{n}, N, k)$ as the pre-experiment probability of at least 1 misleading result at level k amongst N comparisons with n_i observations each, then

$$M_G(\mathbf{n}, N, k) = 1 - \prod_{i=1}^N (1 - M_G(n_i, k)) \leq \sum_{i=1}^N M_G(n_i, k) \leq N/k.$$

So the global probability of misleading evidence increases with the number of comparisons in the same fashion that the family wide type I error does. As the local probability of misleading evidence of a comparison is always less than or equal to the global probability of misleading evidence for the comparison, the local family wide probability of misleading evidence will also be less than the global family wide probability of misleading evidence.

Although multiple comparisons lays a burden on evidential analysis similar to that laid on a classical error statistical analysis, the evidential approach has more flexible ways of mitigating this burden. The ways family wide misleading evidence can be controlled depends on whether sample size is constrained or if it can be increased, either before or after the initial experiment is conducted. If the sample sizes in the comparisons are fixed, then the only control strategy is to increase the strong evidence threshold k , in direct analogy to the test size adjustment of classical multiple comparisons. This will decrease $M_G(n, N, k)$, but with the cost that the probability of weak evidence (W) will increase for all comparisons, similar to the classical decrease in power resulting from test size adjustment.

However, if sample size is flexible then several alternative strategies become available. Strug and Hodge [2006] give a clear description of three scenarios for controlling the global family wide misleading evidence in multiple comparisons by adjusting sample size. Strategy 1: Increase sample size in all comparisons before the experiment. $M_G(n_i, k)$ can be brought to any desired level without changing the strong evidence threshold k for each comparison by increasing sample size. Consequently, $M_G(n, N, k)$ can also be brought to any desired level. This strategy has the advantage that W will be simultaneously decreased for all comparisons, but the cost in terms of increased sample size is high. Strategy 2: Increase sample size for only those comparisons exhibiting strong evidence. This requires an analysis of

interim data, but we have seen that has little untoward influence in an evidential analysis. $M_G(n, N, k)$ can be brought to any desired level, but W will remain unaltered. The sample size cost is less than Strategy 1. Finally, in Strategy 3, the scientist would increase sample size for comparisons with interim strong or weak evidence, but not strong opposing evidence. $M_G(n, N, k)$ is controllable at any desired level, W is reduced, and sample size costs are intermediate between Strategies 1 and 2.

8.4 Multiple candidate models

One of the problems of multiplicities that the evidential paradigm is most susceptible to is the difficulty caused by too many candidate models. One of the great strengths of the evidential paradigm is that it allows and encourages the comparison of multiple models. This allows a more nuanced and accelerated investigation of nature. However, the risk is that, if too many models are considered with a single data set, a model that is not really very good will be favored by chance alone. This has been called model selection bias [Zucchini, 2000; Taper, 2004].

The problem of model selection bias has led Burnham and Anderson and their acolytes strongly and repeatedly argue against “data dredging” and for compact candidate model sets defined by *a priori* scientific theory (e.g. [Anderson *et al.*, 2000; Anderson and Burnham, 2002, Burnham and Anderson, 2002]). There is considerable merit to these recommendations, but the cost is that ability to broadly explore model space is reduced. As with multiple comparisons, several alternatives are possible for an evidential analysis, each with costs and benefits. One suggestion made by Taper and Lele [2004] is to increase k , the threshold for strong evidence. This would decrease the probability of misleading evidence over the entire analysis, but at the cost of potentially ending with a large number of indistinguishable models. Another alternative suggested [Bai *et al.*, 1999; Taper, 2004] is to increase the parameter penalty in coordination with the increase in the number of candidate models. If effects are tapered, these procedures select models with large effects of each parameter. Here the ability to detect model fine structure is traded for the ability to investigate a broad number of models.

9 DISCUSSION

Evidentialism is an adolescent statistical paradigm, neither infantile nor mature. It is capable of much scientific work, but with considerable scope for technical improvement. Strict Royallist evidentialism is rapidly gaining adherents in epidemiology and medical genetics, while information criteria based inference is a major force in ecological statistics.

The elevation of evidentialism to a practical statistical approach is due to Royall’s introduction of the concepts of weak and strong evidence and of misleading evidence. The introduction in this paper of local reliability (post data) currently serves to clarify the epistemic standing of evidential practices. The warrant for

post data inference using the likelihood ratio as the strength of evidence is the local reliability of the evidence. The reliability of the evidence is a function of the local probability of misleading evidence, M_L , which is directly linked to LR. One interesting observation is that local reliability is in general much greater than NP error rates indicate. Further, local reliability is in general greater than global probability of misleading evidence, M_G (*a priori* evidential error rate), indicates.

We have also suggested several measures local reliability that do not develop their probability assessments from the explicit alternative models under consideration but instead use a non-parametric bootstrap to estimate local reliability under the unknown true distribution. All of these are valid assessments of post data error probabilities. Which of them proves most helpful in constructing scientific arguments will become clear through time and use. Together with M_L these bootstrap methods provide the evidential approach a rich suite of tools for post-data assessment of error probabilities that are uncoupled from the estimation of the strength of evidence.

Science needs mechanisms for the accumulation of sound conclusions (*sensu* [Tukey, 1960]). A major rival for Evidentialism as a philosophically sound (in our eyes) system for the advancement of science is the “Error Statistical” brand of Neyman-Pearson analysis promoted by Deborah Mayo.

We dismiss Bayesianism for its use of subjective priors and a probability concept that conceives of probability as a measure of personal belief. Bayesianism is held by many philosophers as the most appropriate method of developing personal knowledge. This may be, but is irrelevant to the task at hand. Science depends on a public epistemology not a private one. The Bayesian attempts to bridge the gap between the private and the public have been tortured.

It is not that we believe that Bayes’ rule or Bayesian mathematics is flawed, but that from the axiomatic foundational definition of probability Bayesianism is doomed to answer questions irrelevant to science. We do not care what you believe, we barely care what we believe, what we are interested in is what you can show. Bayesian techniques that successfully eliminate the influence of the subjective prior [Boik, 2004], such as the Bayesian Information Criterion [Schwarz, 1978] or data cloning [Lele *et al.*, 2007; Ponciano *et al.*, 2009], may be useful.

Our difficulties with Mayo’s Error Statistical approach are didactic. Mayo speaks of probing *a* hypothesis. In our radical fallibilist view of science models or hypotheses can only be supported relative to other models or hypotheses. Certainly, there actually is a cryptic alternative hypothesis in Mayo’s calculations, but we believe that the linguistic suppression of the alternative is counterproductive. A probed hypothesis is smugly self-congratulatory where a pair of hypotheses compared evidentially invites scientist to throw new hypotheses into the mix.

Fundamentally, Mayo’s approach represents a fusion of Fisher’s significance test, a post data error calculation, with Neyman-Pearson hypothesis testing. We think that the use of post data error as the strength of evidence and the shift in emphasis from inductive decision to inductive inference are both helpful steps. However, scientists have struggled with the logics of both Fisherian significance tests and

Neyman-Pearson tests. For generations, it has been a cottage industry for statisticians to write white papers trying to explain these concepts to working scientists. Nothing in Mayo's reformulation will ease these difficulties. On the other hand, the evidential paradigm presents scientists with powerful tools to design and analyze experiments and to present results with great clarity and simplicity. This is because the evidential paradigm is designed around the single task that scientist most need to do: That is to objectively compare the support for alternative models. Master statisticians can, with their decades of training in classical statistics, successfully navigate the conceptual pitfalls of Mayo's recasting of the Fisher and Neyman-Pearson methods, but for working scientists such as us, the evidential paradigm should be a great relief.

ACKNOWLEDGEMENTS

We thank the editors Malcolm R. Forster and Prasanta S. Bandyopadhyay for critical comments that have greatly improved this work.

BIBLIOGRAPHY

- [Anderson and Burnham, 2002] D. R. Anderson and K. R. Burnham. Avoiding pitfalls when using information-theoretic methods. *Journal of Wildlife Management* **66**:912-918, 2002.
- [Anderson *et al.*, 2000] D. R. Anderson, K. P. Burnham, and W. L. Thompson. Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management* **64**:912-923, 2000.
- [Bai *et al.*, 1999] Z. D. Bai, C. R. Rao, and Y. Wu. Model selection with data-oriented penalty. *Journal of Statistical Planning and Inference* **77**:103-117, 1999.
- [Barnard, 1949] G. A. Barnard. Statistical Inference. *Journal of the Royal Statistical Society, Series B* **11**:115-149, 1949.
- [Berger, 1985] J. O. Berger. *Statistical decision theory and Bayesian analysis*. 2nd edition. Springer-Verlag, New York, 1985.
- [Berger and Wolpert, 1988] J. O. Berger and R. L. Wolpert. *The Likelihood Principle*. 2nd edition. Springer-Verlag, New York, 1988.
- [Berry, 2007] D. A. Berry. The difficult and ubiquitous problems of multiplicities. *Pharmaceutical Statistics* **6**:155-160, 2007.
- [Blume, 2002] J. D. Blume. Likelihood methods for measuring statistical evidence. *Statistics in Medicine* **21**:2563-2599, 2002.
- [Blume and Peipert, 2003] J. Blume and J. F. Peipert. What your statistician never told you about P-values. *Journal of the American Association of Gynecologic Laparoscopists* **10**:439-444, 2003.
- [Blume, 2008] J. D. Blume. How often likelihood ratios are misleading in sequential trials. *Communications in Statistics-Theory and Methods* **37**:1193-1206, 2008.
- [Boik, 2004] R. J. Boik. Commentary. on Why Likelihood? by Forster and Sober. Pages 167-180 in M. L. Taper and S. R. Lele, editors. *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. University of Chicago Press, Chicago, 2004.
- [Box, 1979] G. E. P. Box. Robustness in the strategy of scientific model building, in R. L. Launer and G. N. Wilkinson (eds.), *Robustness in Statistics* New York: Academic Press, 201-236, 1979.
- [Burnham and Anderson, 2002] K. P. Burnham and D. R. Anderson. *Model Selection and Multi-model Inference: A Practical Information-Theoretic Approach*. 2nd edition. Springer-Verlag, New York, 2002.

- [Cartwright, 1999] N. Cartwright. *The Dappled World. A Study of the Boundaries of Science*. Cambridge University Press Cambridge, 1999.
- [Cox, 2004] D. R. Cox. Commentary on The Likelihood Paradigm for Statistical Evidence by R. Royall. Pages 119-152 in M. L. Taper and S. R. Lele, editors. *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. University of Chicago Press, Chicago, 2004.
- [Davison and Hinkley, 1997] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, UK, 1997.
- [Davison *et al.*, 1992] A. C. Davison, D. V. Hinkley, and B. J. Worton. Bootstrap Likelihoods. *Biometrika* 79:113-130, 1992.
- [Demets and Lan, 1994] D. L. Demets and K. K. G. Lan. Interim analysis - the alpha-spending function approach. *Statistics in Medicine* 13:1341-1352, 1994.
- [Edwards, 1992] A. W. F. Edwards. *Likelihood*. Expanded Ed. Johns Hopkins University Press, Baltimore, 1992.
- [Efron and Tibshirani, 1993] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, London, UK, 1993.
- [Fisher, 1912] R. A. Fisher. On an absolute criterion for fitting frequency curves. *Messeng. Math* 41:155-160, 1912.
- [Fisher, 1921] R. A. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* 1:3-32, 1921.
- [Fisher, 1922] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London Series A* 222:309-368, 1922.
- [Fisher, 1973] R. A. Fisher. *Statistical methods and scientific inference*. 3rd edition. Hafner, New York, 1973.
- [Forster, 2002] M. R. Forster. Predictive accuracy as an achievable goal of science", *Philosophy of Science* 69 (3):S124-S134, 2002.
- [Forster, 2006] M. R. Forster. Counterexamples to a likelihood theory of evidence. *Minds and Machines* 16:319-338, 2006.
- [Forster and Sober, 2004] M. Forster and E. Sober. Why Likelihood? Pages 153-190 in M. L. Taper and S. R. Lele, editors. *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. The University of Chicago Press, Chicago, 2004.
- [Fraser, 1963] D. A. S. Fraser. On the Sufficiency and Likelihood Principles. *J. Am. Stat. Assn* 58:641-647, 1963.
- [Frigg, 2006] R. Frigg. Scientific Representation and the Semantic View of Theories. *Theoria* 55:49-65, 2006.
- [Gemes, 2007] K. Gemes. Verisimilitude and content, *Synthese* 154 (2):293-306, 2007.
- [Giere, 1988] R. Giere. *Explaining Science*. University of Chicago Press, Chicago, 1988.
- [Giere, 1999] R. N. Giere. *Science without laws (Science and Its Conceptual Foundations)*. Chicago.: University of Chicago Press, 1999.
- [Giere, 2004] R. N. Giere. How models are used to represent reality, *Philosophy of Science* 71 (5):742-752, 2004.
- [Giere, 2008] R. N. Giere. Models, Metaphysics, and Methodology, in Stephan Hartmann, Luc Bovens and Carl Hoefer (eds.), *Nancy Cartwright's Philosophy of Science*: Routledge. Goldman, A. I. 1986. *Epistemology and Cognition*. Harvard University Press, Cambridge, MA, 2008.
- [Goldman, 2008] A. I. Goldman. Reliabilism in E. N. Zalta, editor. *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), <http://plato.stanford.edu/archives/fall2008/entries/reliabilism/>. The Center for the Study of Language and Information (CSLI), Stanford University, Stanford.
- [Goodman, 2008] S. N. Goodman. A dirty dozen: Twelve P-value misconceptions. *Seminars in Hematology* 45:135-140, 2008.
- [Hacking, 1965] I. Hacking. *Logic of statistical inference*. Cambridge University Press., Cambridge, 1965.
- [Harris, 1974] J. Harris. Popper's definition of 'Verisimilitude', *The British Journal for the Philosophy of Science*, 25: 160-166, 1974.
- [Hughes, 1997] R. I. G. Hughes. Models and Representation. *Philosophy of Science (Proceedings)* 64 325-336. 1997.
- [Jeffreys, 1961] H. Jeffreys. *Theory of probability*. Third edition. The Clarendon press, Oxford, 1961.

- [Lele, 2004] S. R. Lele. Evidence Functions and the Optimality of the Law of Likelihood. *in* M. L. Taper and S. R. Lele, editors. *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. The University of Chicago Press, Chicago, 2004.
- [Lele *et al.*, 2007] S. R. Lele, B. Dennis, and F. Lutscher. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters* **10**:551–563, 2007.
- [Lindsay, 2004] B. G. Lindsay. Statistical Distances as Loss Functions in Assessing Model Adequacy. Pages 439–488 *in* M. L. Taper and S. R. Lele, editors. *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. The University of Chicago Press, Chicago, 2004.
- [Lindsay *et al.*, 2007] B. G. Lindsay, M. Markatou, S. Ray, K. Yang, and S. Chen. Quadratic Distances On Probabilities: A Unified Foundation. Columbia University Biostatistics Technical Report Series 9, 2007.
- [Linhart and Zucchini, 1986] H. Linhart and W. Zucchini. *Model Selection*. John Wiley & Sons, 1986.
- [Mayo, 1996] D. G. Mayo. *Error and the Growth of Experimental Knowledge*. University of Chicago Press, Chicago, 1996.
- [Mayo, 2004] D. G. Mayo. An error-statistical philosophy of evidence. Pages 79–118 *in* M. L. Taper and S. R. Lele, editors. *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. The University of Chicago Press, Chicago, 2004.
- [Mayo and Cox, 2006] D. G. Mayo and D. R. Cox. Frequentist statistics as a theory of inductive inference. Pages 77–97 *in* *Optimality: The 2nd Lehmann Symposium*. Institute of Mathematical Statistics Rice University, 2006.
- [Mayo and Spanos, 2006] D. G. Mayo and A. Spanos. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science* **57**:323–357, 2006.
- [McCullagh and Nelder, 1989] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. 2nd edn. Chapman and Hall, London. Miller, D. 1972. The Truth-likeness of Truthlikeness. *Analysis*. **33**:50–55, 1989.
- [Miller, 1974a] D. Miller. On the Comparison of False Theories by Their Bases. *The British Journal for the Philosophy of Science* **25**:178–188, 1974.
- [Miller, 1974b] D. Miller. Popper’s Qualitative Theory of Verisimilitude. *The British Journal for the Philosophy of Science* **25**:166–177, 1974.
- [Miller, 2000] D. W. Miller. Sokal & Bricmont: Back to the Frying Pan. *Pli* **9**:156–173, 2000.
- [Miller, 2006] D. W. Miller. *Out Of Error: Further Essays on Critical Rationalism*. Ashgate, Aldershot, 2006.
- [Morgan, 1999] M. Morgan. Learning from Models, *in* M. Morrison and M. Morgan (eds.), *Models as Mediators: Perspectives on Natural and Social Science*, Cambridge: Cambridge University Press, 347–388, 1999.
- [Neyman and Pearson, 1933] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London Series A* 231:289–337, 1933.
- [Niiniluoto, 1998] I. Niiniluoto. Verisimilitude: The third period. *British Journal for the Philosophy of Science* **49**:1–29, 1998.
- [Nisbet, 1980] R. Nisbet. *History of the Idea of Progress*. Heinemann, London, 1980.
- [Oddie, 2007] G. Oddie. Truthlikeness. *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), <http://plato.stanford.edu/archives/fall2008/entries/truthlikeness/>.
- [Pickett *et al.*, 1994] S. T. A. Pickett, J. Kolasa, and C. G. Jones. *Ecological Understanding: The Nature of Theory and The Theory of Nature*. Academic Press, San Diego, 1994.
- [Platt, 1964] J. R. Platt. Strong Inference. *Science* **146**:347–353, 1964.
- [Ponciano *et al.*, 2009] J. M. Ponciano, M. L. Taper, D. Dennis, and S. R. Lele. Inference for hierarchical models in ecology: Confidence intervals, hypothesis testing, and model selection using data cloning. *Ecology* **90**:356–362, 2009.
- [Popper, 1963] K. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge. 1963.
- [Popper, 1935] K. Popper. *Logik der Forschung*. Vienna: Julius Springer Verlag, 1935.
- [Popper, 1959] K. Popper. *The Logic of Scientific Discovery*. London: Hutchinson. Original edition, 1959. (Translation of *Logik der Forschung*.)

- [Popper, 1976] K. Popper. A Note on Verisimilitude, *The British Journal for the Philosophy of Science* 27 (2):147-159, 1976.
- [Roush, 2006] S. Roush. Tracking Truth. Oxford University Press, Oxford, 2006.
- [Royall, 1986] R. M. Royall. The effect of sample-size on the meaning of significance tests. *American Statistician* 40:313-315, 1986.
- [Royall, 1992] R. M. Royall. The elusive concept of statistical evidence Pages 405-418 in J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors. *Bayesian Statistics 4*. Oxford University Press, Oxford, 1992.
- [Royall, 1997] R. M. Royall. *Statistical Evidence: A likelihood paradigm*. Chapman & Hall, London, 1997.
- [Royall, 2000a] R. M. Royall. On the Probability of Observing Misleading Statistical Evidence. *Journal of the American Statistical Association* 95:760-780, 2000.
- [Royall, 2000b] R. M. Royall. On the probability of observing misleading statistical evidence — Rejoinder. *Journal of the American Statistical Association* 95:773-780, 2000.
- [Royall, 2004] R. M. Royall. The Likelihood Paradigm for Statistical Evidence. Pages 119-152 in M. L. Taper, and S. R. Lele, editors. *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. The University of Chicago Press, Chicago, 2004.
- [Savage, 1976] L. J. Savage. On rereading R. A. Fisher (with discussion). *Annals of Statistics* 4:441-500, 1976.
- [Schwarz, 1987] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics* 6:461-464, 1987.
- [Sellke *et al.*, 2001] T. Sellke, M. J. Bayarri, and J. O. Berger. Calibration of p values for testing precise null hypotheses. *American Statistician* 55:62-71, 2001.
- [Strug and Hodge, 2006a] L. J. Strug and S. E. Hodge. An alternative foundation for the planning and evaluation of linkage analysis I. Decoupling 'error probabilities' from 'measures of evidence'. *Human Heredity* 61:166-188, 2006.
- [Strug and Hodge, 2006b] L. J. Strug and S. E. Hodge. An alternative foundation for the planning and evaluation of linkage analysis II. Implications for multiple test adjustments. *Human Heredity* 61:200-209, 2006.
- [Taper, 2004] M. L. Taper. Model identification from many candidates. Pages 448-524 in M. L. Taper and S. R. Lele, editors. *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. The University of Chicago Press, Chicago, 2004.
- [Taper and Gogan, 2002] M. L. Taper and P. J. P. Gogan. The Northern Yellowstone elk: Density dependence and climatic conditions. *Journal of Wildlife Management* 66:106-122, 2002.
- [Taper and Lele, 2004] M. L. Taper and S. R. Lele. The nature of scientific evidence: A forward-looking synthesis. Pages 527-551 in M. L. Taper and S. R. Lele, editors. *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. The University of Chicago Press, Chicago, 2004.
- [Thompson, 2007] B. Thompson. *The Nature of Statistical Evidence*. New York: Springer, 2007.
- [Tichý, 1974] P. Tichý. On Popper's definitions of verisimilitude, *The British Journal for the Philosophy of Science*, 25: 155-160, 1974.
- [Tukey, 1960] J. W. Tukey. Conclusions vs Decisions. *Technometrics* 2:423-433 1960.
- [van Fraassen, 1980] B. van Fraassen. *The Scientific Image*. Oxford University Press, Oxford, 1980.
- [van Fraassen, 2002] B. van Fraassen. *The Empirical Stance*. Yale University Press, New Haven and London, 2002.
- [Quine, 1951] W. V. O. Quine. Two dogmas of empiricism, *The Philosophical Review* 60:20-43, 1951.
- [Zucchini, 2000] W. Zucchini. An Introduction to Model Selection. *Journal of Mathematical Psychology* 44:41-61, 2000.
- [Zwart, 2001] S. D. Zwart. *Refined Verisimilitude*. Springer. 2001.

Akaikean Paradigm

This page intentionally left blank

AIC SCORES AS EVIDENCE: A BAYESIAN INTERPRETATION

Malcolm Forster and Elliott Sober

1 INTRODUCTION

Akaike [1973] helped to launch the field in statistics now known as model selection theory by describing a goal, proposing a criterion, and proving a theorem. The goal is to figure out how accurately models will predict new data when fitted to old. The criterion came to be called AIC, the Akaike Information Criterion:

$$AIC(M) = \log\text{-likelihood of } L(M) - k,^1$$

where the model M contains k adjustable parameters and $L(M)$ is the member of M obtained by assigning to the adjustable parameters in M their maximum likelihood values. Akaike's theorem is that AIC is an unbiased estimator of predictive accuracy, given some assumptions that are widely applicable.² The three parts of this achievement should be kept separate. Not conflating the goal and the criterion is important, since criteria other than AIC might do a better job in some circumstances in achieving the goal. The criterion and the theorem also need to be distinguished, since, in general, the fact that an estimator is unbiased does not suffice to show that it should be used.

The theorem that Akaike proved made it natural to understand AIC as a frequentist construct. AIC is a device for estimating the predictive accuracy of models just as a kitchen scale is a device for estimating the weight of objects. Bayesians assess an estimator by determining whether the estimates it generates are probably true or probably close to the truth. Evaluating how probable it is that a melon weighs two pounds, given that the scale says that it weighs two pounds, requires that one have a prior probability for the melon's weighing two pounds. Akaike's theorem says nothing about prior or posterior probabilities, so there was no reason to think of AIC in Bayesian terms. Rather, what Akaike did was what

¹Notice that AIC scores are negative numbers, with smaller likelihoods being "more negative" (i.e., farther from zero). Thus, here we adopt the convention used in [Forster and Sober, 1994] wherein higher AIC scores (those closer to zero) are better — they indicate a higher degree of predictive accuracy. In the statistics literature the AIC score is usually represented as -2 times the AIC formula we use here, so the opposite convention is followed — higher AIC scores are worse, because these scores indicate that the fitted model is more distant from the truth.

²See [Sakamoto *et al.*, 1986] for a thorough explanation and proof of Akaike's theorem.

frequentists generally do when they assess an estimator. They establish one or more of its “long-run operating characteristics.” The “long-run” average of an unbiased estimator of a quantity is, by definition, centered on the quantity’s true value. If you repeatedly weigh an object on an unbiased scale, the average of all the readings of the scale converges to the object’s true weight. On any given occasion, the scale’s reading may be too high or too low. To say that an estimator is unbiased leaves open what its variance is.

This fact about Akaike’s theorem is sometimes taken to cast doubt on AIC. But here we must be careful — there may be more to AIC than Akaike’s theorem established.

In fact, AIC isn’t just an unbiased estimator. The expected squared error of an unbiased estimator is strictly less than that of any other estimator that differs from it by a constant.³ AIC is in this respect a better estimator of predictive accuracy than BIC, the Bayesian Information Criterion first derived by Schwarz [1977]. BIC is both a biased estimator of predictive accuracy and has a higher expected squared error. To be sure, BIC was not developed as a method for estimating predictive accuracy, but rather was formulated as a method for estimating a model’s average likelihood. Nonetheless, the point is that there is more to be said on behalf of AIC than that it is unbiased. This does not prove that AIC is the best of all possible estimators of predictive accuracy. However, we suggest that estimators should be evaluated in the same way that empirical scientific theories are. Rather than asking whether general relativity is the best of all possible theories, we should ask whether it is better than the alternative theories that have been identified. The same goes for AIC.

The goal of this paper is not to provide a full-scale assessment of AIC, but to show that it is an estimator whose estimates should be taken seriously by Bayesians, its frequentist pedigree notwithstanding.⁴ Frequentists often maintain that the question of how an individual *estimate* should be interpreted is meaningless — that the only legitimate question concerns the long-term behavior of *estimators*. Bayesians reply that both questions are important and that the interpretation of individual estimates is pressing in view of the fact that a given estimate might be produced by any number of different estimation procedures (see, for example, [Howson and Urbach, 1993]). We agree with Bayesians on this point and so we will pose a question about individual estimates that frequentists typically disdain — do AIC scores provide evidence concerning how predictively

³**Proof:** Consider an estimator that differs from AIC by a constant c , and denote the true predictive accuracy a model has by AIC^* . AIC is an unbiased estimate of AIC^* if and only if $E[AIC] = AIC^*$, where E is “expected value.” We may now calculate the expected squared error of the new estimator using well-known properties of expected values: $E[(AIC + c - AIC^*)^2] = E[(AIC - AIC^*)^2] + c^2$. The “cross-term” is zero because AIC is unbiased. Therefore, the expected squared error of AIC is strictly less than the expected squared error of any estimator that differs from it by a non-zero constant. This includes BIC and any other estimator that is equal to the log-likelihood plus some penalty for complexity.

⁴By “frequentists” we are referring to the classical school of statisticians led primarily by Neyman and Pearson, and Fisher.

accurate a model will be?

2 ESTIMATES AS EVIDENCE

When Bayesians seek to explain why a positive pregnancy test is evidence that the person taking the test is pregnant, they inevitably refer to the Law of Likelihood [Hacking, 1965]:

Observation O favors hypothesis H_1 over hypothesis H_2 if and only if $P(O|H_1) > P(O|H_2)$.

This concept of “favoring” ignores the values of prior probabilities and focuses exclusively on the evidential meaning of the present observations, which is said to be captured by the likelihoods of the competing hypotheses. The odds version of Bayes’s theorem makes it clear that, for Bayesians, likelihoods are the sole vehicle by which the present observations can change one’s degree of belief in the hypotheses:

$$\frac{P(H_1|O)}{P(H_2|O)} = \frac{P(O|H_1)}{P(O|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

The ratio of posterior probabilities differs from the ratio of priors precisely when the likelihood ratio differs from unity. And the more the likelihood ratio differs from unity, the greater the transformation that the observations achieve.

The Law of Likelihood applies to point estimates of a continuous parameter just as much as it does to the presence or absence of a dichotomous trait like pregnancy. For example, if a thermometer, applied to object o , yields a reading $R(o) = x$, this is evidence that o ’s temperature, $T(o)$, is y rather than x precisely when

$$(T_1) \quad P[R(o) = y|T(o) = y] > P[R(o) = y|T(o) = x].$$

If AIC scores are like thermometer readings in this respect, an AIC score of x is evidence that the model has a predictive accuracy of x rather than y precisely when

$$(A_1) \quad P[\text{AIC}(M) = y|PA(M) = y] > P[\text{AIC}(M) = y|PA(M) = x],$$

where $PA(M)$ denotes the predictive accuracy of model M . Akaike’s theorem, that AIC is an unbiased estimator of predictive accuracy, does not entail (A_1) any more than a thermometer’s being an unbiased estimator of temperature entails (T_1) . For suppose a thermometer works like this: when an object has a temperature of y , the thermometer has 90% chance of reading $(y - a)$ and a 10% chance of reading $(y + b)$. The thermometer is an unbiased estimator of temperature precisely when

$$y = (0.9)(y - a) + (0.1)(y + b),$$

which is true if and only if

$$a/b = 1/9.$$

Suppose, for example, that $a = 10$ and $b = 90$. Then an object whose true temperature is $+3$ has a 90% chance of having the thermometer say its temperature is -7 and a 10% chance of having the thermometer say its temperature is $+93$. Since the average (expected) reading is $(0.9)(-7) + (0.1)(93) = 3$, the thermometer is unbiased. Yet, if an object's temperature is 3, the probability is zero that the thermometer will say that its temperature is 3. If the thermometer produces a reading of x , the maximum likelihood estimate of the object's temperature is $(x + 10)$. This unbiased thermometer has an asymmetric error distribution; its readings have a higher probability of being too low than too high. This is why (T_1) is false. The following proposition is true instead:

$$(T_2) \quad \begin{aligned} P[R(o) = y | T(o) = y + 10] &= 90\% > P[R(o) = y | T(o) = x], \\ &\text{for all } x, y \text{ such that } y + 10 \neq x \end{aligned}$$

The thermometer's readings provide evidence about temperature because (T_2) is true; the notion of evidence used here is the one sanctioned by the Law of Likelihood.

3 DIFFERENCES IN AIC SCORES

The reason we have labored the point that an unbiased estimator can have an asymmetric error distribution, and that this does not prevent its estimates from providing evidence, is that AIC is both an unbiased estimator and has an asymmetric error distribution. The proposition (A_1) is false, but AIC scores still provide evidence about the predictive accuracies of models.

Consider two models, M_1 and M_2 , where the first is nested in the second, meaning that M_1 is a special case of M_2 . Suppose that the AIC score of the second model is y units larger than the AIC score of the first. It turns out that this difference in AIC scores does not obey the following principle:

$$(A_2) \quad \begin{aligned} P[AIC(M_2) - AIC(M_1) = y | PA(M_2) - PA(M_1) = y] &> \\ P[AIC(M_2) - AIC(M_1) = y | PA(M_2) - PA(M_1) = x], & \\ \text{for all } x, y \text{ such that } x \neq y. & \end{aligned}$$

The falsity of (A_2) is shown by the different curves in Figure 1, which represents two nested models that differ by one in the number of adjustable parameters they have. Each curve in Figure 1 pertains to a difference in AIC scores; differences in AIC scores are observations, akin to the observation that one object has a higher thermometer reading than another. For any one observation, different hypotheses about the difference in the two models' predictive accuracies confer different probabilities on that observation. Each curve in Figure 1 is in this sense a likelihood function. The derivation of these curves for the general case of a pair of nested

models that differ by k in their number of adjustable parameters is given in the Appendix.

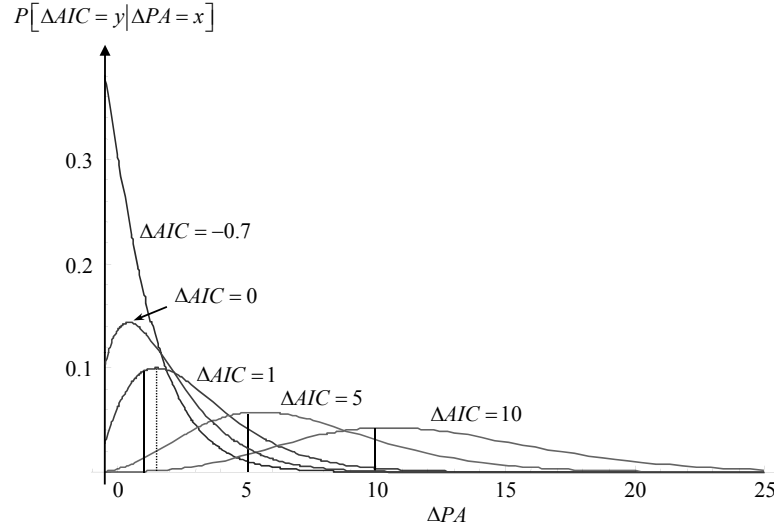


Figure 1. The graphs show the likelihood function for various values of the statistic ΔAIC . The meta-model describes how the difference in predictive accuracy of two nested models (call them object models) will affect the difference in their AIC scores. The meta-model has one adjustable parameter ΔPA , and each curve shows the likelihood of the hypothesis corresponding to the value of ΔPA read from the horizontal axis. In this case, ΔPA is the difference in predictive accuracy of two object models that differ in the number of adjustable parameters by only 1.

Notice in Figure 1 that the peak of the likelihood function for the observation that $AIC(M_2) - AIC(M_1) = y$ does not correspond to the hypothesis that $PA(M_2) - PA(M_1) = y$. Rather, the maximum likelihood hypothesis is that $PA(M_2) - PA(M_1) = x$, for some number x that is greater than y . This is easiest to see in Figure 1 by looking at the case where $\Delta AIC = 1$. This value is made more probable by $\Delta PA = 1.5$ than it is by $\Delta PA = 1$. The same point holds for the curves depicted in Figure 2, which again describes a pair of nested models, but this time one model has 10 more adjustable parameters than the other.

The following generalization holds. If $AIC(M_2) - AIC(M_1) > 0$, then there exists a positive number x such that

$$(A_3) \quad \begin{aligned} &P[AIC(M_2) - AIC(M_1) = y | PA(M_2) - PA(M_1) = x] > \\ &P[AIC(M_2) - AIC(M_1) = y | PA(M_2) - PA(M_1) = z], \text{ for all } z \neq x. \end{aligned}$$

AIC differences therefore favor some hypotheses about predictive accuracy over

others, assuming that the concept of evidential favoring is given by the Law of Likelihood.

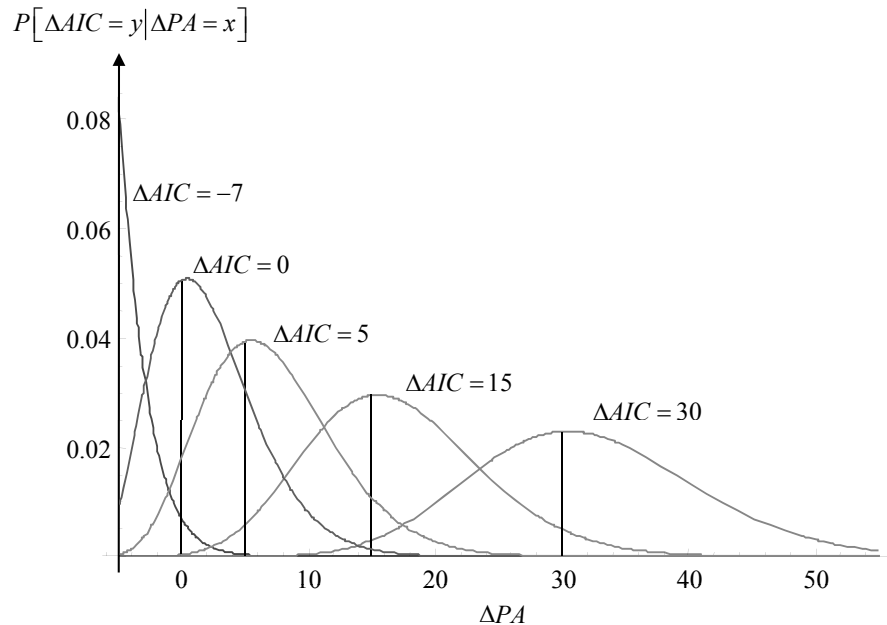


Figure 2. The graphs show the likelihood function for various values of the statistic ΔAIC . As in Figure 1, the meta-model has one adjustable parameter ΔPA , and each curve shows the probability of the statistic for various values of the parameter. In this case, ΔPA is the difference in predictive accuracy of two object models that differ in the number of adjustable parameters by 10.

There is a second interpretation of the results described in Figures 1 and 2 that Bayesians can extract. It is obtained by using Royall's [1997] concept of a likelihood *interval*. Royall supplements the Law of Likelihood by proposing a definition of what it means for an observation to provide *strong* evidence favoring H_1 over H_2 ; he suggests that this is true precisely when the likelihood ratio is at least 8. Royall uses this convention to draw an interval around the maximum likelihood point on a likelihood function that has the following property: the maximum likelihood value has a likelihood that is at least 8 times as large as that possessed by any point that is outside the interval. Interpreted in this way, some of the observations described in Figures 1 and 2 provide *strong* evidence favoring the maximum likelihood estimate (which is positive) over *any* negative estimate.⁵

⁵In Figure 2, $\Delta AIC = -7$ provides strong evidence that ΔPA is negative, although this does not happen in Figure 1.

Again, this fails to be true when the difference in AIC scores is very small.

A further assimilation of AIC scores to the Bayesian framework can be achieved by adding a prior distribution over differences in predictive accuracies. This, plus the likelihood function associated with the observed difference in AIC scores, allows one to calculate the posterior probability (density) that one model's predictive accuracy exceeds another's by x . One also can calculate the posterior probability that the difference in predictive accuracies is positive.

There is a common puzzle about Bayesianism that has been clarified here. Bayesians often talk about comparing the probabilities of nested models. But, in practice, they compare nested models in terms of the ratio of their likelihoods, the Bayes factor, as in [Schwarz, 1978], which may be viewed as an application of LL. This makes sense because the Bayes factor can favor either model. Unfortunately, this comparison contradicts a more fundamental Bayesian comparison in terms of posterior probabilities if the first model entails the second, for then the probability of the first model can never be greater than the second. We have shown that the problem disappears as soon as Bayesians focus on the predictive accuracy of the models, rather than their truth. For now it is possible for the predictive accuracy of the first model to be higher than the second model, even if it is impossible for the first model to be true while the second model is false. Bayes factors are no longer involved in the comparison; now we are comparing meta-hypotheses about predictive accuracy. This shows why Bayesians need not be threatened, or puzzled, when they see scientists comparing models that are related by strict logical entailment.⁶

4 CONCLUSION

AIC began life with Akaike's [1973] theorem, which established that AIC is an unbiased estimator of predictive accuracy. Because this proof described a long-run operating characteristic of the estimator, and not the evidential meaning of the particular estimates that the estimator might provide, and also because prior and posterior probabilities were not part of the framework used by Akaike and his school (see, for example, [Sakamoto *et al.*, 1986]), AIC came to be viewed as a frequentist construct. However, these facts about earlier defenses of AIC do not establish that AIC scores are meaningless from a Bayesian point of view. We have argued that AIC scores provide evidence concerning the predictive accuracies of models in the sense of "evidence" sanctioned by the Law of Likelihood, which is a central Bayesian principle.⁷ AIC scores are no more essentially tied to frequentism than thermometer readings are.

⁶We are grateful to Prasanta Bandyopadhyay for suggesting this point.

⁷The result reported here pertains to any pair of nested models. The case of non-nested models remains to be investigated.

APPENDIX

Suppose that we are interested in corn crop yields in a field that has been divided into numerous plots. The plots are subjected to different conditions, they have different drainage properties, different exposures to the wind, slightly different soils, and perhaps they are treated and irrigated differently as well. The question is whether these different conditions affect the height of the corn plants. Assume that the height of each plant i is represented by a random variable X_i whose value is denoted by x_i . Also to simplify the mathematics, we assume that the random variables for plants in the same plot are independent and identically distributed with a normal distribution and unit variance; $X_i \sim N(\mu_{j(i)*}, 1)$, where $j(i)$ denotes the plot number in which plant i is found. The $*$ indicates that this is the true value of the mean, not something hypothetical. Sometimes we will just write μ_{j*} .

The various null hypotheses one might consider here are *false* because they falsely assert that two or more plots have the same mean, whereas in fact all the μ_{j*} are different (although only slightly in some cases). A typical hypothesis is false, but it's not false that it has some predictive accuracy. The model under test assigns a degree of predictive accuracy to some model about corn yields. This meta-model attributes a property to the object model. The object model is false, but the meta-model may be true.

Suppose that there are only three plots, with mean values $\mu_1^*, \mu_2^*, \mu_3^*$. An arbitrary hypothesis asserts that the mean values are μ_1, μ_2, μ_3 for some particular triple of numbers. A model will leave these as adjustable parameters, but possibly add constraints such as $\mu_1 = \mu_2$. Each such constraint will reduce the number of adjustable parameters by one. For example, the model that (falsely) asserts that the three plots have the same mean yield has one adjustable parameter. It asserts for all plants, $X_i \sim N(\mu, 1)$, for some unknown parameter μ .

Let's say that there are n_1 plants sampled from plot 1, n_2 plants from plot 2, n_3 plants from plot 3, with a total of $n = n_1 + n_2 + n_3$ plants. The log-likelihood of any particular hypothesis in the object model (picked out by a particular triple of numbers μ_1, μ_2, μ_3) is

$$l(\mu_1, \mu_2, \mu_3) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n_1} (x_i - \mu_1)^2 - \frac{1}{2} \sum_{i=n_1+1}^{n_1+n_2} (x_i - \mu_2)^2 - \frac{1}{2} \sum_{i=n_1+n_2+1}^n (x_i - \mu_3)^2.$$

Since μ_1, μ_2, μ_3 are constants, $-2l(\mu_1, \mu_2, \mu_3) - n \log 2\pi$ is the sum of the squares of n normal random variables of unit variance and mean $\mu_1^* - \mu_1, \mu_2^* - \mu_2$, or $\mu_3^* - \mu_3$, depending on the plot in which the plant is sampled. That means that $-2l(\mu_1, \mu_2, \mu_3) - n \log 2\pi$ is a chi-square random variable with n degrees of freedom, and a non-centrality parameter equal to the sum of the squares obtained by substituting the mean value for each random variable in the quadratic. That is,

$$-2l(\mu_1, \mu_2, \mu_3) - n \log 2\pi \sim \chi^2(n, \lambda),$$

where

$$\lambda = n_1 (\mu_1^* - \mu_1)^2 + n_2 (\mu_2^* - \mu_2)^2 + n_3 (\mu_3^* - \mu_3)^2.$$

Since the mean of a chi-square random variable is equal to the number of degrees of freedom (n in this case) plus the non-centrality parameter, we can calculate the predictive accuracy of any hypothesis, where the predictive accuracy of such a hypothesis is, by definition, its expected log-likelihood.

$$PA(\mu_1, \mu_2, \mu_3) = -\frac{n}{2} (\log 2\pi + 1) - \frac{1}{2} \left(n_1 (\mu_1^* - \mu_1)^2 + n_2 (\mu_2^* - \mu_2)^2 + n_3 (\mu_3^* - \mu_3)^2 \right).$$

Note that this is the predictive accuracy of a “point” hypothesis or non-composite hypothesis.

Now consider the hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu$, which has one adjustable parameter. This is a composite hypothesis, not a point hypothesis. Since this distinction is important, we will mark it by always referring to a composite hypothesis as a *model*, and never referring to point hypotheses as models. This is because the equality above does not specify any particular numerical value for μ . The question at hand is: How should we define the predictive accuracy of a model? If we fit a model to the actual set of data, we get a point hypothesis, namely the maximum likelihood hypothesis that belongs to the model. Its predictive accuracy is already well defined, but we do not want to define the predictive accuracy of the model directly in terms of this number because the actual data set may not be typical. So, we imagine that we repeatedly generate data sets of the same size as the actual data set, and define the predictive accuracy of the model as the average predictive accuracy of the maximum likelihood hypotheses generated in this random way. The predictive accuracy of a model is therefore the predictive accuracy of a “typical” maximum likelihood hypothesis.

From a mathematical point of view, things just got complicated because there is now a double expectation involved in calculating the predictive accuracy of a model. First, we take the general formula for $PA(\mu, \mu, \mu)$, which is itself defined as an expected value of a particular point hypothesis, $\mu_1 = \mu_2 = \mu_3 = \mu$, where we are thinking of μ as a fixed number. Recall that the predictive accuracy, PA , is defined here as the expected log-likelihood of the point hypothesis relative to some newly generated data set, which we may call test data. How well the hypothesis fits this new data is naturally thought of as a measure of its predictive accuracy. The next step is to put μ equal to the maximum likelihood estimate of μ determined by the actual data set, which is denoted by $\hat{\mu}$. Thus, we get $PA(\hat{\mu}, \hat{\mu}, \hat{\mu})$. But $\hat{\mu}$ may not be typical, so we want to average $PA(\hat{\mu}, \hat{\mu}, \hat{\mu})$ over all possible values of $\hat{\mu}$ that we would obtain if we were to draw data from the same (unknown) probability distribution. This average is defined as the expected value of $PA(\hat{\mu}, \hat{\mu}, \hat{\mu})$ as determined by the value of $\hat{\mu}$ that would be obtained from any data set that could be used to initially fit the model. We think of this data set as a *calibration data set* because it is used to fix the values of adjustable parameters. The calibration data is conceptually quite different from a test data set. To define the notion of

prediction, these data sets must be kept separate, and it is therefore essential that we define the predictive accuracy of a model as a double expectation.

The maximum log-likelihood of the model is found by finding the maximum likelihood estimate of the adjustable parameter μ , which we denote by $\hat{\mu}$, and then writing down the log-likelihood of the particular hypothesis corresponding to this value of μ . The answer is, clearly,

$$\text{maximum log-likelihood} = l(\hat{\mu}) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^n (x_i - \hat{\mu})^2.$$

Now think of $\hat{\mu}$ as a quantity that varies randomly as we consider different sets of calibration data. It is well known (e.g., see [Hogg and Craig, 1978, p. 279]) that

$$\sum_{i=1}^n (x_i - \hat{\mu})^2 \sim \chi^2(n-1, \lambda^*)$$

is a non-central chi-square distribution with $n-1$ degrees of freedom, whose non-centrality parameter is calculated by substituting the mean values for each random variable in the quadratic. The non-centrality parameter is therefore

$$\lambda^* = n_1 (\mu_1^* - \mu^*)^2 + n_2 (\mu_2^* - \mu^*)^2 + n_3 (\mu_3^* - \mu^*)^2,$$

where

$$\mu^* \equiv \frac{n_1}{n} \mu_1^* + \frac{n_2}{n} \mu_2^* + \frac{n_3}{n} \mu_3^*.$$

Notice what went on here. We began with an object model, and considered its log-likelihood relative to the actual data. Then we introduced a meta-hypothesis about how the log-likelihood would behave statistically if it were determined by a new data set of the same size. In other words, we have treated $\hat{\mu}$ as well as $l(\hat{\mu})$ as random variables, and we have constructed a hypothesis about its statistical behavior. The meta-hypothesis and the object model are different (one may be true, while the other is false).

Now let's consider the general case in which there are n plants sampled from m plots ($n > m$) in numbers $n_1, n_2, \dots, n_m$, where the hypothesis under consideration asserts that the means of two or more plots are equal. We can think of the plots being partitioned into clusters; sometimes there may be one plot in a cluster, or two plots in a cluster, and so on. Obviously, there are fewer clusters than plots, except in the special case in which there is only one plot in any cluster. In that case, the model has m adjustable parameters. In the other extreme case, in which all the plots are in one cluster, there is one adjustable parameter. In general, there are k adjustable parameters, where k is the number of clusters introduced by the hypothesis; clearly $\{1 \leq k \leq m\}$. Let's denote the maximum log-likelihood of the model with k adjustable parameters by $l(\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_k)$. Then

$$-2l(\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_k) - n \log 2\pi \sim \chi^2(n-k, \lambda^*),$$

where the non-centrality parameter λ^* is calculated in the same way as before. The $*$ reminds us that this is a meta-hypothesis about how the log-likelihood of the object model behaves.

According to Akaike's theorem, $\text{AIC} = l(\hat{\mu}_1, \hat{\mu}_1, \dots, \hat{\mu}_k) - k$ is an unbiased estimate of the predictive accuracy of the model M . That is, $E^*[\text{AIC}(M)] = \text{AIC}^*(M)$, where the expectation is taken with respect to the true distribution. But we have just learned that

$$\text{AIC}(M) \sim -\frac{1}{2}\chi^2(n-k, \lambda^*) - \frac{1}{2}n \log 2\pi - k,$$

where the mean value of $\chi^2(n-k, \lambda^*)$ is $n-k+\lambda^*$. Therefore,

$$PA(M) = -\frac{1}{2}(n+k+\lambda^*) - \frac{1}{2}n \log 2\pi.$$

Observe that $PA(M)$ becomes infinitely large as the number of data, n , tends to infinity. It is not standard in statistics to estimate a quantity that grows infinitely large as the sample size increases, and this has led to considerable confusion about the properties of AIC (see [Forster, 2000; 2001], for detailed discussions). But, since this issue is not being discussed here, we shall refer to the *per datum* predictive accuracy when the need arises, and reserve PA to refer to the per datum predictive accuracy times the number of data n .

Therefore, our result is written

$$PA(M) = -C - \frac{1}{2}\lambda^* - \frac{1}{2}k,$$

where $C = \frac{1}{2}(\log 2\pi + 1)n \simeq 1.41894n$. So defined, predictive accuracy is the expected log-likelihood for newly sampled data that are generated by the true hypothesis, since the true hypothesis has no adjustable parameters and zero non-centrality parameter (the postulated mean values are the true values).⁸ Thus, $-C$ is the highest predictive accuracy that could be achieved by any *point* hypothesis in the model.

The constant term is unimportant for our purposes because it is the same for every model. The other terms are negative because the predictive accuracy is expected log-likelihood, and the log-likelihood is negative whenever the likelihood is less than 1. Nevertheless, it is important to understand that higher predictive accuracies (less negative) are better than lower predictive accuracies (more negative).

Firstly, the higher the value of $\frac{1}{2}\lambda^*$, the lower the predictive accuracy. To interpret this fact, notice that $\frac{1}{2}\lambda^*/n$ is independent of the total number of data because λ^* grows proportionally to n . In our simple case, we could write

$$\lambda^*/n = \frac{n_1}{n}(\mu_1^* - \mu^*)^2 + \frac{n_2}{n}(\mu_2^* - \mu^*)^2 + \frac{n_3}{n}(\mu_3^* - \mu^*)^2,$$

so if the proportions n_1/n , n_2/n , and n_3/n are held fixed, then λ^*/n is fixed. Therefore $-(C + \frac{1}{2}\lambda^*)/n$ is the per datum predictive accuracy achieved by the

⁸Equivalently, if the non-centrality parameter is not zero, then the hypothesis is false.

model as the number of data tends to infinity. Alternatively, we could say that $-(C + \frac{1}{2}\lambda^*)$ is the predictive accuracy that could be obtained if the parameter estimation were free of error. It is convenient to refer to $(C + \frac{1}{2}\lambda^*)$ as the *bias of a model*. Bias is bad and predictive accuracy is good, so the smaller the bias of a model, the higher the predictive accuracy that it could potentially achieve.

For small data sets especially, the term $\frac{1}{2}k$ is important because it measures the effect of the sampling errors involved in estimating the parameters of the model. The fewer the adjustable parameters, the greater the amount of data that can be used to estimate the value of each parameter. The negative effect that estimation errors has on the predictive accuracy of the maximum likelihood hypothesis is referred to as the *variance* of the model. Variance is bad and predictive accuracy is good, so they have opposite signs.

In summary, the predictive accuracy of a model can be defined in terms of its bias and variance, thus

$$PA = -(\text{bias} + \text{variance}),$$

where bias and variance are positive quantities. The minus sign reflects the fact that both bias and variance are bad for predictive accuracy. Bias measures the predictive accuracy of the best hypothesis in the model (the hypothesis that would be selected if there were no variance) while variance measures the loss in predictive accuracy that comes about by estimation errors. Variance is reduced in simpler models, but the bias is greater. So, maximizing predictive accuracy (if that is the goal) involves a tradeoff between bias and variance.

The predictive accuracy of a model is unknown because the model bias is unknown, and that's because the non-centrality parameter of the generating distribution is unknown. But hypotheses about its value can be tested, its value can be estimated, and confidence intervals can be inferred.

It may seem that our work is done, but we also want to know how *differences* in AIC scores provide evidence about *differences* in predictive accuracy. This is achieved by showing that differences in the maximum log-likelihoods are also governed by a non-central chi-square distribution. Or at least, we can show this when the models are nested (i.e., when one model is a special case of the other). First, we need the following theorem (which is a special case of a well known theorem proved, for example, in Hogg and Craig, 1978, p. 279).

THEOREM 1. *Let $Q_1 = Q_2 + Q$, where Q_1 , Q_2 , and Q are 3 random variables that are quadratic forms in n mutually stochastically independent random variables that are each normally distributed with means $\mu_1^*, \mu_2^*, \dots, \mu_n^*$ and unit variance. Let Q_1 and Q_2 have chi-square distributions with degrees of freedom r_1 and r_2 , respectively, and let Q be non-negative. Then:*

- (a) Q_2 and Q are mutually stochastically independent, and
- (b) Q has a chi-square distribution with $r = r_1 - r_2$ degrees of freedom.

To apply the Theorem, note that

$$Q_2 = -2l(\hat{\mu}_1, \hat{\mu}_2) - n \log 2\pi = \sum_{i=1}^{n_1} (x_i - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2} (x_j - \hat{\mu}_2)^2$$

is a chi-squared random variable with $n - 2$ degrees of freedom. Then note that

$$Q_1 = -2l(\hat{\mu}) - n \log 2\pi = \sum_{k=1}^n (x_k - \hat{\mu})^2$$

is a chi-squared random variable with $n - 1$ degrees of freedom. A long, but straightforward, calculation shows that

$$\Delta l = l(\hat{\mu}_1, \hat{\mu}_2) - l(\hat{\mu}) = \frac{n_1}{2} (\hat{\mu} - \hat{\mu}_1)^2 + \frac{n_2}{2} (\hat{\mu} - \hat{\mu}_2)^2.$$

From this, we prove that

$$Q_1 = Q_2 + 2\Delta l.$$

From the Theorem, it follows that $2\Delta l$ is a chi-square random variable with one degree of freedom, which has a non-centrality parameter, call it $\Delta\lambda$, found by substituting the mean values for each variable in the quadratic. That is,

$$\Delta\lambda = n_1 (\mu^* - \mu_1^*)^2 + n_2 (\mu^* - \mu_2^*)^2,$$

where, by definition,

$$\mu^* = \frac{n_1}{n} \mu_1^* + \frac{n_2}{n} \mu_2^*.$$

So, the evidential problem reduces to something simple. The statistic $2\Delta l = 2(l_2 - l_1)$ provides evidence about the predictive accuracy of M_2 compared to M_1 . Recall that models with lower numbers of adjustable parameters have higher biases, which confirms that $\Delta\lambda$ has the same sign as Δbias . However, in order to keep the difference in the variances positive, we need to define

$$\Delta\text{variance} = \text{variance}(M_2) - \text{variance}(M_1) = k_2 - k_1 = \Delta k,$$

because the more complex model has a greater variance than the simpler model.

Now define the difference in predictive accuracy as the advantage that the complex model has over the simpler model. That is,

$$\Delta PA = PA(M_2) - PA(M_1).$$

Then the tradeoff between bias and variance is expressed by the formula

$$\Delta PA = \Delta\text{bias} - \Delta\text{variance} = \frac{1}{2}\Delta\lambda - \frac{1}{2}\Delta k.$$

Since the difference in AIC scores is straightforwardly calculated from Δl , it too provides evidence for hypotheses about the difference in predictive accuracy. All

we require is that any particular hypothesis about the value of ΔPA is associated with a particular chi-square distribution, which requires that we know the degrees of freedom and find the non-centrality parameter. The degree of freedom is Δk , which is the difference in the number of adjustable parameters, whereas the non-centrality parameter is given by

$$\Delta\lambda = \Delta k + 2\Delta PA.$$

From this, one can make statistical inferences about differences in predictive accuracy from differences in AIC scores, for we know that $2\Delta l \sim \chi^2(\Delta k, \Delta\lambda)$.

Note that, since $\Delta\lambda \geq 0$, $\Delta PA \geq -\Delta k/2$. That is, simpler models are limited in the advantage in predictive accuracy they can have over more complex models. More interestingly, the advantage is constrained to become closer and closer to zero as the number of data increases. Thus, in the large sample limit, there is no difference in predictive accuracy.

In summary, given any value of ΔPA , Δl is the value of a random variable with the distribution $\frac{1}{2}\chi^2(\Delta k, \Delta\lambda)$. So, ΔAIC is a random variable with the distribution $\frac{1}{2}\chi^2(\Delta k, \Delta k + 2\Delta PA) - \Delta k$, whose expected value is ΔPA . Or, in other words, ΔAIC is a random variable with a known distribution whose expected value is ΔPA , which is the content of Akaike's theorem. But this result goes far beyond Akaike's theorem by providing the probability distribution that one can use whatever method of statistical inference about ΔPA one may wish to deploy.

Here is a final remark about what it means for AIC to be an unbiased estimator of predictive accuracy. It means that if we fix ΔPA and resample the calibration data, we will get values of the statistic ΔAIC whose mean value is equal to ΔPA . It does not mean that for any value of ΔAIC , the mean value of ΔPA will be equal to ΔAIC . Not only would such a statement depend on the assignment of a Bayesian prior distribution over values of ΔPA , but in some cases there is no prior that could even make it true. To see this, suppose ΔPA has its lowest possible value of $-\frac{1}{2}\Delta k$. The statistic ΔAIC will be higher in value sometimes *and lower in value sometimes*, which means that sometimes it will be lower than the lowest possible value of ΔPA , even though statisticians will still say that ΔAIC is an unbiased estimator of ΔPA .

ACKNOWLEDGMENTS

We thank Jim Hawthorne and Prasanta Bandyopadhyay for useful suggestions.

BIBLIOGRAPHY

- [Akaike, 1973] H. Akaike. Information Theory as an Extension of the Maximum Likelihood Principle. In B. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 267-281, 1973.
- [Forster, 2000] M. R. Forster. Key Concepts in Model Selection: Performance and Generalizability, *Journal of Mathematical Psychology* **44**: 205-231, 2000.

- [Forster, 2001] M. R. Forster. The New Science of Simplicity. In A. Zellner, H. A. Keuzenkamp, and M. McAleer (eds.) *Simplicity, Inference and Modelling*. Cambridge University Press, pp. 83-119, 2001.
- [Forster and Sober, 1994] M. R. Forster and E. Sober. How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions. *British Journal for the Philosophy of Science* **45**: 1-36, 1994.
- [Hacking, 1965] I. Hacking. *The Logic of Statistical Inference*. Cambridge: Cambridge University Press, 1965.
- [Howson and Urbach, 1993] C. Howson and P. Urbach. *Scientific Reasoning – the Bayesian Approach*. Peru, IL: Open Court, 1993.
- [Royall, 1997] R. Royall. *Statistical Evidence – a Likelihood Paradigm*. Boca Raton: Chapman and Hall, 1997.
- [Sakamoto *et al.*, 1986] Y. Sakamoto, M. Ishiguro, and G. Kitagawa. *Akaike Information Criterion Statistics*. New York: Springer, 1986.
- [Schwarz, 1978] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics* **6**: 461-465, 1978.

This page intentionally left blank

Part IV

The Likelihood Principle

This page intentionally left blank

THE LIKELIHOOD PRINCIPLE

Jason Grossman

1 INTRODUCTION

The likelihood principle (LP) is a normative principle of very great generality for evaluating statistical inference procedures. The LP can be proved from arguably self-evident premises¹; indeed, it can be proved to be logically equivalent to these premises, as we will see. It tells us, to give a rough first pass, that *inferences from data to hypotheses should depend on how likely the actual data are under competing hypotheses*, not on how likely imaginary data would have been under a single “null” hypothesis nor on any other properties of merely possible data.²

The plan of this chapter is to motivate and prove a precise version of the LP, with a number of caveats; briefly mention some alternative versions; and reply to an interesting objection.

Let’s start with an example (see Table 1).

A standard significance test would reject hypothesis 1 at the 5% level, reasoning that if hypothesis 1 were true then the observation, with a probability of only 3%, would be rather unlikely.

The sense in which 3% is to be considered unlikely, on this view, is of course that it is much less than 100%, which is the total probability of the top row of the table (since all possible observations are supposed to be represented, with the final cell representing a catch-all case).

The importance of the likelihood principle is that it tells us that this is not the relevant comparison. Instead, we should be comparing what the various hypotheses say about what actually occurred. In this case, hypothesis 1 says that what occurred was unlikely (probability of 3%), but the other hypotheses say that it

¹the weak sufficiency principle and the weak conditionality principle, defined below

²A more precise statement of the LP is as follows. (I unpack this statement later in this chapter.) When analysing the result of an observation, provided that the *choice* of observations is not informative about our hypotheses, and provided that the loss (equivalently, utility) function (if we have one) is independent of the observation, *inference procedures which make inferences about simple hypotheses should not be justified by appealing to probabilities assigned to observations which have not occurred*, except for the trivial constraint that those probabilities place on the probability of the actual observation under the rule that the probabilities of exclusive events cannot add up to more than 1. Therefore the only component of $p(x|h)$ which should be used to justify such inferences is the likelihood function $p(x_{\text{obs}}|h)$, where x ranges over the sample space, h ranges over the hypothesis space and x_{obs} is the actual observation.

	actual observation	merely possible observation 1	merely possible observation 2	other possible observations
hypothesis 1	0.03	0.2	0.5	0.27
hypothesis 2	0.001	0.01	0.95	0.039
catch-all hypothesis	0.001	0.001	0.001	0.997

Table 1.

was even *more* unlikely (total probability of 0.2%). The LP does not tell us how to compare these options — for example, whether to use other knowledge we might have about the hypotheses — and it is possible that a statistical method compatible with the LP will give the same answer as the standard significance test (reject hypothesis 1). But it cannot do so by using any of Table 1 except for the first column. And in general it is very common for methods compatible with the LP (most notably Bayesianism) to give different results from methods incompatible with the LP.

A very informal argument in favour of the likelihood principle is that if the observation is much more likely on hypothesis 1 than it is on all the other hypotheses put together, as it is in this example, then that is *prima facie* evidence in favour of hypothesis 1. But on a significance test the observation counts against hypothesis 1. Something seems to be wrong here, and the LP puts its finger on exactly what is wrong: that the significance test takes into account observations which would have been very relevant to hypothesis 1 had they occurred, but which did not occur.

All of the methods incompatible with the LP which are in common use are frequentist methods, which is to say that they all analyse tables such as the example table by considering whole rows. (It does not follow that all frequentist methods are incompatible with the LP. More on this below.) As we have seen, a significance test uses the single row representing the null hypothesis (see Table 2).

Other frequentist methods may use more than one row, but they always use *whole* rows. In contrast, a method compatible with the LP always justifies its inferences by appeals to the single column representing the data actually observed (see Table 3).

Before I state the mathematical apparatus which I need in order to present the LP properly, I would like to discuss informally the very basic contrast which the LP is meant to formalise. In the next two sections, I will consider two different things which we might want to do when we evaluate a statistical inference procedure: we might want to count the number of times (in different situations) it is right, on

	actual observation	possible observation 1	possible observation 2	...
hypothesis H_0	$p_{0,a}$	$p_{0,1}$	$p_{0,2}$	...
hypothesis H_1	$p_{1,a}$	$p_{1,1}$	$p_{1,2}$	...
hypothesis H_2	$p_{2,a}$	$p_{2,1}$	$p_{2,2}$	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 2. Data analysed by a frequentist method will use whole rows in some way

	actual observation	possible observation 1	possible observation 2	...
hypothesis H_0	$p_{0,a}$	$p_{0,1}$	$p_{0,2}$	...
hypothesis H_1	$p_{1,a}$	$p_{1,1}$	$p_{1,2}$	...
hypothesis H_2	$p_{2,a}$	$p_{2,1}$	$p_{2,2}$	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 3. Analysis compatible with the LP

the assumption that some hypothesis or other is true (frequentism); or we might want to compare what it says about various hypotheses in the same situation (as the LP advises).

1.1 *One option: frequentism*

One way to evaluate a statistical inference procedure is to see how often it gets the right answer. But what does “how often” mean? It might mean that we should work out the number of times we should expect a given inference procedure to get the right answer, in some set of hypothetical test cases. If we do this in the same way we would for a deductive inference procedure, we will start with some known true premises and see how often the inference procedure infers true (and relevant) conclusions from them. Now, before we can embark on such an evaluation, we have to decide what types of conclusions we want the statistical inference procedure to infer. Perhaps, if it is going to be a useful procedure, we want it to infer some general scientific hypotheses. We might then evaluate it by asking how often it correctly infers the truth of those hypotheses, given as premises some other general hypotheses and some randomly varying observational data. We can imagine feeding into the inference procedure random subsets of all the possible pieces of observational data, and we can calculate the proportion of those subsets on which it gets the right answer. This method is the frequentist or “error-rate” method.

1.2 *Another option: the LP*

It might seem as though the frequentist method were the only way of finding something loosely analogous to the logician’s method for testing deductive inferences. In order to see whether it is the only way, consider what information is available to us when we are getting ready to use a statistical inference procedure in real life. Some of our premises at that time will be general statements about the way the world is, of the nature of scientific hypotheses. The rest of our premises will be statements about specific observed phenomena. The distinction between these two — fuzzy though it inevitably is — is fundamental to the nature of statistical inference. The most common epistemological goal of science is to make inferences from the latter to the former, from observations to hypotheses. (Not that this is the only goal of science.) And in order for this to be *statistical* inference, the hypotheses must be probabilistic (and not deductively entailed by the premises). In short, when we need a statistical inference procedure it is because we have collected some data and we want to infer something from the data about some hypotheses.

What we want to know in such a situation is how often our candidate statistical inference procedure will allow us to infer truths, and we want to calculate this by comparing its performance to the performance of other possible procedures in the same situation, with the *same* data as part of our premises. This is what we do when we follow the LP.

If frequentist methods were always compatible with the LP then we would have a large constituency of people who agreed on how to evaluate statistical inference procedures. Perhaps they would be right, and if so we could pack up and go home. In fact, it is possible for the two options to agree numerically in some specific cases. But I hope I have made it clear that they disagree about what the norms of inference are. The frequentist method is to evaluate the performance of an inference procedure *only* on (functions of subsets of) all the possible pieces of observational data, while the LP tells us to evaluate its performance *only* on the data actually observed. So the very concept of “the performance of an inference procedure” is a completely different animal according to these two options.

2 CALCULATION, JUSTIFICATION AND CLASSIFICATION

To repeat, the likelihood principle and frequentism are two doctrines about the justification of inference procedures. The LP may appear to rule out frequentist inference procedures by denying that they are justified. But there is a back door by which some of them can creep back in. Given a specific frequentist inference procedure, we can ask: can this inference procedure, which hitherto has been justified by frequentist means, *also* be justified by an argument compatible with the LP? It sometimes transpires that specific brands of frequentism and specific brands of LP-compatible inference, in specific contexts, *give the same answers as each other*. For example, many frequentist estimation procedures are compatible with many Objective Bayesian analyses (see the chapters on Objective Bayesianism in this volume and [Bayarri and Berger, 2004, pp.63]). (All Bayesian methods obey the LP.³)

Unfortunately, since frequentism and LP inference are both broad churches there is little more that can be said concisely about the extent to which they agree.

Although the LP does not tell us how to do statistical inference, and although it does not completely rule out frequentist inference, it does rule out many specific inferences. Also, it allows us to categorise methods of statistical inference in a very natural and powerful way: a way which is more abstract and more general than the usual ways of classifying statistical theories. In particular, it divides theories of statistical inference into just two interesting, exhaustive and mutually exclusive categories. Note that the traditional polarisation of statistical theories into the Bayesian and the frequentist does not do this, since there are methods which are neither Bayesian nor frequentist, such as maximum likelihood estimation [Fisher, 1930a, p.531] and pivotal inference [Barnard, 1985].

³All strictly Bayesian methods obey the LP [Jeffreys, 1961, p.57], [Lindley, 1965, p.59], [Good, 1983, pp.35–36, p.132], although a few methods with “Bayes” in their names contravene both Bayes’s Theorem and the LP (to some extent at least) by counting the same data twice. Examples of this are Empirical Bayes methods [Robbins, 1955; Lindley, 1983] and Objective Bayesian methods with priors chosen as a function of the likelihood function [Berger, 2006, pp.399–400].

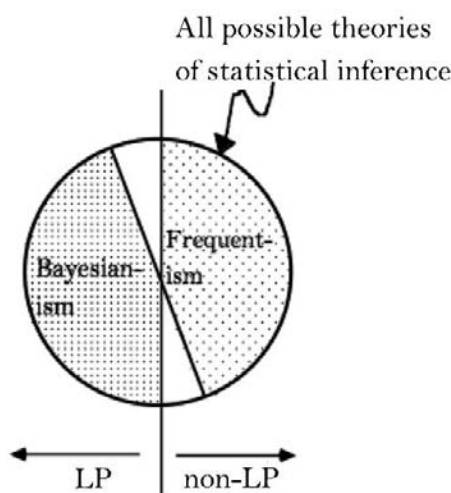


Figure 1. All possible theories of statistical inference

This diagram includes future, as yet unthought of, theories of statistical inference, as well as existing ones, assuming that future theories have roughly the same form as existing theories (i.e., that they are based on probabilistic equations formulated in terms of data and hypotheses).

The LP also captures some of the most attractive features of Bayesianism, while leaving open the question of whether a subjective prior should be used (and leaving open the question of exactly what “subjective” means). Since it provides a lot of common ground between factions of Bayesians, the LP is a good, irenic starting point for agreement between them, and possibly between factions of philosophers of statistics in general.

3 TERMINOLOGY AND CAVEATS

In this section I define the terms I will need to state the LP properly, and state some important caveats.

First of all, what do I mean by “inference procedures”? I will assume that we all know what a procedure (simpliciter) is. There are perhaps issues about ad hoc or ill-defined procedures, and about the difference (if any) between a deliberate experiment and a mere (perhaps accidental) observation, but I will leave those issues to one side for lack of space. The remaining question is what counts as an inference.

Because the LP does not take into account a utility or loss function (see discussion of this below), the LP does not give us a decision theory. Because it does not take into account prior probabilities, it does not even give us a full theory of statistical inference. The LP is about the *evidence* afforded to hypotheses by data,

and only indirectly about our epistemic actions based on that evidence. [Royall, 2004] makes an important distinction between the questions

What should I believe?

What should I do?

How should I interpret this body of observations as evidence?

[Royall, 2004, p.122]

The LP answers only the third question. It is about evidential inferences, or about how inferences should be made insofar as they rely on the evidence provided by data. Therefore in this chapter I use “inferences” in a narrow sense, to refer to any beliefs and partial (probabilistic) beliefs which are held or followed, and any actions which are taken, *as a direct result of the evidence supplied by an observation*.

3.1 Notation

As is standard, I use X and Y to denote random variables⁴ and $p(F(x) \mid G(x))$ as shorthand for $p(F(X) = F(x) \mid G(Y) = G(x))$.

H is the set of hypotheses under active consideration by anyone involved in the process of inference.⁵

Θ is a set (typically but not necessarily an ordered set) which indexes the set of hypotheses under consideration. I will always treat θ as an index on the *whole* set of hypotheses. In other words, $(\forall h \in H) (\exists \theta \in \Theta : H_\theta = h)$.⁶

⁴**Random variable** is standard terminology in discussions of statistics, but it is slightly misleading. Fortunately, I will be able to do without discussing random variables most of the time; but not quite all the time. A random variable such as X is (famously) neither random nor a variable: it is a function which associates a real number (but not generally a unique one) with each possible observation. Typically, it is considered to be subject to the constraint that for all $x \in \mathbb{R}$ the set $\{y : X(y) \leq x\}$ is measurable according to a standard measure on $\mathbb{R}$. Although X , a random variable, is not a variable, x , a possible value of X , *is* a variable. I write the set of possible values of x — in other words, the range of the random variable X — as $\mathcal{X}$.

⁵If we omitted this condition that H contain all the hypotheses under consideration, the likelihood principle could mistakenly require that we treat two observations as evidentially equivalent even though we know that one supports an important but unmentioned hypothesis more strongly than the other one does. Detailed examples illustrating this point are given in [Berger and Wolpert, 1988, pp.36–38] and elsewhere in the statistical literature.

⁶Under what general circumstances can we be sure that there is an index set on H ? First of all, if H is finite or countably infinite then of course it can be indexed. (Strictly speaking it should be provably countable by a constructive proof, but this is an unimportant detail.) Interestingly, if there is a countable number of observables, each with a countable number of possible states, H can be any subset of the set of *all* probability distributions over the possible observations, which is countable and thus can be indexed. Alternatively, if we can fully describe an uncountable but continuous distribution (either in natural language or in mathematics) then we can still count it as being indexed by parameters, the parameters in this case being whatever lexical tokens are used to describe the function (possibly a countably infinite number of them, if the definition contains terms like $(\forall i \in \mathbb{Z})$). So H can be indexed in the discrete case and

$\mathcal{X}$ is a space of possible observations.

x_{obs} is an actual observation. It is either the result of a single experiment, or the totality of results from a set of experiments which we wish to analyse together. When x_{obs} is the only observation being used to make inferences about a hypothesis space H , I will refer to x_{obs} as *the* actual observation. I assume that it represents all observations considered relevant to any of the hypotheses in some set H .

3.2 Four caveats

Caveat: I only discuss simple hypotheses.

Simple hypotheses are ones which give probabilities to potential observations. The contrast here is with *complex* hypotheses, also known as *models*, which are sets of simple hypotheses such that knowing that some member of the set is true (but not which) is insufficient to specify probabilities of data points.

See for example [Forster, 2006] for an extended discussion of the problems introduced by complex hypotheses.

Caveat: I only discuss inference from data to hypotheses.

I will not be concerned with experimental design (in which hypotheses but not data are typically known), nor with hypothesis generation (if there is such a thing).

Caveat: I ignore the relative desirability of finding one hypothesis or another to be true.

To rephrase in statistical terminology: I will be ignoring utility functions, or (equivalently) loss functions. (A loss function is just a utility function multiplied by -1 .)

Since my focus is on the scientific uses of inference, this may seem like a reasonable assumption. Sadly, it is not clear to me that it is. It seems to me that the best way to do inference is often to weight conclusions about hypotheses — each of which is a possible error — according to the desirability of avoiding the amount of error represented by each hypothesis. On the other hand, it is only sometimes possible to do this, since in general there may be agreement in a community as to non-normative facts, including probabilistic facts, but not as to the desirability of hypotheses. Moreover, even when there is agreement as to the desirability of hypotheses among the people directly concerned with a statistical inference, they are likely to need to justify their results to a wider public which does not share their values. So in many cases my caveat will be an appropriate simplifying assumption, even if not in all cases.

Caveat: When I need to consider epistemic decision makers at all, I assume there is only one of them.

in all describable continuous cases. In most systems of pure mathematics there are, provably, indescribable functions; but as philosophers of applied mathematics we need not worry about them too much. Even so-called “non-parametric statistics” can be indexed in practically all cases. Usually this is trivial, as non-parametric models are usually discrete and so have a countable number of hypotheses which are easy to order.

See [Kadane *et al.*, 1999] for a number of deep arguments about the complications which multiple decision makers introduce into Bayesian theory. I do not consider such complications here.

4 THE LIKELIHOOD PRINCIPLE: PRECISE STATEMENT

We want the likelihood principle to say something like the following:

We should analyse Table 1, and any similar table, using only the numbers in the single column corresponding to the data which were actually observed.

Or, more mathematically:

The only permissible contribution of a space of possible observations $\mathcal{X}$ to inferences about a set of simple hypotheses $\{h_i\}$ is via the probabilities those hypotheses give to the actual observation, namely $p(x_{\text{obs}}|h_i)$.

These probabilities, $p(x_{\text{obs}}|h_i)$, are known as the **likelihood function** of x_{obs} . The likelihood function is a function from hypotheses to probabilities. The likelihood function of an observation need not sum to 1. Moreover, it is convenient to allow the identity conditions of likelihood functions to be different from those of functions in general. Two likelihood functions are considered the same iff they are proportional: i.e., iff $L_1(h) = c \times L_2(h)$ for some $c > 0$.⁷

However, the LP is only true under the following conditions.

Conditions of applicability

1. We cannot infer anything about the relative importance of the various possible inferential errors from the observation (i.e., the loss function, or equivalently the utility function, is either independent of the observation or unimportant). ([Good, 1981], reprinted in Good 1983, p.132)
2. The *choice* of observations is not informative about the hypotheses, although of course the *outcome* of the observations hopefully will be informative about the hypotheses [Hill, 1987; Berger and Wolpert, 1988]. The distinction here is analogous to the distinction in quantum theory between the *setting* on a measurement apparatus and the *outcome* of a measurement.

⁷A further condition on equality of likelihood functions is that $L_1(h)$ and $L_2(h)$ are only considered equal if their variables have the same physical interpretations. This is because we use the meanings of likelihood functions, not merely their shapes. This is important in the literature on the LP in general (although not in this chapter) because some proofs of the LP use relabellings of variables, which are valid only as long as this condition is observed. The first (sketch) proof of the LP [Pratt, 1961, p.166] was of this form; others, such as [Birnbaum, 1962, pp.277–278], arguably break the condition; see [Pratt, 1962, pp.315–316].

It is easy to construct cases in which the LP fails if these conditions are broken. [Gelman *et al.*, 1995] provide several particularly nice illustrations of the importance of the second condition, which they call *ignorability* of the method of data collection. For example:

Suppose for instance that we, the authors, give you, the reader, a collection of the outcomes of ten rolls of a die and all are 6's. Certainly your attitude toward the nature of the die would be different if we told you (i) these were the only rolls we performed, versus (ii) we rolled the die 60 times but decided to report only the 6's, versus (iii) we decided in advance that we were going to report honestly that ten 6's appeared but would conceal how many rolls it took, and we had to wait 500 rolls to attain that result. [Gelman *et al.*, 1995, p.192]

This is a case in which the choice of observations is informative about the hypotheses, and so the likelihood principle does not apply. The experimenter *should* report the sampling scheme in such cases, so one option is to dismiss such cases as inaccurate reporting of data.⁸ However, while the experimenter may be misreporting the experiment, they are (arguably) reporting the *observation* accurately, so they would might (arguably) be covered by the LP. For the sake of clarity it is important to exclude such cases explicitly. Fortunately, the likelihood principle makes sense and is useful even when such cases are excluded.

In addition, the statement of the LP itself needs to be refined a little. It is difficult to make precise the notion that the analysis should proceed *via* the likelihood function, or should *depend only* on the likelihood function. The only way to be really clear about this notion, it seems, is to say what the analysis should *not* take into account. Looking back at tables 1 to 3, this makes sense: saying that the analysis should not take into account anything outside column 1 is a precise way of saying that the analysis should only use column 1, without having to say anything about *how* the analysis should use column 1.

These considerations lead to the definition which I foreshadowed earlier:

The likelihood principle (LP)

Under the conditions of applicability mentioned above, **inference procedures which make inferences about simple hypotheses should not be justified by appealing to probabilities assigned to observations which have not occurred**, except for the trivial constraint that those probabilities place on the probability of the actual observation under the rule that the probabilities of exclusive events cannot add up to more than 1. Therefore the only component of $p(x|h)$ which should be used to justify such inferences is the likelihood function $p(x_{\text{obs}}|h)$, where x ranges over the sample space, h ranges over the hypothesis space and x_{obs} is the actual observation.

⁸Thanks to an anonymous referee for this point.

It is worth remembering two points here.

- (i) Strictly speaking, the second sentence is superfluous; what is doing the work is the emphasised clause. The point of the second sentence is to remind us why (historically) the LP is called the *likelihood* principle.
- (ii) The LP is a constraint on statistical *inference*; that is, on the analysis of observations. The LP does not mention experimental design, and it does not apply to other statistical problems which fall outside the category I am calling *statistical inference*, such as calculating long-term averages over repeated actual experiments (so-called “quality control”) [Hacking, 1965].

5 A PROOF OF THE LIKELIHOOD PRINCIPLE

The proof given here is largely based on [Berger and Wolpert, 1988] which in turn is largely based on the seminal proof of [Birnbaum, 1962].

5.1 *Premise: The weak conditionality principle (WCP)*

The following well-known problem adapted from [Cox, 1958] introduces the main premise of the proof.

Suppose that we are doing an experiment to test a hypothesis h_0 and that we decide to go along with the frequentist idea that we should imagine repetitions of the experiment and make sure that at most 5% of them give the wrong answer, on the assumption that h_0 is true. An almost realistic thought experiment which sheds light on our options involves sending blood to one of two pathology laboratories according to which of them sends the next pick-up courier, or according to the toss of a coin. One laboratory is known to send back an estimated haemoglobin count with a large amount of random error; the other lab always sends back a count that’s almost exactly correct. To achieve an overall 5% error rate as defined above we need to take into account *both* error rates. So if the blood actually went to the accurate laboratory, we need to adjust the error rate on the grounds that it could have gone to the inaccurate one.

This is unsatisfactory, and of course it is not what any practising statistician would do. What she would do is take into account only the characteristics of the laboratory the blood actually went to. In this particular case at least, one should treat the coin toss and the laboratory measurement as two separate experiments, regardless of whether they were planned together. Taking the result of the coin toss into account is called *conditioning* on the data. Cox himself, in his [Cox, 1958, (1958)], came to the conclusion that we should perform a conditional calculation (i.e., take into account only the characteristics of the laboratory the blood actually went to) in this particular case. He did not have available to him the proof of the LP, which shows (on very mild assumptions) that this case generalises to practically all statistical inferences.

Cox — like all non-Bayesian statisticians at the time — held that one should condition only under special circumstances, but was unable to work out exactly what those circumstances were. His example was therefore seen as a paradox. More recently, some frequentists have developed rather sophisticated conditional frequentist theories for conditioning on parts of the data (see, for example, [Casella and Berger, 1987; Berger *et al.*, 1994]). However, we will see that the necessity of conditioning in Cox's example, combined with a very plausible axiom of sufficiency, is enough to prove rigorously that one should follow the likelihood principle, which in turn entails that one should condition on *all* the available data.

The LP is a normative principle, and the idea of proving a normative principle may seem strange in the light of Hume's insistence that we can't derive an ought from an is. What makes it possible is that the conditioning premise drawn from the Cox example is also normative (although very, very weak): it says that we *must* take into account the properties of one laboratory and *must* ignore the properties of the other. The other premise which we will need for the proof, the weak sufficiency principle, is also normative. It is possible to prove a normative principle from other normative principles without violating Hume's rule. The chain of reasoning from these premises to the likelihood principle is purely mathematical.

By proving the likelihood principle from a premise about conditioning which mentions only a single coin toss, we can see that if we are squeamish about conditioning on observed data in some cases then we must bite the bullet and disown conditioning in all cases, even in Cox's case; because if we condition in Cox's case then we have the likelihood principle, and that in turn entails that conditioning is *always* necessary (assuming the other premise of the proof, and the conditions of applicability which I've spelled out above).

The Cox example hardly needs to be generalised at all to form the first premise we need. The laboratories will be generalised to arbitrary experiments. The coin toss can stay as it is.

Informal statement:

If one of two possible experimental measurements is chosen by the toss of a fair and indeterministic coin, no inference procedure should require information about the measurement that was *not* performed.

Formal statement:

The weak conditionality principle: Consider two experimental measurements $M_1 = (\mathcal{X}_1, H, p_1)$ and $M_2 = (\mathcal{X}_2, H, p_2)$. (By this I mean that M_1 has sample space $\mathcal{X}_1$, hypothesis space H and probability or probability density function p_1 , and similarly for M_2 .)

Note that the set of hypotheses is the same for each. This is a deliberate restriction which entails that this principle does not apply to hypotheses about alchemy compared with hypotheses about chemistry, although it does apply to comparing statistical models that each encompass *both* alchemy and chemistry. This condition is useful for the

proof, and it is guaranteed to be satisfied if we stick to the condition (stated earlier) that H contain all the hypotheses under consideration.

Now consider an observation from a new experiment, M^* , which consists of using a fair, epistemically indeterministic coin to select one of M_1 and M_2 at random with probability $1/2$ each. Formally, $M^* = (J, \mathcal{X}_J), H, p_j(J, \mathcal{X}_J)$. By “epistemically indeterministic” I mean simply that no deterministic pattern in the behaviour of the coin has been noted or is expected. Some people are known to be able to toss a coin so as to yield a pre-determined outcome. Our coin-tosser must not be one of those people.

Suppose M^* is performed, and turns out to consist of M_1 . Then any inference procedure should derive the same inferences about H from this instance of M^* as it would have derived from M_1 alone.

The weak conditionality principle is called “weak” because logically stronger conditionality principles are sometimes used for similar purposes [Evans *et al.*, 1986, p.185].

5.2 Premise: The weak sufficiency principle (WSP)

The simplest definition of sufficiency is as follows:

sufficiency definition 1

$T(x)$ is a sufficient statistic for h iff $p(x|T(x))$ is independent of h .

A sufficient statistic for h , $T(x)$, typically contains much less information about the world than x does, but the same amount of information (in a sense which I will make precise) about h . For example, if x is a vector of the heights of a sample of people then, under the Normal or log-Normal models most often used for human heights, the average (mean) height of the sample, $T(x) = \frac{\sum_{i=1}^n x_i}{n}$, is a sufficient statistic for the average (mean) height of the population.

The reason for the name “sufficient” is that if $T(x)$ is sufficient for h (in the technical sense above) then it is all we need to know about x , if our sole purpose is to infer things about h , and so it is sufficient information in the lay sense. Anything else we know about x , over and above $T(x)$, is epistemically redundant. For example, if we’re sure that all we care about is the average height of a population (a big if), there is no point in recording more than the average height of the test sample; any other information about the test sample can be thrown away.

An equivalent definition of sufficiency is:

sufficiency definition 2

$T(x)$ is a sufficient statistic for h iff we can find a function T' which allows p to be factorised in the following way:

$$(\forall h \in H) \quad p(x|h) = T'(T(x), h) \times p(x|T(x)).$$

The factorization theorem shows that this definition is equivalent to the earlier definition.⁹

Seeing that statistical sufficiency is epistemically satisfactory is even easier using definition 2 than it was using definition 1. For definition 2 shows that all functions of $p(x|h)$ can be calculated from $T(x)$ and h , when T is sufficient for h . This point may look superficially as though it assumes the likelihood principle, but it does not. That all inferences about h depend on $p(x_{\text{obs}}|h)$ is, more or less, the likelihood principle; but that all such inferences depend on $p(x_i|h)$ for some set $\{x_i\}$, is an unrelated, trivial claim, and that is what I am relying on here.

For example, in most experiments on coin tossing, the number of heads and the number of tails are jointly sufficient for all inferences; the order in which we observe the heads and tails is irrelevant. $(x_1, x_2, \dots)$ are jointly sufficient iff the ordered tuple $\langle x_1, x_2, \dots \rangle$ is sufficient.)

All statistical models have sufficient statistics, trivially, since the identity function on $\mathcal{X}$ is a sufficient statistic for h according to the above definitions. Of course, such trivial sufficient statistics are not very useful. In addition, a model may have many sufficient statistics.

The weak sufficiency principle: If $T(x)$ is a sufficient statistic for h , and if $T(x_1) = T(x_2)$, then inference procedures should not derive different inferences about H from x_1 and x_2 . (adapted from Basu [1975, p.9])

The weak sufficiency principle was named thus by Dawid [1977] because it is weaker (claims less) than other similar principles.

The WSP is essentially the LP limited to a single experiment. As Leslie [2008] puts it, the difference between the WSP and the LP is like the difference between a State law and a Federal law in a Federal country like Australia or Switzerland (see also [Welsh, 1996, p. 79].) To make this point obvious, suppose that an experiment has no obvious sufficient statistic for a parameter θ except for the trivial one consisting of the whole data ... but suppose further that two data points x_1 and x_2 have the same probabilities under all hypotheses. Then the LP says that we should treat x_1 and x_2 equally in inferences about θ . But it also follows that θ does have a non-trivial sufficient statistic after all, namely the (rather ungainly) statistic consisting of the whole data with all occurrences of x_1 replaced by x_2 . Applied to this sufficient statistic, the WSP *also* says that we

⁹One part of the factorization theorem is easy to prove. It is easy to see that if this equation holds then $T(X)$ is sufficient for h on definition 1, thus: if we know $T(x)$ then we know the right-hand side as a function of h (bearing in mind that we can calculate $p(x|T(x))$, because it does not depend on h); hence, we know the left-hand side, which establishes that $T(x)$ is sufficient for h . The converse is more long-winded to prove, and I will not be relying on it (since my working definition of sufficiency will be the second version, and all I need show is that whatever fits my definition also fits the other one, not vice versa), so I will omit the proof.

should treat x_1 and x_2 equally in inferences about θ . This is hardly a rigorous argument, but I hope it makes the equivalence of the WSP and the LP in a single experiment clear.¹⁰

I do not think that any statistician ever deliberately breaks the WSP. If a conflict with the WSP ever arises, the only reasonable conclusion is that $T(x)$ is not a sufficient statistic for h after all. I would like to give three arguments for this. I do not claim that the three arguments are independent of each other; only that one might convince where the others fail.

Firstly, the WSP follows directly from the claim (defended above) that statistical sufficiency entails epistemic sufficiency.

A second argument for the WSP, adapted from [Basu, 1975, p.9], is as follows. Let us imagine that we have observed x_{obs} in a two-step procedure: we have first conducted an experiment with sample space $\mathcal{X}$ but noted only the value of $T(x)$, not the precise value of x . Then we conduct a further, separate experiment with sample space $T(x)$, noting this time the exact value x_{obs} which we obtain. Since T is sufficient for h , the second experiment is “statistically trivial” (Basu’s term) and tells us nothing about h . Hence, the outcome of the second experiment can make no difference to our inferences about h . Hence values x_1 and x_2 which are possible outcomes of the second experiment (i.e., such that $T(x_1) = T(x_2)$) should lead to the same inferences about h .

Thirdly, here is a Bayesian argument for the WSP. It can be proved that if $T(x)$ is sufficient for h , as defined above, then $(\forall x) p(h|T(x)) = p(h|x)$. So knowing $T(x)$ allows us to know the entire function $p(h|x)$ (as a function of h).

6 A PROOF OF THE LIKELIHOOD PRINCIPLE: CONTINUATION FROM THE WSP AND THE WCP

Consider two statistical measurements $M_1 = (\mathcal{X}_1, H, p_1)$ and $M_2 = (\mathcal{X}_2, H, p_2)$.

Next, consider the mixed experiment M^* (which was defined in the statement of the weak conditionality principle as follows: a fair, epistemically indeterministic coin is tossed; according to its outcome, one of the experiments M_1 and M_2 is performed). Now suppose that whichever experiment hasn’t been performed yet is also performed. At this stage we have an outcome x_1 from M_1 , an outcome x_2 from M_2 , an outcome j indicating which experiment was performed first ($j = 1$ for M_1 and $j = 2$ for M_2), and an outcome from M^* . The outcome from M^* is $J = 1$ or 2 and $x^* = x_1$ or x_2 . The possible outcomes are denoted (j, x_j) .

¹⁰I give a rigorous proof of the equivalence of the WSP and the LP in the presence of a further axiom (the WCP) later in this chapter, and it will be obvious that the WCP is trivially true in the context of a single experiment, so that proof will establish formally that the WSP is equivalent to the LP in the context of a single experiment.

Then let t_0 be the arbitrarily chosen data point $(0, 0)$ and consider the statistic

$$\begin{aligned} T(j, x_j) &= t_0 \text{ if } (j, x_j) = (1, x_1) \text{ or } (2, x_2) \\ &= (j, x_j) \text{ otherwise.}^{11} \end{aligned}$$

Is T a sufficient statistic for h ? Generally, no. Recall that T is a sufficient statistic for h iff p can be factorised as

$$(\forall h \in H) \quad p(j, x_j|h) = T'(T(j, x_j), h) \cdot p(j, x_j|T(j, x_j)).$$

There need not, in general, exist a suitable T' to match our choice of T . But suppose that the likelihoods of x_1 and x_2 are equal (i.e., $p_1(x_1|h) \propto p_2(x_2|h)$, or $(\exists k > 0)(\forall h \in H)(p_1(x_1|h) = k \cdot p_2(x_2|h))$). To prove the likelihood principle, we require to show that T is now sufficient for h . Let T' be as follows:

$$\begin{aligned} T'((j, x_j), h) &= \frac{1}{2}p_1(X = x_1|h) + \frac{1}{2}p_2(X = x_2|h), \quad \text{if } (j, x_j) = t_0 \\ &= p(j, x_j|h) \text{ otherwise.} \end{aligned}$$

Then

$$\begin{aligned} T'(T(j, x_j), h) &= \frac{1}{2}p_1(X = x_1|h) + \frac{1}{2}p_2(X = x_2|h), \quad \text{if } (j, x_j) = (1, x_1) \text{ or } (2, x_2) \\ &= p(j, x_j|h) \text{ otherwise.} \end{aligned}$$

To calculate $p(j, x_j|T(j, x_j))$ (the final term in the sufficiency equation), note:

$$\begin{aligned} p((1, x_1)|T = t_0, h) &= p^*(J = 1|T = t_0, h) \cdot p_1(X_1 = x_1|T = t_0, h) \\ &= \frac{1}{2}p_1(X_1 = x_1|T = t_0, h) \\ &= \frac{\frac{1}{2}p_1(X=x_1|h)}{\frac{1}{2}p_1(X=x_1|h) + \frac{1}{2}p_2(X=x_2|h)} \\ p((2, x_2)|T = t_0, h) &= \frac{\frac{1}{2}p_2(X=x_2|h)}{\frac{1}{2}p_1(X=x_1|h) + \frac{1}{2}p_2(X=x_2|h)}, \text{ by symmetry} \\ p((j, x_j)|T &= (j, x_j), h) = 1, \quad (j, x) \neq t_0. \end{aligned}$$

¹¹The *otherwise* clause can never represent an actual outcome, since I have defined the indexing of the experiments in such a way that only $(1, x_1)$ can occur if M_1 is performed first and only $(2, x_2)$ can occur if M_2 is performed first. If we already knew that the likelihood principle was true, we might not be interested in non-actual outcomes. But as this is a proof of the LP, we can consider at least the mathematical properties of such outcomes.

Now we can check the sufficiency equation:

If $J = 1, X_1 = x_1$ then

$$\begin{aligned}
 & T'(T(j, x_j), h) \cdot p(j, x_j | T(j, x_j)) \\
 &= \frac{1}{2} p_1(X = x_1 | h) + \frac{1}{2} p_2(X = x_2 | h) \times \frac{\frac{1}{2} p_1(X = x_1 | h)}{\frac{1}{2} p_1(X = x_1 | h) + \frac{1}{2} p_2(X = x_2 | h)} \\
 &= \frac{1}{2} p_1(X = x_1 | h) \\
 &= p(j, x_j | h).
 \end{aligned}$$

By symmetry, if $J = 2, X_2 = x_2$ then

$$T'(T(j, x_j), h) \cdot p(j, x_j | T(j, x_j)) = p(j, x_j | h).$$

And for all other (J, X_J) ,

$$T'(T(j, x_j), h) \cdot p(j, x_j | T(j, x_j)) = p(j, x_j | h) \times 1.$$

This establishes that T is sufficient for h .

Given this sufficiency of T for h , and given the weak sufficiency principle (applied to the fact that $(T(1, x_1) = T(2, x_2))$), it follows that no inference about h is valid on observation $(1, x_1)$ in the mixed experiment unless it is also valid on $(2, x_2)$.

Now recall that j is chosen by a fair, indeterministic coin toss. Consequently, the weak conditionality principle applies. It tells us that no inference about h is valid on $(1, x_1)$ unless it is also valid on x_1 alone. (x_1 corresponds to M_1 in my formal statement of the weak conditionality principle above.) In other words, the observations $(1, x_1)$ and x_1 are equivalent in terms of the inferences they license. Similarly, the observations $(2, x_2)$ and x_2 are equivalent in terms of the inferences they license. And we determined in the previous paragraph that the observations $(1, x_1)$ and $(2, x_2)$ are also equivalent to each other. Hence, the observations x_1 and x_2 license the same observations as each other. To summarise this paragraph: if we write “ $\equiv$ ” for “license the same observations as each other”, we have just shown that

$$x_1 \equiv (1, x_1) \equiv (2, x_2) \equiv x_2.$$

The relation $\equiv$ is transitive, so $x_1 \equiv x_2$. Consequently, no inference is valid on x_1 (regardless of the value of j) unless it is also valid on x_2 .

We have proved this for *any* x_1 and x_2 with equal likelihoods under the models under consideration. It follows that *any* two observations which share likelihood functions must license the same inferences.

This is enough to establish some of the versions of the LP found in the literature. To establish the version I have given, though, a little more work must be done.¹²

¹²Recall that the version to be established is as follows: Under the conditions of applicability

Consider any experiment $M = (\mathcal{X}, H, p)$, let the outcome of M be labelled x , and based on M and x define a new experiment $M^\odot = (Y, H, p_Y)$ where Y is 1 or 0 according to whether $X = x$ or not, thus:

$$(1) \quad Y = \begin{cases} 1 & \text{if } X = x \\ 0 & \text{if } X \neq x \end{cases}$$

so that:

$$(2) \quad p(Y = 1|h) = p(x|h) \quad \text{and} \quad p(Y = 0|h) = 1 - p(x|h).$$

So $(\forall h) \ p_Y(Y = 1|h) \propto p(x|h)$: the observations $Y = 1$ and x share likelihood functions. Hence no inference is valid on the observation of x in M unless it is also valid on the observation of $Y = 1$ in $M^\odot$. (All this is a trivial consequence of the part of the likelihood principle already proved above.) Now we should ask which inferences are valid on the observation of $Y = 1$ in $M^\odot$.

If our only observation is that $Y = 1$, whatever we can infer about H from $\mathcal{X}$ and x must be a function of the functions of $\mathcal{X}$ and x that appear in the description of $M^\odot$. The only such functions are $p(x|h)$ and $1 - p(x|h)$ (from (2), or directly from (1) if you prefer). But these are just the likelihood function of x and 1 minus the likelihood function of x . And, in particular, no mention of any part of $\mathcal{X}$ except x is made in the description of $M^\odot$. So all inferences from $M^\odot$ and hence from M must depend functionally on x only via the likelihood function, and in particular no inferences from M may use probabilities of any part of $\mathcal{X}$ except the part which was actually observed.

This establishes the full likelihood principle.

The premises used in this proof form a minimal set of premises for proving the likelihood principle (although not of course the only minimal set), as can be shown by proving the premises of the proof from the conclusion of the proof (the likelihood principle). I will do this by showing that the likelihood principle implies each premise separately. Since it is implied by both premises jointly, it must then be logically equivalent to their union, since if $(a \wedge b) \vdash c$ and $c \vdash a$ and $c \vdash b$ then $c \equiv (a \wedge b)$.

It follows directly from the likelihood principle that the correct conclusion in Cox's example is to ignore the characteristics of the laboratory not used. To prove the weak conditionality principle given above (the formal version of that solution to Cox's paradox), we note that in experiment M^* ,

$$p(j, x_j|h) = \frac{1}{2} p_j(x_j|h) \propto p_j(x_j|h).$$

given above, **inference procedures which make inferences about simple hypotheses should not be justified by appealing to probabilities assigned to observations which have not occurred**, except for the trivial constraint that those probabilities place on the probability of the actual observation under the rule that the probabilities of exclusive events cannot add up to more than 1. Therefore the only component of $p(x|h)$ which should be used to justify such inferences is the likelihood function $p(x_{\text{obs}}|h)$, where x ranges over the sample space, h ranges over the hypothesis space and x_{obs} is the actual observation.

So M^* and M_j have proportional likelihood functions, where M_j is the measurement chosen by the coin toss. Hence M^* and M_j licence identical inferences. Hence only M_j matters.

To prove the weak sufficiency principle from the likelihood principle, note that if T is sufficient for H then (by definition) $p(X|T(X))$ is independent of h . If $T(x_1) = T(x_2)$ (as in the premises of the weak sufficiency principle) then $p(x_1|T(x_1), h) = p(x_2|T(x_2), h)$. Then $p(x_1|h) = p(x_2|h)$ — x_1 and x_2 have identical likelihood functions. So, by the likelihood principle, any inference procedure should draw the same conclusions from x_1 as from x_2 .

This completes the proof that the likelihood principle is logically equivalent to the conjunction of the weak conditionality principle and the weak sufficiency principle.

7 OTHER VERSIONS OF THE LIKELIHOOD PRINCIPLE

R. A. Fisher seems to have made the first statements that could be considered to represent the likelihood principle. For example,

Likelihood serves all the purposes necessary for the problem of statistical estimation. ([Fisher, 1930b], reprinted in [Aldrich, 2000, p.171])

A fuller and more precise statement of the principle, and the first version to be picked up and developed by other authors, was by George Barnard in 1947:

The connection between a simple statistical hypothesis H and observed results R is entirely given by the likelihood, or probability function $L(R|H)$. If we make a comparison between two hypotheses, H and H' , on the basis of observed results R , this can be done only by comparing the chances of, getting R , if H were true, with those of getting R , if H' were true. [Barnard, 1947, p.659]

The most influential discussions of the LP have been by Allan Birnbaum and (jointly) by James O. Berger and Robert Wolpert. [Birnbaum, 1962] provided the first full proof of the LP in 1962 (essentially the same proof as I give in this chapter), while [Berger and Wolpert, 1984] and [Berger and Wolpert, 1988] gave a more general proof and a much more complete discussion, also answering many objections both to the LP and to the proof. [Berger and Wolpert, 1988] is still by far the most complete reference on the LP to date.

Both [Birnbaum, 1962] and [Berger and Wolpert, 1988] state the LP in terms of the *evidence* or *information* which the data give us about our hypotheses. To do this they both introduce an “evidence function” Ev , which they define in interestingly different ways. In [Birnbaum, 1962], Ev is defined primarily in

terms of the premises used in his proof of the principle.¹³ In [Berger and Wolpert, 1988], Ev is left deliberately vague in some places, and in others replaced with “information”, as in the following definition of the LP:

Two likelihood functions for θ (from the same or different experiments) contain the same information about θ if they are proportional to one another ... [where] θ represents only the unknown aspect of the probability distribution of X A second qualification for the LP is that it only applies *for a fully specified model* $\{f_\theta\}$. If there is uncertainty in the model, and if one desires to gain information about which model is correct, that uncertainty must be incorporated into the definition of θ A third qualification is that, in applying the LP to two different experiments, it is imperative that θ be the same unknown quantity in each. [Berger and Wolpert, 1988, pp.19–21.2]

The LP has perhaps always been obvious to at least some Bayesians. Explicit statements by influential Bayesians have included the following:

The prior probability of the hypothesis has nothing to do with the observations immediately under discussion, though it may depend on previous observations. Consequently the whole of the information contained in the observation that is relevant to the posterior probabilities of different hypotheses is summed up in the values that they give to the likelihood [Jeffreys, 1961, p.57]

If two sets of data, x and y , have the following properties: (i) their distributions depend on the same set of parameters; (ii) the likelihoods of these parameters for the two sets are the same; (iii) the prior densities of the parameters are the same for the two sets; then any statement made about the parameters using x should be the same as those made using y . The principle is immediate from Bayes’s Theorem because the posterior distributions from the two sets will be equal. [Lindley, 1965, p.59]

Two possible experimental outcomes D and D' —not necessarily of the same experiment—can have the same (potential) bearing on your opinion about a partition of events H_i , that is, $P(H_i|D)$ can equal $P(H_i|D')$ for each i . Just when are D and D' thus evidentially equivalent, or of the same import? ...

$$P(D'|H_i) = kP(D|H_i).$$

... the likelihood principle: Two (potential) data D and D' are of the same import if [this equation] obtains. [Edwards *et al.*, 1963, p.237]

¹³Specifically, the weak sufficiency principle and the weak conditionality principle, defined above.

The last of these is important because it makes explicit that the data need not come from the same experiment, and hence that the likelihood principle is not merely the sufficiency principle in disguise.

As we will see below, some Bayesians have questioned whether the LP really does follow from Bayesianism.

Finally, some authors have distinguished between weak and strong versions of the LP, where the weak version applies to a single experiment and the strong version compares data from two experiments, as follows:

By the term ‘statistical data’ we mean ... a pair (E, x) where E is a well-defined statistical experiment and x the sample generated by a performance of the experiment. ... To begin with, let us agree to the use of the notation

$$\text{Inf}(E, x)$$

only as a pseudo-mathematical short hand for the ungainly expression: ‘the whole of the relevant information about [the world] contained in the data (E, x) ’.

...

(The weak likelihood principle) : $\text{Inf}(E, x') = \text{Inf}(E, x'')$ if the two sample points x' and x'' generate equivalent likelihood functions

...

(The likelihood principle) : If the data (E_1, x_1) and (E_2, x_2) generate equivalent likelihood functions on Ω , then $\text{Inf}(E_1, x_1) = \text{Inf}(E_2, x_2)$. [Basu, 1975, pp.1–11]

As discussed above, Basu’s “weak likelihood principle” is really just a sufficiency principle.

Basu has also given a nice paraphrase of the LP without the arguably ambiguous word “information”:

We are debating about the basic statistical question of how a given data $d = (E, [x_{\text{obs}}])$, where $E = (X, \Omega, p)$ is the model and $[x_{\text{obs}}]$ is the sample, ought to be analysed. ... the likelihood principle ... asserts that if our intention is not to question the validity of the model E but to make relative (to the model) judgements about some parameters in the model, then we should not pay attention to any characteristics of the data other than the likelihood function generated by it. [Basu, 1975, p.62]

Other versions of the LP can be found in [Savage and discussants, 1962, p.17], [Edwards, 1972, p.30], [Savage, 1976, p.474], [Berger, 1980, p.25], [Lindley, 1982, p.432], [Good, 1983, pp.35–36], [Good, 1983, p.132], [Hill, 1987], [Berry, 1987, p.118], [Berliner, 1987], [Berger, 1993], [Stuart *et al.*, 1999, p.438], [Barnett, 1999, p.188], [Casella and Berger, 2002, p.291] and [Royall, 2004, p.126].

7.1 *The law of likelihood $\neq$ the likelihood principle*

It is vital to distinguish between two principles with confusingly similar names. One is the likelihood principle. The other is the law of likelihood, which says something superficially similar but actually very much more ambitious. Here is the law of likelihood:

Law of likelihood: If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x)/p_B(x)$, measures the strength of that evidence [Hacking, 1965]. [Royall, 1997, p.3]

If the defence of the law of likelihood in [Hacking, 1965; Edwards, 1972; Royall, 1997] is successful then its success is inherited by the likelihood principle, because the law of likelihood entails the likelihood principle (provided the two principles are stated with the conditions of applicability spelled out above: no relevant utility functions and noninformative choice of observations). But if the law of likelihood falls, that does not necessarily reflect badly on the likelihood principle, because the likelihood principle is much weaker.

The important difference between the two principles is that the law of likelihood talks about an observation supporting one hypothesis *to a greater extent* than another, while the likelihood principle makes no mention of the *extent* to which an observation supports a hypothesis. This may seem an unproblematic difference, or even no difference at all, since the likelihood principle talks about the conditions under which an observation supports two hypotheses equally. Moreover, both principles imply that the following statement describes a function Ev which in *some sense* tells us the evidential support that x_{obs} provides for h_1 and h_2 :

$$\text{If } p(x_{\text{obs}}|h_1) = p(x_{\text{obs}}|h_2), \text{ then } \text{Ev}(h_1|x_{\text{obs}}) = \text{Ev}(h_2|x_{\text{obs}}).$$

For proponents of the law of likelihood, this statement is straightforwardly true, and Ev is a number. For proponents of the likelihood principle it is also true. But they need not hold that Ev is a number. If they are Bayesian, for example, they hold that Ev is a function (the likelihood function) of arbitrarily high dimension.

Whatever Ev is, presumably it can be reduced to a number for some purposes (e.g. in betting scenarios), although this may mean throwing away information. The law of likelihood asserts that it can be measured by a number under all circumstances. The likelihood principle does not. That is why the likelihood principle is much weaker than the law of likelihood.

The likelihood principle is also not to be confused with the method of maximum likelihood, which was probably invented by Gauss [Fisher, 1930a, p.531] and was popularised by Fisher [1921]. This method starts by considering the likelihood function $p(x_{\text{obs}}|h)$ with x_{obs} fixed (at whatever was actually observed) and h variable. h is then estimated by maximising $p(x_{\text{obs}}|h)$. Although it is by no means

the same as the LP, the method of maximum likelihood uses only the likelihood function and is therefore compatible with the LP.

7.2 The likelihood principle for infinite hypothesis spaces

The proofs above do not go through for arbitrary probability density functions, because of ambiguities in the notion of sufficiency [Basu, 1975] [Evans *et al.*, 1986] [Berger and Wolpert, 1988, pp.28-30]. But they do go through (with very minor modifications) for *continuous* functions. In addition, the following theorem of Berger and Wolpert can be proved from measure-theoretically more sophisticated versions of the weak sufficiency principle and weak conditionality principle. The theorem shows that the likelihood principle applies to some (in a sense, most) non-continuous infinite hypothesis spaces.

Let $\phi : U_1 \rightarrow U_2$ be a Borel bimeasurable one-to-one mapping from $U_1 \subset \aleph_1$ onto $U_2 \subset \aleph_2$, and suppose there exists a strictly positive function c on U_1 such that for all $\theta \in \Theta$,

$$p_\theta(A) = \int_{\phi^{-1}(A)} \frac{1}{c(x_1)} p_\theta(dx_1), \quad A \subset U_2.$$

Then an inference can only be drawn from the observation x if it can also be drawn from the observation $\phi(x)$, for all x except for a set of probability zero (regardless of the value of θ). If it is agreed to ignore the possibility of events of probability zero then inferences about Θ may depend on $\aleph_1$ and $\aleph_2$ only via x and $\phi(x)$.

(Adapted from [Berger & Wolpert, 1988, pp.33-34])

When combined with my version of the LP, this says that the likelihood principle is true for any *finite* set of hypotheses and for any *parametric* infinite set of hypotheses and for many non-parametric infinite sets of hypotheses.

7.3 Bjørnstad's generalisation: Continuation of the likelihood principle

Jan F. Bjørnstad has proved a version of the likelihood principle which applies even when the hypotheses to be examined depend on the observed data. This is a case which is excluded by my general framework, but I will briefly state Bjørnstad's theorem because it holds the prize as the most general version of the likelihood principle to have been proved to date.

Define M as $(\mathcal{X}, h, p)$ as previously, but this time let the quantity about which we wish to draw inferences be not h but λ , and let λ be a function of x , thus:

$\lambda = \lambda(y, \psi)$, where ψ represents the unknown quantities which are being treated as variables. Let θ represent the unknown quantities which are being treated as parameters.

Let $h = (\psi, \theta)$.

Then inferences from $x \in \mathcal{X}$ to λ must be depend on x only via the ordered pair $\langle \lambda, p(x|\lambda, \theta) \rangle$. The first term in this pair is new. The second term is a likelihood function, but not the same likelihood function as in the simpler case proved above (in which it was $p(x|h)$).

See [Bjørnstad, 1996] for a proof of this principle.

In real scientific cases $\lambda(x)$ is often independent of x , and in this common case Bjørnstad's likelihood principle reduces to the simpler likelihood principle proved above.

8 IS THE LIKELIHOOD FUNCTION WELL DEFINED?

A number of objections have been raised to the LP. There is not enough space here to reply to them all; in any case, almost all of them stand or fall with the proof of the LP. See [Basu, 1988], [Berger and Wolpert, 1988] and references cited in the latter for discussions of some of these.

In the space available, I would like to discuss a philosophically interesting objection to the LP which does *not* stand or fall with the proof, because it objects to an aspect of the way in which I have formulated the problem.

Some Bayesians have argued that Bayesianism is effectively incompatible with the likelihood principle, on the grounds that there is no such thing as an isolated likelihood function [Bayarri *et al.*, 1987]. They argue that in a Bayesian analysis there is no principled distinction between the likelihood function and the prior probability function. If that is so, then we will not be able to draw tables like tables 1 to 3, nor apply the definition of the LP, even though the main part of the definition does not explicitly mention the likelihood function.

This objection is motivated, for these Bayesians, by the fact that (they say) we need prior probabilities in order to apply the LP. Once we have admitted the universal necessity of using prior probabilities we will no longer need to separate the likelihood function from the prior [Bayarri *et al.*, 1987; Berger and Wolpert, 1988]. Thus, they accept proofs of the likelihood principle, *conditional* on the assumption that a likelihood function has been specified; but they deny that specifying a likelihood function is necessary, and they deny that it is possible to do so in a principled way. They believe that the likelihood principle is true, if stated carefully, but not straightforwardly applicable.

Despite disputing the applicability of the likelihood principle in this way, Bayesians in this school often see it as a useful weapon with which to combat frequentism. I like to think of this view as Bayesian Hegelianism, as it sees the likelihood principle as an important part of a historical dialectic which will inevitably lead to a

synthesis in which it is no longer required. Such a prediction has been beautifully summarised by Bayarri, DeGroot and Kadane, following a metaphor proposed by Butler [1987]:

The [frequentist] Cheshire Cat vanished quite slowly, first the tail and then the body of frequentist methods. The last visible part was the likelihood [principle] grin, “which remained some time after the rest of it had gone”. But that, too, disappeared. [Bayarri *et al.*, 1987, p.27]

To return to the objection itself: the claim is that there is no principled definition of the likelihood function because there is no principled way of deciding what should be labelled x (data) and what should be labelled h (hypothesis) in the definition of the likelihood as $p(x|h)$.

Bayarri, DeGroot and Kadane’s examples all involve the following set-up.¹⁴

Suppose that the random variable Y is not observed but another random variable X is observed with conditional density $p(x_{\text{obs}}|y, \psi)$. [Then] it is irrelevant which of the factors on the right-hand side [of

$$p(\psi, y|x_{\text{obs}}) = \frac{p(x_{\text{obs}}|y)p(y|\psi)p(\psi)}{\int p(x_{\text{obs}}, y, \psi)}$$

are regarded as part of the [likelihood function] and which are regarded as part of the prior distribution. [Bayarri *et al.*, 1987, pp.6–7]

This is correct, from a Bayesian point of view. (See also [Hill, 1987, p.99].) In contrast, we do have to distinguish the likelihood function in order to apply the likelihood principle. Three natural choices are $p(x_{\text{obs}}|\psi)$, $p(x_{\text{obs}}, y|\psi)$ and $p(x_{\text{obs}}|y, \psi)$ [$\equiv p(x_{\text{obs}}|y)$], but there is no natural way to choose between these three possibilities ... or so Bayarri, DeGroot and Kadane claim.

The problem for the likelihood principle, as thus stated, is very easily solved. One need merely specify what one means by “likelihood function”. I have already done this: for me, the likelihood function is always $p(x_{\text{obs}}|y, \psi)$, where x_{obs} is what is observed. As Berliner [1987] correctly notes, the likelihood principle “applies equally well, though separately, in each of the potential cases [which Bayarri, DeGroot and Kadane] enumerate”, so my solution is perfectly adequate, as would be any other solution which serves to disambiguate the term “likelihood function”.

However, it may appear that a problem remains, since others may disambiguate the likelihood function differently from me. For example, Bayarri, DeGroot and Kadane imagine a case in which two doxastic agents see the same observation, and analyse it using the same mathematical model except that one of them introduces an unobserved variable y into the model while the other does not. This leads the

¹⁴Throughout this section I replace Bayarri, DeGroot and Kadane’s y by x_{obs} , x by y , θ by ψ , and f by p , in order to remain consistent with the terminology of this chapter.

two agents to define the likelihood function in different ways, following which they cannot use the likelihood principle to compare their results.

To see that my version of the likelihood principle still applies, we have only to note that these two agents are using different hypothesis spaces H : for one of them, H includes a specification of an unobserved variable, while for the other it does not. Given a fixed H (which is an explicit precondition of my version of the likelihood principle), only one likelihood function is possible, namely $p(x_{\text{obs}}|h \in H)$. (Note that they agree on x_{obs} ; otherwise no joint analysis of any sort would be possible.) A practical problem remains if neither of the two agents accepts the other's parameterisation of the hypothesis space, but there is no reason why this should happen, since the two parameterisations essentially agree with each other (more precisely, one parameterisation is easily reducible to the other by taking a marginal distribution with respect to y).¹⁵

ACKNOWLEDGEMENTS

Thanks to Prasanta Bandyopadhyay, James O. Berger, Alan Hájek, Claire Leslie, Alison Moore, Max Parmar, Daniel Steel and David Spiegelhalter for discussions which contributed to this chapter, and to an anonymous referee for many helpful suggestions. However, the views given here do not necessarily represent the views of these people, and any errors are my own.

BIBLIOGRAPHY

- [Aldrich, 2000] John Aldrich. Fisher's "Inverse Probability" of 1930. *International Statistical Review*, 68(2):155–172, 2000.
- [Barnard, 1947] G. A. Barnard. Review of Wald's 'Sequential analysis'. *Journal of the American Statistical Association*, 42:658–669, 1947.
- [Barnard, 1985] George Barnard. Discussion of 'In defense of the likelihood principle: axiomatics and coherency'. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, volume 2, pages 57–60. Elsevier, 1985.
- [Barnett, 1999] Vic Barnett. *Comparative Statistical Inference*. John Wiley, New York, 3rd edition, 1999.
- [Basu, 1975] Debabrata Basu. Statistical information and likelihood (with discussion). *Sankhyā Series A*, 37:1–71, 1975.
- [Basu, 1988] Debabrata Basu. *Statistical information and likelihood. A collection of critical essays.*, volume 45 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1988.
- [Bayarri and Berger, 2004] M. J. Bayarri and J. O. Berger. The interplay of Bayesian and frequentist statistics. *Statistical Science*, 19(1):58–80, 2004.
- [Bayarri et al., 1987] M. J. Bayarri, M. H. DeGroot, and J. B. Kadane. What is the likelihood function? (with discussion). In Shanti S. Gupta and James O. Berger, editors, *Statistical decision theory and related topics IV*, volume 1, pages 3–27. Springer-Verlag, New York, 1987.
- [Berger and Wolpert, 1984] James O. Berger and Robert L. Wolpert. *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, California, 1st edition, 1984.
- [Berger and Wolpert, 1988] James O. Berger and Robert L. Wolpert. *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, California, 2nd edition, 1988.

¹⁵See also a similar solution to the problem in [Berliner, 1987, p.19] and alternative solutions in [1987] and [Berger and Wolpert, 1988, p.39].

- [Berger *et al.*, 1994] James Berger, Lawrence D. Brown, and Robert Wolpert. A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics*, 22:1787–1807, 1994.
- [Berger, 1980] James O. Berger. *Statistical Decision Theory: Foundations, Concepts, and Methods*. Springer-Verlag, New York, 1980.
- [Berger, 1993] James Berger. An overview of robust bayesian analysis. Technical Report 93-53C, Purdue University, 1993.
- [Berger, 2006] James Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.
- [Berliner, 1987] Mark Berliner. Discussion of ‘What is the likelihood function?’. In Shanti S. Gupta and James O. Berger, editors, *Statistical decision theory and related topics IV*, volume 1, pages 17–20. Springer-Verlag, New York, 1987.
- [Berry, 1987] D. A. Berry. Interim analysis in clinical trials: the role of the likelihood principle. *The American Statistician*, 41:117–122, 1987.
- [Birnbaum, 1962] Allan Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, June 1962.
- [Bjørnstad, 1996] Jan F. Bjørnstad. On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association*, 91(434):791–806, June 1996.
- [Butler, 1987] R. W. Butler. A likely answer to ‘What is the likelihood function?’. In Shanti S. Gupta and James O. Berger, editors, *Statistical decision theory and related topics IV*, volume 1, pages 21–26. Springer-Verlag, New York, 1987.
- [Casella and Berger, 1987] George Casella and Roger L. Berger. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, 82:106–111, 1987.
- [Casella and Berger, 2002] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, Pacific Grove, 2nd edition, 2002.
- [Cox, 1958] D. R. Cox. Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29(2):357–372, June 1958.
- [Dawid, 1977] A. P. Dawid. Conformity of inference patterns. In J. R. Barra *et al.*, editor, *Recent Developments in Statistics*, pages 245–256. North-Holland, Amsterdam, 1977.
- [Edwards *et al.*, 1963] W. Edwards, H. Lindman, and L. J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242, 1963.
- [Edwards, 1972] A. W. F. Edwards. *Likelihood*. Cambridge University Press, London, 1972.
- [Evans *et al.*, 1986] M. Evans, D. A. S. Fraser, and G. Monette. On principles and arguments to likelihood. *Canadian Journal of Statistics*, 14:181–199, 1986.
- [Fisher, 1921] R. A. Fisher. On the probable error of a coefficient of correlation deduced from a small sample. *Metron.*, 1:3–32, 1921.
- [Fisher, 1930a] R. A. Fisher. Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535, 1930.
- [Fisher, 1930b] R. A. Fisher. Inverse probability (abstract). In *Report of the meeting of the British Association for the Advancement of Science*, page 302, Bristol, 1930. British Association for the Advancement of Science.
- [Forster, 2006] Malcolm R. Forster. Counterexamples to a likelihood theory of evidence. *Minds and Machines*, 16(3):319–338, 2006.
- [Gelman *et al.*, 1995] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, London, 1995.
- [Good, 1981] I. J. Good. Some logic and history of hypothesis testing. In Joseph C. Pitt, editor, *Philosophical Foundations of Economics*, University of Western Ontario Series on the Philosophy of Science, pages 149–174. D. Reidel, Dordrecht, 1981.
- [Good, 1983] I. J. Good. *Good Thinking*. University of Minnesota Press, Minneapolis, 1983.
- [Hacking, 1965] Ian Hacking. *Logic of Statistical Inference*. Cambridge University Press, London, 1965.
- [Hill, 1987] Bruce M. Hill. The validity of the likelihood principle. *American Statistician*, 41(2):95–100, 1987.
- [Jeffreys, 1961] Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, 3rd edition, 1961.

- [Kadane *et al.*, 1999] Joseph B. Kadane, Mark J. Schervish, and Teddy Seidenfeld. *Rethinking the Foundations of Statistics*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge University Press, Cambridge, 1999.
- [Leslie, 2008] Claire Leslie. *Exhaustive Conditional Inference: Improving the Evidential Value of a Statistical Test by Identifying the Most Relevant P-Value and Error Probabilities (PhD Thesis)*. University of Melbourne, Melbourne, 2008.
- [Lindley, 1965] D. V. Lindley. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge University Press, Cambridge, 1965.
- [Lindley, 1982] Dennis V. Lindley. The role of randomization in inference. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Volume Two: Symposia and Invited Papers:431–446, 1982.
- [Lindley, 1983] D. V. Lindley. Response to ‘parametric empirical Bayes inference’ by Morris. *Journal of the American Statistical Association*, page 381?, 1983.
- [Pratt, 1961] J. W. Pratt. Review of Lehmann’s ‘Testing Statistical Hypotheses’. *Journal of the American Statistical Association*, 56:163–166, 1961.
- [Pratt, 1962] J. W. Pratt. Discussion of ‘On the foundations of statistical inference’ by Allan Birnbaum. *Journal of the American Statistical Association*, 57:314–315, 1962.
- [Robbins, 1955] H. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–164. University of California Press, Berkeley, 1955.
- [Royall, 1997] Richard Royall. *Statistical Evidence: A likelihood paradigm*. Chapman and Hall, London, 1997.
- [Royall, 2004] Richard Royall. The likelihood paradigm for statistical evidence. In Mark L. Taper and Subhash R. Lele, editors, *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, pages 119–138. University of Chicago Press, Chicago and London, 2004.
- [Savage and discussants, 1962] L. J. Savage and discussants. *The Foundations of Statistical Inference*. Methuen, London, 1962.
- [Savage, 1976] L. J. Savage. On rereading R. A. Fisher (with discussion). *Annals of Statistics*, 42:441–500, 1976.
- [Stuart *et al.*, 1999] Alan Stuart, J. Keith Ord, and Steven Arnold. *Kendall’s Advanced Theory of Statistics Vol. 2A: Classical Inference and the Linear Model*. Arnold, London, 6th edition, 1999.
- [Welsh, 1996] A. H. Welsh. *Aspects of Statistical Inference*. John Wiley, New York, 1996.

Part V

Recent Advances in Model Selection

This page intentionally left blank

AIC, BIC AND RECENT ADVANCES IN MODEL SELECTION

Arijit Chakrabarti and Jayanta K. Ghosh

OVERVIEW

As explained in e.g., [Ghosh and Samanta, 2001], model selection has somewhat different connotations in Statistics and History or Philosophy of Science. In the latter it has come to mean a major shift in paradigm on the basis of available data, of which one of the most famous examples is the shift from Newtonian Physics to Einstein's relativistic Physics on the basis of data obtained in a famous expedition of Eddington (Example 1). In Statistics it has a useful but much more pedestrian role of distinguishing between two statistical models on the basis of available data. For example, is the data coming from a normal distribution or a Cauchy distribution (see Example 5)? One of the most popular applications of Classical Statistical Model Selection is to determine which variables are important in a regression model (equivalently a linear model) for the dependent response variable y , in terms of the auxiliary variables $\mathbf{x}$ in the model (Example 2). However, the Classical Statistical Model Selection Rules can also be used for problems of paradigm shift (Example 1).

We use the word "Classical Statistics" to distinguish it from "Bayesian Statistics". A standard introduction to Classical Statistics is [Lehmann and Casella, 2001]. A definitive overview of Bayesian Statistics is provided in [Bernardo and Smith, 1994]. The interpretation as well as motivation of Statistical Model Selection Rules depend on whether one believes in the Classical or Bayesian paradigm of Statistics as well as the loss (or utility) function. Of the two most well-known Statistical Model Selection Rules, namely AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), AIC has a classical origin whereas BIC arises as an approximation to a Bayes rule up to $O(1)$ (the exact meaning of this statement will be explained in Section 3). At this level of approximation, one may ignore the prior distribution of the Bayesian.

By recent advances in the title we mean both recent improvements in understanding the properties and performance and relevance of AIC and BIC as well as new model selection rules and other advances. Our review of the new rules and work related to them will be brief since the AIC and BIC are our primary focus.

AIC was proposed by Akaike in a series of papers [Akaike, 1973; 1974]. He seemed to be guided by optimal prediction of a new set of y 's corresponding to a

replicate of the observed $\mathbf{x}$'s. It was proved first by Shibata (see [Shibata, 1981; 1983]) that in certain important problems, for large samples, AIC predicts better than any other model selection rule. He proves it by showing that asymptotically it predicts as well as an Oracle, where the Oracle is a model selection rule which always selects the best model for prediction. The prediction error of the Oracle gives a lower bound to the error committed in prediction for all model selection rules. The exact form of the Oracle considered by Shibata is given in section 2. Shibata's ideas were considerably simplified in [Li, 1987]. Interesting general results based on Li's ideas were obtained by [Shao, 1997]. We discuss more about the predictive properties of AIC in Sections 2 and 4.

BIC was introduced by Schwarz [1978] and can be used for approximating the Bayes Factor corresponding to two models and is discussed in some detail in Section 3. We briefly introduce the Bayes Factor here. Suppose M_1 and M_2 are two models specifying two families of densities $p(x|\boldsymbol{\theta}_1)$, $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1$ and $p(x|\boldsymbol{\theta}_2)$, $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2$ for the given data, where $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are the two parameter spaces corresponding to the two models. The Bayes Factor BF_{21} is the ratio of the marginal (probability) density of the data under M_2 , namely $P(\text{data}|M_2)$, and that under M_1 , namely, $P(\text{data}|M_1)$, with the latter in the denominator and the former in the numerator. If $BF_{21} > 1$, one chooses M_2 . Otherwise one chooses M_1 . A Bayesian assumes a priori probabilities for the two models under consideration and using the data posterior probabilities of the two models, namely $P(M_1|\text{data})$ and $P(M_2|\text{data})$ are obtained. BF_{21} is also equal to the ratio of the posterior odds (of model M_2 with respect to model M_1) and the prior odds. If the prior probabilities of M_1 and M_2 are taken to be half, BF_{21} is the same as posterior odds. The technical definitions of the marginal densities of data and posterior probabilities of models appear in Section 3. indexprior odds'indexHastie, T.

Both AIC and BIC are special cases of penalized likelihood rules, which may be described as follows. Suppose M_2 is a model specifying a family of densities $p(x|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_2$ for the given data and M_1 is a submodel

$$(1) \quad p(x|\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}_1 \subset \boldsymbol{\Theta}_2.$$

This is a case of nested models. (It must noted that in this case $P(M_2|\text{data}) \neq P(\boldsymbol{\theta} \in \boldsymbol{\Theta}_2|\text{data})$. We discuss this in Section 3.) The method of maximum likelihood, so popular in Classical Statistics, would suggest evaluating each model M_i by the maximized likelihood, or equivalently the maximized log-likelihood, of the data x under M_i , namely $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_i} \log p(x|\boldsymbol{\theta})$, where "sup" denotes supremum.

Since $\boldsymbol{\Theta}_1 \subset \boldsymbol{\Theta}_2$,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1} \log p(x|\boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_2} \log p(x|\boldsymbol{\theta}),$$

if we compare only the maximized log-likelihood we would always choose the more complex model. It is intuitively obvious that this may not be a good thing to do for all data. As pointed out in [Hastie *et al.*, 2003; Forster and Sober, 1994], a very complex model fitting the data too well is ignoring the fact that the data is

composed of both signals; i.e., significant, repeatable aspects as well as noise, i.e., random perturbations of the data.

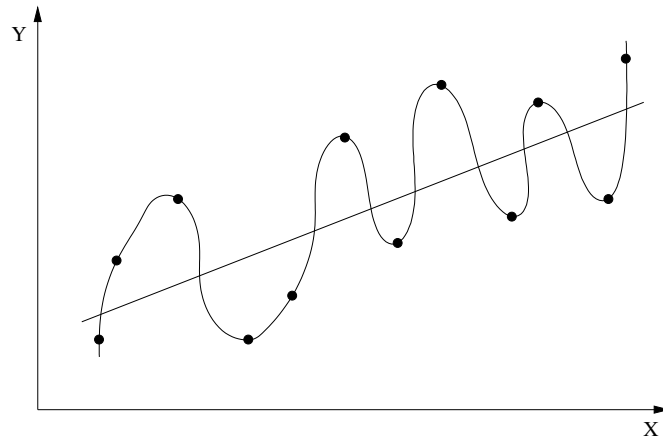


Figure 1.

As in Figure 1 above, a simpler model fitting a simple straight line is likely to be better than the zigzag curve passing through all the observed points.

This is where the so-called principle of parsimony or Ockham's Razor comes in. It suggests one should penalize each model according to its complexity. Consider the simple situation when $\Theta_2 = \mathbf{R}^{d_1}$ and $\Theta_1 = \mathbf{R}^{d_2}$ for some positive integers d_2 and d_1 such that $d_1 < d_2$. In such a case, the usual dimension of the parameter points belonging to each individual model, i.e., d_1 or d_2 is a simple measure of complexity of the model. This is so since larger the dimension, richer is the model in the sense of having a larger number of independently varying parameter components, and more complex it is to make inference on them.

This discussion leads to the penalized log-likelihood

$$\sup_{\theta \in \Theta_i} \log p(x|\theta) - \lambda d_i,$$

where d_i is the dimension of Θ_i and λ is a positive constant specifying the penalty per unit dimension. If $\lambda = 1$, we get AIC. If $\lambda = \frac{\log n}{2}$, we get BIC. We choose the model which maximizes the penalized log-likelihood. For the rest of the paper, we will use the notations p or p_i to denote model dimensions.

For a long time it has been unclear which of AIC and BIC or some other penalized log-likelihood criterion is appropriate. Till very recently, there has not been much effort to understand the penalized likelihood methods in a unified manner, which would provide formal theoretical arguments in favor of each type of penalty for certain specific purposes and tell how and where each penalty comes from a particular mathematical principle. The situation is somewhat clearer now

(see Section 4) with respect to AIC and BIC. Rapid progress in understanding is taking place for other rules also (see Section 5). This understanding in the case of AIC is similar to that of [Forster and Sober, 1994] and we also don't believe that AIC is a solution to all model selection problems. It is worth mentioning in this context that in a conference held in Oberwolfach in 2005 (which is also mentioned in Section 5), serious research aimed at a unified understanding of penalized likelihood methods were presented.

In Section 4 we suggest AIC and BIC are appropriate in somewhat different problems, the difference is due to different purposes and different loss functions.

Section 5 contains a brief review of some recent advances.

1 EXAMPLES

In this section we present several examples where model selection techniques can be applied to answer scientific or statistical questions. Examples 1 through 4 appeared in [Ghosh and Samanta, 2001], with appropriate references to sources of the data.

EXAMPLE 1 Eddington's experiment. According to Einstein's theory of gravitation, light gets deflected by gravitation and the amount of such deflection can also be specified. More specifically, Einstein famously predicted that under the gravitational attraction of the Sun, the light emanating from nearby stars will get deflected, but such an effect would only be visible during a total solar eclipse (when such deflection can be measured through apparent change in a star's position). To verify this prediction, a famous experiment was conducted by a team led by British astrophysicist Eddington immediately after the first world war. Four observations were collected on the amount of angular deflection (measured in seconds) by Eddington's team and other groups (spread over a period of 10 years) and those turned out to be $X_1 = 1.98$, $X_2 = 1.61$, $X_3 = 1.18$ and $X_4 = 2.24$. Einstein predicted that the amount of deflection would be 1.75. Suppose we assume that the X_i 's are independently normally distributed about the unknown mean μ , i.e. $X_i \sim N(\mu, \sigma^2)$, where σ^2 is unknown. To statistically test if Einstein's conjecture were true based on the observed data, we can consider choosing between the two models $M_1 : \mu = 1.75$ and $M_2 : \mu \neq 1.75$ using some model selection techniques. It is also possible to formulate this as a problem of choosing one of two nested models, in which case M_2 would permit μ to have all possible real values. Since that leads to some subtle logical questions, we postpone discussion of nested models for section 3. (However examples 3 and 4 in this section are formulated as nested models). Even though σ^2 is unknown here, just to illustrate how to use these two methods, let us treat σ^2 as known and equal to the sample variance s^2 . Then making the transformation $X' = \frac{X - \mu}{s}$ of the original data X , the model selection problem becomes the same as choosing between the two models $M_1 : \mu = 0$ and $M_2 : \mu$ is a non-zero real number, where μ now denotes the mean of the transformed observations. BIC is the appropriate model selection rule here since we

Table 1. Hald's Cement hardening data

x_1	x_2	x_3	x_4	y
7	26	6	60	78.6
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

want to select the true model. The values of the BIC criterion for M_1 and M_2 are -5.675 and -6.37 respectively and so BIC selects M_2 . Although AIC is not really appropriate for this purpose, it is a curious fact that AIC also selects the model M_2 in this problem. So, according to both these criteria, Einstein's prediction of $\mu = 1.75$ is supported by the observed data, although the evidence is not very strong. This particular data is now only of historical importance. Much stronger confirmation of Einstein's theory has come from other experiments, see [Gardner, 1997].

EXAMPLE 2 Hald's regression data. Table 1 below presents data on heat evolved during the hardening of Portland cement and four variables (all or a subset of which) may be able to explain the amount of heat evolved. These four variables are called the explanatory variables denoted by x_1, x_2, x_3 and x_4 , which measure respectively the percentage weight of four chemical compounds (which together constitute cement) and the response variable is denoted by y , which measures the total calories given-off during hardening per gram of cement after 180 days. This data set has been analyzed several times before, e.g., [Berger and Pericchi, 1995; Burnham and Anderson, 2003].

One traditional approach in statistics to deal with such data sets is to represent y in a normal linear regression model as

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i, i = 1, \dots, n,$$

where $x_1, \dots, x_k$ generically denote the regressors, which according to this particular model, explain the repeatable aspect of the variation in the y values and ϵ_i are independent and identically distributed as $N(0, \sigma^2)$. Here β_i 's are the unknown parameters. The different models in such situations correspond to the different pos-

sible choices of regressor variables from the pool of all potential regressors. For example, for the data in Table 1, one can consider a total of $2^4 - 1 = 15$ possible models with the following choices of regressors $\{x_1\}, \dots, \{x_4\}, \{x_1, x_2\}, \dots, \{x_3, x_4\}, \{x_1, x_2, x_3\}, \dots, \{x_2, x_3, x_4\}$ and $\{x_1, x_2, x_3, x_4\}$. The purpose of choosing a model or a set of models here would be to pinpoint which regressors x 's seem to have the most significant causal relationship with the amount of heat evolved y . Here AIC chooses the model with regressors x_1, x_2 and x_4 while BIC chooses the model with regressors x_1 and x_2 . The AIC and BIC values for each model are reported in [Ghosh and Samanta, 2001], expressed as a difference from the value for the model, which is selected by that criterion. Thus the AIC value for model $\{x_1, x_2\}$ is -0.225, while that for model $\{x_1, x_2, x_4\}$ is 0. On the other hand, the BIC value for these two models are 0 and -1.365 respectively. So AIC favors $\{x_1, x_2, x_4\}$ while BIC favors $\{x_1, x_2\}$.

EXAMPLE 3 Nested Model selection Problem and hypothesis testing. Suppose $X_1, \dots, X_n$ are independent but identically normally distributed with mean μ and variance 1. Suppose the question to a statistician is whether the mean μ is 0 or not. The statistician formulates this problem as the testing problem with the null hypothesis $H_0 : \mu = 0$ versus the alternative hypothesis $H_1 : \mu \neq 0$. Note that H_0 and H_1 specifies two disjoint subsets of $\mathbf{R}$ for the unknown μ . The same problem can also be formulated as one of selecting between two models $M_0 : \mu = 0$ and $M_1 : \mu \in \mathbf{R}$. Note that here M_0 is nested within M_1 . If the data supports $\mu = 0$, then it is consistent with both models but one will choose the simpler model on grounds of parsimony, which requires that of all models which explain the data equally well, one should choose the simplest model. As explained earlier in the Overview, simplicity is defined in terms of the dimension of the model. A more detailed discussion of nested models appears in Section 4.

EXAMPLE 4 ANOVA type problems and High dimensional Setup. Suppose one has p similar normal populations, each population i having a potentially different mean $\mu_i, i = 1, \dots, p$. We have r observations from each population and let $n = rp$. One might be interested in knowing whether these p means are the same or not. This question be thought of as a question of choice between the two models

$$M_1 : \mu_1 = \dots = \mu_p \text{ vs } M_2 : \mu_i \text{'s are arbitrary.}$$

This is again a nested model problem. This is called the one-way ANOVA (Analysis of Variance) problem. Stone [1979] considered the situation where r is fixed but $p \rightarrow \infty$ and hence $n = pr$ also tends to infinity and critically studied the performance of AIC and BIC. Stone showed that AIC performs better than BIC in identifying the true model in the sense that BIC chooses M_1 with probability tending to 1 under M_2 if the μ 's satisfy certain conditions and p grows sufficiently fast, while AIC chooses the correct model M_2 with probability tending to 1. Data sets for which the p = dimension = number of parameters is large is called high dimensional.

High dimensional data sets appear often these days in many applications. A

prime example of this is when one wants to test gene expression levels simultaneously for thousands of genes (p in this form) but has a rather small set of data points (r in this form) on each gene, each data point corresponding to an individual. The level of expression of a gene, for example could be related to its effect on some tumor or character of the individual. It has great medical significance.

EXAMPLE 5 Normal vs. Cauchy. Suppose the following observations (when placed in increasing order of magnitude) are obtained in an experiment and the question is which distribution do the data come from :

$\{-6.56759, -3.14456, -1.19043, -0.64666, -0.64624, -0.54472, -0.43171, -0.34207, -0.32573, -0.31834, -0.29348, -0.19512, -0.14658, -0.12093, -0.0328, 0.075277, 0.131894, 0.187061, 0.199137, 0.214316, 0.226209, 0.471883, 0.654485, 0.719648, 0.788128, 0.911007, 0.946036, 1.28061, 1.676115, 7.986715\}$.

Looking at the data, it seems very likely that the distribution from which the data set is generated is symmetric around zero. But we are not sure what the form of the distribution is, but suspect that it might either be a normal or a Cauchy with an unknown scale parameter. The normal and Cauchy distributions with the same location are often not easy to distinguish in moderate sample sizes, but in this case we have two observations in the data set which seem too far away from the other observations, which is more likely to happen in the case of a Cauchy distribution, because its tails are very thick. But in order to decide more objectively, we want to check statistically whether M_1 : The unknown distribution is a Cauchy distribution with location 0 is true or M_2 : The unknown distribution is a normal distribution with location 0 is true. We use BIC for this purpose. It chooses model M_1 , the difference in BIC criterion value for the two models being 51.7069. In fact BIC chooses the correct model here since the data were simulated from a Cauchy distribution.

2 THE AKAIKE INFORMATION CRITERION (AIC)

In this section we consider the Akaike Information Criterion (AIC) in a few canonical statistical problems and state results of its statistical optimality therein. We also discuss its connection with other model selection criteria and some of the generalizations of it. The optimality is connected with Akaike's original motivation as brought out in [Forster and Sober, 1994] but it does not follow as an immediate consequence. In fact the proofs are quite non-trivial.

We start by introducing the linear model. Consider $\mathbf{Y} = (Y_1, \dots, Y_n)'$, a vector of observations of the response (dependent) variables and let $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ be the $(n \times p)$ matrix of explanatory variables, $\mathbf{X}_j, j = 1, \dots, p$ being the j -th column of $\mathbf{X}$. This is the same as the set up in Example 2 in Section 1, but now written in matrix notation. In the linear model, as the name suggests, one connects the mean vector $\boldsymbol{\mu} = E(\mathbf{Y}|\mathbf{X})$ (assuming that $\mathbf{X}$ is fixed) with the explanatory variables via the relationship $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown parameter

of interest. It is further assumed that

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_{n \times n})$ is the vector of random errors and σ^2 is its variance, which may be known or unknown. In this context a model M specifies a certain subset of the $\boldsymbol{\beta}$ vector to be equal to zero while the others are allowed to be arbitrary. We will, for simplicity, assume that the model space is $\mathcal{M} = \{M_1, \dots, M_p\}$, where under model M_j , $\beta_k = 0$ for $k > j$ while $\beta_1, \dots, \beta_j$ are arbitrary. This is the nested model scenario (as also seen in Examples 3 and 4), since if M_j is true then so are $M_{j+1}, \dots, M_p$, for any $j \in \{1, \dots, p\}$. Model M_j postulates that only the first j explanatory variables are potentially responsible for the variability in the repeatable aspect of the Y 's (measured by the mean) while the others do not contribute anything. Assuming that σ^2 is known, the Akaike Information Criterion (AIC) for model M_j is

$$\text{AIC}(j) = \log L(\hat{\boldsymbol{\beta}}_j) - j,$$

where $\hat{\boldsymbol{\beta}}_j$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ under model M_j and $\log L(\hat{\boldsymbol{\beta}}_j)$ is the joint density of the data under model M_j evaluated at $\hat{\boldsymbol{\beta}}_j$. AIC chooses M_j which maximizes $\text{AIC}(j)$ among $j \in \{1, 2, \dots, p\}$. (Under the assumption that $\mathbf{X}'\mathbf{X} = I_{p \times p}$, one can easily derive this criterion using direct calculation in this setup.) Upon simplification, this criterion can be written equivalently as (ignoring constants depending on n but independent of j)

$$\text{AIC}(j) = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_j\|^2 + 2j\sigma^2,$$

and one minimizes $\text{AIC}(j)$ over $j \in \{1, 2, \dots, p\}$ to choose the best model, according to this criterion. If σ^2 is unknown, $\text{AIC}(j)$ becomes (with σ^2 estimated by its maximum likelihood estimator under model M_j),

$$\frac{n}{2} \log(2\pi) + \frac{n}{2} + \frac{n}{2} \log \left(\frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_j\|^2}{n} \right) + (j+1).$$

An alternative method estimates σ^2 by either the maximum likelihood estimator of σ^2 under the largest model or some estimator which is consistent under all models. It can be shown that the difference between the AIC for unknown σ^2 and this form of AIC with a plug-in estimator of σ^2 is, for large sample size n , approximately a constant depending on n but independent of the j (i.e the model under consideration), under pretty mild conditions. If, instead of the nested model scenario, one considers $\mathcal{M}$ to be any collection of models (each of which specifies a certain subset of the coordinates of $\boldsymbol{\beta}$ to be zero), the definition of $\text{AIC}(M)$ (upto constants independent of models) for a generic $M \in \mathcal{M}$ with number of free parameters p_M , becomes $\text{AIC}(M) = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_j\|^2 + 2p_M\sigma^2$ if σ^2 is known and $\text{AIC}(M) = \log \left(\frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_j\|^2}{n} \right) + 2\frac{p_M}{n}$ if σ^2 is unknown.

One very important problem where AIC can be used as a model selection rule is the problem of nonparametric regression, where the functional form of dependence between the dependent variable and the regressor is not expressible in terms of finitely many unknown parameters, as, for example, in the usual polynomial regression problem where the regression function may be known to be a polynomial of degree five in the regressor, with the six coefficients of the polynomial being the unknown parameters. Instead, the nonparametric regression model states that the expected value of Y given x is some unknown $f(x)$ where f can belong to a pretty large class of functions, e.g, say, the class of all functions which are square integrable. The reason for assuming that f might belong to a very large class of functions is the general perception that the relationship could be pretty complex.

To illustrate our point, we describe a practical example where one does not really have much clue about the relationship between x and Y to start with and the use of nonparametric regression is much more appealing intuitively than the usual parametric regression. The cosmic microwave background (CMB) radiation data from the Wilkinson Microwave Anisotropy Probe (WMAP) is analyzed in several chapters of [Wasserman, 2006] using different approaches to nonparametric regression. The basic data is a temperature map obtained by the WMAP, showing the temperature in different points of the sky 13 billion years ago. The fluctuations in the temperature map, measured through the strength of temperature fluctuations $f(x)$ (called power spectrum) at each frequency x (or “multipole moment”), provide information about the early universe. So estimation of $f(\cdot)$ is the most interesting thing to cosmologists. Through an appropriate procedure, the temperature map can be transformed to a scatterplot of *estimated* power Y versus frequency x , given by $(x_1, Y_1), \dots, (x_n, Y_n)$. The goal of nonparametric regression will be to estimate the function f , based on the Y ’s with very little assumption on its functional form.

In general, based on observations $Y_i, i = 1, 2, \dots, n$ of the dependent variable at regressor values $\{x_i, i = 1, \dots, n\}$, one writes the nonparametric regression model as

$$Y_i = f(x_i) + \epsilon_i, i = 1, \dots, n,$$

where f is assumed to belong to given (large) class of functions and ϵ_i are i.i.d. errors with zero mean and finite variance. If f is square integrable, one can represent the function uniquely as an expansion which is an infinite linear combination of certain sine and cosine functions (called the basis functions), appearing in a specific order. This, in mathematical parlance, is known as the Fourier expansion of the function. Each function is determined by the coefficients (called the Fourier coefficients) that multiply the basis functions in its expansion, i.e., if the Fourier coefficients of two functions are the same, then the two functions must be identical almost everywhere. So estimating f becomes same as estimating its Fourier coefficients. Since one only has finite amount of data, infinitely many Fourier coefficients can’t be estimated. Noting that as one goes further down the expansion, the Fourier coefficients become increasingly negligible, one natural solution of estimating f then is to approximate it by an appropriately chosen partial sum of

this expansion and the problem of choosing a partial sum becomes one of variable selection in linear regression with the basis functions as the regressors (variables to choose from), and the Fourier coefficients as the regression coefficients. This way, the problem of estimation of f also becomes one of model selection, where each model specifies which basis functions will be used to describe f . *Note that this way each model is a false model, being a finite sum of sine and cosine terms, approximating the true infinite sum.* A popular choice of models is by taking them as nested ones, with model M_k considering the partial sum involving the first k basis functions and Fourier coefficients. AIC can be defined here exactly as in the linear model setup considered earlier, and by choosing a model here one simply wants to select the correct order of the partial sum for the given sample size with the goal of good estimation of f and hence good predictive performance with the selected model. Chakrabarti and Ghosh [2006a] show that in this nested model scenario, the model selected by AIC achieves this goal by proving that the estimate (using least squares estimates of the Fourier coefficients under the model chosen by AIC) of the unknown function converges to the truth very fast, at the so-called minimax rate. We will not delve into the technical details of this statement in this paper.

Many authors have studied asymptotic optimality properties of AIC in terms of predictive performance. To sum up, this line of research shows that under some conditions, AIC is able to predict as well as an Oracle asymptotically. The Oracle provides a lower bound for predictive performance which may possibly be attained only asymptotically but can not be implemented for finite sample size since the Oracle depends on the unknown value of the parameter. For example, in the nonparametric regression problem, Shibata [1983] defined an Oracle as

$$M_n^* = \operatorname{argmin}_{M \in \mathcal{M}_n} E \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}_j\|^2,$$

where the $\mathbf{X}$ in the above expression denotes the $n \times \infty$ design matrix involving all the basis functions, $\boldsymbol{\beta}$ is the true Fourier coefficient of the unknown function, $\hat{\boldsymbol{\beta}}_j$ is the least squares estimate of $\boldsymbol{\beta}$ under model M_j and $\mathcal{M}_n$ is the model space which varies with sample size. It is easy to see that M_n^* depends on the true $\boldsymbol{\beta}$ for each n . Shibata [1983] showed that the ratio of the risk of the model selected by AIC and M_n^* tends to 1 as $n \rightarrow \infty$, for each true $\boldsymbol{\beta}$.

We are now in a position to briefly indicate Akaike's original rationale or motivation behind the definition of the criterion named after him, without going deep into the technical details. Suppose f is the true unknown density from which i.i.d sample observations $Y_1, \dots, Y_n$ are generated. The job of the statistician is to mimic the truth based on sample data. In parametric inference one considers a set of models $\mathcal{M}$ and each model consists of densities indexed by a finite number of parameters. The goal is to find a model which is closest to the truth in the sense that it contains a density which is closest (among all densities included in all the candidate models) to the true density in terms of some appropriate measure of divergence. As a measure of divergence between the true density f and an approximating density g , Akaike considered the Kullback-Leibler divergence, given

by

$$K(f, g) = \int f(x) \log \left(\frac{f(x)}{g(x)} \right) dx.$$

(Note that for measuring closeness (or lack of it) between two densities, the Hellinger distance and the Kullback-Leibler divergence are standard measures used by the statisticians and have been used by philosophers also in a slightly different context (see [Joyce, 1999]). For each model M_k in $\mathcal{M}$, one can find the maximum likelihood estimator $\hat{\theta}_k$ based on $Y_1, \dots, Y_n$ under that model. Then, letting $g_k(\cdot|\hat{\theta}_k)$ as the representative density from model M_k , Akaike considered minimizing over $M_k \in \mathcal{M}$, the criterion

$$(2) \quad \frac{1}{n} E \left\{ \int f(\mathbf{Y}^{\text{new}}) \log \left(\frac{f(\mathbf{Y}^{\text{new}})}{g(\mathbf{Y}^{\text{new}}|\hat{\theta}_k)} \right) d\mathbf{Y}^{\text{new}} \right\},$$

where the expectation is taken with respect to the true density f and $\mathbf{Y}^{\text{new}}$ denotes an independent sample of size n from f . This quantity in (2) measures the divergence per observation, between the predicted density and the truth. (Criterion (2) is the most general one, suitable for all situations, e.g. the linear regression setup described above. But if the Y_i 's are i.i.d as in our case the criterion reduces to $E \left\{ \int f(Y^{\text{new}}) \log \left(\frac{f(Y^{\text{new}})}{g(Y^{\text{new}}|\hat{\theta}_k)} \right) dY^{\text{new}} \right\}$, where Y^{new} is a sample of size one from f .) Minimizing (2) is again equivalent to maximizing

$$\frac{1}{n} E \left\{ \int f(\mathbf{Y}^{\text{new}}) \log(g(\mathbf{Y}^{\text{new}}|\hat{\theta}_k)) d\mathbf{Y}^{\text{new}} \right\},$$

with respect to k . It was shown by Akaike that in large samples, an approximately unbiased estimator of the above quantity in (2) is given by

$$\frac{1}{n} (\log(L(\hat{\theta}_k)) - \dim(M_k)),$$

where $\dim(M_k)$ is the number of free estimable parameters in model M_k .

In the linear model example, under the assumption of orthogonal design matrix, a simple calculation shows that an exactly unbiased estimator of $\frac{1}{n} E \left\{ \int f(\mathbf{Y}^{\text{new}}) \log(g(\mathbf{Y}^{\text{new}}|\hat{\theta}_k)) d\mathbf{Y}^{\text{new}} \right\}$ is given by $\frac{1}{n} (\log(L(\hat{\theta}_k)) - k)$.

Now we will briefly mention the connections of AIC with some other model selection criteria. Although the Akaike Information Criterion (AIC) is considered mainly as a non-Bayesian i.e., classical statistical criterion, there has been some studies which show that in certain normal linear model problems, as in Example 4, it has an Empirical Bayes interpretation (where at least some part of the prior is estimated from the data, see e.g., [Ghosh *et al.*, 2006, chapter 9]), in the sense that under some conditions, Empirical Bayes Model selection rule and AIC choose the same model either for each sample size or asymptotically. In all these problems a model is chosen so as to minimize the expected posterior loss (using least squares estimates) in prediction of a new replicate of the dependent variable at a fixed

predictor value. See [Mukhopadhyay and Ghosh, 2003; Chakrabarti and Ghosh, 2007] for further details on this. It is also worth mentioning here that AIC can be also related to a relatively recent model selection criterion DIC of [Spiegelhalter *et al.*, 2002]. Spiegelhalter *et al.* [2002] define, what they call, a Bayesian measure of model complexity or the effective number of parameters in a model using information theoretic considerations. DIC is then defined as a penalized version of a Bayesian measure of fit, the penalty being the model complexity. This is similar in spirit to the usual penalized likelihood model selection criteria, where a measure of fit is often measured by the (minus) twice-maximized log-likelihood. As observed in [Spiegelhalter *et al.*, 2002; Chakrabarti and Ghosh, 2006b], under the assumption of posterior normality of the parameters in the model, DIC coincides with AIC asymptotically.

Last but not the least, we would like to point out the connection of AIC with cross-validation. The cross-validatory way of model selection, as the name suggests, keeps a part of the sample for the estimation of the parameters in candidate models and uses the remaining part of the data for validation and hence choice of the appropriate model, the idea being not to use the same data twice for two different purposes. In its simplest form, namely the leave-1-out cross validation, one just keeps one observation away at a time and calculates for each candidate model, the sum of squared error prediction losses in predicting one observation based on all remaining observations by estimating the parameters in the model using them. One chooses that model which minimizes this cross-validatory criterion. Using certain regularity conditions, Stone [1977] argued that this form of cross-validation and AIC are equivalent, in the sense that the difference between the criteria values for a give model becomes negligible in large samples. But as observed in [Chakrabarti and Ghosh, 2007], this fact seemed to have been “overinterpreted” by people. It turns out that if the model under consideration is not the true model (which by the way is one of the crucial assumptions in [Stone, 1977]), then the widely believed equivalence fails to hold. Some concrete examples of this phenomenon and theoretical explanations are given in [Chakrabarti and Ghosh, 2007].

Finally, for non-nested (or subsets) model selection scenario a modification of AIC with an additional penalty seems to help choose a more parsimonious model, as studied in [Chakrabarti and Ghosh, 2007]. The modification of AIC is

$$\text{AIC}(M) = \log(\hat{\theta}_M) - p_M + p_M \log\left(\frac{w}{1-w}\right),$$

where $0 < w < 1$. The use of such a modification can be partially motivated by seeing it as a penalty for the complexity of the model space, since more complex the model space, there are more comparisons to be made and there is more chance of choosing a larger dimensional model when in fact a smaller model is also true.

3 BAYES FACTOR AND BIC

Suppose $X_1, \dots, X_n$ are independent and identically distributed (i.i.d.) random variables and the models $M_1, M_2, \dots, M_k$ specify k parametric families of densities. The model M_j specifies the density as $p(x|\theta_j)$ where θ_j is a parameter in Θ_j , the parameter space corresponding to model M_j . The Bayesian analyst has to provide a prior $\pi_j(\theta_j|M_j)$ for θ_j conditional on the assumption that M_j is true. For notational simplicity, we will henceforth drop the subscript j and denote this prior as $\pi(\theta_j|M_j)$. Then the probability density of the data under M_j is defined, letting $\mathbf{x} = \{x_1, \dots, x_n\}$, as

$$m_j(\mathbf{x}) = p(\mathbf{x}|M_j) = \int_{\Theta_j} \prod_{i=1}^n p(x_i|\theta_j) \pi(\theta_j|M_j) d\theta_j.$$

It is known that the true model belongs to this class of k models, but it is not known which one is true. The Bayesian would assign prior probability π_j for M_j to be true, where $\sum_{j=1}^k \pi_j = 1$. We consider 0-1 loss, i.e, the loss is 0 if a true model is chosen and is 1 if a false model is chosen. The Bayes rule is to choose the model M_j , for which the posterior probability given by

$$P(M_j|\mathbf{x}) = \frac{m_j(\mathbf{x})\pi_j}{\sum m_j(\mathbf{x})\pi_j}$$

is the largest for the given $\mathbf{x}$. If the k models are assumed to be equally likely, i.e $\pi_j = \frac{1}{k}$, for $j = 1, \dots, k$, the Bayes rule reduces to choosing M_j that maximizes $m_j(\mathbf{x})$ for the given $\mathbf{x}$.

Consider now k nested models $M_1, M_2, \dots, M_k$ specifying $\theta \in \Theta_j$, where $\Theta_1 \subset \Theta_2 \subset \dots \subset \Theta_k$. We assume M_k is a true model, but we do not wish to choose it if a more parsimonious model is also true. It is assumed that under the most complex model there is a given density $p(\mathbf{x}|\theta)$, $\theta \in \Theta_k$. This leads to a similar assumption for any M_j ($j < k$), namely, that under M_j , there is a density $p(\mathbf{x}|\theta)$, $\theta \in \Theta_j$.

We now discuss the usual assignment of probabilities to these models M_j . We distinguish logically between a model M_j and its associated Θ_j . Though the Θ_j 's are nested in the usual set theoretic sense, we do not take the models as logically equivalent to their associated parameter sets. Rather, models are hypotheses about values of θ , often they are scientific hypotheses about natural phenomena. We consider a historical example. We imagine we are contemporaries of Galileo and are speculating on

$$M_1 : \theta = 0 \quad \text{and} \quad M_2 : \theta \text{ is an arbitrary real number,}$$

where θ is the true difference in falling times of two objects in vacuum. Notice that M_1 is Galileo's scientific hypothesis backed by his knowledge and intuition,

and $\pi(\theta|M_1)$ is Galileo's conditional probability for θ , given by $\pi(\theta = 0|M_1) = 1$. On the other hand, the conditional probability that θ takes any particular value under $\pi(\theta|M_2)$, is an assignment based on no particular knowledge and so small for all θ .

The Bayesian may choose these conditional probabilities as his given that he tries to put himself in the same frame of mind as that of Galileo and others, and the background of each model. Alternatively, for each j , he may pretend that he believes in M_j and in that state of mind assign a conditional subjective probability distribution on Θ_j . This is admittedly difficult, if not impossible, but all we are trying to say is that $\pi(\theta|M_1)$ and $\pi(\theta|M_2)$ are logically unrelated, and that it makes sense to take, as Bayesians usually do, $\pi(\theta = 0|M_1) = 1$ and $\pi(\theta|M_2)$ is some suitable low information prior suggested by Jeffreys.

We wanted to make two points above through an illustrative example. The first is that $\pi(\theta|M_j)$ for different M_j 's are not logically determined by the assignment for the most complex model. The second is that we need the assignment $\pi(\theta|M_j)$ to complete our Bayesian set up for selecting one of a set of nested models.

A Bayesian would also assign prior probability π_j to M_j , where $\sum_{j=1}^k \pi_j = 1$ and, usually, because of parsimony, $\pi_1 \geq \pi_2 \geq \dots \geq \pi_k$. The common (so called objective) choice is $\pi_j = \frac{1}{k}$ for all $j \in \{1, \dots, k\}$. In our historical example, this would be the Bayesian's choice if he didn't want to favour either Galileo or the general public. We emphasize again that the Bayesian does not logically identify the model with the associated parameter space and thus can still assign equal probability to two nested models indicating his degree of belief about the truth of two competing models. The conditional density of θ given M_j usually assigns much more probability to $\Theta_j \cap \Theta_{j-1}^c$ than to Θ_{j-1} . In fact usually, the conditional probability of Θ_{j-1} is zero in most actual problems, because Θ_{j-1} has usually lower dimension than Θ_j .

Combining all the components we can calculate the conditional or posterior probability of M_j given x as

$$p(M_j|x) = \frac{\pi_j \int_{\Theta_j} p(x|\theta) \pi(\theta|M_j) d\theta}{\sum_{j=1}^k \pi_j' \int_{\Theta_j'} p(x|\theta) \pi(\theta|M_j') d\theta}.$$

The Bayes rule is to choose the model with maximum posterior probability. In the usual case where $\pi_j = \frac{1}{k}$, $j = 1, \dots, k$, this is the same as choosing a model that maximizes

$$\int_{\Theta_j} p(x|\theta) \pi(\theta|M_j) d\theta$$

which is usually called the marginal density of x obtained by integrating out θ . Note that since the $\pi(\theta|M_j)$'s are different in the sense that one is not logically

derived from the other, there is no reason to expect that the method will always choose the most complex model. Under regularity conditions (see e.g. [Schwarz, 1978; Ghosh *et al.*, 2006, Chapter 4]), logarithm of the above marginal likelihood is

$$\text{BIC} + O(1)$$

where BIC (Bayes Information Criterion) = $\log L(\hat{\theta}_j) - \frac{p_j}{2} \log n$, where p_j is the dimension of Θ_j , $\hat{\theta}_j$ is the maximum likelihood estimator of θ under model M_j and $L(\hat{\theta}_j)$ is the joint density of $x_1, \dots, x_n$ under model M_j for $\theta = \hat{\theta}_j$.

One often assigns probability in a different way in the non-nested case. Let $\pi(\theta)$ be a probability density over $\Theta = \Theta_1 \cup \dots \cup \Theta_k$ and $p(\mathbf{x}|\theta)$ the density of $\mathbf{x}$ under θ , $\theta \in \Theta$. In this case M_j 's do correspond to just the subset Θ_j specified by M_j and the p_j 's are determined automatically from this identification as $\pi_j = \int_{\Theta_j} \pi(\theta) d\theta$. It obtains from a simple algebra that the posterior probability $P(M_j|\mathbf{x}) = P(\theta \in M_j|\mathbf{x})$ is given by

$$P(M_j|\mathbf{x}) = \frac{\int_{\Theta_j} p(\mathbf{x}|\theta)\pi(\theta) d\theta}{\int_{\Theta} p(\mathbf{x}|\theta)\pi(\theta) d\theta}.$$

In this assignment of probabilities, $P(M_{j'}|\mathbf{x}) \leq P(M_j|\mathbf{x})$ if for some pair $\Theta_{j'} \subset \Theta_j$. For nested models with Θ_j 's of different dimension, any assignment of probabilities of this kind to lower dimensional sets would lead to zero prior and posterior probability. More importantly, this assignment ignores the possibility of the density of θ depending on the model. In the nested case as when θ =effect of a drug the probability distribution of θ under $M_1 : \theta = 0$ and $M_2 : \theta \in \mathbf{R}$ are often expected to be very different. We present below a concrete example to illustrate some of our points in this discussion.

EXAMPLE 6. Let θ =effect of a drug on, say, the (systolic) blood pressure. It is common to test $M_1 : \theta = 0$ (no effect) against $M_2 : \theta \leq 0$ (most likely some good effect). This is question of selecting one from two nested models. A typical choice of priors is

$$\begin{aligned} \pi(\theta = 0|M_1) &= 1 \\ \pi(\theta|M_2) &= \frac{1}{2\tau\sqrt{2\pi}} e^{-\frac{\theta^2}{2\tau^2}} \text{ if } \theta \leq 0, \\ &= 0 \quad \text{otherwise} \end{aligned}$$

and the density of a single observation x is defined as

$$\begin{aligned} p(x|\theta) &= N(\theta, \sigma^2), \theta = 0 \text{ under } M_1 \\ p(x|\theta) &= N(\theta, \sigma^2), \theta \leq 0 \text{ under } M_2. \end{aligned}$$

For simplicity we assume σ^2 is known and equal to one while $\tau^2 = 2$ (say). Let $p_1 = p_2 = \frac{1}{2}$. Given data $\mathbf{x} = x_1, \dots, x_n$ on n persons, one has

$$P(M_1|\mathbf{x}) = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum_{i=1}^n x_i^2}}{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum_{i=1}^n x_i^2} + \int_{\theta \leq 0} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2} \frac{1}{2\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}} d\theta}.$$

Similarly, $P(M_2|\mathbf{x})$ is given by

$$P(M_2|\mathbf{x}) = \frac{\int_{\theta \leq 0} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2} \frac{1}{2\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}} d\theta}{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum_{i=1}^n x_i^2} + \int_{\theta \leq 0} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2} \frac{1}{2\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}} d\theta}.$$

If on the other hand we wish to test a new model $M_1 : \theta \leq 0$ (bad or no effect) against $M_2 : \theta \geq 0$ (good or no effect) (i.e a non-nested model choice scenario) and $\theta \sim N(0, \tau^2 = 2)$ (untruncated normal) one has

$$\begin{aligned} P(M_1|\mathbf{x}) &= \frac{\int_{\theta \leq 0} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2} \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}} d\theta}{\int_{\theta \in \mathbf{R}} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2} \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}} d\theta} \text{ and} \\ P(M_2|\mathbf{x}) &= \frac{\int_{\theta \geq 0} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2} \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}} d\theta}{\int_{\theta \in \mathbf{R}} \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2} \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}} d\theta}. \end{aligned}$$

The BIC model selection rule, as defined above, chooses the M_j that maximizes BIC. It can be shown rigorously that under suitable regularity conditions, as $n \rightarrow \infty$ with k and p_j 's held fixed, the BIC rule is asymptotically the same as Bayes rule. Both the Bayes rule and BIC are consistent in the sense that they choose the true model with probability tending to one.

In the nested case, the Bayes rule and BIC choose the true model with smallest dimension. Note that in the nested case if M_j is true then $M_{j+1}, \dots, M_k$ are also true, in the sense that if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_j$, then it also belongs to $\boldsymbol{\Theta}_{j+1}$ and so on. Thus BIC is consistent with the generally accepted principle that model chosen should be as simple as possible if there are more than one model which are true or which explain data equally well. It is easy to show with examples that AIC need not do so. In fact, with $k = 2$ and 0-1 loss, model selection is equivalent to testing $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ vs $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_2$ (or $\boldsymbol{\theta} \in \boldsymbol{\Theta}_2 \cap \boldsymbol{\Theta}_1^c$). In this context AIC has a type

1 error probability α which is asymptotically greater than 0.05 in many simple examples. This makes it a very non-conservative test. This is one reason why Bayesians don't seem to like AIC. However, AIC has some good properties like predictive optimality not possessed by BIC, vide Sections 1 and 4. In short, no model selection rule will serve all purposes.

Before we end, we quote in the context of nested models, [Forster and Sober, 1994, paragraph 3 of section 7]. In this remark "LIN" refers to the class of all straight lines and "PAR" refers to the class of all parabolic curves of $\mathbb{R}^2$. "The key element of any Bayesian approach is the use of Bayes' Theorem, which says that the probability of any hypothesis H given any data is proportional to its prior probability times its likelihood : $p(H/\text{Data})p(H) \times p(\text{Data}/H)$. However, it is an unalterable fact about probabilities that (PAR) is more probable than (LIN), *relative to any data you care to describe*. No matter what the likelihoods are, there is no assignment of priors consistent with probability theory that can alter the fact that $p(\text{PAR}/\text{Data}) \geq p(\text{LIN}/\text{Data})$. The reason is that (LIN) is a special case of (PAR). How, then, can Bayesians explain the fact that scientists sometimes prefer (LIN) over (PAR) ?"

Rather than trying to respond to the criticism in [Forster and Sober, 1994], we have tried to explain in this section the Bayesian practice and point of view as well as we can. Hopefully, this will help to some extent.

4 COMPARISON OF AIC AND BIC THROUGH AN EXAMPLE.

In this section, we consider AIC and BIC from a comparative point of view. Much research has been done on these two criteria. We try to summarize here (with minimum technicality) the knowledge about where these two criteria are suitable (or otherwise), with the aid of an illustrative example.

It is quite clear now that AIC and BIC are appropriate for different purposes and have different motivations. BIC came out as an approximation (in large samples) to the Bayes rule using a 0-1 loss function when the dimensions of the models under study remain fixed. On the other hand, AIC was derived and proposed by Akaike as a rule, which, in large samples, does well in predicting an independent future set of observations based on data at hand, the prediction accuracy being achieved by minimizing the expected Kullback-Leibler divergence between the true model and candidate models. Note that the Bayes rule for 0-1 loss corresponds to choosing the model having the largest a posteriori probability, while minimizing the expected Kullback-Leibler divergence using AIC, for normal linear models, is equivalent to minimizing the expected squared error prediction loss (in predicting an independent future set of observations). It can be said, as Ghosh [2007] points out, that the penalty of BIC is appropriate and arose from the use of the 0-1 loss while that of AIC corresponds to the squared error loss. It indeed turns out, as intuitively somewhat expected from the above discussion, that AIC is a good rule if prediction is the sole purpose and BIC is good if the scientist wants to select the correct model. AIC excels in those problems where all models in the model space

are incorrect or there is at most one correct model in the model space, while BIC excels if the correct model is in the model space. If more than one model is true, AIC may not choose the simplest true model.

We will now expand on thoughts of the last paragraph with a simple example. Let us consider the model

$$y_i = \mu + \epsilon_i, i = 1, \dots, n,$$

where ϵ_i are identically and independently distributed $N(0, 1)$ errors. We have two models $M_1 : \mu = 0$ vs $M_2 : \mu \in \mathbb{R}$. Once a model M_α ($\alpha = 1$ or 2) is selected, μ has to be estimated by the maximum likelihood estimator under that model, namely $\hat{\mu}(\alpha)$. Suppose one wants to minimize with respect to α the quantity $E_\mu(\frac{1}{n} \sum_{i=1}^n (z_i - \hat{z}_i(\alpha))^2)$, where $z_i, i = 1, \dots, n$ is a future set of n independent observations from the same distribution, and $\hat{z}_i(\alpha) = \hat{\mu}(\alpha)$ is the estimated value of z_i based on the y 's and model M_α , $\alpha = 1, 2$. It can be shown that minimizing this is the same as minimizing with respect to α the quantity $E_\mu(\mu - \hat{\mu}(\alpha))^2$ or $(\mu - \hat{\mu}(\alpha))^2$ for each given $y_1, \dots, y_n$. We will work with the quantity $A(\alpha, \mu) = E_\mu(\mu - \hat{\mu}(\alpha))^2$. Note that if the value of μ were known, one would know which one of M_1 and M_2 will minimize $A(\alpha, \mu)$ with respect to α by simply evaluating $A(1, \mu)$ and $A(2, \mu)$. Minimizing $A(\alpha, \mu)$ with respect to α in this way yields the optimal predictive rule, called the Oracle, as

$$M_{\text{Ora}}(\mu) = \begin{cases} M_1 & \text{if } 0 \leq \mu^2 \leq \frac{1}{n}, \\ M_2 & \text{if } \mu^2 > \frac{1}{n}. \end{cases}$$

This way the Oracle, prescribes for each value of μ , the model which will be good for predicting that μ . But this rule depends on the unknown (true) μ and hence can't be used in practice. The expected loss for the Oracle is μ^2 if $0 \leq \mu^2 \leq \frac{1}{n}$ and $\frac{1}{n}$ if $\mu^2 > \frac{1}{n}$. The statistician wants to find a model selection rule which does almost as well as the Oracle for a moderate amount of data, and as well as the Oracle in the limit asymptotically for infinite amount of data. Upon simple calculations, it turns out that AIC chooses M_2 if $n\bar{y}^2 > 2$ and M_1 otherwise, while BIC chooses M_2 or M_1 based on whether not $n\bar{y}^2$ exceeds $\log n$. For AIC, therefore, the loss is $(\bar{y} - \mu)^2$ if M_2 is chosen (i.e. if $n\bar{y}^2 > 2$), since $\hat{\mu}(2) = \bar{y}$, while it is μ^2 if M_1 is chosen since $\hat{\mu}(1) = 0$. The loss can be expressed simply as $(\bar{y} - \mu)^2 I_{n\bar{y}^2 > 2} + \mu^2 I_{n\bar{y}^2 \leq 2}$. Taking expectations, the expected loss for AIC is

$$E_\mu\{(\bar{y} - \mu)^2 I_{n\bar{y}^2 > 2}\} + \mu^2 P_\mu(n\bar{y}^2 \leq 2).$$

Similar arguments gives that for BIC the expected loss is

$$E_\mu\{(\bar{y} - \mu)^2 I_{n\bar{y}^2 > \log n}\} + \mu^2 P_\mu(n\bar{y}^2 \leq \log n).$$

Consider first the case when $\mu = 0$, i.e., M_1 is true and hence M_2 is also true. Then $M_{\text{Ora}}(0)$ selects M_1 , AIC chooses M_1 if $n\bar{y}^2 \leq 2$ and BIC chooses M_1 if

$n\bar{y}^2 \leq \log n$. It is clear, noting that under $\mu = 0$, $n\bar{y}^2 \sim \chi_{(1)}^2$, that AIC chooses model 2 with a fixed non-zero probability (given by the probability of a chi-square random variable with 1 degree of freedom to be less than or equal to 2) for all n , while BIC chooses M_1 with probability tending to 1 as $n \rightarrow \infty$ (since the probability that a chi-square random variable with one degree of freedom being less than $\log n$ tends to 1 as $n \rightarrow \infty$). This reinforces a general fact that if two fixed dimensional nested models are true, AIC often times fails to choose the smaller model while BIC excels in choosing the parsimonious true model by penalizing the larger model heavily. For all n , the expected loss of the Oracle rule is 0 while that of AIC is $\frac{c}{n}$ where $0 < c < 1$ is a constant. The expected loss of BIC is always of a strictly larger order than $\frac{1}{n^{3/2}}$ and simultaneously of a strictly smaller order than $\frac{1}{n(\log n)^\delta}$ for any fixed positive δ , for all large enough n . While talking about orders, we are here following the convention that for two real sequences a_n and b_n , a_n is of a strictly larger order than b_n (and hence b_n of a strictly smaller order than a_n) if $a_n/b_n \rightarrow \infty$ as $n \rightarrow \infty$. Thus, in this situation the predictive performance of BIC is slightly better than that of AIC.

Now consider the case when $\mu \neq 0$, i.e only M_2 is true. Here for any large enough n the Oracle chooses M_2 and with probability tending to 1, both AIC and BIC choose M_2 . It is easy to show that the ratios of the expected oracle loss and that of BIC and AIC tend to 1 as $n \rightarrow \infty$. But, in the absence of any knowledge of the true model and noting that the sample size is never infinite, a conservative approach to choose between AIC and BIC would be to see which rule minimizes, for any fixed n , the worst possible expected loss, i.e. which $\alpha \in \{\alpha_{\text{BIC}}, \alpha_{\text{AIC}}\}$ minimizes $\sup_{\mu \in \mathbb{R}} A(\alpha, \mu)$, where α_{BIC} and α_{AIC} denote, respectively, the model selected by BIC and AIC for the given sample. For the Oracle, this quantity is $\frac{1}{n}$. A simple calculation shows that for AIC this quantity is, for all large n , of the form $\frac{K_1}{n}$ for some constant $1 < K_1 < 2$ while that for BIC is of the form $\frac{K_2 \log n}{n}$ where $K_2 > 0$ is finite. Hence for BIC the worst possible rate is a factor of magnitude higher than that of AIC and the Oracle. This pathology is caused by the fact that for a given n , if μ is of the order of $\frac{\sqrt{\log n}}{\sqrt{n}}$, then BIC still chooses M_1 with non-zero probability, making the second term in the expression for the expected loss large. BIC chooses a lower dimensional model unless the increase in log-likelihood for adding a new parameter is at least $\frac{\log n}{2}$. So, for a fixed large n , for a small $\mu \sim \frac{\sqrt{\log n}}{\sqrt{n}}$, the expected loss of BIC for that μ will be a factor of magnitude larger than that of AIC.

To sum up, this example shows that the overall predictive performance of AIC is better than that of BIC in this problem, while BIC does a better job of selecting the correct model, notwithstanding its affinity to choose the smaller dimensional model because of its large penalty.

5 RECENT ADVANCES IN MODEL SELECTION

Although AIC and BIC are probably the most popular model selection criteria with specific utility (as described in detail) above, they are not the only solutions to all types of model selection problems. In recent years many other penalized likelihood model selection criteria have been proposed. For example, in cubic spline model fitting, the complexity of the model is taken care of by penalizing lack of smoothness of fitted curve. In LASSO (which penalizes the least squares criterion or the log-likelihood criterion for normal linear models by the absolute values of the regression coefficients), one wants to select an optimum model in the presence of sparsity (i.e. when most regression coefficients are zero or close to zero). This is particularly useful in high dimensional problems. Fan and Li [2001] propose three desirable properties of a penalty function and choose a non-concave penalty which possesses optimal properties. However non-concavity makes the rule computationally inefficient. indexBIC

One can also see [Abramovich *et al.*, 2006] for another approach to handling sparse sequences, by connecting simultaneous testing of hypotheses and optimal rate of convergence. They conjectured in their paper that the modified BIC with per-parameter penalty of the form $\frac{\log(n/p)}{2}$ instead of $\frac{\log n}{2}$ would result in a good model selection rule providing optimal estimators in sparse sequences. Here p is the number of non-zero parameters in the model. It is worth pointing here that such a change in the BIC will reduce the difference between AIC and BIC criteria somewhat. This is so since in many real life problems the number of parameters p increases with sample size n (since complex data is typically modelled with more complex models, i.e. those with higher dimension) and if both n and p are large, the difference between the (per-parameter) penalties of AIC and (modified)BIC is expected to be smaller. (It is worth mentioning that the same penalty has been used earlier in [Pauler, 1998; Berger *et al.*, 2003], with the argument that the effective sample size per parameter is $\frac{n}{p}$ if the model under consideration has p free parameters.)

Much of the above theory for model selection is tied up with use of least squares estimates. If one uses Stein type shrinkage estimates, then also the notion of complexity changes a lot. This aspect has not been studied in the literature.

In the field of machine learning, particularly in classification problems, people have considered penalized empirical risk minimization, and used deep results from the theory of empirical processes to derive the penalty function which is sort of an upper bound of the error of estimation, defined to be the error in estimating the best approximating model (i.e. the member in the model space closest to the true model). Vapnik [1998, Chapter 6] recommends minimizing the structural risk which is essentially a penalized risk, with the penalty arising in a somewhat different way. Vapnik [1998, Chapter 6] also provides a good introduction to model selection based on Kolmogorov complexity or some approximation to it as in [Rissanen, 1978]. Other recent contributors to these aspects in machine learning are van de Geer, Koltchinskii, Bartlett, Lugosi, etc. The interested reader

may search the rich literature on this. A good starting point would be the abstracts of talks presented in a conference on model selection at Oberwolfach in 2005, which are available from the internet at the link <http://www.ems-ph.org/journals/abstract/owr/2005-02-04/2005-02-04-05.pdf>. It seems that some form unification of the theory of model selection may be possible in future integrating all these apparently disconnected facts through some more fundamental considerations.

This paper has been devoted mainly to discussing several model selection criteria, most of the emphasis being given to AIC and BIC. Typically, once a model is selected, one uses that model to do further inference on the data at hand or future data obtained from similar experiments. Instead one may use a Bayesian model average for estimation or prediction by combining the Bayes estimates under different models with weights proportional to marginal likelihoods of models. Marginal likelihood is defined in Section 3. See [Raftery *et al.*, 1997; Hoeting *et al.*, 1999] for more details. A similar idea in the classical statistical context has been proposed by Hjort and Claeskens [2003], Breiman [2001], etc. These ideas have become very popular among both Bayesians and non-Bayesians.

SUMMING UP

In the context of model selection, we discussed AIC and BIC (briefly mentioning some of their proposed modifications) along with the original motivation behind them. We studied concrete applications and theoretical results which reinforce our current big picture summary captured in the text. The BIC is more useful in selecting a correct model while the AIC is more appropriate in finding the best model for predicting future observations. Each of these facts hold under suitable conditions mentioned in the text. We also discussed some recent advances in model selection some of which are of interest to researchers in diverse fields of the scientific spectrum.

ACKNOWLEDGEMENT

The authors would like to thank Prasanta Bandyopadhyay and an anonymous referee for valuable suggestions leading to substantial improvement of the paper. In the process, it has clarified our own understanding of the subject even though we have not been able to obtain the version of Bayesian model selection that will be wholly satisfactory to philosophers. The authors would also like to thank M. Prem Laxman Das and Sumanta Sarkar for help with latex in drawing Figure 1.

BIBLIOGRAPHY

- [Abramovich *et al.*, 2006] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate, *The Annals of Statistics*

- 34: 584-653, 2006.
- [Akaike, 1973] H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle, in B. N. Petrov and C. Csaki (eds.), *Second International Symposium of Information Theory*. Budapest: Akademiai Kiado, 267-281, 1973.
- [Akaike, 1974] H. Akaike. A new look at the statistical model identification, *IEEE Transactions on Automatic Control* AC 19: 716-723, 1974.
- [Berger and Pericci, 1995] J. O. Berger and L. R. Pericchi. The Intrinsic Bayes Factor for Linear Models, in Bernardo, J. M. et al. (eds.), *Bayesian Statistics 5* London: Oxford University Press, 23-42, 1995.
- [Berger et al., 2003] J. O. Berger, J. K. Ghosh, and N. D. Mukhopadhyay. Approximations and consistency of Bayes factors as model dimension grows, *Journal of Statistical Planning and Inference* 112: 241-258, 2003.
- [Bernardo and Smith, 1994] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory* Wiley: Chichester, 1994.
- [Breiman, 2001] L. Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author), *Statistical Science* 16: 199-231, 2001.
- [Burnham and Anderson, 2003] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer: New York, 2003.
- [Chakrabarti and Ghosh, 2006a] A. Chakrabarti and J. K. Ghosh. Optimality of AIC in Inference about Brownian Motion, *Annals of the Institute of Statistical Mathematics* 58: 1-20, 2006.
- [Chakrabarti and Ghosh, 2006b] A. Chakrabarti and J. K. Ghosh. A generalization of BIC for the general exponential family, *Journal of Statistical Planning and Inference* 136: 2847-2872, 2006.
- [Chakrabarti and Ghosh, 2007] A. Chakrabarti and J. K. Ghosh. Some aspects of Bayesian model selection for prediction (with discussion), in Bernardo, J. M. et al. (eds.), *Bayesian Statistics 8*. Oxford University Press, 51-90, 2007.
- [Fan and Li, 2001] J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association* 96: 1348-1360, 2001.
- [Forster and Sober, 1994] M. Forster and E. Sober. , How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions, *The British Journal for the Philosophy of Science* 45: 1-35, 1994.
- [Gardner, 1997] M. Gardner. *Relativity Simply Explained*, Dover: New York 1997.
- [Ghosh, 2006] J. K. Ghosh. Different Role of Penalties in Penalized Likelihood Model Selection Rules - abstract of talk presented at a workshop on Multivariate Statistical Methods at the Indian Statistical Institute, 2006.
- [Ghosh and Samanta, 2001] J. K. Ghosh and T. Samanta. Model selection - An overview, *Current Science* 80: 1135-1144, 2001.
- [Ghosh et al., 2006] J. K. Ghosh, M. Delampady, and T. Samanta. *An Introduction to Bayesian Analysis: Theory and Methods* Springer: New York, 2006.
- [Hastie et al., 2001] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, New York: Springer, 2001.
- [Hjort and Claeskens, 2003] N. L. Hjort and G. Claeskens. Frequentist Model Average Estimators, *Journal of the American Statistical Association* 98: 879-899, 2003.
- [Hoeting et al., 1999] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky. Bayesian Model Averaging, *Statistical Science* 14: 382-401, 1999.
- [Joyce, 1999] J. Joyce. *The Foundations of Causal Decision Theory*, Cambridge: Cambridge University Press, 1999.
- [Lehmann and Casella, 2001] E. L. Lehmann and G. Casella. *Theory of Point Estimation* Springer-Verlag: New York, 2001.
- [Li, 1987] K. C. Li. Asymptotic optimality for c_p , c_l , cross validation and generalized cross validation: Discrete index set, *The Annals of Statistics*, 15: 958-975, 1987.
- [Mukhopadhyay, 2000] N. D. Mukhopadhyay. *Bayesian Model Selection for High Dimensional Models with Prediction Error Loss and 0-1 Loss*, Ph.D. Thesis, Purdue University, 2000.
- [Mukhopadhyay and Ghosh, 2003] N. D. Mukhopadhyay and J. K. Ghosh. Parametric Empirical Bayes Model Selection - Some Theory, Methods and Simulations, in K. B. Athreya et al. (eds.), *IMS lecture notes in honor of Rabi Bhattacharya*, 2003.

- [Pauler, 1998] D. K. Pauler. The Schwarz criterion and related methods for normal linear models, *Biometrika* 85: 13-27, 1998.
- [Raftery *et al.*, 1997] A. Raftery, D. Madigan, and A. J. Hoeting. Bayesian Model Averaging for Linear Regression Models, *Journal of the American Statistical Association*, 92: 179-191, 1997.
- [Rissanen, 1978] J. Rissanen. Modeling by shortest data description, *Automatica* 14: 465-471, 1978.
- [Schwarz, 1978] G. Schwarz. Estimating the dimension of a model, *The Annals of Statistics* 1978: 461-464, 1978.
- [Shao, 1997] J. Shao. An asymptotic theory for linear model selection, *Statistica Sinica* 7: 221-264, 1997.
- [Shibata, 1981] R. Shibata. An optimal selection of regression variables, *Biometrika* 68: 45-54, 1981.
- [Shibata, 1983] R. Shibata. Asymptotic mean efficiency of a selection of regression variables, *Annals of the Institute of Statistical Mathematics* 35: 415-423, 1983.
- [Spiegelhalter *et al.*, 2002] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* 64: 583-649, 2002.
- [Stone, 1977] M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society, Series B, Methodological* 39: 44-47, 1977.
- [Stone, 1979] M. Stone. Comments on model selection criteria of Akaike and Schwarz, *Journal of the Royal statistical Society, Series B, Methodological* 41: 276-278, 1979.
- [Vapnik, 1998] V. Vapnik. *Statistical Learning Theory*, Wiley: New York, 1998.
- [Wasserman, 2006] L. Wasserman. *All of Nonparametric Statistics*, Springer, 2006.

This page intentionally left blank

POSTERIOR MODEL PROBABILITIES

A. Philip Dawid

INTRODUCTION

We consider the problem of Bayesian inference about the statistical model from which the data arose. By examining the asymptotic dependence of posterior model probabilities on the prior specifications and the data, we refute the conventional wisdom that such problems of model choice exhibit more sensitivity to the prior than is the case for standard parametric inference.

Where improper priors are used, the definition of Bayes factor has been regarded as problematic. We argue that many of the supposed difficulties can be avoided by specifying a single overall prior as a weighted sum of measures on the various model spaces (the whole being defined up to a single arbitrary scale factor), and focusing attention directly on the typically proper posterior distribution across models that this implies. We discuss how to select the weights, both using subjective inputs and using formal rules. For the latter case, we demonstrate the importance of using the Jeffreys prior within each model — a choice that goes a long way towards resolving many of the perceived problems connected with arbitrary scaling constants. We illustrate the general theory by constructing ‘reference posterior probabilities’ for normal regression models, and by analysis of an ESP experiment.

1 MARGINAL LIKELIHOOD

Suppose we are uncertain about our statistical model. Conditional on the validity of any candidate M , we have a specified statistical model for the observable X , depending on a parameter θ , having density $p(x \mid M, \theta)$. The specification, meaning, and dimensionality of θ will typically be different for different models.

The prior specification is generally given in two parts:

1. A prior distribution over the various models, model M having prior probability $p(M)$; and
2. Within each model M , a prior density $p(\theta \mid M)$ for its parameter θ , conditional on its validity.

We shall assume, without explicit comment, appropriate regularity conditions on the model densities $p(x \mid M, \theta)$ and the prior densities $p(\theta \mid M)$. In particular we assume that $p(\theta \mid M) > 0$ everywhere.

On observing data x , the posterior probability of model M is given by

$$(1) \quad p(M | x) \propto p(M) \times L(M),$$

where $L(M)$ is the *marginal likelihood* of a model M for data x :

$$(2) \quad L(M) \propto \int p(x | M, \theta) p(\theta | M) d\theta$$

$$(3) \quad = p(x | M),$$

the *marginal density* at x under M . The proportionality signs in (1) and (2) indicate an unspecified multiplier which does not depend on M ; in (1), this is such as to ensure that the right-hand side sums to unity over all the different models considered.

The effect of using equations (1) and (2) is to separate out the rôles of the two ingredients of the prior specification, with only the within-model priors entering $L(M)$.

The *Bayes factor* for comparing any two models is just the ratio of their marginal likelihoods; then the posterior odds for this comparison will be just the prior odds multiplied by the Bayes factor. However, in general we will be interested in a larger, finite or countable, set of models, and then the more appropriate thing to look at is the complete *marginal likelihood* function (2), defined, up to an irrelevant proportionality constant, as a function of the model label. A Bayes factor is just the ratio of two terms in this likelihood function. We remark in passing that a compelling minimal criterion for any ‘pseudo-Bayes factor’ is that it be related to some full ‘pseudo-marginal likelihood function’ in just this way. This criterion is not satisfied by many suggestions that have been made, including those of Spiegelhalter and Smith [1982] and most of the ‘intrinsic Bayes factors’ of Berger and Perricchi [1996a; 1996b; 1997].

2 ASYMPTOTICS

2.1 Consistency

An important general property of the posterior distribution (1) is *consistency* [Dawid, 1992, §6.4]. Suppose that the observables $(X_1, X_2, \dots)$ arrive in sequence. We do not need to assume these observables to be independent and identically distributed. We entertain a finite or countable collection of parametric models $\{M_j\}$. Then, under very weak conditions, the prior probability is unity that the posterior distribution given by (1) will converge to a point mass on the true model (or, if there is more than one model containing the true distribution, on the true model of smallest dimension). Under still weaker (virtually vacuous) conditions, using a Bayesian model mixture with weights given by (1) (and with the appropriate posterior distributions within each model) will, with prior probability 1, yield probability forecasts asymptotically indistinguishable from those derived from the true generating distribution. Note that the expression “with prior probability 1” here means, essentially, under any data-generating distribution belonging to any of the models under consideration.

2.2 Behaviour of marginal likelihoods

We now describe some further refinements of the above consistency property, through a none too rigorous account of the behaviour of the marginal likelihood $L(M)$ as the sample size increases. We have to make some more specialised assumptions. For simplicity, we suppose that the observables $(X_1, X_2, \dots)$ are generated independently and identically from some (unknown) distribution Q . Let us fix on an arbitrary one of the models under consideration, M say, and drop it from the notation. We further suppose that, for each value of the parameter θ , which ranges over a Euclidean space of dimension d , the (X_i) are modelled as independent and identically distributed, with distribution P_θ . We denote the Fisher information for a single observation by $I(\theta)$. Let $q^n(x^n)$, $p^n(x^n | \theta)$ be the (product) densities for $X^n := (X_1, \dots, X_n)$ generated by Q and P_θ , and $l^n(\theta) = \log p^n(X^n | \theta)$. Define $K(Q, M) := \inf_\theta K(Q, P_\theta)$, where $K(Q, P)$ denotes the Kullback-Leibler distance between Q and P . We suppose that this infimum is attained, at say θ^* . In particular, if $Q \in M$ then $K(Q, M) = 0$ and $Q = P_{\theta^*}$. Defining the marginal density $p^n(x)$ as in (3), we can then show the following:

THEOREM 1. Under Q , as $n \rightarrow \infty$:

(i) If $Q \notin M$,

$$(4) \quad \log p^n(X^n)/q^n(X^n) = -n K(Q, M) + O_p(n^{\frac{1}{2}}).$$

(ii) If $Q \in M$,

$$(5) \quad \log p^n(X^n)/q^n(X^n) = -\frac{1}{2} d \log \left(\frac{n}{2\pi e} \right) + \log \rho(\theta^*) + U_n + O_p(n^{-\frac{1}{2}}),$$

where $\rho(\theta) = p(\theta)/\{\det I(\theta)\}^{\frac{1}{2}}$. The random variable U_n does not depend on the prior density $p(\theta)$, and asymptotically $U_n \sim \frac{1}{2}(\chi_d^2 - d)$. In particular, the asymptotic expectation of U_n is 0.

The proof of Theorem 1 is given in Appendix A.

We note that the function $\rho(\cdot)$ is the density of the prior distribution with respect to the Jeffreys invariant measure μ_M satisfying $d\mu_M(\theta) = \{\det I(\theta)\}^{\frac{1}{2}} d\theta$. In particular, $\rho(\cdot)$ is unaffected by non-linear reparametrisation. We shall term $\rho(\cdot)$ the *invariantised prior density* over M .

2.3 Behaviour of posterior model probabilities

As we vary the model M , differences in the left-hand side of (4) or (5) give the differences in the log marginal likelihood $l(M) = \log L(M)$ of the models. These are thus approximated, for large n , by differences of expressions of the form of the right-hand sides of these equations, as appropriate. Incorporating the prior model probabilities $p(M)$, as in (1), we can thus obtain the posterior probabilities. In particular, suppose we are comparing two models, M_1 and M_2 , of respective dimensions d_1 and d_2 . Then:

- (a) If $K(Q, M_2) > K(Q, M_1)$ (so that, in particular, model M_2 does not hold, whereas M_1 might or might not), we have

$$(6) \quad \log \frac{p(M_1 | x)}{p(M_2 | x)} = n\{K(Q, M_2) - K(Q, M_1)\} + O_p(n^{\frac{1}{2}}),$$

so that the data will overwhelmingly favour M_1 , the posterior odds ratio being exponential in the sample size n .

- (b) If both M_1 and M_2 hold, with say $q(x) \equiv p(x | M_1, \theta_1^*) \equiv p(x | M_2, \theta_2^*)$,

$$(7) \quad \log \frac{p(M_1 | x)}{p(M_2 | x)} = \frac{1}{2}(d_2 - d_1) \log \frac{n}{2\pi e} + \log \frac{p(M_1)\rho(\theta_1^* | M_1)}{p(M_2)\rho(\theta_2^* | M_2)} + V,$$

where $V = O_p(1)$ with asymptotic expectation 0, and, by (??6), the dependence of V on the prior specification is $O_p(n^{-\frac{1}{2}})$. In particular, if $d_1 < d_2$ then the data will again overwhelmingly favour M_1 (although the posterior odds will now only grow as a power of n). If $d_1 = d_2$, the data will not be able to make a definitive choice between the two models. However, it is usual to entertain a set of models which is closed under intersection. In this case, $M_3 = M_1 \cap M_2$ will be a true model of typically smaller dimension than either M_1 or M_2 , and it will then be M_3 that attracts the overwhelming log marginal likelihood.

2.4 Sensitivity to the prior

It is generally believed that sensitivity to prior specification is a bigger problem for comparing models than for prior-to-posterior inference within a model, and this belief has motivated various attempts to replace ‘true’ marginal likelihoods, or Bayes factors, with supposedly more ‘objective’ variants. However, I consider this conclusion to be seriously misleading. The full posterior distribution over models depends on both prior ingredients (between and within models) and on the data. We need to examine its relative sensitivity to all these inputs. We have already seen that the contribution of each of the prior ingredients remains of order 1 as $n \rightarrow \infty$.

In case (a) of §2.3 above, when comparing two models at least one of which is false, (6) shows that both ingredients of the prior are completely negligible in comparison with the two leading terms. Hence the choice of prior distribution, both between and within models, is utterly immaterial here.

In case (b), of comparing two true models, the prior contributions are still negligible in comparison with the leading term, but now enter into the next, order 1, term in (7). There they are, however, intertwined, both with each other and with the ‘random’ term V , depending only on the data. (If M_1 is a submodel of M_2 , then asymptotically $V \sim \frac{1}{2}\{\chi_{d_2-d_1}^2 - (d_2 - d_1)\}$, by Wilks’ Theorem).

Suppose now that M_1 is the simplest true model, and M_2 the next simplest, both supposed to exist and to be unique. The posterior probability of M_1 is then seen to have the form

$$(8) \quad p(M_1 | X) \approx 1 - kn^{-\frac{1}{2}(d_2-d_1)}$$

where the term k , of order 1, involves the order 1 terms in (7) — and so, in particular, the contributions of the prior.

How does this degree of sensitivity to the prior compare with that of asymptotic within-model posterior distributions? Consider for simplicity the univariate case, and let θ_κ be the κ -quantile of the approximating $\mathcal{N}(\hat{\theta}, \{-l''(\hat{\theta})\}^{-\frac{1}{2}})$ posterior distribution of θ . Then it can be shown [Johnson, 1970] that the true posterior probability that $\theta \leq \theta_\kappa$ differs from κ by an expression with leading term of the form $\gamma n^{-\frac{1}{2}}$; here γ is an order 1 quantity involving the data and the first derivative of the log prior density, evaluated at $\hat{\theta}$. From this point of view, we see that the behaviour of the posterior probability distribution over models is in no way more sensitive to the choice of the prior than is the posterior distribution within a model; indeed, whenever $d_2 - d_1 > 1$, it is strictly less sensitive. (It is interesting, but of no great significance, that it is the actual value of the log prior density at $\hat{\theta}$ that affects the distribution over models, whereas it is its derivative there that affects the distribution within models. The point remains that in each case the posterior will be sensitive to the prior specification, to the order described).

If we consider, as seems reasonable, the total posterior probability of all true models (assuming that there is at least one), we find that this behaves as $1 - e^{-nK}$, with K the minimum, over all false models M , of the Kullback-Leibler distance $K(Q, M)$. (Here we assume $K > 0$ — things become much more delicate if $K = 0$, as can happen when we entertain a countable collection of models.) Not only is the contribution of the prior specification completely negligible, but this probability approaches the ‘true’ value 1 at a much faster rate than that governing the convergence of within-model posterior probabilities.

It follows from the above considerations that we can determine a set of distributions for X whose posterior probability of containing the unknown true data-generating distribution differs from κ by only $O(n^{-\frac{1}{2}})$, as follows. First select the most probable model, then construct, within it, the usual approximate κ -region based on asymptotic normality. We do not even have to use the true posterior model probabilities; the wrong prior, or a crude approximation such as BIC, effectively the leading term of (7), will produce the same asymptotics.

3 IMPROPER PRIORS

The dependence of the posterior odds (7) on the prior specifications, though of smaller order than the dominant data-driven term, is nevertheless often of interest. We have

$$(9) \quad \log \frac{p(M_1 | x)}{p(M_2 | x)} = \frac{1}{2}(d_2 - d_1) \log \frac{n}{2\pi e} + \log R_p + V$$

with

$$(10) \quad R_p := \frac{p(M_1)\rho(\theta_1^* | M_1)}{p(M_2)\rho(\theta_2^* | M_2)},$$

so that asymptotically the contribution of the priors is completely confined to the single quantity R_p . Correspondingly the Bayes factor, which is the ratio of posterior to prior odds, is governed by

$$(11) \quad R_l := \frac{\rho(\theta_1^* | M_1)}{\rho(\theta_2^* | M_2)}.$$

Each of these terms is highly sensitive to the ordinate of the invariantised prior density ρ at the maximum likelihood estimate, for each model. This has worried many workers, who, seeking an ‘objective’ solution, would like to be able to use improper prior densities. However, as these are generally only specified up to an arbitrary scale factor, they do not determine a well-defined ordinate value. This has led to a variety of tricks and techniques, leading to revised definitions [Berger and Pericchi, 1996a; Berger and Pericchi, 1996b; Berger and Pericchi, 1997], some of which may even fail to follow the (prior-insensitive) leading term asymptotics of (7) [Aitkin, 1991; O’Hagan, 1995].

Now it seems a little strange to concentrate attention on the Bayes factor, or the related expression (11), since this incorporates just one of the two equally relevant ingredients of the prior. Instead we shall find it more useful and revealing to focus directly on the posterior odds, and on the associated constituent (10) of expression (9).

For a generic model M and value θ of its parameter, let us write P for the particular data-distribution these determine, and define the measure Γ_M over distributions in M by: $d\Gamma_M(P) = p(M)p(\theta | M)d\theta$. The overall prior distribution is thus

$$(12) \quad \Gamma = \sum_M \Gamma_M,$$

which is a measure over the full model-space $\mathcal{M} = \{P : P \in M, \text{ some } M\}$, the union of the various simple model-spaces under consideration.

We also introduce

$$(13) \quad \gamma(P) = \gamma(M, \theta) := p(M)\rho(\theta | M).$$

Confined to distributions $P \in M$, γ is thus the (invariantised) density of Γ_M with respect to its Jeffreys measure μ_M . (Note that, although μ_M was originally introduced as a measure on the parameter-space for M , it can equally be regarded as defined directly over M itself, a space whose ‘points’ are probability distributions; we shall move between such equivalent representations of μ_M , Γ_M , *etc.* without further comment.) As P ranges over $\mathcal{M}$, then γ is the mixed discrete-continuous prior density of Γ with respect to the underlying measure

$$(14) \quad \mu := \sum_M \mu_M.$$

The measure μ is the simple unweighted sum of the Jeffreys measures on the various models, and is a well-defined σ -finite measure over $\mathcal{M}$. This is a ‘natural’ base measure to use in our model-mixture problem. We shall term μ the *meta Jeffreys measure* for the model collection $\mathcal{M}$.

An important property of the meta Jeffreys measure μ is that its value, for any set, is a dimensionless real number. Consequently the density γ must also take such values, since its integral with respect to μ yields probabilities.

In terms of the above notation, we have, simply,

$$(15) \quad \frac{p(M_1 | x)}{p(M_2 | x)} = \frac{\int p(x | M_1, \theta) d\Gamma_{M_1}(\theta)}{\int p(x | M_2, \theta) d\Gamma_{M_2}(\theta)},$$

and

$$(16) \quad R_p = \gamma(P_{M_1}^*)/\gamma(P_{M_2}^*),$$

where P_M^* is the distribution which maximises the likelihood over model M .

Now

$$(17) \quad \Gamma(\mathcal{M}) = \sum_M \Gamma_M(M),$$

$$(18) \quad i.e., \quad \int_{\mathcal{M}} \gamma(P) d\mu(P) = \sum_M \int \gamma(M, \theta) d\mu_M(\theta).$$

For the case of a proper prior, $\Gamma(\mathcal{M}) = 1$, so that the various summands on the right-hand side (in which each integral is confined to a single model M) are finite and add to 1. We can, if we like, extract the implied prior model probabilities $p(M) = \Gamma_M(M) = \int \gamma(P) d\mu_M(P)$, and the within-model- M invariantised prior density $\rho(\theta | M) = \gamma(M, \theta)/p(M)$, but there is no necessity to do so.

An ‘improper’ prior distribution is a σ -finite measure Γ on $\mathcal{M}$, determined up to a single arbitrary scale factor, and such that $\Gamma(\mathcal{M}) = \infty$. In the cases of interest we will in fact have $\Gamma_M(M) = \infty$ for some or all models M . The formal posterior odds is again given by (15), which is perfectly well-defined so long as both integrals in it converge, which will typically be the case so long as the sample size n is large enough. Likewise the quantity R_p which enters into the asymptotic expression (9) for the posterior odds is well-defined by (16).

In contrast, we cannot sensibly define either the prior odds or the Bayes factor. For the prior odds could only be $p(M_1)/p(M_2) = \Gamma_{M_1}(M_1)/\Gamma_{M_2}(M_2)$, both terms in which will typically be infinite, so that their ratio is meaningless. Correspondingly, we cannot normalise Γ_M , for either model M , to obtain a meaningful probability distribution for θ conditional on M , and so cannot determine $p(x | M) = \int p(x | M, \theta) d\Gamma_M(\theta_M)$, even up to a common scale factor. The same problem affects the expression (11) entering into the asymptotic Bayes factor. However, the problematic infinities cancel when we combine the ingredients to form the posterior odds (15), or the related expression (16).

This analysis suggests that many of the perceived difficulties of defining Bayes factors with improper priors may be due to an inappropriate formulation of the problem. If we concentrate attention instead on the posterior distribution across models, which is after all what we are interested in, then this can be perfectly well-defined, even when Bayes factors are not. This is exactly what we are used to for within-model inference, where we can readily determine a formal posterior odds that θ belongs to a set S , even though it may not be possible sensibly to define either the prior odds on this, or a relevant likelihood ratio. In this within-model case, we are happy to specify ‘ignorance’ by means of a measure, rather than a ‘likelihood’, over the full set of possible distributions. Why then should we try to do otherwise when the model is also uncertain?

4 SUBJECTIVE MODEL WEIGHTS

Suppose that we have decided that the prior measures Γ_M should be of the form $w(M)\Lambda_M$, for some specified (perhaps improper) σ -finite measures Λ_M on the various models M . It remains to choose the prior weights $w(M)$. This can be regarded as a ‘subjective/objective Bayes compromise’, in that we may be using ‘objective’ criteria to choose the within-model base prior measures Λ_M , but allow subjective input into determining the weights.

Note that the weights $w(M)$ are *not* prior probabilities, unless the Λ_M are proper and integrate to 1; nevertheless, they affect posterior odds in essentially the same way as do prior probabilities in the proper case. Note also that, if the values that the (Λ_M) take are not dimensionless, then the weights $w(M)$ will have to be appropriately dimensioned to adjust for this. For example, in a normal model for a length X , with unknown mean μ and standard deviation σ , both these parameters must also have the dimensions of length. The standard (right-invariant) improper prior measure Λ has density element $d\Lambda(\theta, \sigma) = \sigma^{-1} d\mu d\sigma$. If we change our units from, say, centimetres to inches, the values taken by Λ will scale in the same way: they themselves have the dimension of length. Consequently, any weight attached to Λ must have the dimensions of reciprocal length, in order to ensure that resulting integrals are dimensionless.

Although the improper overall prior distribution does not assign finite ‘probabilities’ to the different models, it does so to bounded subsets of the model spaces. In particular, for two models M_1 and M_2 , and bounded subsets A_1 and A_2 of their respective parameters spaces, we have:

$$(19) \quad \frac{p(M_1 \text{ holds and } \theta_1 \in A_1)}{p(M_2 \text{ holds and } \theta_2 \in A_2)} = \frac{w(M_1)}{w(M_2)} \frac{\Lambda_{M_1}(A_1)}{\Lambda_{M_2}(A_2)}.$$

By adjusting the ratio $w(M_1)/w(M_2)$, we can thus ensure that this agrees with subjective assessments as to the relative plausibility of the two events “ M_1 holds and $\theta_1 \in A_1$ ” and “ M_2 holds and $\theta_2 \in A_2$ ”.

In principle we could make many such assessments, varying the pairs of models and the sets in the spaces which we consider. However, we should strive to make

these self-consistent. If, for fixed models M_1 and M_2 , we consider various different choices for the pair (A_1, A_2) , we should want our assessments to agree with a common choice for the ratio $w(M_1)/w(M_2)$. Further, for any three models M_1, M_2 and M_3 , we should require the separate assessments of their pairwise ratios to satisfy $\{w(M_1)/w(M_2)\} \times \{w(M_2)/w(M_3)\} = w(M_1)/w(M_3)$. While ensuring such self-consistency is a demanding task, it is not in principle different from any other case where we need to establish a prior distribution which will be a satisfactory approximation to real subjective beliefs. If it does not seem possible to accomplish this, it may be necessary to rethink the initial specifications of the Λ_M .

A related idea is to consider the implied (typically still improper) predictive distributions for some future observation or observations Y . Thus we might address attention to comparisons of the form

$$(20) \quad \frac{p(M_1 \text{ holds and } Y \in A_1)}{p(M_2 \text{ holds and } Y \in A_2)} = \frac{w(M_1)}{w(M_2)} \frac{\int p(Y \in A_1 \mid M_1, \theta_1) d\Lambda_{M_1}(\theta_1)}{\int p(Y \in A_2 \mid M_2, \theta_2) d\Lambda_{M_2}(\theta_2)},$$

again adjusting the ratio $w(M_1)/w(M_2)$ to bring about agreement with prior assessments (and again striving for self-consistency). This method has the advantage that the sets A_1 and A_2 belong to the same space. Indeed, the method will typically be most intuitive and useful when we choose $A_1 = A_2 = A$ say. In the limit, as A shrinks to a single point y , the left-hand side of (20) reduces to the posterior odds for comparing the two models consequent to observing $Y = y$:

$$(21) \quad \frac{p(M_1 \mid Y = y)}{p(M_2 \mid Y = y)} = \frac{w(M_1)}{w(M_2)} \frac{\int p(y \mid M_1, \theta_1) d\Lambda_{M_1}(\theta_1)}{\int p(y \mid M_2, \theta_2) d\Lambda_{M_2}(\theta_2)}.$$

We may thus evaluate $w(M_1)/w(M_2)$ by choosing a suitable hypothetical future observation $Y = y$ and assessing the resulting posterior odds directly.

5 ‘FULLY OBJECTIVE’ FORMAL PRIORS

Suppose now that we wish to exclude all prior input, and so choose the weights, as well as the within-model measures, by purely formal means. That is, we seek a ‘reference’ measure Γ^{ref} , over the union $\mathcal{M}$ of all the models under consideration. Note that we use the terms “reference prior” or “reference measure” in a generic sense, rather than in the more specialised technical sense of Berger and Bernardo [Bernardo, 1979; Berger and Bernardo, 1996].

5.1 Choice of reference measure

Let γ^{ref} denote the invariantised density of Γ^{ref} , with respect to the meta Jeffreys measure μ . How might we select γ^{ref} ? Recall that any such density takes values which are pure dimensionless numbers. It is difficult to conceive of any argument that could justify taking $\gamma^{\text{ref}}(M, \theta) \neq \gamma^{\text{ref}}(M, \theta')$, for two parameter-values θ, θ'

in the same model M . Consequently we shall restrict attention to the case that $\gamma^{\text{ref}}(M, \theta)$ is a constant, $w^{\text{ref}}(M)$ say, for all such θ . This corresponds to using a prior measure of the form

$$(22) \quad \Gamma^{\text{ref}} = \sum_M w^{\text{ref}}(M) \mu_M,$$

with $w^{\text{ref}}(M)$ serving to weight the Jeffreys prior μ_M for model M . The formal posterior model probabilities will then be

$$(23) \quad p(M | x) \propto w^{\text{ref}}(M) \bar{p}_M(x)$$

where

$$(24) \quad \bar{p}_M(x) := \int p(x | M, \theta) d\mu_M(\theta)$$

is the (typically improper) ‘reference marginal density’ of X under M . By (9) and (16), we have the asymptotic behaviour

$$(25) \quad \log \frac{p(M_1 | x)}{p(M_2 | x)} = \frac{1}{2}(d_2 - d_1) \log \frac{n}{2\pi e} + \kappa + V$$

when both models M_1 and M_2 hold, where now $\kappa = \log\{w^{\text{ref}}(M_1)/w^{\text{ref}}(M_2)\}$.

5.2 Reference model probabilities

How should we set the weights $w^{\text{ref}}(M)$ — recalling that these are dimensionless real numbers?

The simplest proposal is just to take equal weights, $w^{\text{ref}}(M) \equiv 1$ for all M , equivalent to using the meta Jeffreys measure μ itself as the reference prior distribution. This leads to the formula

$$(26) \quad p^{\text{ref}}(M | x) \propto \bar{p}_M(x),$$

with $\bar{p}_M(x)$ given by (24), μ_M being the Jeffreys measure on model M . We shall (somewhat arbitrarily) fix on this as our basic definition, thus calling posterior probabilities constructed according to (26) the *reference posterior model probabilities*. Note that, with this choice, we have $\kappa = 0$ in (25).

5.3 Adjusting the weights

An important property of the reference posterior distribution (26) over models is that it is genuinely easy to use it for *reference*, rather than as a solution in itself. If, instead of the implicit weights $w^{\text{ref}}(M) \equiv 1$ in (22), we wish to use alternative weights $w(M)$, we can readily adjust the reference probabilities:

$$(27) \quad p(M | x) \propto w(M) \bar{p}_M(x)$$

$$(28) \quad \propto w(M) p^{\text{ref}}(M | x).$$

We have already discussed subjective methods of assessing the $w(M)$. We can also consider ‘formal’ choices other than $w \equiv 1$. For example, it might be thought appropriate to make a ‘correction for dimensionality’, so taking the weights to have the form

$$(29) \quad w(M) = g(d_M)$$

for some function $g(\cdot)$, where d_M is the dimensionality of model M . An appealing choice for the function g is

$$(30) \quad g(d) = k^d$$

for some dimensionless constant k . This leads to

$$(31) \quad p(M \mid x) \propto k^{d_M} \bar{p}_M(x),$$

with asymptotics

$$(32) \quad \log \frac{p(M_1 \mid x)}{p(M_2 \mid x)} = \frac{1}{2}(d_2 - d_1) \log \frac{n}{2\pi k^2 e} + V.$$

The choice $k = (2\pi)^{-\frac{1}{2}}$ has been suggested by Wasserman [2000] as that producing closest agreement with the Jeffreys-Schwarz Bayesian Information Criterion (BIC), which is itself based on the approximation $-\frac{1}{2} d \log n$ for A in (??1) (compare (??9)). However, there seems to be no compelling reason to regard one choice rather than another for the omitted $O(1)$ terms as more ‘objective’; and, as always when considering sensitivity to the prior, in (25) the effect κ of the choice of weights is asymptotically unimportant in comparison with the driving term $\frac{1}{2}(d_2 - d_1) \log n$ (at any rate, for comparing models of different dimensions).

5.4 Reference marginal likelihood

Although the prior weights $w(M)$ in (27) are *not* prior probabilities, they behave much as if they were, and can be given a similar interpretation. Likewise, although $\bar{p}_M(x)$ is not a proper marginal likelihood for M , it too behaves much as if it were. Thus we may reasonably refer to $\bar{L}(M) \propto \bar{p}_M(x)$ as the *reference marginal likelihood* over models M , based on data x .

Now suppose that we have a single parametric model M , with parameter θ , say, and are interested in a function ϕ of θ . Each hypothetical value for ϕ defines a submodel M_ϕ of M , and so we can use the above analysis to construct a reference marginal likelihood for ϕ . This is thus defined by:

$$(33) \quad \bar{L}(\phi) \propto \int p(x \mid \theta) d\mu(\theta \mid \phi),$$

where $\mu(\cdot \mid \phi)$ is the Jeffreys measure over the model M_ϕ . (In forming $\mu(\cdot \mid \phi)$, we must not, of course, discard any ‘proportionality constant’ that depends on ϕ .) Then (33) is fully invariant under reparametrisation.

How are we to use such a marginal likelihood? Naturally, by combining it with a set of prior weights $w(M)$, according to (27). But we must be careful in assigning these weights, taking into account that they should not be interpreted as probabilities. The suggestions of §4 and §5.3 could be useful here, but the problem of interpretation of the weights deserves still further attention.

5.5 *Invariant and non-invariant priors*

The absolute invariance of the Jeffreys measure ensures that it is dimensionless. In particular, and essential for our purposes, it is not plagued by the problem of an unspecified multiplier: twice the Jeffreys measure is not a Jeffreys measure. In this respect at least, even though it may not be normalisable, it behaves like a proper probability distribution.

Consider first the model, for an observable X , which is normal with unknown mean θ and known variance σ^2 . The Jeffreys measure element for this model is

$$(34) \quad \sigma^{-1} d\theta,$$

which is dimensionless, both θ and σ being measured in the same units as X . This density element is often described simply as “proportional to $d\theta$ ”, with the scaling constant σ^{-1} omitted, but this is misleadingly imprecise; the problem of determining an appropriate overall scale to apply to the measure element $d\theta$ would then be exacerbated by the need to use a divisor with the same dimensions as X — a constraint which is all too rarely appreciated.

If we now suppose the variance to be unknown also, the Jeffreys measure element becomes

$$(35) \quad \sqrt{2} d\theta d\sigma / \sigma^2,$$

which is again (as it must be) dimensionless. Note also the fully determined numerical multiplier $\sqrt{2}$. Other location or location-scale families would yield Jeffreys measures essentially the same as (34) or (35), but with possibly different numerical multipliers.

We can decompose the bivariate Jeffreys measure $\sqrt{2} d\theta d\sigma / \sigma^2$ in a natural way, as the product of the (uniquely defined) conditional Jeffreys measure $\sigma^{-1} d\theta$ for θ given σ , and the remaining term $\sqrt{2} d\sigma / \sigma$, which term may thus be thought of as defining the ‘marginal Jeffreys measure’ for σ ; note that the expression $\sqrt{2}(d\sigma/\sigma^2) \int d\theta$ which would formally define such marginalisation involves an infinite integral, and is thus meaningless. This should be contrasted with the more ‘obvious’, but non-invariant, decomposition into $d\theta$ and $\sqrt{2} d\sigma / \sigma^2$. (We remark, however, that such ‘natural’ decomposition of the Jeffreys measure is not always possible. Even in the normal case, we cannot decompose the bivariate Jeffreys measure into the conditional Jeffreys measure for σ given θ , which is $\sqrt{2} d\sigma / \sigma$ irrespective of θ , and any marginal measure for θ .)

By contrast, other improper prior distributions (*e.g.* right Haar measure, Berger–Bernardo reference priors, ...) which have been suggested — even though they

may be ‘relatively invariant’ under scale or other relevant transformations — are typically not fully invariant. In particular, they are not dimensionless, and so can not come ready supplied with a natural overall scale. For instance, the improper prior commonly suggested for the above two-parameter normal model is that with element “proportional to $d\theta d\sigma/\sigma$ ”. However, we recall that $d\theta d\sigma/\sigma$ has the same dimensions as X . Without any further quantity in the problem which could set a yardstick for defining ‘unit length’, there is nothing we can use to eliminate this dimensional dependence. Correspondingly, there can be no universal formal prescription for choosing the arbitrary multiplier, since this has to be a dimensioned quantity. Likewise, there can be no natural decomposition of such a bivariate measure into (say) a conditional measure for θ given σ and a marginal measure for σ .

It is for the above reasons that we do not consider it as *generally* appropriate to use other improper priors than the Jeffreys measure for purposes of ‘fully objective’ formal model comparison. There may be some cases where this constraint can be relaxed. For example, if all our models are location-scale models, with location parameter θ and scale parameter σ , we might choose to use the common measure $k d\theta d\sigma/\sigma$, where the choice of the (dimensioned) constant k will not affect model comparisons. (Even here, there is no obvious reason not to allow the scaling constant k to vary across such models, similar to the case for the Jeffreys prior.) In some other problems there may be, in each model M , a function of its parameter θ_M which corresponds to a common externally defined quantity ϕ . One might then combine an essentially arbitrary common measure for ϕ in each model with a suitable reference measure for θ_M given ϕ — which, for the reasons discussed above, will generally be the conditional Jeffreys measure, and which may in any case incorporate a ‘proportionality constant’ depending on ϕ .

6 PROBLEMS WITH THE JEFFREYS MEASURE

There are nevertheless some difficulties associated with the use of the Jeffreys measure.

6.1 Problems of sample size

We have concentrated attention on problems where the observations are independent and identically distributed. In this case, the Jeffreys measure appears well determined, independent of the sample size. Even in this special problem, however, this is not really the case, as the following example illustrates.

Example. Consider the model:

$$(36) \quad M_1 : X_i \sim \mathcal{N}(\theta, \sigma^2), \quad i = 1, \dots, n,$$

independently, σ^2 being known. We wish to compare this with an alternative model M_2 , which postulates $\theta = 0$ in (36). Suppose first that $n = 1$. The Jeffreys

prior for M_1 , $\sigma^{-1} d\theta$, leads to $\bar{p}_1(x) = \sigma^{-1}$, for any x , and so the reference posterior odds in favour of M_2 is

$$(37) \quad \Omega := \frac{p(x \mid \theta = 0)}{\bar{p}_1(x)} = (2\pi)^{-\frac{1}{2}} \exp(-(x^2/2\sigma^2)).$$

If we now have a sample of size n , it seems that we should, without loss, be able to reduce the data to the single sufficient statistic $\bar{X} \sim \mathcal{N}(\theta, \tau^2)$, where $\tau^2 = \sigma^2/n$, and where again $\theta = 0$ under M_2 . Applying (37) to this now 1-dimensional case would then yield the reference odds

$$\Omega_n := (2\pi)^{-\frac{1}{2}} \exp(-(n\bar{x}^2/2\sigma^2)).$$

Now, under M_2 , $n\bar{X}^2/\sigma^2 \sim \chi_1^2$, for any n . So the distribution of Ω_n does not depend on n . This result also follows directly from the complete invariance of the problem under a change of scale, which is exactly the effect that increasing the sample size n has on the model for $\bar{X}$. However, this seems to contradict (25), which would require $\Omega_n = \frac{1}{2} \log n + O_p(1)$. How can this be?

The above result does not require the initial reduction by sufficiency, and would also ensue if we used the model for all the data —together with the Jeffreys prior for this model. And there's the rub: the Jeffreys prior based on the model for $(X_1, \dots, X_n)$ (or, equivalently, on that for $\bar{X}$), is not independent of n : it is in fact $n^{\frac{1}{2}} \sigma^{-1} d\mu$. And, if we change the prior with n in this drastic way as the data accrue, of course we will upset the asymptotics (which assume a fixed prior). So we need to find a way to adjust for this effect of sample size on the Jeffreys measure: for example, by basing it only on the model for a single observation — always supposing we can give meaning to this phrase.

In general, the Jeffreys measure μ_M^s associated with a sequence of s independent and identically distributed observations from a model M , of dimension d_M , satisfies

$$(38) \quad \mu_M^s = s^{d_M/2} \mu_M.$$

Now if our observables $(X_i : i = 1, \dots, n)$ form a random sample of size n , we can just as well think of them as a random sample $(Y_j : j = 1, \dots, n')$, of size $n' = n/s$, of 'clusters', with Y_1 denoting the cluster $(X_1, \dots, X_s)$, *etc.*. From this point of view, the appropriate Jeffreys measure would appear to be that associated with the model for a single cluster, *i.e.* μ_M^s , rather than μ_M . When we are comparing models of differing dimensions, the choice of s will affect the definition of our base measure; our previous construction now yields, rather than μ or anything proportional to it, a different meta Jeffreys measure $\mu^s := \sum_M \mu_M^s = \sum_M s^{d_M/2} \mu_M$. This change will in turn affect any proposed formal construction of a reference prior (such as the prescription: take $\gamma^{\text{ref}} \equiv 1$).

Note that the density γ of μ^s with respect to μ is of the form (30), which provides additional interpretation of, and justification for, that family of formal densities.

Now we have

$$(39) \quad \mu^s = \sum_M \left(\frac{s}{n}\right)^{d_M/2} \mu_M^*,$$

where $\mu_M^* = \mu_M^n$ denotes the Jeffreys measure based on the model M for *all* the data. Thus, to construct the meta Jeffreys measure relevant to taking observations in clusters of size s , we could first find the $\{\mu_M^*\}$, and then apply (39). This is a better way of thinking of things, since the definition of μ^* is clear-cut, independent of any arbitrary choice as to cluster size. We can also extend (39), without difficulty, to apply to models in which the variables need not be taken as independent and identically distributed.

There remains the choice of cluster size s . The ‘natural’ choice is $s = 1$, *viz.*

$$(40) \quad \mu^1 = \sum_M n^{-d_M/2} \mu_M^*,$$

which, in the independent and identically distributed case, recovers the original definition (14), and can also be applied more generally. We shall fix on this choice as our general reference prior distribution, and again refer to formal posterior model probabilities constructed from it as the *reference posterior model probabilities*, $p^{\text{ref}}(M | x)$. These are thus given by:

$$(41) \quad p^{\text{ref}}(M | x) \propto n^{-d_M/2} \int p(x | M, \theta) d\mu_M^*(\theta),$$

with μ_M^* the Jeffreys prior, for model M , based on all the data.

Even this definition is not unproblematic, however. In more complex problems, such as hierarchical models, it may be far from clear how the ‘sample size’ n should be defined. This problem is related to, but distinct from, that of defining the dimensionality d_M in such problems [Pauler, 1998]. Even in the independent and identically distributed case, the definition of n depends on the identification of “a cluster of size one”, which is itself not wholly unambiguous. More difficulties arise when the sample size, instead of being fixed in advance, is determined by the application of a stopping rule to sequentially generated data: this complicates both the interpretation of n and the appropriate definition of the Jeffreys measure. See §7.2 below.

6.2 Marginal likelihood

Consider again the normal model of (36), with now both θ and σ unknown. The Jeffreys prior for θ given σ has element $d\mu(\theta | \sigma) = \sigma^{-1} d\theta$, so that (33) gives the following reference marginal likelihood for σ :

$$(42) \quad \bar{L}(\sigma) \propto \sigma^{-n} e^{-R/2\sigma^2},$$

with R the usual residual sum-of-squares, $\sum_i (x_i - \bar{x})^2$. This is identical to the ‘profile likelihood’ [Kalbfleisch, 1986], and suffers from the same problems: for example, its maximum is at $\hat{\sigma}^2 = R_0/n$, and so it does not take account of the

reduction in the degrees of freedom from estimating θ . This equivalence to profile likelihood also holds for more general normal regression problems, where the ‘incorrect’ degrees of freedom can be more bothersome; and extends to problems such as that of Neyman and Scott (Example 3 of Berger and Bernardo [1996]), where the number of mean parameters increases with the sample size, leading to inconsistent estimation of σ^2 . Another difficulty with reference marginal likelihood is its susceptibility to marginalisation paradoxes [Dawid *et al.*, 1973]. (We recall that it is not in general appropriate to combine the reference marginal likelihood (42) with a prior *probability* distribution over ϕ , but rather with a *weight* distribution. However, this does not affect the logic of the marginalisation paradox.)

In particular problems we may be able to justify integrating out the unwanted components of the parameter θ with respect to some non-Jeffreys conditional distribution for θ given ϕ , the parameter of interest. This might, at its simplest, simply involve using a different choice of ‘normalising constant’. In the above example, we could consider such adjusted priors as $k(\sigma/\sigma_0)\sigma_0^{-1}d\theta$, where σ_0 is a ‘natural scale’ in the problem, having the same dimensions as X ; and $k(\cdot)$ is some dimensionless positive real function of a dimensionless real argument. For the special case that k has the form $k(x) \equiv x^\alpha$, for some real α , the choice of σ_0 will not affect the analysis. This would then allow other popular improper priors, for example right-Haar measure, or Berger-Bernardo reference priors, some of which might exhibit better behaviour. We remark, however, that in general it is not possible to eliminate all marginalisation paradoxes in this way.

6.3 Proper Jeffreys priors

In some problems, the Jeffreys measure turns out to be finite. However, that does not mean that it is a probability distribution, since its total mass need not be 1.

Consider, for example, the case of a Binomial random variable X , with index n and unknown probability parameter $\theta \in (0, 1)$. The Jeffreys measure (reduced to $n = 1$) has density element $\theta^{-\frac{1}{2}}(1 - \theta)^{-\frac{1}{2}}d\theta$, and thus total mass π ; it is proportional to, but *not* identical with, the Beta distribution $\beta(\frac{1}{2}, \frac{1}{2})$.

Should we use the actual Jeffreys measure in this case, or replace it by the associated probability distribution? I would argue that, for the purposes of ‘objective’ model comparison, there is nothing to be gained by rescaling to unit mass (a possibility that does not even exist for most models), and that the actual Jeffreys measure should be used in (24), (26), *etc.*

6.4 Dilution and vagueness

The problem of ‘dilution’ [George, 1999] arises when similar or identical models are present in our model collection $\mathcal{M}$. For example, model M_1 might express the observables (X_i) as jointly normal, with common mean, variance, and correlation (between any two distinct variables); M_2 might model them as $X_i = Z + Y_i$, with Z and the (Y_i) independent normal, the (Y_i) being identically distributed. These

two models induce exactly the same family of distributions for the observables. If they are both included in $\mathcal{M}$, it would, in effect, lead to ‘double counting’. Cases with similar rather than identical models arise, say, when we consider which independent variables to include in a regression problem with highly correlated regressors. The dilution problem is not confined to use of the Jeffreys measure, but can affect any ‘automatic’ method of assigning posterior probabilities over models. Suitable additional adjustment for this effect thus needs to be made.

A specific difficulty with the Jeffreys measure arises if we try to allow for some ‘vagueness’ in a model. Thus, instead of a model M under which the observables X_i are independent and identically normally distributed with unknown mean μ and known standard deviation $\sigma = 1$, we might consider M' , under which σ is allowed to range between 0.99 and 1.01. Under M the Jeffreys measure element for θ is just $d\theta$; but under M' , which could reasonably be expected to have higher weight, it is $0.028 d\theta$ (where $0.028 = \int_{0.99}^{1.01} \sqrt{2} d\sigma/\sigma^2$). Again, some further adjustment is required.

7 EXAMPLES

We now present some simple examples to help clarify some of the properties — both good and bad — of reference posterior probabilities.

7.1 Regression model

Consider a normal regression model M for observable Y :

$$(43) \quad Y \sim \mathcal{N}(X\theta, h^{-1}\Sigma),$$

where Y is $(n \times 1)$, θ is $(p \times 1)$, and the precision h is unknown, along with θ , whereas Σ , of full rank, is known (*e.g.* $\Sigma = I_n$). Thus the dimensionality of this model is $d = p + 1$.

It is readily verified that the Jeffreys measure μ^* , based on the model for all the data, has density element

$$(44) \quad d\mu^* = \left(\frac{n}{2}\right)^{\frac{1}{2}} |X^T \Sigma^{-1} X|^{\frac{1}{2}} d\theta h^{\frac{1}{2}p-1} dh.$$

We note that formula (44) is invariant under non-singular linear transformations both of response and of predictor variables, as well as under rescaling of Σ .

Adjusted to a sample size of 1, we thus get Jeffreys measure

$$(45) \quad \begin{aligned} d\mu^1 &= \left(\frac{1}{n}\right)^{\frac{1}{2}(p+1)} d\mu^* \\ &= 2^{-\frac{1}{2}} (\det A)^{\frac{1}{2}} d\theta h^{\frac{1}{2}p-1} dh, \end{aligned}$$

where $A := n^{-1} X^T \Sigma^{-1} X$ will typically be of order 1 as n increases. Using μ^1 in (24) yields

$$(46) \quad \bar{p}(y) = c_n \left(\frac{2\pi}{n} \right)^{\frac{1}{2}p} (\det \Sigma)^{-\frac{1}{2}} R^{-\frac{1}{2}n},$$

where $c_n = \Gamma(\frac{1}{2}n)/\sqrt{2\pi^n}$ is the same for all such regression models, and $R := y^T \{\Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\}y$ is the residual sum of squares. For this to be finite (although still not normalisable), we require $R > 0$, and thus $n > \text{rank } X$. From (46) (or equivalently directly from (41)), we then obtain the reference posterior model probabilities:

$$(47) \quad p(M | y) \propto \left(\frac{2\pi}{n} \right)^{\frac{1}{2}p_M} (\det \Sigma_M)^{-\frac{1}{2}} R_M^{-\frac{1}{2}n}.$$

For comparing two regression models with the same dispersion structure Σ but different mean-structure, the reference posterior odds is given by

$$(48) \quad \frac{p(M_1 | y)}{p(M_2 | y)} = \left(\frac{2\pi}{n} \right)^{\frac{1}{2}(p_1 - p_2)} \left(\frac{R_2}{R_1} \right)^{\frac{1}{2}n}.$$

If M_2 is a submodel of M_1 , we can rewrite this in terms of the F -statistic for testing M_1 within M_2 , using

$$\frac{R_2}{R_1} = 1 + \frac{(p_1 - p_2)}{(n - p_1)} F.$$

When, further, both models are true, F will be bounded in probability, having a limiting $\chi^2_{p_1 - p_2}/(p_1 - p_2)$ distribution. Then $(R_2/R_1)^{\frac{1}{2}n} \sim \exp\{\frac{1}{2}(p_1 - p_2)F\}$ as $n \rightarrow \infty$, and the agreement with the asymptotic form (25), with $\kappa = 0$, is readily confirmed.

Known precision

A similar analysis may be conducted in the simpler case that the dispersion matrix of Y is fully specified in each model, so that we can take $h = 1$. In this case, for $p > 0$ the appropriate Jeffreys measure is $(\det A)^{\frac{1}{2}} d\theta$, and the reference marginal density becomes

$$(49) \quad \bar{p}(y) = (\det 2\pi\Sigma)^{-\frac{1}{2}} \left(\frac{2\pi}{n} \right)^{\frac{1}{2}p} e^{-\frac{1}{2}R}.$$

No free parameters

An interesting feature now arises when we consider the case $p = 0$, so that the model specifies a completely fixed distribution, P_0 say, with no unknown parameters at all. The standard routine for determining the Jeffreys measure is not now applicable. Clearly, any measure contemplated must confine its mass to P_0 — but what mass should it put there? The answer “1” appears natural and desirable, but by no means obviously correct (particularly in the light of §6.3).

Now if we formally put $p = 0$ in (49), we obtain

$$(50) \quad \bar{p}(y) = (\det 2\pi\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}R}.$$

We see that this is, not merely proportional to, but *identical* with, the joint density $p_0(y)$ of Y under P_0 . That is, it corresponds to integrating over the ‘model’ $\{P_0\}$ with respect to the measure putting mass 1 on P_0 . This roundabout argument thus justifies taking the appropriate Jeffreys measure in this case to be the degenerate *probability* measure at P_0 . Arguing by analogy (perhaps a dangerous enterprise), we might propose more generally that, for any model M_0 which fully specifies a single distribution P_0 for the data, the associated ‘Jeffreys measure’ be taken as the unit mass on that distribution, thus yielding $\bar{p}_{M_0}(x) \equiv p_0(x)$ in (24), and, in (41), $p^{\text{ref}}(M_0 | x) \propto p_0(x)$.

7.2 ESP Experiment

In an experiment to investigate the existence of extra-sensory perception [Jahn *et al.*, 1987; Jefferys, 1990], a sequence of $n = 104,490,000$ trials, taken as Bernoulli with parameter θ , resulted in a data-sequence $\mathbf{x}$ containing $r = 52,263,471$ successes. The maximum likelihood estimate of θ is $\hat{\theta} = 0.500177$. The two models considered are $M_1 : 0 < \theta < 1$, with dimension $d_1 = 1$, and $M_2 : \theta = 0.5$, with dimension $d_2 = 0$ (the null hypothesis of no ESP). Although $\hat{\theta}$ appears very close to the null value, with this huge sample size a traditional hypothesis test leads to rejection of the null hypotheses at conventional levels of significance: the approximate standard normal deviate is 3.614, the corresponding two-tailed P -value being less than 0.0003.

Binomial sampling

We first take the ‘natural’ approach of assuming that the total number n of trials was fixed in advance, so that r is an observation from the binomial distribution $\text{Bin}(n; \theta)$. The corresponding unadjusted Jeffreys measure element under M_1 , based on the full data $\mathbf{x}$, is

$$(51) \quad d\mu^*(\theta) = n^{\frac{1}{2}} \theta^{-\frac{1}{2}} (1 - \theta)^{-\frac{1}{2}},$$

and when standardised to $n = 1$ becomes

$$(52) \quad d\mu^1(\theta) = \theta^{-\frac{1}{2}} (1 - \theta)^{-\frac{1}{2}}.$$

As suggested by §7.1, we take the Jeffreys measure over M_2 to be the unit mass. Then application of (41) (using Stirling’s approximation to the Gamma integral under M_1) leads to reference posterior odds

$$(53) \quad \frac{p^{\text{ref}}(M_2)}{p^{\text{ref}}(M_1)} = n^{\frac{1}{2}} \frac{p_2(\mathbf{x})}{\int p_1(\mathbf{x} | \theta) d\mu_1^*(\theta)}$$

$$(54) \quad \begin{aligned} &= \frac{p_2(\mathbf{x})}{\int p_1(\mathbf{x} | \theta) d\mu^1(\theta)} \\ &= 5.948, \end{aligned}$$

expressing a preference for the null hypothesis M_2 . This is in general accordance with other Bayesian analyses of this problem.

Note that the meta Jeffreys measure for this problem (for $n = 1$) has element $\{\delta(0) + \theta^{-\frac{1}{2}}(1 - \theta)^{-\frac{1}{2}}\}d\theta$, where δ denotes the Dirac delta-function. This has finite total mass $1 + \pi$, and is thus proportional to the *proper* prior probability structure with

$$\begin{aligned} p(M_1) &= \pi/(1 + \pi) \\ p(M_2) &= 1/(1 + \pi) \\ \theta | M_1 &\sim \beta(\tfrac{1}{2}, \tfrac{1}{2}). \end{aligned}$$

Thus in this particular case use of this proper ‘reference prior structure’ would give the same reference posterior odds (54).

Negative Binomial sampling

Now suppose that it was r , rather than n , that was fixed in advance, so that n is regarded as an observation from the negative binomial distribution $\text{NegBin}(r; \theta)$. This affects the Jeffreys measure element which, unadjusted, is now

$$(55) \quad d\mu_1^*(\theta) = r^{\frac{1}{2}}\theta^{-1}(1 - \theta)^{-\frac{1}{2}}.$$

Although (55) differs from (51), we see that they are essentially identical in the region of $\theta = \hat{\theta}$, *i.e.* where the likelihood is non-negligible. (This is a general feature, at least for exponential families: the Fisher information evaluated at the maximum likelihood estimate is always the same as the observed information, and this latter is not affected by the stopping rule.) Consequently the integral in (41) is effectively unchanged — in fact, the reference odds, calculated exactly as in (53), is now 5.951. (However, because (55) is not normalisable, there is now no corresponding proper ‘reference prior structure’.) We see then that the well-known dependence of the Jeffreys prior on the stopping rule, which might seem *prima facie* to be problematic, is of no essential practical importance in our approach.

There is, however, another, quite distinct problem that arises: how should we specify the appropriate scaling factor under M_1 ? After all, we now have observed, not n independent Bernoulli trials, but r independent geometric trials. Slavish application of our general recommendations would appear to imply that we should scale (55) by $r^{-\frac{1}{2}}$ rather than $n^{-\frac{1}{2}}$ as in §7.2, so renormalising it to a single observation which is, however, now geometric rather than Bernoulli. If we did so, the reference posterior odds would be multiplied by $\hat{\theta}^{\frac{1}{2}} = 0.707$, leading to a value of 4.21, corresponding to (slightly) weaker evidence against ESP. Common sense, however, suggests that the stopping rule should not have such an arbitrary influence on the answer, and that we should continue to scale (55) by $n^{-\frac{1}{2}}$ — even though the ‘sample size’ n is data-dependent.

8 CONCLUSION

We have indicated how some, at least, of the difficulties of conducting ‘reference’ analyses, when the model is uncertain, can be eliminated by concentrating attention directly on posterior model probabilities, rather than on marginal likelihood or Bayes factors. A tentative case has been made for the use of the ‘meta Jeffreys’ prior and the associated reference posterior model probabilities. However, we have also pointed out a number of difficulties associated with this proposal. Further work will be needed to try and produce a general method which shares the desirable properties (invariance, *etc.*) of the meta Jeffreys prior, while avoiding its undesirable features (in particular, its problematic dependence on sample size and stopping rule). It remains to be seen whether these desiderata can be made mutually compatible.

ACKNOWLEDGMENTS

I am grateful to Merlise Clyde, Susie Bayarri and Bertrand Clarke for valuable discussions on the material in this paper, to José Bernardo for suggesting that I analyse the example of §7.2, and to an anonymous referee for helpful suggestions.

BIBLIOGRAPHY

- [Aitkin, 1991] Murray Aitkin. Posterior Bayes factors (with Discussion). *Journal of the Royal Statistical Society, Series B*, 53:111–142, 1991.
- [Berger and Bernardo, 1996] James O. Berger and José-Miguel Bernardo. On the development of reference priors (with Discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 35–60, Oxford, 1996. Clarendon Press.
- [Berger and Pericchi, 1996a] James O. Berger and Luis Raul Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91:109–122, 1996.
- [Berger and Pericchi, 1996b] James O. Berger and Luis Raul Pericchi. The intrinsic Bayes factors for linear models (with Discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 5*, pages 25–44, Oxford, 1996. Clarendon Press.
- [Berger and Pericchi, 1997] James O. Berger and Luis Raul Pericchi. On criticisms and comparisons of default Bayes factors for model selection and hypothesis testing. Technical Report 97-43, Institute of Statistics and Decision Sciences, Duke University, 1997.
- [Bernardo, 1979] Jose-Miguel Bernardo. Reference posterior distributions for Bayesian inference (with Discussion). *Journal of the Royal Statistical Society, Series B*, 41:113–147, 1979.
- [Clarke and Barron, 1990] Bertrand S. Clarke and Andrew R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36:453–71, 1990.
- [Clyde, 1999] M. A. Clyde. Model averaging and model search (with Discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 6*, pp. 157–185. Clarendon Press, Oxford, 1999.
- [Dawid *et al.*, 1973] Alexander Philip Dawid, Mervyn Stone, and James Victor Zidek. Marginalization paradoxes in Bayesian and structural inference (with Discussion). *Journal of the Royal Statistical Society, Series B*, 35:189–233, 1973.
- [Dawid, 1992] Alexander Philip Dawid. Prequential analysis, stochastic complexity and Bayesian inference (with Discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 109–125, Oxford, 1992. Clarendon Press.

- [George, 1999] Edward I. George. Discussion of Clyde [1999]. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 175–177, Oxford, 1999. Clarendon Press.
- [Jahn *et al.*, 1987] R. G. Jahn, B. J. Dunne, and R. D. Nelson. Engineering anomalies research. *J. Scientific Exploration*, 1:21–50, 1987.
- [Jefferys, 1990] William H. Jefferys. Bayesian analysis of random event generator data. *J. Scientific Exploration*, 4:153–169, 1990.
- [Johnson, 1970] Richard A. Johnson. Asymptotic expansions associated with posterior distributions. *Annals of Mathematical Statistics*, 41:851–864, 1970.
- [Kalbfleisch, 1986] John D. Kalbfleisch. Pseudo-likelihood. In Samuel Kotz, Norman L. Johnson, and Campbell P. Read, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 324–327. Wiley-Interscience, New York, 1986.
- [O’Hagan, 1995] Anthony O’Hagan. Fractional Bayes factors for model comparison (with Discussion). *Journal of the Royal Statistical Society, Series B*, 57:99–138, 1995.
- [Pauler, 1998] Donna K. Pauler. The Schwarz criterion and related methods for normal linear models. *Biometrika*, 85:13–27, 1998.
- [Spiegelhalter and Smith, 1982] David J. Spiegelhalter and Adrian F. M. Smith. Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society, Series B*, 44:377–387, 1982.
- [Tierney and Kadane, 1986] Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86, 1986.
- [Wald, 1949] Abraham Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20:595–610, 1949.
- [Wasserman, 2000] Larry Alan Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44:92–107, 2000.

A APPENDIX: PROOF OF THEOREM 1

Proof. We can write

$$\begin{aligned}
 \log \frac{p^n(X^n)}{q^n(X^n)} &= \log \frac{p^n(X^n)}{p^n(X^n|\hat{\theta}_n)} + \log \frac{p^n(X^n|\hat{\theta}_n)}{p^n(X^n|\theta^*)} + \log \frac{p^n(X^n|\theta^*)}{q^n(X^n)} \\
 \text{(A.1)} \quad &= A + B + C \quad \text{say,}
 \end{aligned}$$

where $\hat{\theta}_n$ denotes the maximum likelihood estimator of θ in model M based on data X^n . Henceforth we drop n from the notation.

Note that the prior distribution only enters into the term A . The asymptotic form of the term A , based on Laplace expansion, is well-known [Tierney and Kadane, 1986]. We have:

$$\text{(A.2)} \quad A = \frac{1}{2} d \log 2\pi + \log \frac{p(\hat{\theta})}{\{\det -l''(\hat{\theta})\}^{\frac{1}{2}}} + O(n^{-1}).$$

This expansion is essentially algebraic, rather than probabilistic, and will typically hold for essentially all data sequences, whether or not they appear to be generated by some distribution in the model M ; in particular, we can regard the final term as $O_p(n^{-1})$ under Q .

Under Q , $\hat{\theta} \rightarrow \theta^*$ almost surely [Wald, 1949]. Taylor expansion in the neighbourhood of $\hat{\theta}$ yields:

$$\text{(A.3)} \quad l'(\theta) \approx l''(\hat{\theta})(\theta - \hat{\theta}),$$

so that

$$(A.4) \quad \hat{\theta} - \theta \approx -l''(\hat{\theta})^{-1}l'(\theta).$$

Under Q , each of $l'(\theta)$, $l''(\theta)$ is a sum of n independent and identically distributed components; moreover, it may be checked that $E_Q\{l'(\theta^*)\} = 0$. It follows that

$$(A.5) \quad n^{\frac{1}{2}}(\hat{\theta} - \theta^*) \xrightarrow{L} \mathcal{N}\{0, (J_2^*)^{-1} J_1^* J_2^* J_1^* J_2^* J_1^*\},$$

where $J_1^* = J_1(\theta^*)$, $J_2^* = J_2(\theta^*)$, with $J_1(\theta) = \text{var}_Q\{l'_1(\theta)\}$, $J_2(\theta) = E_Q\{-l''_1(\theta)\}$. In particular, under Q , $\hat{\theta} = \theta^* + O_p(n^{-\frac{1}{2}})$, $-l''(\hat{\theta}) = nJ_2^* + O_p(n^{\frac{1}{2}})$. It then follows from (??2) that

$$(A.6) \quad A = -\frac{1}{2} d \log \frac{n}{2\pi} + \log \frac{p(\theta^*)}{\{\det J_2^*\}^{\frac{1}{2}}} + O_p(n^{-\frac{1}{2}}).$$

Note in particular that the asymptotic effect of the within-model prior density $p(\theta)$ on $\log \frac{p^n(X^n)}{q^n(X^n)}$ is entirely captured in its order 1 contribution to (??6).

From this point on we consider the two cases (i) and (ii) individually.

- (i) If $Q \notin M$, then $K(Q, M) > 0$. In this case, under Q , $\log p(X | \theta^*)/q(X)$ is a sum of n independent and identically distributed components with mean $-K(Q, M)$, and thus term C in (??1) has the form

$$(A.7) \quad C = -nK(Q, M) + \left[n \text{var}_Q \left\{ \frac{\log p^1(X | \theta^*)}{q^1(X)} \right\} \right]^{\frac{1}{2}} Z,$$

where $Z \xrightarrow{L} \mathcal{N}(0, 1)$. Equation (??6) shows that $A = O_p(\log n)$. Finally, consider $B = l(\hat{\theta}) - l(\theta^*)$. By Taylor expansion,

$$(A.8) \quad \begin{aligned} B &\approx \frac{1}{2} \{n^{\frac{1}{2}}(\hat{\theta} - \theta^*)\}^T \{-n^{-1}l''(\hat{\theta})\} \{n^{\frac{1}{2}}(\hat{\theta} - \theta^*)\} \\ &\approx \frac{1}{2} \{n^{\frac{1}{2}}(\hat{\theta} - \theta^*)\}^T \{J_2^*\} \{n^{\frac{1}{2}}(\hat{\theta} - \theta^*)\}, \end{aligned}$$

which, by (??5), is $O_p(1)$. Thus both terms A and B are of smaller order than C , and (4) follows.

- (ii) Now suppose $Q \in M$, so that $K(Q, M) = 0$ and $Q = P_{\theta^*}$. In this case, term C in (??1) vanishes. We also now have $J_1^* = J_2^* = I(\theta^*)$. Thus

$$(A.9) \quad A = -\frac{1}{2} d \log \frac{n}{2\pi} + \log \rho(\theta^*) + O_p(n^{-\frac{1}{2}}).$$

Also by (??8) and (??5), or directly from Wilks' Theorem,

$$(A.10) \quad B \xrightarrow{L} \frac{1}{2} \chi_d^2,$$

and in fact the distribution of B will typically differ from that of $\frac{1}{2}\chi_d^2$ by $O(n^{-\frac{1}{2}})$. Then (5) follows. (For a more rigorous treatment, see Clarke and Barron [1990].)

■

Part VI

Attempts to Understand
Different Aspects of
“Randomness”

This page intentionally left blank

DEFINING RANDOMNESS

Deborah Bennett

We all have a commonsense understanding of randomness. Ordinary, everyday definitions of randomness encompass notions of lack of order, periodicity, pattern, aim, or purpose. When we reach for simple examples of a random occurrence, we usually think of a coin toss, a dice throw, or a lottery. If your chance of tossing a head on a coin toss is the same as mine in order to decide who gets to go first or who gets first choice, no one has an unfair advantage. If you have a fair chance of winning a lottery, then you are satisfied that the choice of the winner is random. There is something democratic about randomness. The inability to know in advance what is coming prevents us and, more importantly, prevents others from manipulating or controlling events.

The inability of mere mortals to manipulate the outcome is undoubtedly what led early societies to employ the use of random devices to reveal the gods' decisions in divination. From ancient Babylonian and Hebrew societies, through ancient Greek and Roman uses, the lot revealed divine intentions. The lot may or may not have been controlled by a deity, but the outcome of the lottery was beyond the control of other humans. Its use was somehow *fair*.

Not only were random devices used in divination, but they have also been used in games since ancient times. Practically every society, modern as well as ancient, uses some type of randomizers in games. Today's six-sided cubical dice have been around for over three thousand years, but much earlier dice were two-sided and four-sided. Again, the notion of *fairness* arises, as the introduction of chance into a game seems to level the playing field between unmatched brains or brawn.

While we all have an understanding of the concept of randomness, pinning down a satisfactory definition can prove to be troublesome. What we may agree on is that randomness encompasses the concepts of fairness and unpredictability.

In science, an event is described as *random* if it may or may not occur. It may be more likely to occur, less likely to occur, or equally likely to occur as some other random event. Of course, if we know that an event will occur, it is *certain* and if we know that the event will not occur, it is *impossible*. But a random or chance event defies precise predictability.

Some would prefer randomness to remain undefined—just as the concept of a point or a line remains undefined in geometry. But for others, the definition of randomness is essential since it affects our concept of probability. Randomness is the foundation on which the calculus of probability is based. It may seem odd, but the unpredictability of the random event allows prediction at the macro-level by way of the *laws of chance*. The *unruliness* of chance allows *laws* of probability.

Although there are similarities, randomness is not the same as chaos and it is not anarchy. Nicholas Rescher describes their shared characteristics, saying that “rational forecasting of outcomes is simply infeasible.” In the case of chaos, processes are so sensitive to minute changes in conditions that determination of their instability is beyond our power. In the case of anarchy, there are no laws, only mayhem.

RANDOMNESS FINALLY DEFINED

I suppose no one felt a need to *define* randomness until probability became a *science*. Games of chance that utilized random devices or randomizers had existed since ancient times, but it took many centuries before chances were quantified and probabilities were studied. The calculus of probability arose from the analysis of games of chance, appearing first in 1564 when Girolamo Cardano wrote a book analyzing various games of chance. Cardano’s book was not published until almost a hundred years later, but between 1613 and 1623 Galileo Galilei wrote a small tract, entitled *Thoughts about Dice Games*. In the mid-17th century, serious discussion of probability began with the correspondence between Blaise Pascal and Pierre de Fermat over a gambling question from the Chevalier de Méré. Despite the fact that the Pascal/Fermat correspondence originated from a seemingly trivial gaming problem, the consideration they gave the subject was noticed by other mathematicians and philosophers of the time. The discussions of these distinguished mathematicians caught the attention of Christian Huygens who wrote the first printed text on the calculus of probability in 1692. Even then, no one thought it necessary to *define* probability, much less randomness.

These probability pioneers were followed by eighteenth-century greats, such as Abraham de Moivre whose *The Doctrine of Chance* saw multiple editions from 1711 through 1756. The almost three-hundred year development of the theory of probability culminated in the 19th century with Pierre-Simon Laplace who was first to recognize the need for a precise definition of probability. Laplace’s work represented what came to be known as the classical interpretation of probability.

The nineteenth century enjoyed an explosion of the applications of probability and the initial stages of its use in statistical inference. Scientists became increasingly aware that analogies existed between games of chance and random phenomena in biology and the physical and social sciences. The twentieth century produced even more applications of probability—in physics, economics, insurance, and telecommunications. Classical interpretations of probability gave way to the frequency interpretation of probability: The probability of an event is its relative frequency over a long, long time.

Frequentists defined probability in terms of the relative frequency of objective properties that they could measure and sample. Leading the way for this point of view were Jerzy Neyman and Egon Pearson, who were influenced by John Venn, R. A. Fisher, and Richard von Mises. As they drew random samples for experiments they were performing in order to attain statistical estimates and perform statistical

tests, the frequentists perceived a need to define *random*.

Definitions of randomness began to be established in the 20th century when Richard von Mises attempted to provide a definition of a random sequence that was both mathematically and intuitively satisfying to the frequency view of probability. According to von Mises, a sequence was *random* if it is impossible to predict where in the sequence a particular observation will occur. Von Mises's definition is often referred to as the *impossibility of a gambling system*. If it is impossible for a gambler to invent a system of prediction that will change his long-run odds of winning, then the gambler is playing a game of pure chance. No system and no formula can specify the gambler's outcomes ahead of time thus enabling him to adjust his bets in anticipation of a particular outcome. If we are unable to devise a system for predicting a particular observation without prior knowledge of the sequence, then we are dealing with a random sequence. The intuitively appealing part of von Mises's definition is that it eliminates systematic, patterned sequences since they would be predictable.

The most popular opposing theory to that of the frequentists is that of the subjectivists (or Bayesians) who espouse epistemic probabilities. Subjective probabilities are judgments, perhaps informed judgments. Subjective probabilists express degrees of belief in uncertain propositions. Personal or subjective evaluations of probability are used particularly when events have not occurred often enough to provide long term evidence. Before you object that odds can't be subjective, think about how the odds are set in a horse race. The degree of belief by the bettors influences their bets and the odds for a horse are based on the distribution of bets by those subjective believers. For the subjectivists, their definition of randomness is founded in their *principle of indifference* or *principle of insufficient reason*.

Von Mises's view defines randomness in terms of the inability to reliably predict over the long run. In the toss of a coin, the outcome is random and the sides are equally-likely because the relative frequencies over the long, long run have demonstrated that this is so. For the subjectivist, the toss of a coin is random because we have no reason to believe that one side is more likely to appear than the other, thus forcing us to be indifferent to the choice.

Even experts disagree about the many meanings of randomness. Do you think that the throw of a die is random with the sides equally likely to come up because centuries of dice-throwing has shown us that each side comes up one-sixth of the time? Or, do you base your thinking on the physical uncertainty of the situation? The die appears to be constructed in such a way that any side can come up equally often and, in the absence of further information, you have no sufficient reason to judge otherwise.

As the science of probability and statistics gained widespread application in the twentieth century, inferences needed to be based on random samples, and sources for random numbers began to be explored. A new definition of a random sequence evolved—an algorithmic definition. Definitions of *complexity* in information theory led Andrei Kolmogorov and Gregory Chaitin to define a sequence as random by virtue of the length of the program, algorithm, or rule that is necessary to describe

it. The more compact the rule, the less random the sequence. The longer the rule, the more random the sequence. The ideal random sequence is one that cannot be described by a rule shorter than the length of the sequence itself (any sequence can be described by naming the digits of the sequence one by one).

INADEQUACIES AND DEFICIENCIES

One argument against the long run frequency view of probability has to do with its definition of randomness. The impossibility of a prediction system implies that no rule or formula exists to predict a random sequence. The insistence that “no rule exists” seems a rather stringent requirement. Perhaps a rule exists but if it is unknown to everyone, then a prediction system may prove impossible nevertheless.

Statistical experimentation most often encountered randomness in connection with random samples. Random samples were selected according to some probabilistic criteria and often, in order to select members of a sample randomly, random number generators were needed. A random number generator might involve a simple procedure such as drawing numbered cards (after mixing them) or rolling a many-sided die. On the more sophisticated side, a random number generator might be a fancy computer algorithm. Wait a minute . . . , you say. A computer program written by human beings must produce deterministic output. If this is so, how can a mathematical formula create random digits?

Non-manmade sources for random digits have also been considered. The digits of irrational numbers like π have been examined for randomness. These non-repeating, never-ending decimal numbers have the sense of the unexpected, but they are still determined constants. I may not know the 342nd digit of the decimal expansion of π , but I can assure you that someone does.

Is randomness caused by the process or the outcome? A computer random number generator is not a random process but it may output a sequence that appears random. Is something random by virtue of its disorderly appearance or by virtue of the mechanism or device that created the disorderly appearance? Indeed, a random process does not even guarantee a disorderly appearance.

Suppose we wish to generate a random sequence of 0s and 1s by using some lottery or coin-tossing method. Every possible sequence has an equal chance of being *the* sequence that we create. Therefore, the sequences of all 0s or all 1s or alternating 0s and 1s have an equal chance of being created. Should one of these occur as our random sequence, we might want to throw it out as not *looking* random enough. The *process* was random enough, but the *product*, the sequence we produced, was quite orderly and patterned. If a sequence is random by virtue of the device that created it, then do we toss it out if it doesn’t appear “random enough”?

When we are faced with a supposed random sequence, must we question how it was generated or should we judge its randomness based on how it looks? Some experts have said that it matters not how the sequence was generated, what is important is its disorderliness. If a sequence is random by virtue of its disorderliness

then random sequences must be able to pass objective tests of disorderliness. Numerous measures of orderliness (or disorderliness) have been developed and have become *tests* for randomness. But finding a sequence that passes *all* tests faces the same difficulties as finding a sequence that passes all systems of prediction (the impossibility of a gambling system). We must then decide: If we can't require our sequence to pass all tests of randomness, then what percent of the tests should our sequence pass? Which tests of randomness should our sequence pass? And, what level of significance shall we set for the tests?

The definition of a random sequence emanating from the realm of informational complexity is based on both the process and the output. If a sequence output is complex enough to require a lengthy description, it is a random sequence. For example, a 100-digit sequence of 0s and 1s that requires a simple (or short) description is not very random. Take the examples mentioned earlier. The descriptions, "all 0s" or "all 1s" or "alternating 0s and 1s," are concise descriptions and therefore do not describe sequences complex enough to be defined as being random. This is intuitively pleasing since they do not *feel* very random; they hold no sense of uncertainty for us. The longest possible description of a sequence of 0s and 1s is as long as the sequence itself. In order for the maximally complex sequence to be described, each digit must be named one by one.

For a computer-generated sequence, this "description of a sequence" equates to the computer program or algorithm used to produce the sequence, and program lengths can be measured. However, to create random sequences we need long programs and long computer programs are neither efficient nor practical. On the other hand, measuring the complexity of sequences does provide us with a scale against which the randomness of a sequence can be measured along a continuum. In other words, we can measure *degrees* of randomness.

COMMONALITIES

None of the definitions of randomness are free from contradiction and none of them seem completely satisfactory. They do, however, share some commonalities. They all encompass our uncertainty, and they all hold perfect randomness to be an absolute. The frequency view of randomness is based on our inability to know what is coming in the long run; perfect randomness is found when no system exists that would allow us to predict what is coming. The subjective view of randomness is founded on our ignorance of conditions or the unknowability of outcomes, and this uncertainty guarantees insufficient understanding to predict accurately. Randomness based on the length of the description necessary to identify a given sequence provides us with a way to measure the degree of disorder and defines the sequence that is maximally complex.

If perfect randomness is an absolute, we may be forced to accept the reality that an absolute is often unattainable. Pure chance, the ideal, may be impossible to achieve. The chance outcomes we are so familiar with may not be situations involving pure chance; they may, in fact, just be chancy. If perfect randomness is

beyond our reach, we may be forced to accept relative randomness.

Although the ideal random sequence is maximally complex, a relatively random sequence may be “complex enough.” Although a sequence may not pass all possible statistical tests for randomness, we may set a threshold level of how many and which tests must be passed in order to declare a sequence “random enough.” We may even accept different threshold levels for different purposes. The thresholds we set for randomly choosing components of a nuclear reactor to test might differ from those we set to randomly choose a door prize at a charity event. “Relative” randomness would permit degrees of randomness depending on the situation.

We began by suggesting that randomness may be difficult to define. Even relative randomness may defy definition. However, we can pinpoint qualities that we require. We require that relative randomness fulfill our sense of fairness and equity over the long run. Like absolute randomness, relative randomness may seem fickle and unfair in the short run, but it must appear predictable in the long run. We will accept local patterns but not global ones. We require that relative randomness satisfy our notion of uncertainty and unpredictability. It should be not only impractical for us to predict the relatively random outcome, it should be impossible for us to make rational predictions. We require that relative randomness measure up to ideal randomness as much as is humanly possible. As we reach for perfection, we will accept only the near perfect.

ACKNOWLEDGMENTS

I would like to acknowledge the anonymous referee who offered suggestions for this paper.

BIBLIOGRAPHY

- [Bennett, 1998] D. J. Bennett. *Randomness*. Cambridge, Mass: Harvard University Press, 1998.
- [Chaitin, 2001] G. Chaitin. *Exploring Randomness*. London: Springer-Verlag, 2001.
- [Chaitin, 1975] G. Chaitin. Randomness and mathematical proof. *Scientific American* 232 (5): 47-52, 1975.
- [Church, 1940] A. Church. On the concept of a random sequence. *American Mathematical Society Bulletin* 46: 130-135, 1940.
- [David, 1988] F. N. David. *Games, Gods & Gambling: A History of Probability and Statistical Ideas*. Dover Publications, 1988.
- [Eagle, 2005] A. Eagle. Randomness is unpredictability. *British Society for the Philosophy of Science* 56: 749-790, 2005.
- [Fisher, 1926] R. A. Fisher. On the random sequence. *Quarterly Journal of the Royal Meteorological Society* 52: 250, 1926.
- [Gardner, 1977] M. Gardner. Random numbers. In *Mathematical Carnival*. New York: Vintage Books, 1977.
- [Hacking, 1965] I. Hacking. *Logic of Scientific Inference*. Cambridge University Press, 1965.
- [Humphreys, 1976] P. W. Humphreys. Inquiries in the philosophy of probability: Randomness and independence. Ph.D. diss., Stanford University, 1976.
- [Knuth, 1997] D. E. Knuth. *The Art of Computer Programming. Vol. 3: Seminumerical Algorithms, 3rd ed.* Reading, MA: Addison-Wesley, 1997.

- [Kolmogorov, 1963] A. Kolmogorov. On tables of random numbers. *Sankhya*, ser. A 25: 369-376. 1963.
- [Kolmogorov, 1968] A. Kolmogorov. Three approaches for defining the concept of information quantity. *Problems of Information Transmission* 1: 4-7, 1968.
- [Martin-Löf, 1969] P. Martin-Löf. The literature on von Mises Kollektivs revisited. *Theoria* 35: 12-37, 1969.
- [Mises, 1939] R. von Mises. *Probability, Statistics and Truth*. 2nd ed. Trans. J. Neyman, D. Sholl, and E. Rabinowitsch. New York: Macmillan, 1939.
- [Ore, 1953] O. Ore. *Cardano, the Gambling Scholar*. Princeton University Press, 1953.
- [Porter, 1988] T. M. Porter. *The Rise of Statistical Thinking 1820-1900*. Princeton University Press, 1988.
- [Rescher, 1995] N. Rescher. *Luck: The Brilliant Randomness of Everyday Life*. New York: Farrar Straus Giroux, 1995.
- [Stigler, 1990] S. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press, 1990.

This page intentionally left blank

MATHEMATICAL FOUNDATIONS OF RANDOMNESS

Abhijit Dasgupta

This article is dedicated to the centenary of the Borel Strong Law.

1 INTRODUCTION

1.1 A random blackbox?

Imagine a “blackbox” which supposedly produces its outcomes “randomly” according to some fixed finite probability distribution.

$$\boxed{\text{BLACKBOX}} \rightarrow \text{Outcome}$$

Thus there are a finite number of possible outcomes, say $\omega_1, \omega_2, \dots, \omega_n$, and each outcome ω_i has a fixed non-trivial probability p_i , so that $P(\omega_i) = p_i$ with $0 < p_i < 1$ for $i = 1, 2, \dots, n$, and $\sum_{i=1}^n p_i = 1$. The outcomes may be generated either automatically from a continuously running process, or on demand, say by pressing a button on the blackbox.

We think of this as an abstract model representing a process, a machine, or an experiment. Familiar examples are the flip of a coin, the turn of a casino gambling wheel, the roll of an “electronic die” on a handheld video game device, the snapshot of weather data, the time between two successive clicks of a Geiger counter detecting radioactive decay, the stock market index value, etc.

Here, we approach the problem in a *purely extensional way*. This means that except for the given knowledge of the probability values of for each possible outcome, we are only able to *observe* the outcomes, and do not have any access to, or information about, the internal workings of the machine. Hence the term *blackbox*.

Gambling houses and forecasters of weather and stock market as well as philosophers of probability and statistics have found the following question to be of considerable interest.

Question A. Is the blackbox a *random* device? Does it produce its outcomes *randomly* (while obeying a fixed probability distribution)?

1.2 Sequences

We cannot hope to answer the last question by observing a single outcome of the blackbox. In fact, no finite amount of observation of outcomes can fully confirm that a process is random (or not). On the other hand, by repeating the process a sufficiently large number of times and observing the resulting *sequence of outcomes*, we may hope to gain enough information for answering the question with a desired level of confidence. (Such a sequence is denoted by $\langle x_1, x_2, \dots, x_k, \dots \rangle$, where each x_k , the k -th term of the sequence, equals one of the outcome values $\omega_1, \dots, \omega_n$.)

Given a sequence of outcomes of the blackbox, we want to determine whether or not it was produced randomly. Our extensional approach means that other than a knowledge of the underlying probability distribution for the possible outcome values, we only have the completed sequence of outcomes available, with all information about its origination permanently removed, so the answer must depend solely on the sequence itself and not on how it was produced. This necessitates the consideration of a different but more precise question:

Question B. When is a given sequence of outcomes random (relative to a fixed probability distribution for the outcome values)?

This article deals entirely with this last question (Question B, randomness of sequences), and not with the earlier question (Question A, randomness of processes). It is one of the fundamental questions in the philosophy of probability and statistics. The longer the sequence, the more information we have for determining if it is random or not. And in the ideal case, the sequences will be infinite sequences. We will consider both cases — finite and infinite sequences — of the question in detail.

It was von Mises who first treated this question rigorously (for infinite sequences), and considered its answer to be the very foundation of probability, known as the *frequentist* interpretation of probability.

1.3 Pseudo-randomness: A Galilean Dialogue

In his celebrated work *Gödel, Escher, Bach*, Douglas Hofstadter quotes the following “beautiful and memorable passage” from *Are Quanta Real? — a Galilean Dialogue* by J. M. Jauch [Jauch, 1990]:

SALVIATI: Suppose I give you two sequences of numbers, such as

78539816339744830961566084...

and

1, $-1/3$, $+1/5$, $-1/7$, $+1/9$, $-1/11$, $+1/13$, $-1/15$, ...

If I asked you, Simplicio, what the next number of the first sequence is, what would you say?

SIMPLICIO: I could not tell you. I think it is a random sequence and that there is no law in it.

SALVIATI: And for the second sequence?

SIMPLICIO: That would be easy. It must be $+1/17$.

SALVIATI: Right. But what would you say if I told you that the first sequence is also constructed by a law and this law is in fact identical with the one you have just discovered for the second sequence?

SIMPLICIO: This does not seem probable to me.

SALVIATI: But it is indeed so, since the first sequence is simply the beginning of the decimal fraction [expansion] of the sum of the second. Its value is $\pi/4$.

SIMPLICIO: You are full of such mathematical tricks . . .

The dialogue illustrates an aspect (among many) of the problem of defining randomness for sequences. An apparently random sequence of digits may really be *pseudo-random*: While it may appear to be “statistically random” and unpredictable, there may be a (hidden) rule or arithmetical method for generating the entire infinite sequence purely deterministically.

Our intuition tells us that if there is a deterministic and effective rule for computing every term of an infinite sequence, then, despite appearance, the sequence cannot be genuinely random, as von Neumann famously cautioned. But how do we precisely define the class of infinite sequences that are “genuinely random”?¹

1.4 A Laplacian problem

Imagine our blackbox to be a computer program simulating the flip of a fair coin, with 1 denoting heads, and 0 denoting tails. Suppose that we run it to generate fifty flips and observe the resulting outcome sequence. Such an outcome sequence is a binary string of length 50, and there are 2^{50} possible outcome sequences. If we observe the outcome sequence to be

10010001001100111011001010100000110100001100011100,

we are not surprised. We regard it as a “random binary string”, and consider the program to be running normally.

But if the outcome sequence is

010101010101010101010101010101010101010101010101,

we consider it to be an extraordinary event, and justifiably suspect malfunction, perhaps a bug in the program. But according to simple probability, this binary

¹Note that it would be an error to define random infinite sequences as those which cannot be generated deterministically by a specified rule or arithmetical method. This is because there are only countably many such methods, while there are uncountably many sequences which are not random in any sense of the word. For example, there are uncountably many infinite binary sequences for which the bits at even positions are all set to 0 but the bits at odd positions are allowed to be arbitrary, and none of these sequences is random.

string of perfectly alternating 0s and 1s is as likely to be produced as the first supposedly “random-looking” string, and so should not be regarded any more special than the first one. Yet, in a clear intuitive sense, the “regularity” of the second string makes it much less random compared to the first string.

P. S. Laplace (1749–1827) was aware of this problem and pointed out the following reason why, intuitively, a regular outcome of a random event is unlikely:

“We arrange in our thought, all possible events in various classes; and we regard as *extraordinary* those classes which include a very small number. In the game of heads and tails, if heads comes up a hundred times in a row, then this appears to us extraordinary, because the almost infinite number of combinations that can arise in a hundred throws are divided in regular sequences, or those in which we observe a rule that is easy to grasp, and in irregular sequences, that are incomparably more numerous.” [de Laplace, 1819/1952]

In addition to the rarity of regular patterns, Laplace points out that certain strings may have a “cause”, making them unlikely to be random:

“The regular combinations occur more rarely only because they are less numerous. If we seek a cause whenever we perceive symmetry, it is not that we regard the symmetrical event as less possible than the others, but, since this event ought to be the effect of a regular cause or that of chance, the first of these suppositions, is more probable than the second. On a table, we see letters arranged in this order: **C o n s t a n t i n o p l e**, and we judge that this arrangement is not the result of chance, not because it is less possible than others, for if this word were not employed in any language we would not suspect it came from any particular cause, but this word being in use among us, it is incomparably more probable that some person has thus arranged the aforesaid letters than this arrangement is due to chance.” [de Laplace, 1819/1952]

We therefore observe (still using our vague and imprecise language) that among all finite strings of a fixed large size (say length 100), there are only a relatively small number of “non-random” strings — strings which possess “regularity” or more generally have some “cause” behind them. We may call the other strings to be random.

Thus, it appears that for finite strings of a fixed large size (say length 100), there is some attribute that corresponds to the intuitive notion of randomness. But how do we precisely define it?

This intuitive attribute of randomness can be *partially approximated* by formulating certain events described in the ordinary language of classical probability. E.g., by requiring that the standard deviation of the run lengths of digits be within certain limits one can avoid such regular sequences as the one above with alternating 0 and 1. In fact, such events are designed and formulated as statistical tests

for estimating the “randomness confidence” of a finite sequence of digits. Some examples are restrictions on the distribution of run-lengths, autocorrelation, serial correlation, comparison with standard test distributions such as χ^2 -tests, etc.²

However, all such tests appear to be only partial approximations, and no event formulated in the usual language of classical probability seem to precisely capture this attribute of randomness in an intuitively satisfactory way.

1.5 The main problems. What this article is about

The considerations above lead us to two classic problems of mathematical and statistical philosophy:

PROBLEM 1 Randomness for Infinite Sequences. When is a given infinite sequence of digits random?

PROBLEM 2 Randomness for Finite Strings. When is a given finite string of digits random?

As it turned out, most of the early mathematical results concerning randomness were about infinite sequences.³ Work on defining randomness for finite sequences started later.⁴ As Ulam had noted [MacHale, 1993]: “The infinite we shall do right away. The finite may take a little longer.”

It also turned out that the two concepts are closely connected in a remarkable way with “algorithm” or “effective computability” playing a central role in their definitions. Quite satisfactory answers to both problems emerged almost simultaneously in mid 1960s, together with the birth of the new fields of *Algorithmic Randomness* and *Algorithmic Complexity*.

The primary goal of this article is to present these celebrated solutions of the above two classic philosophical problems as a brief introduction to algorithmic randomness and complexity. It is, however, important to note that a vigorous effort to further sharpen, refine, and calibrate the definition of randomness is currently being pursued through a large and growing body of lively research activity in algorithmic randomness [Nies, 2009; Downey and Hirschfeldt, 2010; Li and Vitanyi, 2008; Shen *et al.*, 20??]. Therefore the topic is best viewed as a part of continuing research, and there is still ample scope for debate if the answers obtained in the 1960s are final.

²See Knuth [1998] for such statistical tests of randomness. Chris Wetzel of Rhodes College has interactive web pages illustrating such tests of randomness (<http://faculty.rhodes.edu/wetzel/random/intro.html>).

³Borel’s 1909 work on normal numbers and the Weyl equidistribution theorem of 1916 are preliminary forms of randomness for infinite sequences, but it was von Mises who in 1919 directly focused on the concept and gave a fundamental definition. Church’s introduction of “algorithm” in 1940 made von Mises’ definition mathematically precise and also made the subject permanently “algorithmic”. For a truly satisfactory answer, one had to wait another quarter century until Martin-Löf found it in 1965.

⁴Early work was done in the 1960s by Solomonoff [1960; 1964], Kolmogorov [1963; 1965], and then Chaitin (see [Chaitin, 1992] for references).

Throughout this article, we will consider only strings and sequences, and make the further simplifying assumption that the only digits (possible outcomes) are 0 and 1 with equal probability of $\frac{1}{2}$ each — the “fair coin model”. In other words, the blackbox always represents the flip a perfectly fair coin. In particular, all strings and sequences will be *binary* strings and sequences. This restriction is not really as stringent as it may appear to be (see subsection 2.5).

1.6 The solutions. Algorithmic randomness

There are three key approaches for defining randomness in sequences: *unpredictability*, *typicality*, and *incompressibility* [Downey and Hirschfeldt, 2010]. We now briefly and informally introduce these approaches. They will be described in detail using mathematically precise language later in this article.

- *Unpredictability*. This can be understood in terms of impossibility of successful gambling strategies. According to *randomness as unpredictability*, an infinite binary sequence is random if, roughly speaking, it is impossible to effectively specify a gambling strategy which can make long run gains for the gambler when played against this sequence as the outcomes. A *frequentist* version of randomness as unpredictability was used by von Mises; see the comments of Feller quoted in subsection 5.1.
- *Typicality*. A property or attribute of infinite binary sequences is called *special* if the probability that the property holds is zero, and is called *typical* if the probability that the property holds is one. An attribute is special if its complement (negation) is typical, and vice versa. For example, consider the property of having no run of zeros of length seven. It can be shown that the probability is zero that an infinite binary sequence has no run of zeros of length seven, thus this property is special. The intuition here is that if a sequence has a special property or attribute, then it cannot be random, and randomness is equivalent to the complete lack of special attributes. According to *randomness as typicality*, an infinite binary sequence is random if, roughly speaking, it is impossible to effectively specify a special property that the sequence possesses. This is essentially the definition of Martin-Löf-randomness. What Martin-Löf did was to provide a mathematically precise notion of effective specifiability in this context.
- *Incompressibility*. Some finite strings can be effectively specified by descriptions much shorter than the string itself. For example, the string

0101010101010101010101010101010101010101010101010101010101010101,

can be specified as “01 repeated fifty times”. The short description here exploits the regularity of pattern. The idea is also utilized by data compression programs: Strings which compress well are the ones having regularity of pattern or redundancy of information, and highly irregular or “random” strings

do not compress well. This approach, known as *Kolmogorov Complexity*, was introduced by Solomonoff, Kolmogorov, and Chaitin, and yields a definition of randomness for finite strings: A string is random if its shortest description has length equal to the length of the string itself. This can also be carried over to infinite strings: According to *randomness as incompressibility*, an infinite binary sequence is random if none of its initial segments compress “very much”.

These three approaches to defining randomness may appear to be independent of each other, and we might expect them to lead to different definitions of randomness. It is therefore a remarkable fact that (with appropriate choices for the notion of effective specifiability) all three definitions turn out to be equivalent! Schnorr’s Theorem, a celebrated result, establishes the equivalence of randomness as typicality (i.e. Martin-Löf randomness) with randomness as incompressibility.

This surprising equivalence between randomness as unpredictability, randomness as typicality, and randomness as incompressibility, is strong evidence that this common equivalent definition is satisfactory, and that algorithmic randomness provides an adequate mathematical foundation for randomness.

1.7 What this article is not about

Let us now clearly state a disclaimer. This article’s scope is restricted to discussions of only Problems 1 and 2 above, dealing with randomness of sequences and strings from a purely extensional view. Our coverage of randomness is thus limited to what may perhaps be called “the randomness of bit patterns”, while the word randomness as used in ordinary language is overloaded with meanings many of which do not necessarily involve bit patterns (binary strings and sequences).

Randomness may be discussed in a more general setting than just bit patterns. This has been done, e.g., by Eagle, who introduces the idea of randomness as *maximal unpredictability*, and has written a critique of algorithmic randomness. One may also regard *random and probabilistic processes* as the main framework for investigating randomness. See Eagle [2005], where further references can be found.

Also, there are popular books on randomness (e.g. Aczel [2004], Beltrami [1999], and Bennett [1998]) with accessible approaches to various aspects of randomness. The books by Taleb [2005] and by Mlodinow [2008] include various financial and social implications of randomness as well.

Even within the limited case of “bit-pattern randomness”, there are other important issues, such as the following broad question of considerable philosophical and practical interest: *How can random sequences be physically generated?* Another example is the relation between quantum mechanics and (algorithmic) randomness. Several interesting discussions and further references are in [Calude, 2005; Yurtsever, 2000; Svozil, 1993; Longo, 2009; Bailly and Longo, 2007; Penrose, 1989; Stewart, 2002].

Because of our extensional approach, we have omitted all such issues and limited ourselves only to mathematical definitions of randomness for sequences, i.e., *precise criteria to distinguish random sequences from non-random ones*.

But even for mathematical definitions based on the extensional view, algorithmic randomness is not the only approach possible. For example, in [van Lambalgen, 1990], Van Lambalgen introduces an axiomatic approach to randomness. We will not discuss such approaches either.

Other than reasons of space limitation and cohesion, there is another point which has compelled us to select only the topic of algorithmic randomness for discussion,

This year marks the 100th anniversary of Borel's strong law (1909), which first showed the importance of the concept of limiting frequency, and is the precursor of many ideas based on it, including von Mises' idea of randomness. We believe that in the one hundred years since Borel's strong law, algorithmic randomness stands out as the crowning achievement in the study of random sequences. In addition to providing deep philosophical insight, it has unraveled fundamental connections of randomness with many diverse areas that has been truly spectacular.⁵

1.8 How this article is organized

Section 2 sets up notation for strings and sequences, the Cantor Space, which forms the basic framework for carrying out further discussions, and contains a brief and informal introduction to Lebesgue measure on the unit interval and the Cantor space, which we treat as equivalent.

Discussion of mathematical randomness begins in Section 3, where we define a series of classical stochastic or frequency properties in an effort to distill out randomness, but conclude by noting how this forces upon us notions like definability and effective computability.

In Section 4, an informal but precise definition of effective computability is given.

Sections 5, 6, 7, 8, and 9 form the core material on algorithmic randomness: Von Mises randomness, Martin-Löf randomness, Kolmogorov complexity and randomness of finite strings, application to Gödel incompleteness, and Schnorr's theorem.

Sections 10 and 11 form an overview of somewhat more advanced and recent topics, including definitions of various other forms of related randomness notions that have been studied, and indicating the current state of affairs.

⁵Algorithmic randomness is intimately related to the creation and development of several other fields such as Algorithmic Probability (originated by R. J. Solomonoff) and Universal Search (originated by L. A. Levin), all of which together comprise a broader (and vast) area now generally known as *Algorithmic Information Theory*. In addition to the creation of such fields, there has been wide and fundamental impact on many areas in mathematics, philosophy, statistics, and computer science: We mention recursion theory and Hausdorff dimensions; inductive and statistical inference and prediction; classical information theory; and complexity theory, machine learning, and artificial intelligence.

Finally, in Section 12 we express our view on the Martin-Löf-Chaitin thesis.

Proofs of mathematical results are provided or outlined whenever they are fairly straightforward. Involved proofs are omitted, almost all of which can be found in one of the books [Feller, 1968; Knuth, 1998; Li and Vitanyi, 2008; Downey and Hirschfeldt, 2010; Nies, 2009; Chaitin, 1992; Calude, 1994; Odifreddi, 1992].

2 STRINGS, SEQUENCES, CANTOR SPACE, AND LEBESGUE MEASURE

We use the term “string” (if not further qualified) to be a synonym for “finite sequence”, and the term “sequence” (if not further qualified) to be a synonym for “infinite sequence”.

To describe sequences and strings, we will use a finite *alphabet* Σ consisting of a finite number of letters or symbols. E.g., the *decimal* alphabet $\Sigma = \{0, 1, 2, \dots, 9\}$ consists of ten *digits*, and the binary alphabet $\Sigma = \{0, 1\}$ consists of just two *bits*. A *string* from Σ is simply a finite sequence of members of Σ . The empty string is denoted by Λ .

The set of all strings (finite sequences) from Σ is denoted by Σ^* . For a string $\sigma \in \Sigma^*$, $|\sigma|$ denotes the length of σ .

$\mathbf{N}$ denotes the set of all natural numbers. We deliberately leave it ambiguous if zero is considered a natural number or not. It can always be understood from context, and this ambiguity does not cause any problem. The set of all infinite sequences from Σ will be denoted by $\Sigma^{\mathbf{N}}$. For $x \in \Sigma^{\mathbf{N}}$, we will often write $x = \langle x_n \rangle_{n=1}^{\infty} = \langle x(n) \rangle_{n=1}^{\infty}$, where $x_n = x(n) \in \Sigma$ for each $n = 1, 2, \dots$.

2.1 Binary strings and sequences. The Cantor space

To simplify the discussion, we will almost always limit ourselves to binary sequences and strings, i.e. the case where the alphabet is $\Sigma = \{0, 1\}$. The set $\{0, 1\}^*$ is the set of all *binary strings*. The set $\{0, 1\}^{\mathbf{N}}$ is the set of all infinite binary sequences, and is called the *Cantor space*.

If $\sigma \in \{0, 1\}^*$ is a finite binary string, we will let $N(\sigma)$ to denote the set of all infinite binary sequences beginning with σ . E.g., $N(101)$ consists of all infinite binary sequences $\langle x_n \rangle_{n=1}^{\infty}$ for which $x_1 = x_3 = 1$, $x_2 = 0$, and x_k is arbitrary for all $k \geq 4$.

The subsets of the Cantor space $\{0, 1\}^{\mathbf{N}}$ having the form $N(\sigma)$ (for some binary string σ) will be called the *basic intervals* of the Cantor space. (By defining open sets as finite or infinite unions of basic intervals, the Cantor space becomes a topological space which is compact and metrizable.)

2.2 Lebesgue measure over the unit interval

The problem of Lebesgue measure over the unit interval $[0, 1] = \{x \in \mathbf{R}: 0 \leq x \leq 1\}$ is essentially a geometric one: Given a subset $E \subseteq [0, 1]$, we want to assign it a number $\mu(E)$ which represents its “size” or “length”. For very simple subsets

such as an interval, its measure is simply its length: If $J \subseteq [0, 1]$ is an interval with endpoints $a \leq b$ (i.e., J is one of the intervals (a, b) , $[a, b)$, $(a, b]$, or $[a, b]$), the measure of J , denoted by $\mu(J)$, is defined as the length of the interval, $\mu(J) = b - a$.

The next step is to define the length of any *open subset* of $[0, 1]$. A subset G of $[0, 1]$ is said to be *open* if it is a union of open intervals. A standard fact about this linear continuum is that any open set can be expressed *uniquely* as a disjoint union of (possibly infinitely many) intervals. This allows us to naturally and uniquely define the measure $\mu(G)$ of an open set $G \subseteq [0, 1]$ to be the sum (possibly as an infinite series) of all the constituent disjoint intervals.

A key idea here is that a set of “small measure” can be covered by an open set of “small measure”: A set E is said to be *measure-zero* if E can be covered by open sets of arbitrarily small measure, i.e., for any $\epsilon > 0$ there is an open set G containing E with $\mu(G) < \epsilon$. Slightly more constructively, E has measure-zero if there is an infinite sequence of open sets $G_1, G_2, G_3, \dots$ with each G_n covering E and with $\mu(G_n) < 1/n$.

A subset $E \subseteq [0, 1]$ is defined to be (*Lebesgue*) *measurable* if for any $\epsilon > 0$ there is an open set G containing E and an open set H containing the difference $G \setminus E$ with $\mu(H) < \epsilon$. Thus a measurable set is one which can be approximated from outside by open sets arbitrarily closely. If E is measurable, it can be shown that the measure of the open set G above approaches a unique limit as $\epsilon \rightarrow 0$, and we denote this limit by $\mu(E)$. This defines the Lebesgue measure for every measurable subset of $[0, 1]$. If $E \subseteq F \subseteq [0, 1]$ are measurable sets, then we have $0 \leq \mu(E) \leq \mu(F) \leq 1$.

The class of measurable sets form a vast collection. If $E \subseteq [0, 1]$ is measurable, so is its complement $[0, 1] \setminus E$ with $\mu([0, 1] \setminus E) = 1 - \mu(E)$. If $\langle E_n \rangle$ is a sequence of measurable sets, then their union $\cup_n E_n$ and intersection $\cap_n E_n$ are also measurable. If the sequence $\langle E_n \rangle$ consists of disjoint measurable sets, then the measure of their union is the sum of the measures of the individual sets: $\mu(\cup_n E_n) = \sum_n \mu(E_n)$.

2.3 Lebesgue measure on $[0, 1]$ as probability

Consider the experiment of choosing a member x of the unit interval $[0, 1]$ in such a fashion that for any two subintervals $J_1, J_2 \subseteq [0, 1]$ of equal length, it is equally likely for x to be in J_1 as to be in J_2 . (This is referred to as the uniform distribution over $[0, 1]$.) For a subset $E \subseteq [0, 1]$, the problem of determining the “probability that x is in E ”, denoted by $P(E)$, is identical to Lebesgue’s problem of finding the geometric measure (or length) of E .

We therefore identify, for this experiment, the notion of “events” with the class of measurable sets, and the probability of an event E with the Lebesgue measure of E : $P(E) = \mu(E)$.

2.4 The Cantor space: Infinite sequence of flips of a fair coin

An infinite sequence of flips of a coin, with 1 representing heads and 0 representing tails, can be naturally represented by an infinite binary sequence, i.e., by a member of the Cantor Space $\{0, 1\}^{\mathbf{N}}$, the set of all possible outcomes in an infinite sequence of flips of a coin. Subsets of $\{0, 1\}^{\mathbf{N}}$ correspond to events, e.g., the basic interval $N(101)$ represents the event that the first and third flips are heads and the second is tails.

We stipulate that the coin is fair and the flips are independent by requiring that for each $n \in \mathbf{N}$ the 2^n possible outcomes in the first n flips are equally likely, or equivalently that the probability of the event $N(\sigma)$ equals $1/2^{|\sigma|}$. Just as we extended Lebesgue measure on $[0, 1]$ from intervals to open sets and then to arbitrary measurable sets, a similar process can be carried out to obtain define a “probability measure” for the “measurable subsets” of $\{0, 1\}^{\mathbf{N}}$, a process which we now briefly describe.

The open sets in $\{0, 1\}^{\mathbf{N}}$ are defined to be arbitrary unions of basic intervals of the form $N(\sigma)$. A basic interval $N(\sigma)$ is said to be maximally contained in a set A if $N(\sigma) \subseteq A$ but $N(\tau) \not\subseteq A$ for each proper initial segment τ of σ . Every open set G then decomposes *uniquely* into a disjoint union of component basic open intervals maximally contained in G . So we can now naturally and uniquely define the measure of G , denote by $\mu(G)$, to be the sum of the lengths of these components. A key idea is again that of “small sets”: If $\epsilon > 0$ and if we can cover a set E by an open set of measure less than ϵ , then we can expect the measure of E to be less than ϵ as well. A set E is said to have *measure-zero* if there is an infinite sequence $G_1, G_2, G_3, \dots$ of open sets each containing E with $\mu(G_n) < 1/n$ for all n .⁶ Finally, as before in the case of the unit interval, define $E \subseteq \{0, 1\}^{\mathbf{N}}$ to be (*Lebesgue*) *measurable* if E can be approximated from outside by open sets arbitrarily closely, i.e., if for any $\epsilon > 0$ there is an open set G containing E and an open set H containing the difference $G \setminus E$ with $\mu(H) < \epsilon$.

It can then be shown that the measurable sets form a comprehensive collection including the open sets (and so all the basic intervals) and each measurable set E gets naturally assigned a unique measure $\mu(E) \in [0, 1]$. Also, the complement of any measurable set E is measurable with $\mu(\{0, 1\}^{\mathbf{N}} \setminus E) = 1 - \mu(E)$, and for any sequence $\langle E_n \rangle$ of measurable sets, their union $\cup_n E_n$ and intersection $\cap_n E_n$ are also measurable, with $\mu(\cup_n E_n) = \sum_n \mu(E_n)$ whenever the sequence $\langle E_n \rangle$ consists of pairwise disjoint sets. Also, if $E \subseteq F \subseteq \{0, 1\}^{\mathbf{N}}$ are measurable, then $0 \leq \mu(E) \leq \mu(F) \leq 1$.

Thus, starting with the simple method of assigning the probability (or measure) $1/2^{|\sigma|}$ to each basic interval $N(\sigma)$, we are then able to naturally extend this assignment procedure to assign probabilities (or measure) to vastly more general types of infinite coin-flip events (measurable subsets of $\{0, 1\}^{\mathbf{N}}$). We call this as-

⁶It is easy to see that a singleton is measure-zero. Using convergent infinite geometric series, it now follows that a countable union of measure-zero sets is also measure-zero. Thus all countable sets are measure-zero. A less trivial fact is that there are uncountable measure-zero sets.

signment (μ) to be the *Lebesgue (or uniform) probability measure on the Cantor Space* $\{0, 1\}^{\mathbb{N}}$.

In fact, by mapping an infinite binary string $x = \langle x(n) \rangle_{n=1}^{\infty}$ to the real number $\sum_n 1/2^{x(n)} \in [0, 1]$, we can naturally identify the uniform probability measure on the Cantor space with the Lebesgue measure on $[0, 1]$: The interval $[1/2, 1]$, e.g., corresponds to the event “the first flip is a heads”. This correspondence, as a mapping, is not quite one-to-one since the dyadic rationals of the form $m/2^n$ (where m, n are positive integers with $0 < m < 2^n$) have two different binary expansions, but these form only a countable set of exceptional points. All other reals in $[0, 1]$ have a unique infinite binary expansion. So this mapping between the Cantor space $\{0, 1\}^{\mathbb{N}}$ and the unit interval $[0, 1]$ is an almost one-to-one correspondence satisfying $\mu(E) = P(E')$ for any measurable subset $E \subseteq [0, 1]$ with E' being the set of infinite binary sequences which are binary expansions of the members of E .

To summarize, the Lebesgue measure on $[0, 1]$ (i.e., the uniformly distributed probability measure on $[0, 1]$) and the uniform probability measure on the Cantor space $\{0, 1\}^{\mathbb{N}}$ are essentially the same thing.

The zero-one law. This important result (used later) asserts the following for any measurable $E \subseteq \{0, 1\}^{\mathbb{N}}$. Suppose that whenever $x, y \in \{0, 1\}^{\mathbb{N}}$ are sequences differing only at a finite number of places (i.e., $(\exists m)(\forall n > m)(x(n) = y(n))$), then $x \in E \iff y \in E$. Then E has measure either zero or one.

2.5 More general probability distributions

We stress the following: *The notion of randomness must be understood relative to an a priori fixed probability distribution for the outcome values.* Each probability distribution for the outcome values determines a specific set of random sequences, and a sequence which is random relative to a given probability distribution for the outcomes will not be random relative to a different probability distribution for those outcomes. Without an underlying fixed a priori probability distribution for the possible outcome values, the notion of a “random sequence” would not even make sense. It is thus useful to fix a specific probability distribution on a specific set of outcome values (a probability model) when discussing random sequences.

As mentioned earlier, in this article we will restrict our attention to the case where the underlying probability distribution for the blackbox is *the fair coin model*: Only two equiprobable outcomes 0 and 1 with $P(0) = P(1) = \frac{1}{2}$. We will then try to find out which sequences are random (i.e. investigate Question B of Section 1.2 and Problems 1 and 2 of Section 1.5) relative only to this fair coin model. This may at first appear to be too severe a restriction, but the extra generality obtained by considering more general probability distributions, while adding verbiage, would not provide much extra insight into the question of randomness. For the reader who is still worried about our restriction to the fair coin model alone, we now mention some technical results which show how this simple case can represent, at least for sequences, other more general distributions.

Suppose that our blackbox represents a more general probability distribution,

with n outcomes $\omega_1, \omega_2, \dots, \omega_n$, and corresponding probabilities $P(\omega_i) = p_i$, $i = 1, 2, \dots, n$, where $0 < p_i < 1$ and $\sum_{i=1}^n p_i = 1$. Then the probability space $P^{\mathbb{N}}$ of all infinite sequences of outcomes is still essentially identical to the Cantor space with Lebesgue measure. In particular, there is a measurable bijection between $P^{\mathbb{N}}$ and $\{0, 1\}^{\mathbb{N}}$ which preserves measure. This is a technical result on finite Borel measures on complete separable metric spaces. If $p_1, p_2, \dots, p_n$ are computable real numbers, then this bijection can be assumed to be an effective one.

If the probabilities $p_1, p_2, \dots, p_n$ are dyadic rational numbers (as is the case, e.g., in computer representations of numbers), then one can use an especially simple effective coding by which any sequence (finite or infinite) of outcomes of the blackbox can be represented effectively by a corresponding binary sequence from the fair coin model, while still preserving probabilities of all events.

As another example, we mention the so called *von Neumann trick*, a simple method by which one can “turn a biased coin into an unbiased one”, or, more precisely, simulate sequences of flips of a perfectly unbiased coin using sequences of flips of a biased coin (whose probability of heads differs from probability of tails). For every two consecutive flips, record the outcome as a single 0 if the (ordered) pair of flips is HT, record it as a single 1 if the pair of flips is TH, otherwise discard the pair (the cases HH and TT), and move to the next pair of flips of the coin.

We will see later that Martin-Löf’s definition of randomness is general and flexible enough to be applied directly to very general probability distributions. However, the technical results mentioned here show that there are mathematically sound ways to restrict our attention to the simplified case of the fair coin model, without imposing any serious limitation to the study of random sequences. We adopt this simplification, which equivalently means that the underlying sequence space will always be the Cantor Space with the (uniform) Lebesgue measure.

3 CLASSICAL STOCHASTIC RANDOMNESS IN INFINITE SEQUENCES

We now begin the discussion of the fundamental question stated in Problem 1: *Given an infinite binary sequence $x = \langle x_1, x_2, \dots, x_n, \dots \rangle \in \{0, 1\}^{\mathbb{N}}$, how do we determine if it is random?*

3.1 Key points

We start by noting several key points of randomness in infinite binary sequences. Recall that the Cantor space with Lebesgue measure is the underlying probabilistic model throughout. Our observations here will be heuristic and informal but of fundamental importance.

- (a) *The probability that a sequence is random equals one*, i.e., the random sequences form a set of measure one (full-measure set). A basic intuition here is that the omission or addition of one single bit has no effect on the randomness of an infinite sequence: If x is the sequence $x_1 x_2 x_3 \dots$, and x' is the

sequence $x_2x_3x_4\dots$ obtained by dropping only the first bit of x , then x' is random if and only if x is random. It follows by mathematical induction that the randomness of an *infinite* sequence should depend only on its “eventual behavior” and no amount of finite part can determine the randomness of the entire sequence.⁷ Thus if two infinite sequences x and y agree on all but a finite number of places (i.e. $\exists m \forall n > m (x_n = y_n)$), then the randomness of x is equivalent to the randomness of y . This implies that the set of random sequences satisfies the condition for the zero-one law, and therefore must be either a measure-zero set or a full-measure set. As pointed out by Laplace, it seems natural that the random sequences should form the vast majority of sequences: If we “randomly pick” a sequence in $\{0,1\}^{\mathbb{N}}$, or equivalently, generate one by “randomly flipping” a fair coin infinitely many times, the probability that the result is random should be high, and so non-zero. It follows that the random sequences must form a set of measure one.

- (b) Second, *if the sequence of outcomes of a gambling wheel with two equiprobable outcomes is random then no successful betting strategy can be devised against it.* Note that, here the bits of an infinite sequence x are thought to be generated by the independent turns of the gambling wheel, and that the randomness of x would imply a strong form of unpredictability for the future bit values of x (from previously observed bit values). More precisely, suppose that a gambling house is generating the infinite binary sequence x and offering the following fair game: A gambler can bet an amount b of money predicting the next bit of x ; if the prediction is correct, the gambler wins an amount b , otherwise loses the same amount b . (This rule is tantamount to our underlying assumption of a fair coin model.) We say that the gambler is able to *devise a successful gambling system against x* , or *beat the house against x* , if by using a suitable strategy the gambler can start with a finite initial capital and win an arbitrarily large fortune without going bankrupt. By a strategy we mean a finitely specifiable rule which determines how much (and whether) to bet on particular turn based on the outcomes of the previous turns. (These notions will be made more precise later in subsections 5.1 and 11.2.) From the point of view of the house, the randomness of the sequence x of outcomes must imply that the bits of x should be so unpredictable that no gambler would be able to beat the house against x . We carefully note that this *impossibility of a successful gambling system is a fundamental necessary condition for randomness of an infinite sequence*, (a fact first recognized by von Mises, see subsection 5.1 for quoted comments of Feller).⁸

⁷This is quite similar to the notion of *the limit* of an infinite numerical sequence found in elementary calculus: No amount of alternation of the values of any finite number of terms of a sequence can affect its limit.

⁸Later, during the attempts in subsections 5.1 and 11.2 to find a suitable definition of randomness, things will be turned around to postulate this impossibility condition as also a sufficient condition for randomness.

- (c) On the other hand, *randomness cannot be identified with complete and absolute lawlessness*. We may try to think of a sequence to be random if it satisfies no law whatsoever (“absolutely lawless”). But, as pointed out by Calude in [Calude, 2000] (see also [Volchan, 2002]), no such sequence can exist, since *every* digit-sequence satisfies the following Ramsey-type combinatorial law first proved by van der Waerden [1927]: The positions (indices) for at least one digit-value will contain arbitrarily long arithmetic progressions. Thus we have to abandon such ideas of “complete lawlessness”.
- (d) In fact, *random sequences will necessarily satisfy certain limiting or stochastic properties*. For example, given a sequence $x = \langle x_k \rangle$, let $\mathbf{S}_n[x]$ denote the number of 1s (or **S**uccesses) in the first n terms of the sequence x :

$$\mathbf{S}_n[x] = \sum_{k=1}^n x_k = \text{Number of 1s in first } n \text{ terms of } x,$$

so that the quantity $\frac{1}{n}\mathbf{S}_n[x]$ represents the *proportion* of 1s in the first n terms of x . This proportion is called the *relative frequency* or simply the *frequency* (of successes). If for a sequence x this proportion exceeds $\frac{2}{3}$ (say) infinitely often (i.e., $\frac{1}{n}\mathbf{S}_n[x] > \frac{2}{3}$ for infinitely n), then x cannot be random (under the fair coin model), because a gambler would then be able to exploit this “bias within x ” to devise a strategy which (starting with a finite initial capital) can return an arbitrarily large fortune without going bankrupt.

More precisely, if x is to be random so that no gambling system would be successful against it, then it can be shown that the following condition must hold:

*For no positive number p should the proportion $\frac{1}{n}\mathbf{S}_n[x]$ exceed $\frac{1}{2} + p$ (or fall below $\frac{1}{2} - p$) infinitely often.*⁹

This condition on x is equivalent to requiring that the relative frequency $\frac{1}{n}\mathbf{S}_n[x]$ approach the value $\frac{1}{2}$ in the limit as n approaches infinity.

Thus, for every random x we have $\lim_{n \rightarrow \infty} \frac{1}{n}\mathbf{S}_n[x] = \frac{1}{2}$. This exemplifies that *in order to be random, a sequence, instead of being “totally lawless”, must actually satisfy certain stochastic laws of “unbiasedness”*. In the rest of this section we will discuss further stochastic laws that should be satisfied by every random sequence, starting with the *Borel strong law*, whose basic condition is the same as the one in the above example.

⁹If $\frac{1}{n}\mathbf{S}_n[x]$ exceeds $\frac{1}{2} + p$ infinitely often for a positive p , then the gambler can beat the house by betting a constant fraction r of his remaining capital at each turn predicting a bit outcome of 1, where r is a constant fraction with $0 < r < p/(1+p)$. We omit the details of calculation.

3.2 The Borel Strong Law

In 1909, Émile Borel [1909] proved a remarkable fact about infinite binary sequences which was later generalized by many mathematicians (from Cantelli to Kolmogorov) and became a fundamental result of probability theory called *The Strong Law of Large Numbers*. Borel established that with probability one, the proportion of 1s among the first n terms ($\frac{1}{n}\mathbf{S}_n$) approaches the value $\frac{1}{2}$ in the limit.

THEOREM 1 The Borel strong law. *For independent infinite sequences of flips of a fair coin, let B denote the event that the proportion of successes among the first n flips, $\frac{1}{n}\mathbf{S}_n$, approaches the limit $\frac{1}{2}$ as $n \rightarrow \infty$, or formally, put*

$$B = \left\{ x \in \{0, 1\}^{\mathbf{N}} : \lim_{n \rightarrow \infty} \frac{\mathbf{S}_n[x]}{n} = \frac{1}{2} \right\}.$$

Then the probability of the event B is 1, i.e. the set B has Lebesgue measure 1.

We saw that due to impossibility of successful betting strategies, a random sequence must satisfy the condition of the Borel strong law. That, together with our earlier observation that the random sequences form a set of measure one, provides an informal proof for the Borel strong law. (For a formal proof, see [Feller, 1968].)

Informally, the Borel strong law asserts that if we randomly pick a member x from $\{0, 1\}^{\mathbf{N}}$, the probability is one that the relative frequency of 1s among initial parts of x approach a limiting value called *the limiting frequency*, and this limiting frequency is equal to $\frac{1}{2}$, so that “randomly picked sequences are unbiased”.

Conversely, if for a sequence x this limiting frequency exists but is not equal to $1/2$, then, in view of our underlying fair coin model, x would clearly be biased, not random. And if the limiting frequency does not exist then either $\limsup_n \frac{1}{n}\mathbf{S}_n[x] > \frac{1}{2}$ or $\liminf_n \frac{1}{n}\mathbf{S}_n[x] < \frac{1}{2}$, and in either case arbitrarily large segments of x would be biased with arbitrarily great statistical significance, so x again would be non-random (successful gambling systems could be devised against x in these cases).

Thus, it is natural to view this “stochastic law of unbiasedness” as a “stochastic law of randomness”.

Note that satisfying this law is only a basic necessary condition for being random. For example the sequence $0101010101 \dots$ of alternating 0s and 1s satisfies the condition of the strong law but this sequence is clearly not random. We therefore look for stronger stochastic laws which would exclude such simple examples.

3.3 Borel Normality

Let σ be a fixed binary string with length $|\sigma| = k$ (e.g., if $\sigma = 00110101$ then its length is $|\sigma| = k = 8$). If a fair coin is flipped k times, the probability of obtaining σ as the resulting outcome equals $1/2^k$. For an infinite binary sequence x , consider a block of k bits starting at bit position n : This is the block segment of x given by $x_n x_{n+1} \dots x_{n+k-1}$. It is not hard to see that with probability 1 the string σ

must occur infinitely many times in an infinite binary sequence.¹⁰ Adding this condition (that every finite binary string must occur in x with infinite frequency) as a stochastic law certainly excludes simple sequences satisfying the Borel strong law, such as $01010101\dots$, as being random.

In [Borel, 1909], Borel had established an even stronger fact. Call an infinite binary sequence x to be *Borel normal in base 2* if for each fixed binary string σ of length $|\sigma| = k$, the frequency of occurrences of σ among the first n bits of x (as a fraction of n) approaches $1/2^k$ as $n \rightarrow \infty$. (For $|\sigma| = k = 1$, this reduces the condition of the Borel strong law.) Borel proved that *the probability that x is normal in base 2 is one, i.e. almost all infinite binary sequences are normal in base 2*. Any infinite binary sequence satisfying this property will clearly be “much more random” than simple sequences like $01010101\dots$.

An immediate consequence of the Borel normality is another form of unpredictability for random sequences: *The observation of any particular bit pattern in a random sequence does not influence the value of the next bit*. More precisely, for almost all infinite binary sequences x and any fixed finite binary string σ (bit-pattern), the limiting frequency that σ is immediately followed by 0 equals the limiting frequency that σ is followed by 1.

An explicit example of an infinite binary sequence which is normal in base 2 is the *Champernowne binary sequence* obtained by concatenating the binary representation of every non-negative integer (taken in their natural order):

$$01101110010111011110001001101010111100110111101111\dots$$

As far as randomness is concerned, this is a big improvement over the simple $01010101\dots$, but it is impossible to call this sequence random, since it is also generated by a simple effective procedure.

Recall that one can identify $\{0,1\}^{\mathbb{N}}$ with the unit interval $[0,1]$ by viewing infinite binary sequences as binary expansions of reals in $[0,1]$ (after disregarding a negligible countable subset of $\{0,1\}^{\mathbb{N}}$). This identification preserves measurable sets and the measure of every such set. Moreover, given an integer base $b > 1$, every real $x \in [0,1]$ can be expanded in base b as:

$$x = \sum_{k=1}^{\infty} x_k b^{-k}, \quad \langle x_k \rangle \in \{0,1,\dots,b-1\}^{\mathbb{N}}.$$

The terms of the infinite sequence $\langle x_k \rangle \in \{0,1,\dots,b-1\}^{\mathbb{N}}$ above are known as the *b-ary digits of x* . E.g., $b = 2$ for binary and $b = 10$ for decimal expansion.

For $x \in [0,1]$ with base b expansion $\langle x_k \rangle$, we say that x is *Borel normal in base b* if for each fixed finite string $\sigma \in \{0,1,\dots,b-1\}^*$ of length $|\sigma| = k$, the

¹⁰Proof: Divide x into consecutive blocks of size k each. For any n , the probability that σ is not the n -th block equals the constant $r = 1 - 1/2^k$. So the probability that σ does not occur as one in a run of m consecutive blocks is r^m . Since $r < 1$, $r^m \rightarrow 0$ as $m \rightarrow \infty$, and thus for any n the probability that σ does not occur after the n -th block is zero. It follows that the probability that σ occurs in only finitely many blocks is also zero, QED.

frequency of occurrences of σ among the first n b -ary digits of x (as a fraction of n) approaches $1/b^k$ as $n \rightarrow \infty$. Finally, x is *absolutely normal* if x is Borel normal in every base $b > 1$. Note that by the identification of $\{0, 1, \dots, b-1\}^{\mathbf{N}}$ with $[0, 1]$, these definitions apply to infinite sequences as well.

Two examples numbers Borel normal in base 10 are the following reals shown in decimal expansion:

0.12345678910111213... (Decimal Champernowne number),

0.235711131719232931... (Decimal Copeland-Erdős number).

For the first number above, the decimal digits after the decimal point are formed by concatenating the positive integers written in decimal notation (in their natural order), while the second one has decimal digits formed by concatenating the positive primes written in decimal notation. While both these are Borel normal in base 10, it is not clear if they are Borel normal in any other base.

Borel's result implies that almost all reals in $[0, 1]$ (almost all infinite binary sequences) are actually absolutely normal. But it is harder to come up with examples of absolutely normal numbers, as they have an even higher degree of randomness compared to the Champernowne number or the Copeland-Erdős number. It is an old open question if the number π (or $\sqrt{2}$, or e) is absolutely normal, or even normal in some base whatsoever.

Following early work of Sierpinski [1917] (also Turing [1992]), Becher and Figueira [2002] have constructed absolutely normal numbers using an effective procedure (these are real numbers which are computable, although somewhat complicated to define).

3.4 Laws of random walk (“wandering drunkard” laws)

One can identify the infinite binary sequences (or, equivalently, all possible outcomes of an infinite sequence of coin flips) with the set of all random walks. Consider the number line infinitely extended in both directions and indexed by the integers $\mathbf{Z}$, and a person starting at 0 taking a sequence of steps determined by $x \in \{0, 1\}^{\mathbf{N}}$ as follows: The first step is one unit to the right (in the positive direction) if $x_1 = 1$ and one unit to the left if $x_1 = 0$; and the n -th step of the person is similarly determined by the value of x_n . If x is random, then this results in a *random walk* (sometimes called the drunkard's walk). If $\mathbf{S}_n[x]$ denotes the number of 1s (Successes) in the first n bits of x and $\mathbf{F}_n[x] = n - \mathbf{S}_n[x]$ the number of 0s (Failures) in the first n bits of x , then the position of the person on the number line after step n is given by:

$$\mathbf{S}_n[x] - \mathbf{F}_n[x] = 2\mathbf{S}_n[x] - n.$$

It is not hard to see that with probability 1, the person must move away an arbitrarily large distance to the right of the starting point, and also an arbitrarily

large distance to the left of the starting point.¹¹ Formally, in terms of infinite binary sequences, the set

$$\left\{ x \in \{0, 1\}^{\mathbf{N}} : \sup_n \mathbf{S}_n[x] - \mathbf{F}_n[x] = +\infty \text{ and } \inf_n \mathbf{S}_n[x] - \mathbf{F}_n[x] = -\infty \right\}$$

must have measure (probability) one.

In particular, this stochastic law implies that for a walk to be random, the person must oscillate about the origin with “arbitrarily large amplitudes”, and must “cross the origin” from right to left and from left to right infinitely many times:

THEOREM 2 Law of Symmetric Oscillation in Random Walks. *If $x \in \{0, 1\}^{\mathbf{N}}$ is random, then we must have:*

$$\frac{\mathbf{S}_n[x]}{n} > \frac{1}{2} \text{ for infinitely many } n, \text{ as well as: } \frac{\mathbf{S}_n[x]}{n} < \frac{1}{2} \text{ for infinitely many } n.$$

Thus a walk in which the person eventually stays on one side of the origin (eventually to the right or eventually to the left) is “biased” and cannot be “random”.

3.5 The Law of Iterated Logarithms and Strong Normality

The Borel strong law says that $\frac{1}{n}\mathbf{S}_n$ converges to the expected value $\frac{1}{2}$ with probability 1, but it does not say how the variance (or standard deviation) of $\mathbf{S}_n$ behaves asymptotically.

A much more precise and stronger theorem (than the Borel strong law) is the *Law of Iterated Logarithms*, which gives an exact asymptotic bound on the deviation of $\mathbf{S}_n$: *With probability one, the values of $\mathbf{S}_n$ spread around its mean to an asymptotic distance of $\sqrt{2\log\log n}$ times its standard deviation.* To state it more precisely, note that the mean of $\mathbf{S}_n$ is $\mu_n = n/2$, and its standard deviation $\sigma_n = \sqrt{n}/2$. Then the Law of Iterated Logarithms asserts the following: For any $\lambda > 1$, the probability is one that $\mathbf{S}_n < \mu_n + \lambda\sqrt{2\log\log n}\sigma_n$ for all but finitely many n , and for $\lambda < 1$, the probability is one that $\mathbf{S}_n > \mu_n + \lambda\sqrt{2\log\log n}\sigma_n$ for infinitely many n ; and similarly for the lower bounds of $\mathbf{S}_n$. (See Feller [1968] for proof).

The Law of Iterated Logarithms has both the Borel strong law and the Law of Symmetric Oscillation in 3.4 as immediate corollaries.

¹¹Proof (outline): Computing binomial probabilities, the probability that $|\mathbf{S}_n[x] - \mathbf{F}_n[x]| \leq k$ for all but finitely many n equals zero for each k , and hence the probability that the sequence $\langle \mathbf{S}_n[x] - \mathbf{F}_n[x] : n \in \mathbf{N} \rangle$ is unbounded equals one. Now the events $\sup_n \mathbf{S}_n[x] - \mathbf{F}_n[x] = +\infty$ and $\inf_n \mathbf{S}_n[x] - \mathbf{F}_n[x] = -\infty$ are equiprobable and satisfy the conditions of the zero-one law. So either both events have probability zero, or both have probability one. But it is impossible for both these events to have probability zero, as that would imply boundedness of the sequence $\langle \mathbf{S}_n[x] - \mathbf{F}_n[x] : n \in \mathbf{N} \rangle$ with probability one (a contradiction), and the result follows.

Belshaw and Borwein [2005;] have used a slightly weaker version of the Law of Iterated Logarithms to define the notion of *Strong Normality*. They show that requiring a sequence to be strongly normal makes it more random than requiring it to be simply Borel normal. This is established both by graphic empirical evidence and by proving that the Champernowne binary sequence is not strongly normal. Since every sequence satisfying the Law of Iterated Logarithms is strongly normal, it follows that the Champernowne binary sequence violates the Law of Iterated Logarithms.

3.6 Equidistribution laws and the Ergodic Frequency Theorem

For any statement P , we use the notation $\llbracket P \rrbracket$ to denote the binary truth value of the statement P , i.e.,

$$\llbracket P \rrbracket = \begin{cases} 1 & \text{if } P \text{ is true} \\ 0 & \text{if } P \text{ is false} \end{cases}.$$

Let $\langle x_n \rangle$ be an infinite sequence of real numbers in $[0, 1)$. We say that $\langle x_n \rangle$ is *equidistributed* (or *uniformly distributed*) if for all $0 \leq a < b \leq 1$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \llbracket x_k \in [a, b) \rrbracket = \text{measure of } [a, b) \quad (= b - a).$$

In other words, $\langle x_n \rangle$ is equidistributed if the limiting frequency with which x_n enters the interval $[a, b)$ equals the size of $[a, b)$.

Similarly, a sequence $\langle x_n \rangle$ of infinite binary sequences (i.e. each $x_n \in \{0, 1\}^{\mathbb{N}}$) is equidistributed if for each finite binary string σ ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \llbracket x_k \in N(\sigma) \rrbracket = \text{measure of } N(\sigma) \quad \left(= \frac{1}{2^{|\sigma|}} \right),$$

where $N(\sigma)$ is the set of all infinite binary sequences having σ as an initial segment.

Equidistribution can also be viewed as a form of unbiasedness: Every subinterval asymptotically gets its “proper share” of the sequence.

A great deal of classical mathematical literature exists on equidistribution (see [Kuipers and Niederreiter, 1974]). We mention a theorem of Weyl: *If $\langle a_n \rangle$ is a sequence of distinct integers, then the sequence $\langle \text{FRAC}(a_n x) \rangle$ is equidistributed for almost all x in the unit interval.* (Here $\text{FRAC}(x)$ denotes the fractional part of $x \in \mathbf{R}$: $\text{FRAC}(x) = x - \lfloor x \rfloor$, where $\lfloor x \rfloor$ is the floor of x , or the greatest integer not greater than x .)

Borel normality can be viewed as a special case of the Weyl equidistribution theorem just mentioned. Taking $a_n = 2^{n-1}$, we see that $\langle \text{FRAC}(2^{n-1} x) \rangle$ is equidistributed over the unit interval for almost all x . Moving to $\{0, 1\}^{\mathbb{N}}$, this implies that for almost all $x = \langle x_n \rangle \in \{0, 1\}^{\mathbb{N}}$, the sequences $\langle x_1, x_2, x_3, \dots \rangle$, $\langle x_2, x_3, x_4, \dots \rangle$, $\langle x_3, x_4, x_5, \dots \rangle$, etc, are equidistributed over $\{0, 1\}^{\mathbb{N}}$.

Let $L: \{0, 1\}^{\mathbb{N}} \rightarrow \{0, 1\}^{\mathbb{N}}$ denote the left-shift operator:

$$L(\langle x_1, x_2, x_3, \dots \rangle) = \langle x_2, x_3, x_4, \dots \rangle.$$

Then we may restate this equidistribution by saying that for almost all $x \in \{0, 1\}^{\mathbb{N}}$, the sequence $\langle x, Lx, L^2x, \dots \rangle$ is equidistributed. In other words, for every finite binary string σ ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{I}[L^k x \in N(\sigma)] = \text{measure of } N(\sigma) \left(= \frac{1}{2^{|\sigma|}} \right),$$

for almost all $x \in \{0, 1\}^{\mathbb{N}}$. But this is precisely the statement that almost all infinite binary strings are normal in base 2.

The left-shift operator L is an example of an ergodic operator, and the last displayed equation is a special case of the Birkhoff Ergodic Theorem.

In subsection 11.6, we further discuss how this approach can be used as a stochastic law for specifying randomness.

3.7 General probabilistic laws for specifying randomness

In general, suppose that a specific law L for sequences (i.e. a specific property L of sequences) is satisfied with probability one, i.e. satisfied by almost all sequences. We have seen many examples such laws (“laws of large numbers”): The Borel strong law, Borel normality, Symmetric Oscillations, Iterated Logarithms, etc, form an increasingly stringent sequence of such laws. Let us call any such law to be a *probabilistic law of randomness*. Formally, a *probabilistic law of randomness*, or simply a *randomness law*, is an *explicitly defined* set L of binary sequences having full-measure. Given such a law L , the probability that a random sequence will satisfy it is 1, so we can make L to be another law for randomness.

By specifying more and more stringent randomness laws, we can try to verify that a given sequence is sufficiently random in this sense. In fact, any algorithm for generating pseudo-random sequences needs to be subjected to a series of such theoretical stochastic tests¹² to estimate the “quality of randomness” for sequences generated by the algorithm.

After a sufficient number of such stages of refinement, we can hope to arrive at the “right” definition of randomness for infinite sequences — a definition which best matches our intuition. This is the approach taken by Knuth in his fascinating article *What Is a Random Sequence?* (Section 3.5 in [Knuth, 1998]). Knuth gives a series of definitions, **R1–R6**, and claims (or hopes, as he clarifies in a later footnote) that the final refinement (definition **R6**) is an appropriate definition of randomness.

¹²This should not be confused with statistical tests for randomness in a finite sequence of digits. As mentioned earlier, for an infinite sequence, no amount of finite part can determine if the entire sequence is random. Statistical tests, in particular the χ^2 -tests, are important for estimating “randomness confidence” for a finite set of data. See Knuth [Knuth, 1998] for a list of statistical tests of randomness, and p. 80 of [Knuth, 1998] for theoretical tests. John Walker distributes a suite of statistical tests for randomness.

3.8 Is absolute probabilistic randomness possible?

What happens if we define a sequence to be random if it satisfies *all* explicitly defined probabilistic randomness laws? Provisionally, let us call such a sequence *absolutely random*.

Unfortunately, such a definition causes problems.

First, we have an apparent paradox: For an arbitrary binary sequence y , consider the randomness law $L_y := \{x \in \{0, 1\}^{\mathbb{N}} : x \neq y\}$. If x satisfies all randomness laws, then $x \in L_y$ for all $y \in \{0, 1\}^{\mathbb{N}}$, but this implies $x \in L_x$, so $x \neq x$, a contradiction! It follows that no binary sequence can satisfy all randomness laws!

However, this is not a real paradox. Since every law can be written using a finite sequence of symbols from a finite alphabet, there are at most countably many laws that can be *explicitly defined*. For all but a countably many members $y \in \{0, 1\}^{\mathbb{N}}$, the law L_y cannot even be explicitly stated, and hence will not count.

But, second, we run into a more serious technical metamathematical problem: The notion of “all randomness laws”, which makes good intuitive sense, is not formalizable in the standard system of axiomatic mathematics (ZFC). The reason for this is that the notion of a randomness law really refers to a *definable* subset of $\{0, 1\}^{\mathbb{N}}$ of full-measure, but the notion of *definability* itself can only be defined in terms of *satisfaction*, or *truth*. By a classic result of Tarski, it is impossible to formalize the notion of truth in a formal system within the system.

One way out from this metamathematical difficulty is to not talk about the entire class of all definable subsets of $\{0, 1\}^{\mathbb{N}}$, but restrict ourselves to those which can be defined using formulas with a limited number of quantifiers. By restricting the number and scope of the quantifiers in the defining formulas, various hierarchies of definable subsets of $\{0, 1\}^{\mathbb{N}}$ are obtained, such as the *arithmetical hierarchy* (scope of quantifiers limited to the set of natural numbers), and the *analytical hierarchy* (scope of quantifiers limited to natural and real numbers).

This means that we must be satisfied with relative degrees of randomness, and Absolute Randomness, like truth in formal systems, must remain elusive forever.

This is the approach taken by modern research. Very roughly, the n -th level of the arithmetical hierarchy consists of sets defined by formulas with n alternating quantifiers ranging over natural numbers. For each n , a notion of n -randomness can be defined appropriately. The higher the value of n , the stronger is the randomness. The lowest level of this hierarchy, which defines 1-randomness by suitably considering algorithmic randomness laws, is especially important. It captures the random sequences by defining them as the ones which satisfy all “algorithmic stochastic laws converging algorithmically.” But we first need to rigorously define the concept of algorithm to talk about these notions precisely.

4 ALGORITHMS AND POST MACHINES

Like randomness, the intuitive notion of algorithm (or effective computation) was not easy to capture. During the early part of the 20th century, in response to

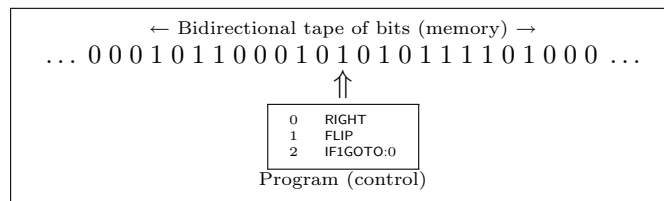
Hilbert’s program, there was a great deal of effort by mathematicians to come up with a precise definition of the term “algorithm”. During the 1930’s decade, which began with Gödel’s celebrated negative answer to Hilbert’s program, several mathematicians independently came up with such definitions. These included, in addition to Gödel, Herbrand, Church, Kleene, Post, and of course, Turing.

After sorting out more restrictive definitions (such as those now known as primitive recursive computation), mathematicians converged on a definition that Church (in 1936) announced as the appropriate definition for algorithm. Church’s assertion is known as *Church’s Thesis* or the *Church-Turing thesis*. This is the definition that is accepted today, and so the Church-Turing thesis turned out to be highly successful. Most remarkably, almost all independent definitions coincided and produced the same characterization of the notion of algorithm!

We present this definition as a variant of Post’s original version (see also [Uspensky, 1983; Davis, 1980]).

4.1 Post machines and programs

A *Post Machine* consists of a control part containing a *program* (a finite list of *instructions* described below), a bidirectional infinite *tape* of memory bits, and a *head* (shown as $\uparrow\uparrow$) which at any time is located at some unique bit position (the “current bit”) of the tape.



The Post Machine

The head can read the bit at the current position as either 0 or 1 and inform this value to the control, and at the direction of control it can either change the bit value at the current position, or move to one bit position to the left, or to one position to the right. The value of every tape bit is always either 0 or 1 (there are no blank symbols).

Instruction	Code	Function
FLIP	0	Complement current bit: 0 becomes 1 and 1 becomes 0
LEFT	1	Move the head one bit position to the left
RIGHT	2	Move the head one bit position to the right
IF1GOTO: <i>n</i>	<i>n</i> + 3	Conditional jump: If current bit = 1, GO TO instruction # <i>n</i>

Table 1. Post Machine Instruction set and their numeric encodings

The Post Machine has four¹³ types of instructions, as listed in the table.

A Post Machine *program* is simply a finite list of instructions as given in Table 1. Equivalently, since every instruction corresponds uniquely to a natural number code in Table 1, a program can be defined as a finite sequence of natural numbers. In addition, every instruction in the program is assumed to be labeled serially by its “line number”, starting with 0 for the first instruction. Program execution starts at line number 0 and sequentially proceeds to the following line, except possibly for the IF1GOTO:*n* instruction. There is no explicit STOP or HALT instruction, and the program terminates whenever it is unable to execute the “next instruction”.

Here is an example program and its code as a finite sequence of natural numbers:

0	RIGHT
1	IF1GOTO:0
2	FLIP
3	LEFT
4	IF1GOTO:3

$$= \langle 2, 3, 0, 1, 6 \rangle.$$

We encode natural numbers as strings by the correspondence $n \leftrightarrow 1^{n+1}0$:

$$0 \leftrightarrow 10, 1 \leftrightarrow 110, 2 \leftrightarrow 1110, 3 \leftrightarrow 11110, \dots$$

Given a program P and a positive integer k , we will now define a partial function φ of k natural number arguments. To find $\varphi(n_1, \dots, n_k)$, we start the program with the input string $1^{n_1+1}0 \dots 1^{n_k+1}0$ on the tape and the head, as in:

$$\dots 0 \underset{\substack{\uparrow \\ P}}{1} \overbrace{1 \dots 1}^{n_1} 0 \underset{\substack{\uparrow \\ P}}{1} \overbrace{1 \dots 1}^{n_2} 0 \dots \dots \underset{\substack{\uparrow \\ P}}{1} \overbrace{1 \dots 1}^{n_k} 0 0 \dots \quad (\text{other tape bits are } 0).$$

There are now two possibilities, and we define $\varphi(n_1, \dots, n_k)$ accordingly:

Case 1 *The program P terminates when started as above.* We then define $\varphi(n_1, \dots, n_k) = m$, where m is the length of the run of 1s to the right of the head, before the first 0 to the right of the head, as in:

$$\dots * \underset{\substack{\uparrow \\ P}}{*} \overbrace{1 \ 1 \ \dots \ 1}^m 0 * \dots$$

(We express the situation in this case by saying that *the program P halts on the inputs $n_1, \dots, n_k$ with output m .*)

Case 2 *The program P started as above does not terminate (“loops forever”).* In this case we leave $\varphi(n_1, \dots, n_k)$ undefined.

¹³One can combine the first two instructions into a single one to have an adequate version of the Post Machine with only three types of instructions.

Thus, for every program P and positive integer k , there is a unique k -ary partial function φ determined as above, and we express this by saying φ is the k -ary partial function computed by P , or P computes the k -ary partial function φ .

For example, the example program above computes the 1-ary function $f(n) = n + 2$, but it also computes the 2-ary function $f(m, n) = m + n + 3$. As simpler examples, note that the single line program with the only instruction **LEFT** computes the (1-ary) successor function $s(n) = n + 1$, and the empty program computes the identity function $f(n) = n$.

We can now formally define the notion of “algorithm” or *effective computability*:

DEFINITION 3. A k -ary partial function $f = f(n_1, n_2, \dots, n_k)$ is *effectively partial computable* or simply *partial computable* if there is Post machine program P such that P computes f . If f is total, we say that f is *effectively computable*, or simply *computable*.

A subset $R \subseteq \mathbf{N}^k$ (i.e., a k -ary relation R , which is simply a set of natural numbers if $k = 1$) is called *effectively computable* or *algorithmically decidable* or simply *computable* if its characteristic function is a computable function.

$R \subseteq \mathbf{N}^k$ is called *computably enumerable* (or *c.e.*) if it equals the domain of some k -ary partial computable function.

(In older literature the term “recursive” is used in place of “computable”.)

It can be shown that a set A is c.e. iff it equals the range of a partial computable function f . If the set A is non-empty, the function f can be assumed to be total. This explains the terminology: A set E is computably enumerable if its elements can be enumerated by a computable function f , as in $E = \{f(1), f(2), f(3), \dots\}$.

Another important fact is that a set is computable iff both the set and its complement are c.e.

Most importantly, we want to note that the notions of effectively computable functions and computable sets, c.e. sets, etc, are all independent of the particular model of computation. In particular, if a function can be computed by some other computer, however powerful, it will be also computable by some Post machine program.

Since there are only countably many programs but uncountably many sets and functions of natural numbers, it follows that most functions are not computable and most sets are not decidable (not even c.e.). We will see some specific examples soon.

4.2 Gödel numbering Post machine programs

Since a program P is a finite sequence of natural numbers, say $P = \langle p_1, p_2, \dots, p_m \rangle$, each program is easily coded into a single natural number in an effective manner.

One way to do this would be to build an integer $e(P)$ from $P = \langle p_1, p_2, \dots, p_m \rangle$ in binary notation as follows: Write 1 followed by p_m zeros, followed by another 1 and p_{m-1} more zeros, and so on, ending with 1 followed by p_1 zeros. Finally,

convert this binary string into an integer in the usual way. In other words,

$$e(P) = 2^{p_1} + 2^{p_1+p_2+1} + 2^{p_1+p_2+p_3+2} + \dots + 2^{p_1+p_2+\dots+p_m+(m-1)}.$$

$e(P)$ is called the *Gödel number of P* . For example, the Gödel number of the example program $P = \langle 2, 3, 0, 1, 6 \rangle$ is 66244 (in decimal). Note that Gödel number of the empty program (which computes the identity function) is 0.

There is an equally effective procedure to convert any natural number to the corresponding Post machine program. For example, given the integer 140 (in decimal), we first write it in binary as 10001100, and then read off the number of consecutive zeros after each 1, starting from the rightmost 1 and moving to the left. This gives us the sequence $\langle 2, 0, 3 \rangle$, which decodes into the program

$$\langle 2, 0, 3 \rangle = \begin{array}{|l|l|} \hline 0 & \text{RIGHT} \\ 1 & \text{FLIP} \\ 2 & \text{IF1GOTO:0} \\ \hline \end{array}.$$

It is another remarkable fact that there is Post machine program U which, given a finite sequence of numbers as input arguments, treats the first argument as the Gödel number of a program P , and then (by decoding the first argument) is able to simulate P on the remaining arguments. Such programs U are called *universal*.

THEOREM 4 Universal Programs and Computable Functions. *There is a (universal) Post machine program U which computes a partial function $\Phi(m, n)$ of two variables with the following property: For every Post machine program P with Gödel number $e = e(P)$, the 1-ary function φ computed by P equals the function Φ_e defined by $\Phi_e(n) = \Phi(e, n)$, i.e. $\varphi = \Phi_e$. In particular, every single-variable partial computable function f equals Φ_e for some e .*

Now put $W_e = \text{dom } \Phi_e$, and recall that a set is called c.e. if it equals the domain of some partial computable function. Since the sequence $\langle \Phi_e : e \in \mathbf{N} \rangle$ contains all partial computable functions, so the sequence $\langle W_e : e \in \mathbf{N} \rangle$ forms a list of all c.e. sets.

We define the special set $\emptyset'$ as $\emptyset' = \{e : e \in W_e\}$. This is an example of a c.e. set whose complement is not c.e. To see this suppose the complement of $\emptyset'$ is a c.e. set so that for some e we have $n \in W_e \iff n \notin \emptyset'$ for all n . Taking $n = e$ we get

$$e \in W_e \iff e \notin \emptyset' \iff e \notin W_e,$$

a contradiction. It follows that $\emptyset'$ is not computable.

The domain of the function Φ is called HALT, since $\Phi(m, n)$ is defined iff the program with Gödel number m halts on input n . HALT is a c.e. set by definition, with

$$n \in \emptyset' \iff (n, n) \in \text{HALT, for all } n.$$

Therefore if HALT were computable, so would be $\emptyset'$, which we have seen to be non-computable. Hence HALT itself is non-computable. This is often expressed by saying that *the Halting Problem is uncomputable*.

We thus have examples of c.e. sets HALT and $\emptyset'$ which are not computable, and so their characteristic functions would be examples of non-computable functions.

4.3 Computation on strings

We have been using numbers for computability notions. We now fix an effective one-to-one correspondence between the natural numbers and the binary strings so that computability notions can be extended to strings.

DEFINITION 5. We fix the following one-to-one correspondence between the natural numbers and the binary strings:

$$0 \leftrightarrow \mathbf{\Lambda}, \quad 1 \leftrightarrow 0, \quad 2 \leftrightarrow 1, \quad 3 \leftrightarrow 00, \quad 4 \leftrightarrow 01, \quad 5 \leftrightarrow 10, \quad 6 \leftrightarrow 11, \quad 7 \leftrightarrow 000, \dots$$

Given a number n , the corresponding binary string is denoted by $\mathbf{str}(n)$. Given a binary string σ , the corresponding number is denoted by $\mathbf{num}(\sigma)$.

This correspondence is effective: An algorithm for converting n to $\mathbf{str}(n)$ is obtained as “write $n + 1$ in binary notation (without leading zero) and then drop (erase) the leading 1”. An algorithm for converting σ to $\mathbf{num}(\sigma)$ is obtained as “prefix σ with an additional 1, regard the resulting string as an integer m written in binary notation, and have the final result to be $m - 1$.”

With this effective correspondence, we can extend every computability notion for numbers into one for strings, *by converting (translating) between the two types back and forth as needed.*

In particular, we will say “the program P halts on input string δ with output string σ ” to really mean “the program P halts on input $\mathbf{num}(\delta)$ with output $\mathbf{num}(\sigma)$ ”.

Another example: We will say that a function $f: \{0, 1\}^* \rightarrow \{0, 1\}^*$ is effectively computable to really mean that the function $g: \mathbf{N} \rightarrow \mathbf{N}$ defined by the “translations” $g = \mathbf{num} \circ f \circ \mathbf{str}$ (i.e., $g(n) = \mathbf{num}(f(\mathbf{str}(n)))$) is effectively computable.

We will say that a set of strings is c.e., if the corresponding set of numbers is c.e.

We could also as easily define computability notions for functions from numbers to strings and vice versa, and in general for subsets of $\mathbf{N}^m \times (\{0, 1\}^*)^n$.

4.4 Effective topological notions

The idea of effective computability has been extended to topological notions and is used extensively in an area called *descriptive set theory*. Many classical notions of analysis get refined this way. Here we describe two such notions, *effective open* and *uniformly effective open*.

Recall that an open set G in the Cantor space is a union of basic intervals, $G = \bigcup_{\sigma \in S} N(\sigma)$ for some set S of strings. We refine this definition by requiring the index set over which the union is taken to be computably enumerable:

DEFINITION 6 Effective Open Sets. A subset G of $\{0, 1\}^{\mathbb{N}}$ is called *effective open* or *computably enumerable* iff there is a c.e. set S of strings such that

$$G = \bigcup_{\sigma \in S} N(\sigma).$$

A c.e. set of strings is one whose members can be listed by a program (or a partial computable function), so a set G is effective open iff there is a program which prints a list of basic intervals whose union equals G . More precisely, G is *effective open* iff there is a partial computable function $\sigma: \mathbb{N} \rightarrow \{0, 1\}^*$ such that

$$G = \bigcup_{n=1}^{\infty} N(\sigma(n)),$$

where we take $N(\sigma(n)) = \emptyset$ if $\sigma(n)$ is not defined.

Finally we define what it means to say that the sets $G_1, G_2, G_3, \dots$ are *uniformly effective open*. To be uniformly effective open, it is not enough that each set G_n in the sequence be individually effective open, but the entire sequence must be listed together in an effective way, i.e., there should be a single program listing a double-sequence of basic intervals whose unions form these sets. More precisely:

DEFINITION 7 Uniformly Effective Open Sets. A sequence of sets $G_1, G_2, \dots$ is *uniformly effective open* if there is a partial computable function $\sigma: \mathbb{N} \times \mathbb{N} \rightarrow \{0, 1\}^*$ such that

$$G_n = \bigcup_{m=1}^{\infty} N(\sigma(m, n)),$$

where we take $N(\sigma(m, n)) = \emptyset$ if $\sigma(m, n)$ is undefined.

5 VON MISES' DEFINITION OF RANDOM SEQUENCE

Undoubtedly, one of the most successful achievements of twentieth century mathematics was the measure-theoretic axiomatization of probability. Proposed by Kolmogorov in 1933, it soon became the essential basis for the study of mathematical probability theory. The generality and elegance of this abstract axiomatic approach — a hallmark of modern formalism mathematics — found wide applicability in almost every situation in probability theory. However, this method with its formalist nature does not directly involve the notion of randomness in any fundamental way, and the rather intuitionistic problem of defining randomness did not arise in its course of development.

Richard von Mises was the first person to focus clearly and deeply on the mathematical essence of randomness in sequences. His pioneering work early in the twentieth century on developing a *frequentist theory of probability* ([Von Mises, 1919; Von Mises, 1981], already shed considerable light on the heart of the matter of

randomness, as Kolmogorov himself has remarked [Li and Vitanyi, 2008, p.50]. For von Mises, random sequences, which he called “collectives”, formed the essential basis of his frequentist theory. Subsequently, his work led other mathematicians to carry out further investigations that have clarified the notion of randomness significantly. The far reaching implications of his work is still deeply influencing current research on randomness. Kolmogorov is known as the father of modern mathematical probability theory, and it is perhaps appropriate to call von Mises the founder of the modern mathematical theory of randomness in infinite sequences.

Here we only briefly outline the central ideas of von Mises concerning random sequences. For more details, see his own works [Von Mises, 1919; Von Mises, 1981] and works of van Lambalgen [1987a; 1987b; 1996; 1990].

5.1 Von Mises’ definition

The condition of the Borel strong law can be stated as saying that the relative frequency of successes in initial parts of the infinite binary sequence under consideration should have a limiting frequency of $1/2$. Recall that this is a condition for *unbiasedness* (subsection 3.2).

The fundamental intuition of von Mises is often summarized as *the invariance of limiting frequency under (admissible) place selections*.

To understand what this means, consider an infinite sequence of turns of a gambling wheel — turn 1, turn 2, and so on — with each turn having two equiprobable outcomes 0 or 1. Let the sequence $x = \langle x_n \rangle$ denote the outcomes of the turns (if the n -th turn of the wheel produces a 0, then $x_n = 0$, else $x_n = 1$).

Suppose that a gambler is observing the outcomes of the turns, and before every turn the gambler decides whether to bet on that turn or not, *perhaps basing the decision on the finite past history of earlier outcomes of the turns*.

Example 1: The gambler chooses to bet on every third turn (turns 3, 6, 9, etc), disregarding earlier history of outcomes altogether (“lucky-third” rule).

Example 2: The gambler chooses to bet after any run of five consecutive 0s.

In any case, bets may not get placed on every turn, but rather on selected turns determined by the gambler’s strategy. This results in a subsequence of turns selected for placing bets, say turns $n_1 < n_2 < \dots < n_k < \dots$, as shown below:

1	2	...	$n_1 - 1$	n_1	$n_1 + 1$	$n_1 + 2$	...	$n_2 - 1$	n_2	...
skip	skip	...	skip	bet	skip	skip	...	skip	bet	...

Using the strategy, the gambler selects turn n_1 for placing his first bet, turn n_2 for the second bet, etc. In the first example, the subsequence of selected turns for betting is $3 < 6 < 9 < \dots$, independent of the outcome of the turns. In the second example, the turns on which the gambler bets will depend on the particular infinite sequence of turn outcomes.

Once the turns for betting are all selected, suppose we restrict to the outcome values at only these selected turns — discarding (erasing) the outcome values of those turns on which bets are not placed — and then compute the limiting

frequency for this new restricted subsequence of outcomes. According to von Mises, if the original sequence of outcomes were random, then this new limiting frequency would still equal $1/2$, regardless of the gambling strategy being used! Moreover, this crucial property is the essence of randomness, and therefore characterizes it:

DEFINITION 8 Von Mises Randomness, Initial Version. An infinite binary sequence $x \in \{0,1\}^{\mathbb{N}}$ is random if whatever be the gambler's strategy and the resulting turns $n_1 < n_2 < \dots$ selected for placing bets, the subsequence of x obtained by restricting to these selected turns still has limiting frequency $1/2$, i.e.:

$$\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{k=1}^m x_{n_k} = \frac{1}{2}.$$

In von Mises' terminology, a "turn" (on which the gambler may or may not bet) is called a *place*, and a "strategy" by which the gambler selects which turns to bet on, is called a *place selection rule*.

The "invariance of limiting frequency under admissible place selections" can now be understood as a form of *unpredictability* arising from *unbiasedness*: *No betting strategy of place selections can succeed by improving predictability within a random sequence, since such selections will leave unbiasedness intact (identical limiting frequency for the resulting subsequence)*. In other words, not only is the entire sequence unbiased (limiting frequency of $1/2$), but there is no hidden biased or unstable subsequence that can be found (by a gambler) using a suitable strategy of place selection.¹⁴

We quote some relevant remarks of Feller:

"The painful experience of many gamblers have taught us the lesson that no system of betting is successful in improving the gambler's chances." [Feller, 1968, VIII.2, p. 198]

... "[U]nder any system the successive bets form a sequence of Bernoulli trials with unchanged probability of success. ... The importance of this statement was first recognized by von Mises, who introduced the impossibility of a successful gambling system as a fundamental axiom." [Feller, 1968, VIII.2, p. 199]

"Taken in conjunction with [the] theorem on impossibility of gambling systems, the law of the large numbers implies the existence of the [limiting frequency] not only for the original sequence of trials but also for all subsequences obtained in accordance with the rules of [place selection]." [Feller, 1968, VIII.4, p. 204]

¹⁴For simplicity we are restricting (by using the von Neumann trick if necessary) only to the special limiting frequency value of $1/2$ instead of the general value p ($0 < p < 1$) used by von Mises. For us, this does not cause much loss of generality as we are focusing only on randomness. Von Mises' chief objective was to develop *the frequentist theory of probability*.

5.2 Mises-Wald-Church randomness

We first formalize the earlier definition of von Mises Randomness.

A *place selection rule* or *betting strategy* is a partial function $\varphi: \{0, 1\}^* \rightarrow \{0, 1\}$. (It tells the gambler when to bet: For a binary sequence $\langle x_1, x_2, \dots, x_n, \dots \rangle$ of outcomes, the n -th turn is selected for betting according to the strategy φ if and only if $\varphi(\langle x_1, x_2, \dots, x_{n-1} \rangle) = 1$.)

Given a place selection rule φ and $x \in \{0, 1\}^{\mathbb{N}}$ such that $\varphi(\langle x_1, x_2, \dots, x_{n-1} \rangle) = 1$ for infinitely many n , let n_1 be the least n for which $\varphi(\langle x_1, x_2, \dots, x_{n-1} \rangle) = 1$, n_2 be the next such n , etc. Then the sequence $\langle x_{n_1}, x_{n_2}, \dots, x_{n_k}, \dots \rangle$ is called *the φ -selected part of x* . (It is the subsequence of x obtained by restricting x to those indexes which are selected for betting according to φ .) Thus we say that *the φ -selected part of x has limiting frequency $1/2$ iff*

$$\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{k=1}^m x_{n_k} = \frac{1}{2}.$$

We can now try to restate the definition of von Mises randomness as follows. A sequence $x \in \{0, 1\}^{\mathbb{N}}$ is random iff for all place selection rules φ for which $\varphi(\langle x_1, x_2, \dots, x_{n-1} \rangle) = 1$ for infinitely many n , the φ -selected part of x has limiting frequency $1/2$.

We run into a problem with this definition with its unrestricted use of the universal quantifier in the clause “for all place selection rules”. Given any sequence $x \in \{0, 1\}^{\mathbb{N}}$, put $A = \{n: x_n = 1\}$ if this set is infinite, otherwise put $A = \{n: x_n = 0\}$. Define φ by the condition that $\varphi(\langle y_1, \dots, y_{n-1} \rangle) = 1$ if $n \in A$ and $= 0$ otherwise. It is easy to see that the φ -selected part of x has a limiting frequency equaling either 0 or 1, violating randomness. It follows that no sequence is random!

However, as von Mises points out, this is not a real problem. The defect of the argument is that the rule φ used in the argument *selects a place n based on the outcome value x_n* , and such rules are of course not allowed [Von Mises, 1981, p. 25]. The place selection rules in the definition of randomness are restricted to only certain *admissible* rules, instead being completely arbitrary, and the problem is resolved.¹⁵

Here is the corrected definition in its original intended form:

DEFINITION 9 Von Mises Randomness. A sequence $x \in \{0, 1\}^{\mathbb{N}}$ is random iff for all *admissible* place selection rules φ for which $\varphi(\langle x_1, x_2, \dots, x_{n-1} \rangle) = 1$ for infinitely many n , the φ -selection of x has limiting frequency $1/2$.

¹⁵From certain sections of von Mises' detailed discussion of the concept [Von Mises, 1981], it is also clear that he wants a place selection rule, or betting strategy, to be “specifiable in some effective manner”. In the modern language of mathematical logic, that could be interpreted as some notion of effective description (say as being effectively computable or at least being definable in some definite language), but von Mises does not precisely specify any such rigorous criterion for defining “admissible”.

Abraham Wald [1936] then showed that whenever the set of admissible place selection rules is countable, random sequences according to the von Mises definition do exist, and form a set of full-measure.

This implies that if “admissible” is taken to mean any form effective specifiability using finite sequence of symbols from a finite alphabet (such as effective computability), then the set of admissible rules remains countable, and so von Mises random sequences would exist.

In 1940, Alonzo Church [1940] proposed the use of *effectively computable* (total) place selection rules for precisely defining von Mises randomness. Such random sequences are now called *Church randomness* or *Church stochastic*.

Using *partial computable* place selection rules, we have the following definition:

DEFINITION 10 Mises-Wald-Church Randomness. A sequence $x \in \{0, 1\}^{\mathbb{N}}$ is *Mises-Wald-Church random* iff for all *partial computable* place selection rules φ for which $\varphi(\langle x_1, x_2, \dots, x_{n-1} \rangle) = 1$ for infinitely many n , the limiting frequency of x under φ -selection is $1/2$.

Thus, in 1940, the first precise and rigorous mathematical definition of randomness for infinite sequences was found.

In recent literature, the term *stochastic* is used for randomness defined using limiting frequency after place selections, and so Mises-Wald-Church random sequences are now also called *Mises-Wald-Church Stochastic*.

As explained in Section 3, any notion of randomness must be subjected to the fundamental stochastic laws such as Borel strong law, Borel normality, Symmetric Oscillation, etc.

It is easy to see that every Mises-Wald-Church random sequence satisfies the condition of the Borel strong law, since the place selection rule φ defined by $\varphi(\sigma) = 1$ for all σ is computable, and the resulting subsequence is simply the entire original sequence.

It can also be shown that Mises-Wald-Church random sequences are Borel normal.

But a big blow to the definition came when in 1939 Ville [1939] proved that there are Mises-Wald-Church random sequences which do not satisfy the Law of Symmetric Oscillations: For certain Mises-Wald-Church random sequences x , the relative frequency satisfies $\frac{1}{n} \mathbf{S}_n[x] > 1/2$ for all n . In terms of random walk, this means the position of the walking person stays always to the right of the origin, a violation of the Law of Symmetric Oscillations.

The definition of Mises-Wald-Church randomness can be viewed as the impossibility of any successful algorithmic betting strategy of place selections. Unfortunately, Ville’s result shows that this condition is *not sufficient* to guarantee the randomness of a sequence (recall from subsection 3.1 that the condition must be necessary for randomness).

The method outlined in subsection 3.8 of capturing the random sequences using “effective stochastic laws converging effectively” has turned out to be more

successful, and we discuss it in Section 6.

If, instead of considering betting strategies of place-selection and the resulting limiting frequency, we consider *capital betting strategies* (martingales), then the corresponding analog of the Mises-Wald-Church definition — namely the impossibility of any successful suitably algorithmic capital betting strategy — turns out to be more well behaved (see Section 11).

6 MARTIN-LÖF AND SOLOVAY RANDOMNESS

6.1 Martin-Löf randomness

Subsection 3.8 outlined the program of defining a sequence x to be random if it satisfies all “effective probabilistic laws of randomness”, where an “effective probabilistic randomness law” is simply an “effective full-measure” set (or, what we called a “typical property” in subsection 1.6). Going to complements, this means that x should not belong to any “effective measure-zero” set (or, in the language of subsection 1.6, that x should not have any “special property”).

The question now, therefore, is how to precisely define “effective measure-zero”. Twenty-five years after the Mises-Wald-Church definition, in 1965, a satisfactory solution to this crucial problem was found by the Swedish mathematician Martin-Löf [Martin-Löf, 1966], which we now describe.¹⁶

Recall that (Section 2) a set E has measure-zero if there is a sequence of open sets $G_1, G_2, G_3, \dots$ with each G_n covering E and $\mu(G_n) < 1/n$. Recall also that such a sequence of sets $G_1, G_2, G_3, \dots$ is uniformly effective open if there is a single program listing the basic intervals whose unions form these sets (subsection 4.4).

Martin-Löf’s fundamental idea was that by simply taking the sequence of covering sets G_n to be uniformly effective open, we get the correct notion of “effective measure-zero”.

A constructive proof of a probabilistic law of randomness (such as the Borel strong law) would usually proceed this way: Given n , one uniformly builds an effective open set G_n of measure less than $1/n$ such that every sequence in the complement of G_n satisfies the law in question, which immediately establishes that the set of sequences satisfying the law has full measure. The strongest probabilistic law of randomness that we have mentioned, the Law of Iterated Logarithms, is known to have such a constructive proof [van Lambalgen, 1987b, p. 733].

We thus have the following definitions of *effective measure-zero* and *effective full-measure* sets:

¹⁶Martin-Löf was visiting Kolmogorov in Moscow during 1964–65 and they were working on randomness and complexity of finite objects (Kolmogorov Complexity). Note also that during the twenty-five year period 1940–65, computability theory (recursion theory) was progressing vigorously and expanding its domain into classical analysis, leading to highly refined development of the notion of effectiveness, including effective open sets and effective Borel sets by Kleene, Addison, Mostowski, and others. See [Moschovakis, 1980] for more details.

DEFINITION 11 Martin-Löf. A set $E \subseteq \{0, 1\}^{\mathbb{N}}$ is *effective measure-zero* iff there is a uniformly effective sequence open sets, say $G_1, G_2, \dots$, such that for all n :

- (a) $E \subseteq G_n$, and
- (b) $\mu(G_n) < 1/n$.¹⁷

A set has *effective full-measure* if its complement is effective measure-zero.

For example, the set of all sequences with bit value 0 at every third position is effective measure-zero.

Another example of an effective measure-zero set is the set of all sequences in which the bit pattern 0110110 does not occur.

One can think of the sets $G_1, G_2, G_3, \dots$ as providing a uniformly effective sequence of statistical tests for randomness with stronger and stronger significance.

Finally, we define *Martin-Löf Randomness*.

DEFINITION 12 Martin-Löf. A sequence $x \in \{0, 1\}^{\mathbb{N}}$ is *Martin-Löf Random* iff x does not belong to any effective measure-zero set, i.e., iff x belongs to every effective full-measure set.

We can think of an *effective probabilistic randomness law* L to be simply an effective full-measure set L , and think of a sequence x to be *satisfying the law* L iff $x \in L$. We can then restate the definition of Martin-Löf randomness as: x is *Martin-Löf Random* iff it satisfies all effective probabilistic randomness laws.

Martin-Löf also established the remarkable fact that the set of all Martin-Löf random sequences itself has effective full-measure, that is, the set U of non-Martin-Löf-random sequences is effective measure-zero. This means that there is a sequence of uniformly effective open sets $U_1, U_2, U_3, \dots$ such that $\mu(U_n) < 1/n$ and $U \subseteq \bigcap_n U_n$. But also $U \supseteq \bigcap_n U_n$ by definition. Hence $U = \bigcap_n U_n$, and thus the sequence $\langle U_n \rangle$ acts as a *universal test* for Martin-Löf randomness: x is *Martin-Löf random* iff $x \notin U_n$ for some n . This universal test condition gives an especially simple characterization of Martin-Löf randomness.

6.2 Solovay's characterization of randomness

The Borel-Cantelli Lemma of probability theory implies that if $G_1, G_2, \dots, G_n, \dots$ is an infinite sequence of events and the sum of their probabilities converges (as an infinite series), then with probability one, only finitely many of these events can occur.

¹⁷Of course, instead of the measure bounds $1/n$ for G_n , one can use any sequence of positive rational numbers ϵ_n so long as ϵ_n can be effectively computed from n and the sequence $\langle \epsilon_n \rangle$ converges to zero. It is not hard to see that if a computable sequence of positive rationals $\langle r_n \rangle$ converges to zero, then given any other computable sequence of positive rationals $\langle s_n \rangle \rightarrow 0$, there is a computable subsequence $\langle r_{n_k} \rangle$ of the original sequence with $r_{n_k} < s_k$ for all k . Therefore, the choice of the bounding sequence is completely arbitrary, so long as it forms a computable sequence of positive rationals converging to zero.

The following remarkable result of Solovay shows that this “Borel-Cantelli condition” characterizes Martin-Löf randomness, provided that we restrict the sequence of open sets to be uniformly effective.

THEOREM 13 Solovay. *An infinite binary sequence is Martin-Löf random iff for every uniformly effective sequence $G_1, G_2, \dots, G_n, \dots$ of open sets,*

$$\sum_{n=1}^{\infty} \mu(G_n) < \infty \implies x \text{ belongs to only finitely many } G_n \text{ s.}$$

See Chaitin [1992] for a proof.

The above condition (in the theorem) for characterizing Martin-Löf randomness is known as *Solovay Randomness*.

Note that in Solovay’s characterization, the infinite series $\sum_n \mu(G_n)$ is simply assumed to converge, there is no need to assume that it converges in any effective way.

6.3 More general probability spaces

Martin-Löf’s and Solovay’s definitions for randomness are so general and flexible that they can be applied to any effective separable complete metric space with an effective probability measure. This includes a very large class of probability spaces, including many (perhaps most) probability spaces arising in practice.

Thus the Martin-Löf definition provides a way of assigning a precise meaning of the word “random” in quite general settings. See [González, 2008; Hoyrup, 2008] for more on this.

7 RANDOMNESS OF FINITE STRINGS: KOLMOGOROV COMPLEXITY

We now turn to the “Laplacian Problem” mentioned in the introduction, namely, that of defining randomness for finite sequences. Laplace’s observation was that among all the sequences of a fixed large size, only a few “regular” ones have a “rule that is easy to grasp” and Laplace attributes this to those sequences having “a regular cause”. The other sequences, “incomparably more numerous”, are irregular and we therefore take them to be the random ones.

If we follow this “Laplacian Program” then our problem reduces to precisely isolating the notion of a sequence having a “regular cause” behind it (or being generated by a “rule that is easy to grasp”). This is precisely the philosophical point missing from classical probability theory, which fails to distinguish the strings which we think of being “regular”.

This problem was resolved quite satisfactorily in the mid 1960s by Solomonoff, Kolmogorov, and Chaitin. Their theory provided a measure for the *information content* or the *complexity* of a binary string (or more generally of a finite object which can be represented by a binary string) by taking it to be the length of the

“shortest possible complete description” of the string, or its *description-complexity*. The idea is based on our intuition that a relatively simple object will have a short complete description, while a highly complex one will lack a short description which can completely specify it.

Moreover, the related notion of *algorithmic probability* invented by Solomonoff assigns a form of a priori universal probability to binary strings. But unlike classical probability, it takes into account the information-complexity of the string when assigning the probability. As a result, its probability assignments sharply discriminates between the regular strings and the random ones, and provides an explanation of Laplace’s intuition into why the regular strings are more likely to have a “cause”.

We have been using the term “description” freely, without much qualification. It is important to precisely specify what is meant by a “description”. Unrestricted use of the term, as done in natural languages, causes problems, as shown by the following.

7.1 The Berry paradox

Consider the definition:

The Berry number is the smallest positive integer that cannot be described in less than eighteen words.

Since only a finite number of positive integers can be described using less than eighteen words, the Berry number is well-defined, and by definition it cannot be described in less than eighteen words. Yet the above definition describes it using only seventeen words. This is the Berry paradox.

The problem here is with the use of the property of “a number being *described* in certain words”, which is not precisely defined, and cannot be defined, as used in the definition of the Berry number, without being circular.

Whatever be its resolution, the Berry paradox reminds us that we have to be careful when talking about the “description” of a number or a string. Once again the concept of algorithm or effective computation allows us to make this precise: We restrict only to “algorithmic descriptions” or “effective descriptions” as defined below. Formally, by an algorithm or program we mean a Post machine program.

DEFINITION 14 Algorithmic Description. Let P be a program and σ be a binary string. We say that a string δ is a P -description of σ , or that δ P -describes σ , if the program P on the input δ halts with output σ .

The idea here is that whenever the program P halts on the input δ with output σ , we think of the string δ as being an *algorithmic description* of the string σ , according to algorithm P . The computation of σ from input δ by P is thought of as P reconstructing σ from its description δ . This is equivalently called the *decompression* of δ by P (into σ). The description δ is intended to be shorter

than the object being described (the string σ), and therefore it can be viewed as a “compressed” version of the string σ .

Given a program P and a string σ , we now look for the *shortest* string(s) P -describing σ , and take the length of such string(s) as a complexity-measure of the string σ with respect to P . Of course for certain programs P , a string σ may not have any P -description, in which case the complexity of σ (with respect to P) is considered to be infinite.

DEFINITION 15 Algorithmic P -Complexity. The *plain algorithmic complexity* of a string σ with respect to the program P , or simply the *P -complexity* of σ , denoted by $C_P(\sigma)$, is the length of the shortest string(s) P -describing σ , provided that there are such strings. If there is no string which P -describes σ , we let $C_P(\sigma) = \infty$.

There are programs P such that the P -complexity $C_P(\sigma)$ is a finite number for all strings σ (i.e., P has the property that $\forall \sigma \exists \delta (\delta \text{ } P\text{-describes } \sigma)$). Let us call such programs P to be *complexity-finite*.

An example of a complexity-finite program is the empty program E (the program with no instructions), which computes the identity function: Every string E -describes itself.

If $C_P(\sigma)$ is much smaller than the length of σ we think of σ being “well-compressed” by P , since there are P -descriptions of σ much shorter than σ . The ratio $|\sigma|/C_P(\sigma)$ is the “compression factor” for the string σ , with respect to P . For the empty program E , $C_E(\sigma) = |\sigma|$ for all σ , and so the compression factor is 1 for all strings, meaning no string is really “compressed” by E .

On the other hand, there are complexity-finite programs P which compress infinitely many strings by arbitrarily large factors.

Example. Let P be the program informally described as follows. If the first symbol of the input string σ is 0, then P erases this leading 0, shortening its length by 1, outputs the resulting string, and halts; else P outputs the string consisting of $2^{|\sigma|}$ 1s and halts. Then $C_P(\sigma) \leq |\sigma| + 1$ for all σ , since 0σ is a P -description of σ , so P is complexity-finite. But for any n , if we put $\delta_n = 1^n$ and $\sigma_n = 1^m$ where $m = 2^n$, then δ_n is a P -description of σ_n , so $C_P(\sigma_n) \leq |\delta_n| = n$, so the compression factor for σ_n is $|\sigma_n|/C_P(\sigma_n) \geq 2^n/n$. Given any number a , we can find n such that $2^n/n > a$, and so the strings $\sigma_n, \sigma_{n+1}, \dots$ are all compressed by a factor more than a . In particular, the string σ_{10} (string of 1024 1s) is compressed by P to the string 111111111, thus by a factor more than 100. More drastically, the string σ_{64} is compressed by a factor of 288230376151711744.

However, a simple but important counting argument (an example of the so-called pigeon-hole principle) shows that no method can compress strings too uniformly.

Given an arbitrary partial function $f: \{0,1\}^* \rightarrow \{0,1\}^*$, we think of f being a general method of string-description, and think of δ being an f -description of σ if $f(\delta) = \sigma$. (The P -descriptions given by programs P are special “algorithmic” case of this.) We say that a string σ is compressed by f if there is an f -description of σ which is shorter than σ . More generally, given a positive integer b , we say that

σ is b -compressed by f if there is an f -description of σ which is shorter than σ by at least b bits (i.e., if $\exists \delta(f(\delta) = \sigma \wedge |\delta| \leq |\sigma| - b)$). Thus σ is compressed by f iff σ is 1-compressed, that is by at least 1 bit.

THEOREM 16 “Only a small minority of strings compress”. *For any string-description method f , less than half of all the strings of length $\leq k$ can be compressed (i.e. 1-compressed) by f . More generally, less than a fraction of $1/2^b$ of all the strings of length $\leq k$ can be b -compressed by f .*

Proof. For any k and b , let A be the set of strings of length $\leq k$, B be the set of strings of length $\leq k - b$, and C be the subset of A consisting of those members which can be b -compressed by f . Note that there is a one-to-one correspondence between C and a subset of B (for each $\sigma \in C$, fix $\delta_\sigma \in B$ with $f(\delta_\sigma) = \sigma$; then the correspondence $\sigma \leftrightarrow \delta_\sigma$ is one-to-one). Also note that $|A|$ (number of members of A) equals $2^{k+1} - 1$ and $|B| = 2^{k-b+1} - 1$, hence the fraction of those strings in A which are b -compressed equals

$$\frac{|C|}{|A|} \leq \frac{|B|}{|A|} = \frac{2^{k-b+1} - 1}{2^{k+1} - 1} < \frac{2^{k-b+1}}{2^{k+1}} = \frac{1}{2^b}.$$

■

For example, among the strings of length not exceeding a thousand (or any other number) bits, more than 99.9% will either not compress at all or compress by at most 9 bits, whatever be the method of string description.

We are of course interested in *shortest possible descriptions* or *best possible compression factor*. Therefore, among the complexity-finite programs we prefer the ones which tend to give *overall* shorter descriptions (better compression) for strings. In other words, given complexity-finite programs P and Q , we regard P to be “better” than Q if the P -complexity of σ , $C_P(\sigma)$ is lower than the Q -complexity $C_Q(\sigma)$ for all strings σ . We can then try to choose a “best” complexity-measure and use it as a standard.

Unfortunately, it is impossible to get such a “best” complexity-finite program in the uniformly strict sense above, because for every complexity-finite program one can find another one which lowers complexity (compresses better) for an arbitrarily large number of strings by an arbitrarily large amount.¹⁸

Therefore, we will compare programs using a “general overall sense” rather than the uniformly strict sense above, relaxing the relation of one program being better

¹⁸More precisely, for every complexity-finite P and every m and n , there is another complexity-finite program Q such that $C_Q(\sigma) < C_P(\sigma) - n$ for at least m strings σ . Proof: Without loss of generality assume that $m = 2^k - 1$ for some k , and fix distinct $\sigma_1, \dots, \sigma_m$ such that $C_P(\sigma_j) > n + k$ for $j = 1, \dots, m$. This can be done since there are only finitely many strings σ with $C_P(\sigma) \leq n + k$. Also let $\alpha_1, \dots, \alpha_m$ be a non-repeating listing of all strings of length less than k . Now a program P' can be so designed that it contains coded copies of $\sigma_1, \dots, \sigma_m$ inside it, and behaves in the following way: If the input is α_j ($1 \leq j \leq m$), then output the string σ_j and halt; otherwise, truncate the input string by removing its first k symbols and emulate the program P with this truncated input.

than another as follows. For programs P and Q , let us define:

$$P \text{ matches } Q \text{ (in terms of complexity)} \iff \exists k \forall \sigma (C_P(\sigma) \leq C_Q(\sigma) + k),$$

i.e., from a complexity viewpoint, the program P is regarded to be as good as the program Q or better (“ P matches Q ”) if the P -complexity of every string is less than its Q -complexity *modulo some constant independent of the string*. Let us also call programs P and Q to be *complexity-equivalent* if each one matches the other, i.e., if the difference between P -complexity and Q -complexity is bounded uniformly by some constant.

It now turns out that in this sense, there is indeed a “best compressing” or *optimal* program, which is also unique in the sense that any other optimal program is complexity-equivalent to it.

DEFINITION 17 Optimal or Universal Programs. A program is called *universal*, or *complexity-optimal*, or simply *optimal* if it matches every program.

THEOREM 18 Solomonoff-Kolmogorov-Chaitin Invariance Theorem. *There is an optimal program U . Moreover, every optimal program is equivalent to U .*

Proof. A program U is defined as follows. Given an input string σ , the program U finds the length m of the longest prefix of σ consisting of 0s ($m = 0$ if σ begins with 1 or is empty), and erases this initial run of 0s in σ . If the remaining string begins with a 1, that symbol is also erased, with the final result being the string ρ . U then runs the program with Gödel number m on the input ρ .

To show that U matches P for any P , let $e = e(P)$ be the Gödel number of P and let $k_P = e + 1$. For any string σ , let δ be a shortest P -description of σ , so that $C_P(\sigma) = |\delta|$. Put $\tau = 0^e 1 \delta$, then τ is a U -description of σ , so

$$C_U(\sigma) \leq |\tau| \leq e + 1 + |\delta| = C_P(\sigma) + k_P,$$

where $k_P = e + 1$ is independent of σ . ■

We now fix a particular optimal U and define the *plain algorithmic complexity* of a string σ , $C(\sigma)$, to be $C_U(\sigma)$.

DEFINITION 19 Universal Plain Algorithmic Complexity C . Define the plain algorithmic complexity function C by $C(\sigma) = C_U(\sigma)$, where U is an optimal program fixed permanently.

Being optimal, U matches E , where E is the empty program with E -complexity $C_E(\sigma) = |\sigma|$. So:

COROLLARY 20. *There is k such that for all σ , $C(\sigma) \leq |\sigma| + k$.*

The plain algorithmic complexity $C(\sigma)$ of a string σ is also known as *Kolmogorov complexity*. The term “Kolmogorov complexity” is used in a wide and general sense as synonym for algorithmic complexity, and so *prefix-free complexity* (described next) is also known as Kolmogorov complexity.

The invariance theorem is quite remarkable as it shows that the concept of plain algorithmic complexity is essentially unique and independent of the particular model of computation being used.

The plain algorithmic complexity $C(\sigma)$ of a string σ , can usefully be viewed as a measure of the (algorithmic) information content of the string σ . We therefore have a formal definition for the somewhat vague notion of *information contained in a finite object*.¹⁹

DEFINITION 21 Randomness and Compressibility for Finite Strings. Let σ be a string and b be a positive integer. We then define:

- (a) σ is *b-compressible* if $C(\sigma) \leq |\sigma| - b$;
- (b) σ is *compressible* if it is 1-compressible, i.e., if $C(\sigma) < |\sigma|$; and
- (c) σ is *random* if σ is not compressible, i.e., if $C(\sigma) \geq |\sigma|$.

THEOREM 22 Existence of Random Strings. *For every n there are random strings of length n . More generally, for any n and $b > 0$ at least $2^n - 2^{n-b+1} + 1$ strings of length n are b -incompressible.*

Proof. Fix n and for each σ with $|\sigma| = n$, pick a string δ_σ of smallest length describing σ . The strings δ_σ are distinct for distinct σ , and there are less than 2^n strings of length less than n , hence by the pigeon-hole principle $|\delta_\sigma| \geq n$ for some σ of length n , so that $C(\sigma) \geq n$, and so σ is a random string of length n .

The second statement is proved similarly. ■

For finite strings, note that we really have relative degrees of randomness. If we have two thousand-bit strings one of which is not compressible and the other, say, 2-compressible but not 3-compressible, then the first one is more random than the second one, but only slightly more so. The complexity measure $C(\sigma)$ therefore provides a measure for the degree of randomness in σ : The smaller the value of $C(\sigma)$ compared to $|\sigma|$, the less random it is. Among all binary strings of a fixed length, the most random are the ones on which the function C achieves its maximum value, and the most non-random ones are the ones on which C is minimized.

Also, short finite strings (such as 111) can be random yet quite simple. This is not surprising when we regard the strings as being produced by a fixed number of flips of a fair coin: If a series of three flips produces 111, there is no reason to suspect the randomness of the process and so it is easy to also accept the outcome 111 as random. However, a long string of all 1s (say a million bits, all

¹⁹Information content as defined by *algorithmic* complexity should be contrasted with the one known as Shannon entropy in “classical” information theory, where it is defined in *probabilistic* terms for random variables. While Shannon’s theory of information focuses on an entire set of strings associated with varying probabilities, the Kolmogorov theory focuses on an individual string. However, the two notions are closely related, see [Li and Vitanyi, 2008, p.603–608].

1s) is dramatically non-random and will cause us to question the randomness of the process generating it.

THEOREM 23. *The complexity function C is not computable.*

Proof. (The proof resembles the argument of the Berry paradox.)

If C were computable then one can define a program P which given any string σ as input computes a string σ^* such that $C(\sigma^*) > 2|\sigma|$.

By the invariance theorem, there is a constant k such that $C(\sigma) < C_P(\sigma) + k$ for all σ .

Let $\sigma = 1^{k+1}$. Then $C(\sigma^*) > 2|\sigma|$, but since σ is a P -description of σ^* so $C(\sigma^*) \leq C_P(\sigma^*) + k \leq |\sigma| + k < 2|\sigma|$, a contradiction. ■

The plain complexity function C also satisfies much of the intuitive concepts relating to “information content of a finite object.” For example, if σ is a long string with significant information content, the information content will not double in the string $\sigma\sigma$, because of redundancy of information: If σ can be described, so can $\sigma\sigma$, with little additional verbiage. To put it differently, every program can be modified by adding only a few lines where the final output is duplicated by “post-processing”. We therefore have:

THEOREM 24. *For some k , $C(\sigma\sigma) \leq C(\sigma) + k$ for all σ .*

Random finite strings also satisfy a very general “stochastic” property. Let R be a property of binary strings, that is $R \subseteq \{0,1\}^*$. We say that a binary string σ *satisfies the property R* if $\sigma \in R$, and we say that *almost all strings satisfy the property R* if the fraction of the strings of length n which satisfies R , $|\{\sigma \in R : |\sigma| = n\}|/2^n$, approaches 1 as n approaches ∞ . For a proof of the following result, see [Sipser, 1997, p.219].

THEOREM 25 General Stochasticity for Finite Random Strings. *If almost all strings satisfy a computable property R , then all except a finite number of random strings satisfy R . The result also holds for b -incompressible strings (for any $b > 0$).*

More properties of general algorithmic complexity will be stated in the next section using a variant of the plain complexity function C that we just described. The new complexity function K will be obtained essentially by restricting the class of strings that are allowed to be algorithmic descriptions: Descriptions must now have a particular form of “unique readability” under some rule.

In either case, we see that by using the concept of algorithm to formalize the fundamental idea of *incompressibility* or *lack of shorter descriptions*, we arrive at a precise and invariant definition of randomness for finite strings that remarkably captures our intuition as described by Laplace. See subsection 8.2 for more discussion on this topic.

We also note that while we used Post-Turing computability (Post Machine programs) as the model of computation, any other model of computation could be used satisfactorily. A particularly recent new approach is to use *Binary Lambda Calculus*, due to John Tromp [2009], to study Kolmogorov Complexity.

The literature of modern research in the subject of Kolmogorov Complexity is vast, see [Li and Vitanyi, 2008].

We end this section with a slight digression by giving a specific example of an application of algorithmic complexity.

7.2 An application to Gödel incompleteness

While most strings are random, only a finite number of them can be proved to be so. For each string σ , let n_σ denote its complexity, i.e., $n_\sigma = C(\sigma)$. In any formalization of mathematics (say ZFC), the relation “the complexity of σ is n ” can be formally expressed. Using a variant of the argument of the Berry paradox and the fact that the set of theorems is computably enumerable, we now show that while the sentences

“the complexity of σ is n_σ ”, $\sigma = \mathbf{A}, 0, 1, 00, 01, 10, 11, 000, 001, \dots$

are all true, only finitely many of them can be proved in the theory. We therefore have an “information theoretic version” of Gödel’s Incompleteness Theorem. It also follows that there is a “maximum provable complexity”: For some m , no string can be proved to have complexity more than m . [Boolos and Jeffrey, 1989; Davis, 1980]

THEOREM 26 Gödel’s Incompleteness Theorem, Information-Complexity Version. *Only finitely many of the sentences of the form “ $C(\sigma) = n$ ”, where σ ranges over binary strings and n over natural numbers, are theorems of mathematics.*

Proof. Let $\mathbf{C}(x, y)$ be the formula expressing “the complexity of x is y ” in the formal theory,²⁰ and also for each string σ and natural number n , let $\ulcorner \sigma \urcorner$ and $\ulcorner n \urcorner$ be their formal names in the theory. Since the set of theorems can be computably enumerated, there is a program (Post machine) P which, on input string δ , searches through all the theorems to check if any of them is of the form $\mathbf{C}(\ulcorner \sigma \urcorner, \ulcorner n \urcorner)$ with $n > 2|\delta|$, and if one such theorem is found, then outputs the string σ and halts. Using the invariance theorem fix k such that $C(\sigma) \leq C_P(\sigma) + k$ for all σ .

Now run the program P on the input string $\delta_k = 1^k$. Since there are infinitely many n for which $\mathbf{C}(\ulcorner \sigma \urcorner, \ulcorner n \urcorner)$ is a theorem, so P eventually finds such a one with $n = n_0 > 2k$ and halts with an output string $\sigma = \sigma_0$. Since $\mathbf{C}(\ulcorner \sigma_0 \urcorner, \ulcorner n_0 \urcorner)$ is true, so $C(\sigma_0) = n_0$. But $C(\sigma_0) \leq C_P(\sigma_0) + k \leq |\delta_k| + k = 2k < n_0$, a contradiction. ■

Interestingly, while only finitely many of the true statements “the complexity of σ is n_σ ” can be proved (where, again, n_σ denotes the unique natural number

²⁰To do this, recall our fixed optimal program U , and let a relation H_U be defined as $H_U(x, n, y) \iff U$ halts with output x in less than n program execution steps on some input string of length not exceeding y . Since H_U is computable, there is a formula $\psi(x, n, y)$ such that $\psi(x, n, y)$ is provable if $H_U(x, n, y)$ is true, else its negation $\neg\psi(x, n, y)$ is provable. Now let $\mathbf{C}(x, y)$ be $\exists n(\psi(x, n, y) \wedge \neg\exists z < y(\exists n(\psi(x, n, z))))$.

equal to the complexity of σ , i.e. $n_\sigma = C(\sigma)$, it is also easy to see that the true statement “the complexity of σ is $\leq n_\sigma$ ” can be proved for every string σ . In other words, while we cannot prove that the complexity of σ *equals* n_σ (except for finitely many strings σ), we can prove, for every string σ , that the complexity of σ *does not exceed* n_σ (its true value), without being able to recognize that n_σ is indeed the true value for the complexity of σ .

For a critical discussion of the information-complexity version of Gödel incompleteness, see [van Lambalgen, 1989].

8 THE PREFIX-FREE COMPLEXITY K

While the plain complexity measure $C(\sigma)$ yields a quite satisfactory theory of randomness for finite strings, the function C has some defects. One such defect is about how it relates to the randomness of infinite sequences. If we look at the initial segments $\sigma_n = \langle x_1, x_2, \dots, x_n \rangle$ of an infinite sequence x , the complexity of the n -th initial segment $C(\sigma_n)$ drops by an undesirable amount for infinitely many n , a phenomenon known as *complexity oscillation* [Li and Vitanyi, 2008, p.143].

Also, for the needs of Solomonoff’s theory of Algorithmic Probability, using the literal value of plain complexity measure was not the correct formulation.

Several ideas were developed for dealing variously with such problems, such as monotone complexity, process complexity, decision complexity, and uniform complexity, but it is *prefix-free complexity*, also called *prefix complexity* in short, due to Levin [Kolmogorov, 1974], Gács [1974] and Chaitin (see [Chaitin, 1992] for references), that has now become the standard for algorithmic (Kolmogorov) complexity, and is denoted by K . It is very much like C , but with a more restricted definition of “description”, where only certain classes of strings having a form of unique readability are allowed to be descriptions.

Suppose that we want to pass a string σ directly to a Post machine program as input by placing the string on the tape (all other bits being zero) and starting the program with its head at the beginning bit of the string. Unfortunately, since the tape consists only of 0s and 1s and no special termination markers, there is no general way for any program to determine the end of the string. For example if the head is started on a single 1 with all other tape bits being 0, how can the machine know be sure that there is not another 1 after a trillion bits? Even simpler, how can the machine know if this is supposed to represent the string 1, or 10, or 10000? We express this problem by saying that the plain representation of a binary string is *not properly delimited*. We circumvented this problem by converting a string σ first to its number code $\mathbf{num}(\sigma) = n$ (say), and then passing the unary coded version of n , namely $1^{n+1}0$, to the machine. This coding is an example of a *prefix-free* or *self-delimiting* code, where no code string is a proper initial segment of another, and the input to the machine is uniquely determined. More generally, a set S of strings is said to be *prefix-free* if no string in S is a prefix of another member of S . The set of unary codes, $1^{n+1}0$, $n = 0, 1, 2, \dots$, do form a prefix-free

set, but is exponentially more inefficient than passing the plain binary string. The following example describes a more efficient prefix-free coding.

Example 1 (The 1-code). Consider the scheme where each binary string σ is coded by $1^{|\sigma|}0\sigma$. The string $1^{|\sigma|}0\sigma$ will be called the *1-code* of σ . For example, the 14-bit string $\tau = 00011101011100$ is the following 1-code:

$$\overbrace{1111111111111110}^{1^{|\tau|}0} \overbrace{00011101011100}^{\tau}$$

Prefix-free 1-coding of τ

The set of 1-codes form a prefix-free set. In fact, by placing the head at the beginning of the 1-code of σ , it can be uniquely decoded back to σ (by first decoding from the initial unary part $1^{|\sigma|}0$). It will take $2|\sigma| + 1$ bits to encode the plain binary string σ into its 1-code (e.g., in the above displayed example $|\tau| = 14$, so the 1-coded string has length 29).

Example 2 (The 2-code). An even more efficient prefix-free coding is obtained by the following scheme. Given a binary string σ , first express the *length* $|\sigma|$ of the string in plain binary notation $\mathbf{bin}(|\sigma|)$ and then prefix σ with the 1-code of $\mathbf{bin}(|\sigma|)$ to get what we will call *the 2-code* of σ . For example, the string $\tau = 00011101011100$ of the previous example has length 14, which in binary notation is 1110. Since 1110 has length 4, its 1-code is 111101110, and we prefix this to τ to get the 2-code of τ :

$$\overbrace{111110}^{1^{|\mathbf{bin}(|\tau|)|}0} \overbrace{1110}^{\mathbf{bin}(|\tau|)} \overbrace{00011101011100}^{\tau}$$

Prefix-free 2-coding of τ

The 2-code gives a prefix-free encoding which will encode the string σ using $|\sigma| + 2\log_2 |\sigma| + 1$ bits. For example, the 2-code of τ shown above consists of 23 bits, a saving of 6 bits over its 1-code.

Further improvements can be made by iterating this method.

Note that all these encodings are effective, i.e., there are simple algorithms for decoding and encoding strings according to any of these schemes. Moreover, the above examples are *length-monotonic*, meaning that longer strings have longer codes.

A real world example of a prefix-free set (over the alphabet of decimal digits) is the set of country dialing codes in the international telephone system.

We will now define prefix-free complexity, which is quite similar to plain complexity and is defined as the length of the “shortest description”. *The main difference is that while for plain complexity any binary string could possibly count as a “description”, for prefix-free complexity only “prefix-free strings” (under an effective prefix-free encoding) are allowed to be descriptions.*

DEFINITION 27 Prefix-free Complexity Functions. A partial function $\psi: \{0, 1\}^* \rightarrow \{0, 1\}^*$ mapping strings to strings is *prefix-free* if its domain is prefix-free. Given

a partial computable prefix-free function ψ , the *prefix-free ψ -complexity function* K_ψ is defined by letting $K_\psi(\sigma)$ to be the length of the shortest string(s) δ for which $\psi(\delta) = \sigma$, and putting $K_\psi(\sigma) = \infty$ if no such strings exist.

An example of a prefix-free complexity function ψ_2 is obtained by coupling the decoding function for the 2-codes with the optimal program U for plain complexity C as follows.

Let D_2 denote the set of all 2-coded strings, so that D_2 is a prefix-free set. The decoding function $d_2: D_2 \rightarrow \{0,1\}^*$ for decoding 2-codes establishes a one-to-one correspondence between D_2 and $\{0,1\}^*$, but we regard it as a partial function from $\{0,1\}^*$ to $\{0,1\}^*$. If a string δ is not in D_2 , then $d_2(\delta)$ is not defined, so that the domain of d_2 is D_2 .

Now define ψ_2 to be the function which, given an input string δ , regards δ as 2-code for some string, decodes it into $d_2(\delta)$ and sends this decoded string to the program U as input. U , in turn, runs with $d_2(\delta)$ as input, and may halt with an output string σ , in which case we put $\psi_2(\delta) = \sigma$. If δ is not in D_2 so that $d_2(\delta)$ is not defined, or if U does not halt on input $d_2(\delta)$, then we leave $\psi_2(\delta)$ as undefined.

Clearly ψ_2 is partial computable and its domain is a subset of D_2 , hence it is a partial computable prefix-free function. Let the corresponding prefix-free complexity function K_{ψ_2} be denoted simply by K_2 .

We thus have an example of a prefix-free complexity function K_2 .

How does K_2 compare with the plain complexity function C ? Given a string σ with plain complexity $n = C(\sigma)$, let τ be a string of length n for which U outputs σ on input τ . Let δ be the 2-code for τ , so that $d_2(\delta) = \tau$, and thus $|\delta| = |\tau| + 2 \log_2 |\tau| + 1 = n + 2 \log_2 n + 1$. Now by definition, $\psi_2(\delta) = \sigma$. Moreover, since encoding-decoding using the 2-code is length monotonic, so there is no string δ' shorter than δ with $\psi_2(\delta') = \sigma$, hence:

$$K_2(\sigma) = |\delta| = n + 2 \log_2 n + 1 = C(\sigma) + 2 \log_2 C(\sigma) + 1,$$

which shows that $K_2(\sigma)$ exceeds $C(\sigma)$ by $2 \log_2 C(\sigma) + 1$.

Of course K_2 is not “optimal” and there are “better” prefix-free partial computable functions ψ giving lower lower complexity values. In order to get a “universal optimal” function, we need the following fundamental result.

THEOREM 28 Invariance Theorem for Prefix-free Complexity. *There is an partial computable optimal prefix-free function ξ . That is, for each partial computable prefix-free function ψ there is a constant k satisfying:*

$$K_\xi(\sigma) \leq K_\psi(\sigma) + k, \quad \text{for all } \sigma.$$

DEFINITION 29 Universal Prefix-free Complexity. The *Prefix-free Algorithmic Complexity* $K(\sigma)$ is defined by taking $K(\sigma) = K_\xi(\sigma)$, where ξ is a permanently fixed optimal partial computable prefix-free function.

How does the prefix-free complexity value $K(\sigma)$ compare with the plain complexity value $C(\sigma)$?

Roughly speaking, since fewer strings are allowed to be descriptions of σ under prefix-free complexity K , we may expect a somewhat higher value for the prefix-free complexity $K(\sigma)$ than $C(\sigma)$ (modulo a constant). This indeed turns out to be true.

On the other hand, we can use 2-codes to convert any plain description δ into a prefix-free description of length $|\delta| + 2 \log_2 |\delta| + 1$. More precisely, for the specific prefix-free complexity function K_2 in the example above, we saw that

$$K_2(\sigma) = C(\sigma) + 2 \log_2 C(\sigma) + 1.$$

For the optimal prefix-free complexity K , we would get an overall lower value for $K(\sigma)$ than $K_2(\sigma)$ (modulo a constant), and so $C(\sigma) + \log_2 C(\sigma) + 1$ is only an upper bound for $K(\sigma)$. The following result gives a standard upper bound for $K(\sigma)$.

THEOREM 30. *Modulo additive constants,*

$$C(\sigma) \leq K(\sigma) \leq C(\sigma) + 2 \log_2 |\sigma|.$$

Proof. The first inequality follows from the invariance theorem for C (optimality of C) since the partial computable function ξ used to define K is just one specific “program”.

To prove the second inequality, note that we had earlier already established that $K_2(\sigma) \leq C(\sigma) + 2 \log_2 C(\sigma)$ and $C(\sigma) \leq |\sigma|$, modulo constants. Combining the two, the result follows. ■

Of course, using more efficient prefix-free encodings, this result can be further sharpened.

K has many nice properties which are lacking in C .

From now on, we will be using the prefix-free complexity function K in place of C . In particular, the definitions for compressibility and randomness for finite strings are redefined as follows.

DEFINITION 31 Randomness and Compressibility for Finite Strings. Let σ be a string and b be a positive integer. We then define:

- (a) σ is *b-compressible* if $K(\sigma) < |\sigma| - b$; otherwise it is *b-incompressible*;
- (b) σ is *compressible* if it is 1-compressible, i.e., if $K(\sigma) < |\sigma|$; otherwise it is called *incompressible*; and
- (c) σ is *random* if σ is incompressible, i.e., if $K(\sigma) \geq |\sigma|$.

Since K assigns a higher complexity value to strings than C (uniformly modulo a constant), strings compress a “little less” under K than under C , and so random strings now become “more numerous”. In particular, the existence theorem for random strings (which holds for any complexity measure) remains valid.

THEOREM 32 Existence of Random Strings. *For every n there are random strings of length n . More generally, for any n and $b > 0$ at least $2^n - 2^{n-b+1} + 1$ strings of length n are b -incompressible.*

8.1 Properties of finite random strings

We mention two more “stochastic” properties of finite random strings which conform to our intuition as evidence that incompressibility is the correct definition of randomness for finite strings.

THEOREM 33. *Long random strings have a “balanced” number of 0s and 1s. More precisely, for any $\epsilon > 0$ there is k such that for all random strings σ of length $n > k$, we have*

$$\left| \frac{S_n[\sigma]}{n} - \frac{1}{2} \right| < \epsilon,$$

where $S_n[\sigma]$ denotes the number of 1s in σ .

THEOREM 34. *Any run of zeros or ones in a random string n is asymptotically bounded above by $O(\log n)$. That is there is k and a constant a such that for every random string σ of length $n > k$, the longest run of 0s in σ is less than $a \log n$.*

8.2 Kolmogorov Complexity as vindication of Laplace

We end our discussion of randomness for finite strings with the position that Kolmogorov Complexity provides a satisfactory solution to Problem 2 of the Introduction. As outlined and evidenced above, the *incompressibility* definition of randomness for finite strings conforms quite well to our intuition. In fact, an important test of randomness for finite strings now is to apply standard computational compression programs to the the string in question and check if it compresses or not.

Kolmogorov Complexity also provides a strong vindication of all of Laplace’s intuitions, by classifying strings according to their complexity: The lower the K value, the more “regular” the strings, and the higher the K value the more “irregular” (or random) they are.

Cause and Regularity Laplace mentions that we perceive a “cause” in strings which are “regular”, “those in which we observe a rule that is easy to grasp.” Kolmogorov complexity provides a precise and objective way to define this idea, using *short effective descriptions*. Given a string σ with small complexity value $K(\sigma)$, its “cause” (or a “rule that is easy to grasp”) is any of its minimal descriptions, i.e. a minimal length strings (of length $K(\sigma)$) which describes σ via the optimal algorithm. If $K(\sigma)$ is not small compared to $|\sigma|$, we regard the string σ as “irregular” or random. This interpretation of cause and regularity is obtained by classifying strings by their complexity, i.e., by measuring how short an effective description is possible. The invariance of

Kolmogorov complexity under all possible methods of effective description (for sufficiently large strings) shows that this is not an arbitrary measure of complexity, but is essentially an objective one.

Rarity of Regularity Laplace mentions that the “irregular sequence ... are incomparably more numerous” compared to the “regular” ones. This is again confirmed by that fact that if cause or regularity is defined as *descriptions which are sufficiently short*, then the irregular sequences automatically become “incomparably more numerous”. We saw this in the theorem which showed that only a small minority of strings compress well.

Probability of Regular Strings Laplace explains that while the regular strings are much less numerous, if we observe a highly regular but long string, “we seek a cause whenever we perceive symmetry”, and it is “*more probable*” that “this event ought to be the effect of a regular cause” than “that of chance”. For example, let σ be the thousand bit long string containing the pattern 01010101... throughout the entire string. If we observe σ , our intuition tells us that in a sense not explained by classical probability theory, it is remarkably different from another random string generated by a thousand coin flips. Classical probability, being information-blind, will assign the same probability $1/2^{|\sigma|} = 1/2^{1000}$ to σ , as well as to all other thousand bit string. But the algorithmic probability (using Kolmogorov complexity) of σ is $1/2^{K(\sigma)}$. If in addition to random coin flips we also consider “effective causes”, then algorithmic probability remarkably explains Laplace’s intuition. As the string σ is regular (has a short description), $K(\sigma)$ will be quite smaller than $|\sigma| = 1000$, and so its algorithmic probability $1/2^{K(\sigma)}$ will be much higher compared to another random string of the same length, since the algorithmic probability of a random string will be, by definition of randomness for finite strings, at most $1/2^{|\sigma|} = 1/2^{1000}$. In other words, the probability that σ was generated by an “effective cause” will be higher than the probability that it was generated randomly, by a factor of at least $2^{|\sigma| - K(\sigma)} = 2^{1000 - K(\sigma)}$. With the conservative estimate of $K(\sigma) = 950$, this factor is 1125899906842624.

At last, the Laplace Program is realized.

9 KOLMOGOROV-CHAITIN RANDOMNESS AND SCHNORR’S THEOREM

We now return back to randomness for infinite sequences.

We will say that an infinite binary sequence x is *b-incompressible* if every initial segment $\langle x_1, x_2, \dots, x_n \rangle$ of x is *b-incompressible* as a finite string, i.e. if

$$K(\langle x_1, x_2, \dots, x_n \rangle) > n - b \quad \text{for all } n.$$

We say that the infinite binary sequence x is *incompressible* if it is b -incompressible for some b , i.e., if *no initial segment of x can be compressed by more than a fixed number of bits*.

This property of incompressibility of an infinite binary sequence x can be regarded as an information-complexity definition of randomness for x .

DEFINITION 35 Kolmogorov-Chaitin Randomness. An infinite binary sequence x is *Kolmogorov-Chaitin Random* if x is incompressible (no initial segment of x can be compressed by more than a fixed number of bits), or in other words, if

$$\text{For some } b: \quad K(\langle x_1, x_2, \dots, x_n \rangle) > n - b \quad \text{for all } n.$$

Remark: The term “Kolmogorov-Chaitin random” is not in standard use. In the literature it is known variously as “Chaitin random”, “Levin-Chaitin random”, “Levin-Chaitin-Schnorr random”, K -incompressible, etc.

The definition of Kolmogorov-Chaitin randomness appears to be significantly different when compared to the definition of Martin-Löf randomness (or Solovay randomness).

The notion of Martin-Löf randomness is based on effective stochastic laws — or predicates (properties) which are satisfied almost surely, i.e. with probability one. Randomness of a sequence x in the Martin-Löf sense is defined not by looking at the individual sequence x alone, but using an entire collection of predicates of x , and the definition appears to be in the form a “second-order” definition, involving universal quantification over predicates of sequences.²¹

On the other hand, the Kolmogorov-Chaitin definition does not directly refer to any external object other than the sequence x itself and the complexity measure K . Instead of taking an “external top down” approach, it looks at x “from inside” in terms of initial segments (a purely internal view), measuring their information-complexity using K , and declares x to be random if none of the initial segments admit any substantially shorter description. The Kolmogorov-Chaitin definition therefore reduces the definition of randomness for infinite sequences to that for finite strings, establishing a fundamental connection between the two notions.

This striking dissimilarity makes the following celebrated theorem of Schnorr truly remarkable.

THEOREM 36 Schnorr’s Theorem. *A sequence is Martin-Löf random if and only if it is Kolmogorov-Chaitin random.*

For a proof, see any of [Nies, 2009; Downey and Hirschfeldt, 2010; Li and Vitanyi, 2008; Chaitin, 1992].

The equivalence of Martin-Löf randomness with Kolmogorov-Chaitin randomness forms the basis of the assertion that Martin-Löf’s definition has truly captured the notion of randomness for infinite sequences, and therefore gives a satisfactory

²¹Of course, since the predicates in question are effectively enumerated, in actuality the universal quantifier is reduced to range over natural numbers, but we are referring to the form of the definition in classical terms.

solution to this classic problem in the philosophy of mathematics and statistics (Problem 1 of the Introduction).

Moreover, as the Kolmogorov-Chaitin definition shows, the notions of randomness for finite and infinite sequences are fundamentally linked, and therefore the solutions to both Problems of the introduction can be given simultaneously in an interconnected fashion. (A characterization of Martin-Löf randomness in terms of plain complexity C has also been obtained, but it is a much more complicated condition compared to the one for K .)

From now on, by a *random infinite sequence* we will mean a Martin-Löf random or equivalently Kolmogorov-Chaitin random sequence.

DEFINITION 37 Randomness for Infinite Sequences, Final Version. An infinite binary sequence x will be called *random* if it is Martin-Löf random or equivalently Kolmogorov-Chaitin random.

The assertion that Martin-Löf randomness or equivalently Kolmogorov-Chaitin randomness captures the “true notion of randomness” conforming to our intuition is sometimes called the *Martin-Löf-Chaitin thesis*. The Martin-Löf-Chaitin thesis, like the Church-Turing thesis for the definition of algorithm, is not a mathematical proposition that can be proved or refuted. We discuss it further in Section 12.

9.1 Properties of infinite random sequences

We list here some regularity properties of infinite random sequences as evidence that we have the correct definition of randomness for infinite sequences. Recall that when we say “random” without qualification, we mean Martin-Löf random, or equivalently Kolmogorov-Chaitin random.

For proofs and further details of the following facts, see [Calude, 1994; Li and Vitanyi, 2008; Nies, 2009; Downey and Hirschfeldt, 2010].

THEOREM 38 Effective Place Selections Preserve Randomness. *Let $\langle x_1, x_2, \dots \rangle$ be a random infinite binary sequence and $\varphi: \{0, 1\}^* \rightarrow \{0, 1\}$ be a partial computable function. Suppose that $\varphi(\langle x_1, \dots, x_{n-1} \rangle) = 1$ for infinitely n , and $n_1 =$ the least n such that $\varphi(\langle x_1, \dots, x_{n-1} \rangle) = 1$, $n_2 =$ the next such n , etc. Then the subsequence $\langle x_{n_1}, x_{n_2}, \dots, x_{n_k}, \dots \rangle$ is also random.*

COROLLARY 39. *Every random sequence is Mises-Wald-Church random.*

COROLLARY 40 Computable Restrictions Preserve Randomness. *If x is random, and $n_1 < n_2 < n_3 < \dots$ form a computable sequence of strictly increasing numbers then the sequence $\langle x_{n_1}, x_{n_2}, x_{n_3}, \dots \rangle$ is also random.*

The above result remains true if $n_1, n_2, \dots$ form a computable sequence of distinct numbers (not necessarily increasing).

COROLLARY 41. *If x is random, then neither the set $\{n: x_n = 1\}$ nor its complement $\{n: x_n = 0\}$ can contain any infinite computably enumerable set (they are immune). In particular, neither these sets nor the sequence x is computable.*

A real number a is called *computable* if the set $\{(m, n) : m/n < |a|\}$ is a computable subset of $\mathbf{N} \times \mathbf{N}$.

A real number $a \in [0, 1]$ is called *random* if there is a random sequence $x \in \{0, 1\}^{\mathbf{N}}$ such that x is the sequence of digits in a binary expansion of a , i.e. $a = \sum_{n=1}^{\infty} x_n/2^n$.

COROLLARY 42. *If $a \in [0, 1]$ is a random real number, then a is not computable. In particular, a is irrational, and in fact transcendental, since all algebraic real numbers are computable.*

The following result shows that like being convergent, being random is an eventual property.

THEOREM 43. *The randomness of a sequence is a “tail” property. In particular:*

- (a) *If the sequence y is obtained from x by altering only finitely many values of x , then x is random iff y is random.*
- (b) *If $x = \langle x_1, x_2, \dots, x_n, \dots \rangle$ and $y = \langle x_{n+1}, x_{n+2}, \dots, x_{n+k}, \dots \rangle$ is obtained by removing the first n terms of x (an n -step shift), then x is random iff y is random.*

THEOREM 44. *If x is random, then:*

- (a) *x satisfies the law of iterated logarithms.*
- (b) *x is absolutely Borel normal: If the real number $\hat{x}$ having x as its binary expansion digits is expanded in base $b > 1$, then the resulting expansion is Borel normal in base b .*

9.2 An example of a specific random sequence: Ω

It is clear that a set S of strings is prefix-free iff the set of basic intervals $N(\sigma)$ (in the Cantor space) indexed by strings from S form a disjoint family, and so the open set formed by their union has a measure equal to the sum of the measures of the basic intervals, i.e., $\mu(\cup_{\sigma \in S} N(\sigma)) = \sum_{\sigma \in S} \mu(N(\sigma)) = \sum_{\sigma \in S} 1/2^{|\sigma|}$. Therefore:

$$\text{For any prefix-free set } S \text{ of strings, we have: } \sum_{\sigma \in S} \frac{1}{2^{|\sigma|}} \leq 1,$$

an important fact known as *the Kraft inequality*.

Recall now the partial computable prefix-free function ξ that used to define the optimal prefix-free complexity K . We define a real number Ω , called the *Halting Probability* or *Chaitin's Omega*, by:

$$\Omega = \sum_{\sigma \in \text{dom}(\xi)} \frac{1}{2^{|\sigma|}}.$$

Since the domain of ξ is a prefix-free set, $\Omega \leq 1$ by the Kraft inequality.

Fix a program Q which computes the partial computable function ξ . Then Q halts on an input binary string δ iff $\xi(\delta)$ is defined. Suppose now that a fair coin is flipped until some initial segment of the sequence of flips is found to be in the domain of ξ , or equivalently until it generates a string having an initial segment on which Q halts. Of course, in many cases no such string will be generated (i.e. we may have an infinite sequence of flips for which there is no initial segment string on which Q ever halts). In this sense, Ω denotes the probability that Q halts if its input is generated by a random sequence of coin flips. This is the reason Ω is known as the Halting Probability. It is a non-computable real number, and so is transcendental.

Alternatively, define an open set by:

$$G_\Omega = \bigcup_{\sigma \in \text{dom}(\xi)} N(\sigma).$$

Then G_Ω is effective open and Ω equals the Lebesgue measure of G_Ω .

The infinite sequence of bits forming the binary expansion of Ω is also denoted by Ω . Ω is then a random infinite sequence, our first example of a specific random infinite sequence.

Ω has many remarkable properties, see [Bennett, 1979; Calude, 1994; Chaitin, 1992].

10 RELATIVE AND STRONGER RANDOMNESS. HIERARCHIES

Given two sequences $x = \langle x_1, x_2, \dots \rangle$ and $y = \langle y_1, y_2, \dots \rangle$, we merge them into a single sequence $x \oplus y$ by intertwining the terms as follows:

$$x \oplus y = \langle x_1, y_1, x_2, y_2, \dots, x_n, y_n, \dots \rangle.$$

We call $x \oplus y$ the *join of x and y* . Since each of x and y can be extracted from $x \oplus y$, we can, from an information content point of view, think of $x \oplus y$ as perfectly combining the *information* contained in x and that in y , without any “information loss”.

From the properties of randomness given earlier, it is immediate that if $x \oplus y$ is random then so are both x and y . However, if x and y are random, it does not follow that $x \oplus y$ is random. As a drastic example, if $x = y$, then $x \oplus y$ cannot be random since every alternate pair of consecutive bits of $x \oplus y$ would be identical, which allows one to devise a simple successful gambling system against it (it also violates Borel normality as neither the pattern 010 nor the pattern 101 occurs in it).

So the question arises: *Under what conditions on x and y do we have $x \oplus y$ random?*

This was answered beautifully by van Lambalgen in terms of the notion of *relative randomness*, which we now discuss. Roughly speaking, x is random relative to

y (abbreviated x is random in y) if even a complete knowledge of y does not improve the predictability of the bits of x . This is exactly opposite of the situation in our drastic example of $x = y$, where knowledge of y allows us to perfectly predict x , or, in other words, the information about x can be obtained (in this case completely) from information of y . We may therefore expect, that the relation of one sequence being random relatively to another is some form of “information-independence”, although it is not a priori clear that this relation should be symmetric (the drastic example shows that the relation must be irreflexive).

We proceed to formalize this idea.

DEFINITION 45 Effective Open, Relative Version. Let $z \in \{0,1\}^{\mathbb{N}}$. A set G is said to be *effective open relative to z* , or simply *effective open in z* , or in symbols G is $\Sigma_1^0(z)$, if there is an effective open set H such that for all $x \in \{0,1\}^{\mathbb{N}}$:

$$x \in G \iff x \oplus z \in H.$$

Notice how, in the formation of G , the information of z becomes available: As before, G is still the union of basic intervals which are enumerated by some computation, but now that computation is also allowed to use any additional information from z as needed.

Thus, if G is effective open, then G is effective open in z , for any z (additional information from z is available, but not used, in the computation which enumerates basic intervals forming G). On the other hand if G is effective open in z and z is computable, then G is already effective open (since z can be computed by some program P , the computation which enumerates basic intervals forming G does not get any additional help by knowing z , since any information about z could also be obtained by calling P as a subprogram).

Similarly, we define a sequence of sets being “uniformly effective open in z ”.

DEFINITION 46 Uniformly Effective Open (Relative). A sequence $G_1, G_2, \dots$ of sets is *uniformly effective open in z* , or *uniformly $\Sigma_1^0(z)$* , if there are sets $H_1, H_2, \dots$, uniformly effective open, such that for all n ,

$$x \in G_n \iff x \oplus z \in H_n.$$

Again, we define a set being “effective measure-zero in z ” just by changing the old definition with “effective open” replaced by “effective open in z ”. (This process, known as *relativization*, can actually be carried out fruitfully throughout most of computability theory.)

DEFINITION 47 Effective Measure-Zero, Relative Version. A set E is *effective measure-zero in z* if there are sets $G_1, G_2, \dots$, uniformly effective open in z , such that $\mu(G_n) < 1/n$ and $E \subseteq \bigcap_n G_n$.

And, finally:

DEFINITION 48 Relative Randomness. x is *random relative to y* (or x is *random in y*) iff x does not belong to any set effective measure-zero in y .

Now we can state van Lambalgen's theorem:

THEOREM 49 Van Lambalgen. *For $x, y \in \{0, 1\}^{\mathbb{N}}$, the following conditions are all equivalent to each other:*

- (a) $x \oplus y$ is random;
- (b) y is random in x and x is random in y ;
- (c) y is random and x is random in y ;

Van Lambalgen's theorem is remarkable, because the apparently weaker third condition in the theorem implies the second. In particular, if x and y are random, then

$$x \text{ is random in } y \implies y \text{ is random in } x,$$

which is surprising, because, as we mentioned earlier, this symmetry is not at all clear a priori.

The existence theorems all remain valid under relative randomness. E.g., for any y , the set of sequences random in y form a full-measure set whose complement is effective measure-zero in y . If y is computable, this does not give a new collection, as then the set of sequences random in y equals the set of random sequences. But if y is not computable, we may have stronger versions of randomness. E.g., if y is the characteristic function of the uncomputable set $\emptyset'$, or if y is Ω (more precisely, the sequence y consists of the digits in the binary expansion of Ω), then the collection of sequences random in y cannot contain Ω anymore, and therefore form a strictly smaller subclass of the random sequences, known as the 2-random sequences. Using a prefix free function ξ_Ω partial computable in Ω which is optimal for all prefix free functions partial computable in Ω , one can now define the *halting probability relative to Ω* as:

$$\Omega_2 = \sum_{\sigma \in \text{dom}(\xi_\Omega)} \frac{1}{2^{|\sigma|}}.$$

Then Ω_2 is 2-random, while Ω is random but not 2-random.

This process can be iterated to stronger and stronger versions of randomness, which we now describe in greater generality.

10.1 The arithmetical hierarchy and n -randomness

We will use a notation where relations are identified with predicates: If R is a 3-place relation, then we abbreviate " $(a, b, c) \in R$ " by simply writing $R(a, b, c)$, etc.

A relation $A \subseteq \mathbb{N}^k \times \{0, 1\}^{\mathbb{N}}$ is called *effective open* if there is a computably enumerable $E \subseteq \mathbb{N}^k \times \{0, 1\}^*$ such that for all $m_1, \dots, m_k \in \mathbb{N}$ and $x \in \{0, 1\}^{\mathbb{N}}$,

$$A(m_1, \dots, m_k, x) \iff E(m_1, \dots, m_k, \sigma) \text{ for some initial segment } \sigma \text{ of } x.$$

A relation is called *effective closed* if its complement is effective open, and it is called *computable* (or *effective clopen*) if it is both effective open and effective closed.

Starting with the computable relations as a basis, we can define relations of higher complexity by adding a series n “alternating quantifiers” ranging over natural numbers as follows. We define a relation A to be Σ_n^0 ($n \geq 1$) if there is a computable relation R such that for all $m_1, \dots, m_k \in \mathbf{N}$ and $x \in \{0, 1\}^{\mathbf{N}}$,

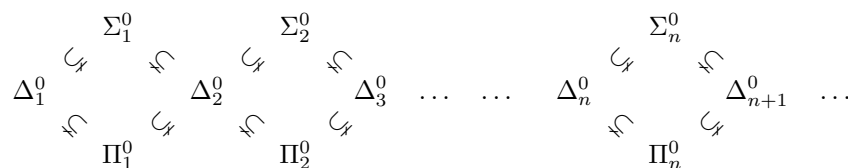
$$A(m_1, \dots, m_k, x) \iff (\exists p_1)(\forall p_2) \dots (Q p_n) R(p_1, \dots, p_n, m_1, \dots, m_k, x),$$

where Q stands for “ $\exists$ ” if n is odd and for “ $\forall$ ” if n is even.

We also define a relation to be Π_n^0 if its complement is Σ_n^0 , and a relation to be Δ_n^0 if it is both Σ_n^0 and Π_n^0 .

It then turns out that the class Σ_1^0 coincides with the class of effective open relations, and Δ_1^0 is same as the class of computable relations.

Moreover, this indeed gives a strict hierarchy of classes of relations defined by their definitional complexity (the number of alternating quantifiers), with the class Δ_n^0 strictly contained in each of Σ_n^0 and Π_n^0 , both of which are strictly contained in Δ_{n+1}^0 , as shown below.



This hierarchy is called the *Arithmetical Hierarchy*. (It is a refinement of the finite levels of the classical hierarchy of Borel sets in analysis, but only *effective* countable unions and intersections are allowed. See [Rogers Jr, 1987; Odifreddi, 1992; Moschovakis, 1980] for more details.)

Finally, we can define n -randomness.

DEFINITION 50. A sequence z is called n -random iff there is no Σ_n^0 -effective measure-zero set containing z , or more precisely if there is no sequence $H_1, H_2, \dots$ of sets such that the relation H defined by $H(n, x) \iff x \in H_n$ is Σ_n^0 , $\mu(H_n) < 1/n$ for all n , and $z \in H_n$ for all n .

We say z is *arithmetically random* iff z is n -random for all $n = 1, 2, \dots$

One can show that it does not make any difference to the definition whether we require the sets H_n to be all open or not.

Thus, 1-random is same as being random (i.e., Martin-Löf-random), and we have a sequence of stronger and stronger notions of randomness, corresponding to the levels of the arithmetical hierarchy. This hierarchy is indeed strict, meaning that $(n+1)$ -randomness implies n -randomness, but for each n there is an n -random sequence which is not $(n+1)$ -random. E.g., Ω is 1-random but not 2-random, and

Ω_2 is 2-random but not 3-random, and so on. Moreover,

$$\begin{aligned} x \text{ is 2-random} &\iff x \text{ is random relative to } \Omega \\ &\iff x \text{ is random and } \Omega \text{ is random relative to } x. \end{aligned}$$

The first of these equivalences follow from standard facts about the arithmetical hierarchy, and the second from van Lambalgen's theorem.

Recall that an infinite binary sequence $x = \langle x_1, x_2, \dots \rangle$ was defined to be Kolmogorov-Chaitin random if for some $b > 0$, $K(\langle x_1, x_2, \dots, x_{n-1} \rangle) > n - b$ for all n . If prefix complexity K is replaced by plain complexity C , then one produces an empty definition, as was proved by Martin-Löf: *There is no sequence $x = \langle x_1, x_2, \dots \rangle$ such that for some $b > 0$, $C(\langle x_1, x_2, \dots, x_{n-1} \rangle) > n - b$ for all n .* However, it was also shown that the sequences x which satisfy the condition:

$$\text{for some } b > 0, C(\langle x_1, x_2, \dots, x_{n-1} \rangle) > n - b \text{ for infinitely many } n,$$

form a full-measure set which is contained in the set of random sequences. Such sequences are sometimes called *Kolmogorov Random*. Remarkably, it was established recently that the Kolmogorov random sequences are precisely the 2-random ones.

10.2 Other stronger notions of randomness

Many other notions of randomness stronger than 1-randomness have been studied. E.g., by relaxing the condition that $\mu(H_n) < 1/n$ in the definition of n -randomness to $\lim_n \mu(H_n) = 0$, one obtains the notion of weak- $(n+1)$ -randomness, which lies strictly between n -randomness and $(n+1)$ -randomness.

One can extend the arithmetical hierarchy into the transfinite using computable ordinals, which results in the *hyperarithmetical hierarchy*. Martin-Löf first suggested the notion of hyperarithmetical randomness. Beyond the hyperarithmetical classes, there is an even more comprehensive hierarchy known as the *analytical hierarchy*. At the first level of this hierarchy are the Π_1^1 relations. A relation A is called Π_1^1 if there is an arithmetical relation (Σ_2^0 is enough) B such that

$$A(n_1, \dots, n_k, x) \iff \forall y B(n_1, \dots, n_k, x \oplus y).$$

(The class of Π_1^1 relations includes all hyperarithmetical sets and more, and hence certainly all arithmetical sets as well.)

The notion of randomness in the sense of Martin-Löf has recently been extended to the class of Π_1^1 sets (and has been named Π_1^1 -ML-randomness) fruitfully by Hjorth and Nies [Hjorth and Nies, 2007], and they have established the analog of Schnorr's theorem and other results.

The *strongest notion of randomness* that appears to have been studied so far is called Π_1^1 -randomness by Hjorth and Nies [Hjorth and Nies, 2007]. The union of all Π_1^1 measure-zero sets is itself a Π_1^1 measure-zero set (the largest Π_1^1 measure-zero set), so Π_1^1 -randomness is defined as membership in the complement of the largest Π_1^1 measure-zero set.

10.3 Reducibility and degrees of computability

Given $A \subseteq \mathbb{N}$ and $z \in \{0, 1\}^{\mathbb{N}}$, we say that A is *computably enumerable in z* , or in symbols $A \in \Sigma_1^0(z)$ if there is a c.e. set $B \subseteq \mathbb{N} \times \{0, 1\}^*$ such that $\forall n(A(n) \iff \exists k B(n, \langle z_1, z_2, \dots, z_k \rangle))$.

We say that A is *computable in z* , or A is *Turing-reducible to z* , in symbols $A \leq_T z$, if both A and its complement are computably enumerable in z . Finally a sequence $x \in \{0, 1\}^{\mathbb{N}}$ is *computable in z* or *Turing-reducible to z* if the set $\{n: x_n = 1\}$ is computable in z . Roughly speaking, $x \leq_T y$ means that x can be computed by a program which has access to the bits of y in order, or even more vaguely y is computationally at least as complex as x .

The notion of Turing-reducibility is reflexive and transitive, and the corresponding equivalence relation, called *Turing equivalence*, $x \equiv_T y \iff x \leq_T y \wedge y \leq_T x$ generates equivalence classes known as *Turing degrees*. The study of Turing-reducibility and degrees has been one of the most important areas of classical recursion theory. There are several other types of computational reducibilities, generally stronger than Turing-reducibility, that are important for the theory of computability.

The interaction between randomness and Turing-reducibility (and other computational reducibilities not introduced here) has also been studied, and has generated fruitful applications in both directions.

Another notion straddling both randomness and computability theory that has been studied extensively is that of K -triviality. A sequence $x = \langle x_1, x_2, \dots \rangle$ is *K -trivial* if $\forall n(K(\langle x_1, \dots, x_n \rangle) \leq K(n) + b)$ for some constant b . This property is quite orthogonal to that of randomness. It is known that x is K -trivial iff every random sequence is random relative to x .

Many other notions that interact with both computability theory and randomness form a part of current research, which is progressing vigorously. We refer the reader to [Nies, 2009; Downey and Hirschfeldt, 2010] where further extensive references can be found.

11 RANDOMNESS VIA MARTINGALES. OTHER FREQUENTIST DEFINITIONS

In the previous section, we considered randomness notions stronger than Martin-Löf randomness. Now we will focus on weaker notions of randomness, which are perhaps more important from a philosophical viewpoint.

11.1 Schnorr randomness

A critique of Martin-Löf randomness by Schnorr was that it yields too strong a notion of randomness as its notion of defining effective measure-zero is not effective enough. In order for a sequence of uniformly effective open sets $G_1, G_2, \dots$ to define an effective measure-zero set via intersection, it is not enough, according

to Schnorr, that their measures effectively approach zero by just having *effective bounding* (e.g. as $\mu(G_n) < 1/n$), but we need the measures $\mu(G_n)$ of the sets themselves to be computable real numbers (uniformly in the index n). Using this stronger criteria for being effective measure-zero, we get a weaker notion of randomness, called *Schnorr randomness*.

Schnorr randomness has been studied extensively, but it fails to have certain regularity properties of Martin-Löf randomness. Two examples are:

- (a) Unlike Martin-Löf randomness, Schnorr randomness does not possess a uniform test, i.e. the class of Schnorr random sequences cannot be defined as the complement of the intersection of a sequence of uniformly effective open sets $G_1, G_2, \dots$ such that $\mu(G_n) < 1/n$ and such that $\mu(G_n)$ is a computable real number uniformly in n .
- (b) The van Lambalgen theorem fails for Schnorr randomness. In fact, there is Schnorr random sequence $z = x \oplus y$ such that that the two halves x and y are Turing equivalent. This does not conform well with the intuitive notion of randomness.

11.2 Randomness defined by martingales

Recall that in von Mises type definitions of randomness, one uses a the concept of a betting strategy, or more precisely a *place selection rule*, to select places on which to bet, and after selections are all done, one checks if the limiting frequency value has become biased or not (subsections 5.1 and 5.2). The definition does not a priori have anything to do with the *amount of bet*. Such properties are called *stochasticity* (as opposed to randomness) in the modern mathematical literature, but note that this is essentially a matter of terminology.

We now introduce a concept of betting strategy involving the amount of bet, or equivalently the *capital* of the gambler (“total money in pocket”), at each stage of betting.

We think of the infinite sequence $x = \langle x_1, x_2, x_3, \dots \rangle$ being revealed to the gambler, bit by bit, in order. Before each bit is revealed, the gambler may bet an amount a predicting the value of the bit (one of “next bit revealed will be 0” or “next bit revealed will be 1”). We will assume the *fairness condition* that if the prediction turns out to be correct, the gambler *gains* an amount of a (capital increases by a), otherwise the gambler *loses* the same amount a (capital decreases by a).

To formalize this type of strategy, we think of a finite binary string σ of length $|\sigma| = n - 1$ as representing the n -th stage of the game, so that σ consists of the bits of the infinite sequence that have been revealed so far (before the n -th bit is revealed), and let $F(\sigma)$ denote the gambler’s capital at this stage.

Suppose that at stage σ , with capital $F(\sigma)$, the gambler bets an amount a predicting the next bit to be 0. If the gambler turns out to be correct, then σ is

extended to $\sigma 0$ and the capital goes up by a to $F(\sigma 0) = F(\sigma) + a$, but if incorrect, then σ is extended to $\sigma 1$ and the capital goes down by a to $F(\sigma 1) = F(\sigma) - a$. Similarly, if the gambler had predicted the next bit to be 1 (with same bet amount a) then we would have σ extending to σ_1 and $F(\sigma 1) = F(\sigma) + a$ if the gambler turns out to be correct, and σ extending to σ_0 and $F(\sigma 0) = F(\sigma) - a$ if the gambler is incorrect. A final case is when the gambler chooses not to bet at this stage, which is expressed by having $a = 0$ and $F(\sigma 0) = F(\sigma 1) = F(\sigma)$ (no change in capital).

All cases can be summarized using a single *zero sum condition*:

$$\overbrace{F(\sigma 0) - F(\sigma)}^{\text{capital change if next bit is 0}} + \overbrace{F(\sigma 1) - F(\sigma)}^{\text{capital change if next bit is 1}} = 0.$$

Therefore we make the following definition.

DEFINITION 51 Martingales. A *martingale* or *capital betting strategy* is a function $F: \{0, 1\}^* \rightarrow \mathbf{R}$ satisfying two conditions:

- (a) $F(\sigma) \geq 0, \forall \sigma$ (finiteness condition); and
- (b) $[F(\sigma 0) - F(\sigma)] + [F(\sigma 1) - F(\sigma)] = 0, \forall \sigma$ (zero-sum or fairness condition).

Given $x \in \{0, 1\}^{\mathbf{N}}$ and a martingale $F: \{0, 1\}^* \rightarrow \mathbf{R}$, we say that the martingale F *succeeds on* x if the capital becomes unbounded on outcome sequence x , i.e., if

$$\sup_n F(\langle x_1, x_2, \dots, x_{n-1} \rangle) = +\infty.$$

An example of a martingale is where the gambler always predicts an outcome of 0 with the bet being a fixed fraction r ($0 < r < 1$) of the available capital. This martingale is the function recursively defined as $F(\sigma 0) = (1+r)F(\sigma)$ and $F(\sigma 1) = (1-r)F(\sigma)$; or, if initial capital is 1, more explicitly as $F(\sigma) = (1+r)^m(1-r)^n$ if σ is a string of length $m+n$ with m zeros and n ones.

Another example is where the gambler always predicts an outcome of 0 betting the entire amount of available capital (“bold play”). If the initial capital is 1, this martingale is given as follows. If 1 does not occur in σ , then $F(\sigma) = 2^{|\sigma|}$, else $F(\sigma) = 0$.

It can be shown that the concept of a martingale is really a generalization of place selection rule. To each place selection rule φ one can assign an especially simple type of martingale F_φ (one which always uses a constant fraction of the existing capital as the next bet), such that the φ -selected part of x has limiting frequency $1/2$ iff F_φ does not succeed on x . A converse association is also possible.

Possible definitions of randomness either using place selection rules (as done by von Mises) or using martingales are both examples of *characterizations of randomness via impossibility of gambling systems*, with the place selection method known as the *frequentist* approach, while the martingale method may be called non-frequentist.

11.3 A martingale characterization of randomness

A martingale F is said to be *computably enumerable* if the relation R defined by

$$R(m, n, \sigma) \iff \frac{m}{n+1} < F(\sigma)$$

is computably enumerable as a subset of $\mathbf{N} \times \mathbf{N} \times \{0, 1\}^*$.

THEOREM 52 Martingale Characterization of Martin-Löf Randomness. *A sequence x is random (i.e. Martin-Löf random) iff no computably enumerable martingale succeeds on x .*

Thus we now have three different but equivalent definitions for randomness. This last characterization in terms of martingales gives a definition of randomness using the *unpredictability* approach.

If, instead of computably enumerable martingales, we require stronger effectiveness conditions on the martingales, we obtain weaker notions of randomness, as we will see now.

In the following, we consider only rational valued martingales.

DEFINITION 53. A *partial computable martingale* is a partial function $F: \{0, 1\}^* \rightarrow \mathbf{Q}$ satisfying, for all σ :

- (a) $F(\sigma) \geq 0$;
- (b) If $F(\sigma)$ is defined, so is $F(\tau)$ for any prefix τ of σ ;
- (c) $F(\sigma 0)$ is defined iff $F(\sigma 1)$ is defined, and if so, then:

$$[F(\sigma 0) - F(\sigma)] + [F(\sigma 1) - F(\sigma)] = 0;$$

- (d) The relation R defined by

$$R(\sigma, m, n) \iff F(\sigma) = m/(n+1)$$

is a computably enumerable subset of $\{0, 1\}^* \times \mathbf{N} \times \mathbf{N}$.

A *computable martingale* is a partial computable martingale which is total (i.e., whose domain is $\{0, 1\}^*$). (It can be seen that the graph of a computable martingale is computable, not just computably enumerable.)

Here are the main notions of randomness arising out of these types of martingales.

DEFINITION 54. A sequence is *partial computably random* if no partial computable martingale succeeds on it.

A sequence is *computably random* if no computable martingale succeeds on it.

Of course, every partial computably random sequence is computably random. It can be shown that

$$\begin{aligned} \text{(Martin-Löf) Random} &\implies \text{Partial Computably Random} \\ &\implies \text{Computably Random} \\ &\implies \text{Schnorr Random,} \end{aligned}$$

but none of these implications can be reversed [Nies, 2009].

11.4 Non-monotonic betting strategies

So far, we have seen two distinct types of betting strategies giving rise to notions of randomness:

- *Martingales*, or capital betting strategies, which lead to definitions of randomness based on the failure of the martingale, such as *partial computable randomness* and *computable randomness* that we just saw.
- *Place selection rules*, which lead to definitions of randomness based on the limiting frequency of the selected part, such as *Mises-Wald-Church stochasticity (or randomness)*, and *Church stochasticity (or randomness)* that we saw in subsection 5.2.

These four types just mentioned all are notions weaker than (Martin-Löf) randomness.

There is a further generalization possible in the type of betting allowed — called *non-monotonic betting* — that actually tightens the notions further, and makes them more robust.

To understand non-monotonic betting, both for martingales and for place selection rules, suppose that the bits of the sequence $x = \langle x_1, x_2, \dots, x_n, \dots \rangle$ lay covered on an infinitely long table (instead of being revealed serially one by one). The gambler now uncovers the bits in a not-necessarily increasing order, and along the way decides which places to select or to bet on.

For example, the gambler may choose to first uncover the ninth place to find the value of x_9 , then uncover the fourth place to find x_4 , and then, based on these two observations, decide to select or bet on the seventeenth position (before uncovering it). After x_{17} is uncovered, the gambler may choose to next uncover either x_3 or x_7 , depending on whether x_{17} turns out to be 0 or 1, and so on.

We omit the formal details and hope that the above informal description makes it intuitively clear what non-monotonic betting is. In particular, it can be applied both to martingales and to place selection rules as follows.

DEFINITION 55. A sequence x is called *Kolmogorov-Loveland Random* if no computable non-monotonic martingale succeeds on it.

A sequence x is called *Kolmogorov-Loveland Stochastic* if for every computable non-monotonic place selection rule, the selected part has limiting frequency $= \frac{1}{2}$.

A nice feature of non-monotonic betting strategies is that it does not matter whether we use “computable” or “partial computable” in the definitions above, as in each case it results in an equivalent notion. In other words, changing the above definitions to “partial computable” will not give us a stronger notion of randomness. Thus the non-monotonic forms have a kind of robustness that was not there for monotonic betting strategies, since partial computable randomness is a strictly stronger notion of randomness compared to computable randomness, and Mises-Wald-Church stochasticity is strictly stronger than Church stochasticity.

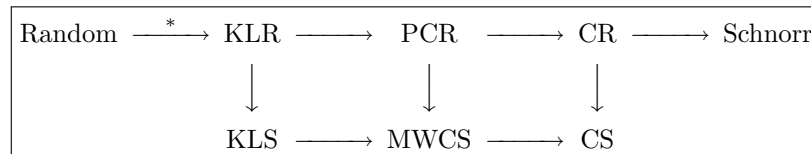
Since martingales are more general than place selection rules, it follows that a martingale version will give a stronger notion of randomness than the corresponding stochastic (i.e limiting frequency after place selections) version. Thus Kolmogorov-Loveland randomness is strictly stronger than Kolmogorov-Loveland stochasticity, partial computable randomness is strictly stronger than Mises-Wald-Church stochasticity, and computable randomness strictly stronger than Church stochasticity.

We summarize these notions in the following table. Each notion in this table implies the one directly below it and also the one to the right of it.

Strategy type	Monotonicity and computability of strategy function		
	Non-monotonic	Monotonic	
	Partial or total	Partial	Total
Martingale	<i>Kolmogorov-Loveland Random (KLR)</i>	<i>Partial Computably Random (PCR)</i>	<i>Computably Random (CR)</i>
Place selection	<i>Kolmogorov-Loveland Stochastic (KLS)</i>	<i>Mises-Wald-Church Stochastic (MWCS)</i>	<i>Church Stochastic (CS)</i>

Six Randomness Notions for Various Types of Betting Strategies

In fact, with “Random” standing for Martin-Löf randomness, “Schnorr” for Schnorr randomness, and the rest of the abbreviations as in the table, we have the following implications.



Implication Diagram for Weak Randomness Notions

Each arrow represents an implication, and almost all the implications above are strict (they cannot be reversed). However, for the implication marked with “*”, it is not known whether the implication can be reversed. In other words we have:

Question. *While every Martin-Löf random sequence is Kolmogorov-Loveland random, is the converse true?*

This is perhaps the biggest open problem in current research on randomness. Many researchers feel that the answer is no, although work of Merkle et al [Merkle *et al.*, 2006] have shown that the two notions are rather close.

Nies and Miller have published a list of open problems [Miller and Nies, 2006] in the area, some of which have been solved since then.

11.5 Can we resurrect von Mises?

While Kolmogorov-Loveland randomness is the only major notion of randomness that remains close to Martin-Löf randomness, it does not have the true spirit of von Mises' idea of randomness, since von Mises's definition, which is based on place selections, is a *truly frequentist* one, that is, it is defined in terms of limiting frequency, while the martingale notions are all defined in terms of capital growth. Therefore, even if it turns out that Kolmogorov-Loveland randomness is equivalent to Martin-Löf randomness, one would still be looking for a characterization of Martin-Löf randomness in terms of a frequentist condition.

In recent literature, the term *stochasticity* is used for randomness defined in terms of a frequentist condition, or more precisely using the limiting frequency of place selections. Among these, the one closest to Martin-Löf randomness is *Kolmogorov-Loveland stochasticity*, as the above diagram of implications indicates. Unfortunately, like Mises-Wald-Church stochasticity, Kolmogorov-Loveland stochasticity gives random sequences which do not satisfy the Law of Symmetric Oscillations, and therefore is rather far from Martin-Löf randomness.

Li and Vitanyi writes in their 2008 book [Li and Vitanyi, 2008, p. 158]:

“[T]he problem of giving a satisfactory definition of infinite Martin-Löf random sequences in the form proposed by von Mises has not yet been solved.”

Thus, the search for a true frequentist characterization of Martin-Löf randomness continues.

11.6 The Ergodic Theorem as a Frequentist Definition

The general version of the Birkhoff Ergodic Theorem is not fully effective (see [Avigad, 2009; Hoyrup, 2008]). However, we consider the version of the theorem which is known as the “law of frequencies”. We restrict to Lebesgue measure on the Cantor space.

Say that a measurable map $U: \{0, 1\}^{\mathbb{N}} \rightarrow \{0, 1\}^{\mathbb{N}}$ is *measure preserving* if $\mu(U^{-1}[A]) = \mu(A)$ for all measurable A , and a map $T: \{0, 1\}^{\mathbb{N}} \rightarrow \{0, 1\}^{\mathbb{N}}$ is *ergodic* if T is measure preserving and for all measurable A , if $T^{-1}[A] \Delta A$ is measure-zero then $\mu(A) = 0$ or $\mu(A) = 1$.

We now state the Birkhoff Ergodic Theorem in a slightly variant form:

THEOREM 56 The Birkhoff Ergodic Theorem as the Law of Frequencies. *If $T: \{0, 1\}^{\mathbb{N}} \rightarrow \{0, 1\}^{\mathbb{N}}$ is ergodic, E is measurable, and $U: \{0, 1\}^{\mathbb{N}} \rightarrow \{0, 1\}^{\mathbb{N}}$ is measure preserving, then (we put $\llbracket P \rrbracket = 1$ if the statement P is true, else $\llbracket P \rrbracket = 0$):*

$$\text{For almost all } x: \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \llbracket T^k(U(x)) \in E \rrbracket = \mu(E),$$

i.e., the frequency with which the T -orbit of $U(x)$ enters E approaches $\mu(E)$.

We also view of this theorem, the Law of Frequencies, as a form of *equidistribution* (as in Weyl equidistribution): The T -orbit of $U(x)$ is equidistributed.

Now if φ is a place selection rule satisfying the property that for almost all x , $\varphi(\langle x_1, x_2, \dots, x_{n-1} \rangle) = 1$ for infinitely many n , then the map $U_\varphi: \{0, 1\}^{\mathbb{N}} \rightarrow \{0, 1\}^{\mathbb{N}}$

$$U_\varphi(x) = \varphi\text{-selected part of } x$$

is defined for almost all x and is a measure preserving (and continuous) map. Finally, take T to the left shift map $T(\langle x_1, x_2, \dots \rangle) = \langle x_2, x_3, \dots \rangle$, and $E = N(1) = \{x \in \{0, 1\}^{\mathbb{N}}: x_1 = 1\}$ = the basic interval consisting of sequences with 1st bit = 1.

With the above setting, the condition in the von Mises definition coincides precisely with the condition in the Birkhoff Ergodic Theorem above, at least in the case when the domain of U_φ has full measure. Furthermore, if E is allowed to range of over the basic intervals, then the condition of Borel normality of is obtained, which we view as equidistribution of the T -orbit of $U(x)$.

Perhaps the main weakness of the von Mises definition is the lack of the requirement of general equidistribution, which is illustrated in the approach of [Knuth, 1998]. The notion of equidistribution is more general than the notion of limiting frequency, while still being in the frequentist spirit.

Suppose we try to strengthen the Mises-Wald-Church definition by requiring Borel normality of the φ -selected part, not just existence of unbiased limiting frequency, i.e., we demand that for x to be random, the φ -selected part of x has to be Borel normal in base 2 for every partial computable place selection rule φ . This does not really change anything, as all Mises-Wald-Church stochastic sequences are Borel normal. However, instead of being limited to limiting frequency, it casts the definition in terms of equidistribution of the T -orbit of $U(x)$, where T is the left-shift map. In other words, in Mises-Wald-Church stochasticity, the condition that the place-selected part y of x must satisfy is equivalent to the equidistribution of the following sequence of sequences obtained from y :

$$\langle y_1, y_2, y_3, \dots \rangle, \langle y_2, y_3, y_4, \dots \rangle, \langle y_3, y_4, y_5, \dots \rangle, \dots$$

We think that a frequentist definition of randomness should allow more general forms of equidistribution, so long as the method of forming the sequence is uniformly effective and ergodic (see [Knuth, 1998] for examples).

For example, suppose that the natural numbers are partitioned into an infinite number of infinite uniformly computable subsets, say into the sets $\{1, 3, 5, 7, \dots\}$, $\{2, 6, 10, 14, \dots\}$, $\{4, 12, 20, 28, \dots\}$, etc. Then for a random x , if y is obtained from x by effective place selection, we expect that the sequences obtained by restricting y to each of these subsets

$$\langle y_1, y_3, y_5, \dots \rangle, \langle y_2, y_6, y_{10}, \dots \rangle, \langle y_4, y_{12}, y_{20}, \dots \rangle, \dots$$

should be equidistributed (which follows from the Ergodic Theorem). It is not a priori clear that Mises-Wald-Church stochasticity guarantees this. (Kolmogorov-Loveland stochasticity allows more general forms of U , but the T operator is still the same left shift.)

Perhaps the following definition can be taken to be an “ergodic generalization” of von Mises’ definition of randomness, where we think of U as “place selection”:

DEFINITION 57. $x \in \{0, 1\}^{\mathbb{N}}$ is *random* iff for all sufficiently effective $E \subseteq \{0, 1\}^{\mathbb{N}}$, sufficiently effective ergodic T , and sufficiently effective measure preserving U , we have:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \llbracket T^k(U(x)) \in E \rrbracket = \mu(E),$$

i.e., the frequency with which the T -orbit of $U(x)$ enters E approaches $\mu(E)$.

The three instances of the phrase “sufficiently effective” are deliberately left open to interpretation. We ask:

Question. *Are there interpretations of “sufficiently effective” in the above definition which characterize Martin-Löf randomness?*

For example, it is known that with “sufficiently effective” interpreted as Δ_1^0 (computable) in all three instances, every Martin-Löf random sequence is random in the sense of the above definition.²² However, we are not aware of any result that positively answers the above question.

12 CONCLUSION. THE MARTIN-LÖF-CHAITIN THESIS

In this article, we introduced the reader to definitions of the notion of a random sequence using the three main ideas described in Section 1.6 that have dominated algorithmic randomness (cf. [Downey and Hirschfeldt, 2010]):

²²See [González, 2008; Hoyrup, 2008] for a proof and a further list of problems. See also [V’yugin, 1997; V’yugin, 1999]. In [González, 2008; Hoyrup, 2008], this is studied in a more general setting (such as effective probability metric spaces), and results have been obtained for Schnorr randomness, but it is stated as an open problem for Martin-Löf randomness. It appears that the problem of characterization of Martin-Löf randomness in this ergodic way is open even in the specific case of the Cantor Space with Lebesgue measure.

- *Randomness as typicality.* According to this stochastic or measure-theoretic idea, randomness of a sequence means its membership in all effective full-measure sets, or equivalently that the sequence “passes” all effective stochastic tests. The first major stochastic laws, the Borel strong law and Borel normality go back to 1909. Martin-Löf randomness (1966), the first definition that is almost universally accepted now as the correct one, is defined using this approach.
- *Randomness as incompressibility.* This is an information-complexity approach that views the sequence “from inside”. According to this idea, randomness means lack of short complete descriptions, or equivalently a high degree of algorithmic complexity, for all initial parts of the sequence.
- *Randomness as unpredictability.* According to this approach, randomness means the impossibility of devising a successful betting strategy against the sequence in question. In particular, this means that knowledge of some part of the sequence does not help to predict any other unknown bit. Two types of definitions of randomness arise from this deeply intuitive approach: The special *frequentist* type definitions first put forward by von Mises as “invariance of limiting frequency under admissible place selections”, and the more general *non-frequentist* type found in the definitions of randomness using martingales.

The ideal definition of randomness would be one which naturally and simultaneously satisfies the criteria given by these three approaches.

12.1 The Martin-Löf-Chaitin thesis

Following Delahaye [Delahaye, 1993], we use the term *Martin-Löf-Chaitin Thesis* for the assertion that *Martin-Löf randomness and equivalently Kolmogorov-Chaitin randomness is the correct formulation of the intuitive notion of randomness for sequences*. In this sense, it parallels the classic Church-Turing thesis, and is not a mathematical proposition to be proved or disproved. The Church-Turing thesis turned out to be highly successful in capturing the intuitive notion of *algorithm*.

Delahaye has carried out a detailed comparison between the Church-Turing thesis and the Martin-Löf-Chaitin thesis, and concludes that in both cases, the resulting precise definitions provide “profound insights to the mathematical and philosophical understanding of our universe.” Delahaye admits that the Church-Turing thesis is “more deeply attested” and that the definition of randomness of sequences is “more complicated” compared to the definition of algorithm, but hopes that with time the Martin-Löf-Chaitin thesis will reach a level of certainty similar to the Church-Turing thesis. We think that overall, Delahaye’s assertions still remain valid.

In the past few decades, there has been a vast amount of research activity in the area of algorithmic randomness. Many definitions of randomness for sequences

have been studied extensively, but none was found to be clearly superior to the Martin-Löf definition. Compared with other notions, it appears to be of optimal strength: Weaker notions turn out to be too weak, and the stronger ones too strong. In this way, the Martin-Löf-Chaitin thesis has gained strength in a slow but steady fashion.

The proliferation of definitions of randomness for sequences makes the field harder for non-experts, but it should not be regarded negatively. It is an indication of the richness of the area, and the associated healthy and lively activity provides refinements and insights deep into the subject. Recall that while we consider the Church-Turing thesis as more satisfying, there was an even larger number of associated notions of computability, both stronger and weaker, that were (and still are) studied fruitfully.

Perhaps the strongest evidence for the Martin-Löf-Chaitin thesis available so far is Schnorr's theorem, which establishes the equivalence between a naturally formulated "typicality definition" (Martin-Löf randomness) and a naturally formulated "incompressibility definition" (Kolmogorov-Chaitin randomness).

Another justification of the Martin-Löf-Chaitin thesis is provided by the simplicity of the definition of Martin-Löf randomness within the arithmetical hierarchy. As seen in Schnorr's theorem,

$$x \text{ is random} \iff \exists b \forall n K(\langle x_1, \dots, x_{n-1} \rangle) \geq n - b.$$

This shows that Martin-Löf randomness has a Σ_2^0 definition (which also follows from the existence of a universal test). Most other definitions of randomness are more complicated, and situated at higher levels of the arithmetical hierarchy. In fact, the definitional complexity of Martin-Löf randomness is at the lowest possible level of the arithmetical hierarchy, assuming that any definition of randomness must satisfy the two basic axioms:

- (a) No random sequence should be computable.
- (b) The set of random sequences has full-measure.

It then follows that no definition of randomness can be Π_2^0 or simpler, as it is a standard "basis theorem" that any Π_2^0 set of full-measure contains computable sequences.

We also doubt if the resolution of the question of whether there are Kolmogorov-Loveland random sequences which are not Martin-Löf random will have much impact on the Martin-Löf-Chaitin thesis.

However, a *purely frequentist* natural characterization of Martin-Löf randomness can substantially increase the strength of the Martin-Löf-Chaitin thesis. While the characterization in terms of computably enumerable martingales is a nice "unpredictability definition", it is not as intuitive nor as frequentist as the von Mises definition. This is perhaps the most unsatisfying gap in the current state of affairs.

To summarize, we believe that while the Martin-Löf-Chaitin thesis is not (yet) as strong as the Church-Turing thesis, the two problems of the introduction, namely defining randomness for sequences and strings that captures our mathematical intuition of these objects, have essentially been solved quite satisfactorily as described in this article. It is perhaps not too surprising that the definition of randomness, which in all cases presupposes the definition of algorithm, has turned out to be more complicated than the definition of algorithm itself.

ACKNOWLEDGMENTS

The author wishes to thank Ananda Sen for help with some of the references. The author is also indebted to Prasanta S. Bandyopadhyay and the anonymous referee for several useful suggestions.

BIBLIOGRAPHY

- [Aczel, 2004] A.D. Aczel. *Chance: A Guide to Gambling, Love, the Stock Market & Just About Everything Else*. Thunder's Mouth Press, New York, 2004.
- [Avigad, 2009] J. Avigad. The metamathematics of ergodic theory. *Annals of Pure and Applied Logic*, 157(2-3):64–76, 2009.
- [Bailly and Longo, 2007] F. Bailly and G. Longo. Randomness and determinism in the interplay between the continuum and the discrete. *Mathematical Structures in Computer Science*, 17(02):289–305, 2007.
- [Becher and Figueira, 2002] V. Becher and S. Figueira. An example of a computable absolutely normal number. *Theoretical Computer Science*, 270(1-2):947–958, 2002.
- [Belshaw and Borwein,] A. Belshaw and P. Borwein. Strong Normality of Numbers. <http://www.cec.m.sfu.ca/personal/pborwein/PAPERS/P211.pdf>. Contains material added and updated after Belshaw's masters thesis.
- [Belshaw, 2005] A. Belshaw. On the normality of numbers. Master's thesis, Simon Fraser University, 2005.
- [Beltrami, 1999] E.J. Beltrami. *What is random? Chance and Order in Mathematics and Life*. Copernicus (Springer-Verlag), New York, 1999.
- [Bennett, 1979] C.H. Bennett. On Random and Hard-to-Describe Numbers. Technical Report RC-7483, IBM Watson Research Center, Yorktown Heights, New York, 1979. reprinted in *Randomness And Complexity, from Leibniz To Chaitin* (C. Calude, ed.), World Scientific 2007, pp 3–12.
- [Bennett, 1998] D.J. Bennett. *Randomness*. Harvard University Press, Cambridge, Mass., and London, England, 1998.
- [Boolos and Jeffrey, 1989] G. S. Boolos and R. C. Jeffrey. *Computability and Logic*. Cambridge University Press, 1989.
- [Borel, 1909] E. Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rend. Circ. Mat. Palermo*, 27:247–271, 1909.
- [Calude, 1994] C. S. Calude. *Information and Randomness: An Algorithmic Perspective*. Springer Verlag, 1994.
- [Calude, 2000] C. S. Calude. Who is afraid of randomness? Technical Report CDMTCS-143, University of Auckland, New Zealand, 2000.
- [Calude, 2005] C. S. Calude. Algorithmic randomness, quantum physics, and incompleteness. In *Proceedings of the Conference on Machines, Computations and Universality(MCU2004)*, volume 3354, pages 1–17. Springer, 2005.
- [Chaitin, 1992] G. J. Chaitin. *Algorithmic Information Theory*. Cambridge University Press, 1992.

- [Church, 1940] A. Church. On the concept of a random sequence. *Bull. Amer. Math. Soc.*, 46:130–135, 1940.
- [Davis, 1980] M. Davis. What Is a Computation? In Lynn Arthur Steen, editor, *Mathematics Today*, pages 241–267. Vintage Books, New York, 1980.
- [de Laplace, 1819/1952] Pierre-Simon de Laplace. *A Philosophical Essay on Probabilities*. Dover, translated from 6th french ed edition, 1819, 1952.
- [Delahaye, 1993] J.P. Delahaye. Randomness, Unpredictability and Absence of Order: The Identification by the Theory of Recursivity of the Mathematical Notion of Random Sequence. *Philosophy of Probability*, pages 145–167, 1993.
- [Downey and Hirschfeldt, 2010] R. G. Downey and D. Hirschfeldt. *Algorithmic Randomness and Complexity*. Springer, 2010.
- [Eagle, 2005] A. Eagle. Randomness is unpredictability. *The British Journal for the Philosophy of Science*, 56(4):749–790, 2005.
- [Feller, 1968] W. Feller. *An Introduction to Probability Theory and its Applications*. Vol. 1. John Wiley & Sons, New York, 1968.
- [Gács, 1974] P. Gács. On the symmetry of algorithmic information. *Soviet Math. Dokl.*, 15:1477–1480, 1974.
- [González, 2008] Cristóbal Rojas González. *Randomness and Ergodic Theory: An Algorithmic Point of View*. PhD thesis, École Polytechnique, Paris, France, and Università di Pisa, Italy, 2008.
- [Hjorth and Nies, 2007] G. Hjorth and A. Nies. Randomness via effective descriptive set theory. *Journal of the London Mathematical Society*, 75(2):495, 2007.
- [Hoyrup, 2008] Mathieu Hoyrup. *Computability, Randomness and Ergodic Theory on Metric Spaces*. PhD thesis, University Paris Diderot, France, and Università di Pisa, Italy, 2008.
- [Jauch, 1990] J. M. Jauch. *Are Quanta Real? A Galilean Dialogue*. Indiana University Press, 1990.
- [Knuth, 1998] D.E. Knuth. *The Art of Computer Programming, Volume 2: Seminumerical Algorithms, 3rd ed.* Addison-Wesley Reading, Mass, 1998.
- [Kolmogorov, 1963] A.N. Kolmogorov. On tables of random numbers. *Sankhyā, Ser. A*, 25:369–376, 1963.
- [Kolmogorov, 1965] A.N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems Inform. Transmission*, 1(1):1–7, 1965.
- [Kolmogorov, 1974] A.N. Kolmogorov. Laws of information conservation (non-growth) and aspects of the foundation of probability theory. *Problems Inform. Transmission*, 10:206–210, 1974.
- [Kuipers and Niederreiter, 1974] L. Kuipers and H. Niederreiter. *Uniform distribution of sequences*. John Wiley & Sons, New York, 1974.
- [Li and Vitanyi, 2008] M. Li and P. Vitanyi. *An introduction to Kolmogorov complexity and its applications*. Springer, 3rd edition, 2008.
- [Longo, 2009] G. Longo. Randomness and Determination, from Physics and Computing towards Biology. In *SOFSEM 2009: Theory and Practice of Computer Science: 35th Conference on Current Trends in Theory and Practice of Computer Science, Špindleruv Mlýn, Czech Republic, January 24–30, 2009. Proceedings*, page 49. Springer, 2009.
- [MacHale, 1993] D. MacHale. *Comic Sections: The book of mathematical jokes, humour, wit, and wisdom*. Boole Press, Dublin, 1993.
- [Martin-Löf, 1966] P. Martin-Löf. The definition of random sequences. *Inform. and Control*, 9:602–619, 1966.
- [Merkle et al., 2006] W. Merkle, J.S. Miller, A. Nies, J. Reimann, and F. Stephan. Kolmogorov–Loveland randomness and stochasticity. *Annals of pure and applied logic*, 138(1-3):183–210, 2006.
- [Miller and Nies, 2006] J.S. Miller and A. Nies. Randomness and computability: open questions. *Bulletin of Symbolic Logic*, pages 390–410, 2006.
- [Mlodinow, 2008] L. Mlodinow. *The Drunkard’s Walk: How Randomness Rules Our Lives*. Pantheon Books, New York, 2008.
- [Moschovakis, 1980] Y. N. Moschovakis. *Descriptive Set Theory*, volume 100 of *Studies in Logic and the Foundations of Mathematics*. North-Holland Publishing Company, Amsterdam, 1980.
- [Nies, 2009] A. Nies. *Computability and Randomness*. Oxford University Press, 2009.

- [Odifreddi, 1992] P. Odifreddi. *Classical Recursion Theory*, volume 125 of *Studies in Logic and the Foundations of Mathematics*. North-Holland (Elsevier), Amsterdam, 1992.
- [Penrose, 1989] R. Penrose. *The Emperor's New Mind*. Oxford University Press, New York, 1989.
- [Rogers Jr, 1987] H. Rogers Jr. *Theory of recursive functions and effective computability*. MIT Press, Cambridge, Mass., 1987.
- [Shen *et al.*, 20??] A. K. Shen, Uspensky V. A., and Vereshchagin N. K. *Kolmogorov Complexity and Randomness*. To appear, 20??.
- [Sierpinski, 1917] W. Sierpinski. Demonstrationelementaire du theoreme de M. Borel sur les nombres absolument normaux et determination effective dun tel nombre. *Bull. Soc. Math. France*, 45:127–132, 1917.
- [Sipser, 1997] M. Sipser. *Introduction to the Theory of Computation*. PWS Publishing Company, 1997.
- [Solomonoff, 1960] R.J. Solomonoff. A preliminary report on a general theory of inductive inference. Technical Report ZTB-138, Zator Company, Cambridge, Mass., 1960. (November, 1960).
- [Solomonoff, 1964] R.J. Solomonoff. A formal theory of inductive inference, part 1 and part 2. *Inform. Contr.*, 7:1–22, and 224–254, 1964.
- [Stewart, 2002] I. Stewart. *Does God Play Dice?: The New Mathematics of Chaos*. Blackwell Publishers, 2002.
- [Svozil, 1993] K. Svozil. *Randomness & undecidability in physics*. World Scientific, 1993.
- [Taleb, 2005] N.N. Taleb. *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. Random House Trade Paperbacks, New York, 2005.
- [Tromp, 2009] J. Tromp. Binary Lambda Calculus and Combinatory Logic. <http://homepages.cwi.nl/~tromp/cl/cl.html>, 2009. PostScript file of paper updated on March 13, 2009.
- [Turing, 1992] A.M. Turing. A note on normal numbers. *Collected Works of AM Turing, Pure Mathematics, North-Holland, Amsterdam*, pages 117–119, 1992.
- [Uspensky, 1983] V.A. Uspensky. *Post's Machine*. Little Mathematics Library. Mir Publishers, Moscow, 1983.
- [van der Waerden, 1927] B. L. van der Waerden. Beweis einer baudetschen vermutung. 15:212–216, 1927.
- [van Lambalgen, 1987a] M. van Lambalgen. *Random Sequences*. PhD thesis, University of Amsterdam, 1987.
- [van Lambalgen, 1987b] M. van Lambalgen. Von Mises' definition of random sequences reconsidered. *Journal of Symbolic Logic*, 4:725–755, 1987.
- [van Lambalgen, 1989] M. van Lambalgen. Algorithmic information theory. *Journal of Symbolic Logic*, pages 1389–1400, 1989.
- [van Lambalgen, 1990] M. van Lambalgen. The axiomatization of randomness. *Journal of Symbolic Logic*, pages 1143–1167, 1990.
- [van Lambalgen, 1996] M. van Lambalgen. Von Mises' axiomatization of randomness reconsidered. In LS Shapley TS Ferguson and JB MacQueen, editors, *Statistics, Probability and Game Theory, papers in honor of David Blackwell*, volume 30 of *IMS Lecture Notes and Monograph series (Hayward, CA)*, 1996.
- [Ville, 1939] J. Ville. *Etude Critique du Concept de Collectif*. Gauthier-Villars, Paris, 1939.
- [Volchan, 2002] Sergio B. Volchan. What is a random sequence? 109:46–63, 2002.
- [Von Mises, 1919] R. Von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Math. Zeitschrift*, 5:52–99, 1919.
- [Von Mises, 1981] R. Von Mises. *Probability, statistics and truth (1957)*. Dover Publications, New York, 1981.
- [V'yugin, 1997] Vladimir V. V'yugin. Effective convergence in probability and an ergodic theorem for individual random sequences. *SIAM Theory of Probability and Its Applications*, 42(1):39–50, 1997.
- [V'yugin, 1999] Vladimir V. V'yugin. Ergodic theorems for individual random sequences. *Theoretical Computer Science*, 207(4):343–361, 1999.
- [Wald, 1936] A. Wald. Sur la notion de collectif dans le calcul des probabilités. *C. R. Acad. Sci.*, 202:1080–1083, 1936.
- [Yurtsever, 2000] U. Yurtsever. Quantum mechanics and algorithmic randomness. *Arxiv preprint: arXiv:quant-ph/9806059v2*, pages 1–8, 2000.

Part VII

Probabilistic and Statistical Paradoxes

This page intentionally left blank

PARADOXES OF PROBABILITY

Susan Vineberg

1 INTRODUCTION

Numerous puzzles have arisen from the application of probability. These range from simple cases that are easily resolved to others that have generated considerable philosophical discussion and remain topics of lively debate. Quite a few classic puzzles arise from basic misapplications of the standard axioms of probability. While the mistakes involved are common, and are indicative of how prone humans are to fallacious probabilistic reasoning, the resolution of such problems, while instructive, is for the most part non-controversial. The solution to some of these involves a bit more than elementary probability theory, but the principles needed are fairly unproblematic. Other cases involve more substantive principles of probability and/or decision. In many instances there are generally accepted resolutions, although the principles involved are more open to question than the basic axioms of probability. As such, these, along with the very simple problems involving clear mistakes of reasoning, do not seem very paradoxical. However, other problems lack a single clear solution, and involve a number of hotly contested issues.

It will be useful in introducing the subject of probabilistic paradoxes to begin by examining a family of puzzles known as Bertrand's paradoxes, as well as some variations that have been proposed by others. These vary in terms of the principles that they may be said to depend upon, and accordingly vary considerably in terms of their ease of resolution. After examining these, the concept of paradox will be considered, which will give rise to a way of categorizing the so-called paradoxes of probability. A wide variety of paradoxes will then be considered by type.

1.1 *Bertrand's Paradoxes*

Each of Bertrand's paradoxes turns on alternative ways of counting up possibilities. This links these paradoxes with the notorious principle of indifference that forms the basis for the classical interpretation of probability. The principle of indifference requires that

Events for which we have no reason to expect (favor) one over the other are to be taken as equipossible.

The classical interpretation of probability then dictates that the probability of an event is the ratio of favorable cases among equipossible ones, as given by the principle of indifference. The alternative ways of counting up possibilities lead to inconsistent assignments of probability, which has generally been taken as posing a problem for the classical interpretation. However, Bertrand's paradoxes vary in the pressure that they put on the principle of indifference. Some of these indicate the need for mild reform, whereas others suggest that the principle, and hence the classical interpretation of probability, may be simply unworkable.

The first case is easily resolved, and does not put serious pressure on the principle of indifference. It involves the following setup involving a box with three drawers, and accordingly is called Bertrand's box paradox:

One drawer contains two gold coins, one contains two silver coins, and one contains one gold and one silver coin. A drawer is chosen at random, and a silver coin is drawn from that drawer. What is the probability that it came from the drawer with two silver coins?¹

There are two ways to count alternatives in this problem, and accordingly we obtain two answers:

Answer 1: Given that a silver coin was drawn, it came from either the drawer with one silver and one gold coin, or the drawer with two silver coins. We have no reason to prefer one over the other, so the probability that the coin came from the drawer with two silver coins is $1/2$.

Answer 2: The silver coin drawn was either the one silver coin in the drawer with one silver and one gold coin, or it was the first silver coin in the drawer with two silver coins or it was the second silver coin in that drawer. There is no reason to favor any of these possibilities over the others, but two are favorable, so the probability is $2/3$.

Here we have two different answers to the question that derive from the principle of indifference. Each answer has some plausibility, as evidenced by the fact that when the puzzle is posed to students and others, it is common to hear each answer defended in the ways given above. The conflicting answers suggest a problem, and the principle of indifference may be thought to be the culprit. Where the principle is read as requiring equal probability for events where there is no recognized reason for favoring one possibility over another, then it leads to both answers. However, while those who propose answer 1 may fail to recognize a reason to favor the drawer with one silver coin over that with two, there is in fact evidence for the latter. This is a consequence of some basic assumptions of the problem.

¹There are various equivalent formulations of Bertrand's box problem. A common version (the three coins) involves a hat containing three quarters. One is an ordinary quarter, one is two-headed and the third has two tails. A quarter falls out, heads up. What is the probability that it has two heads? Yet another version (The three card game) involves three cards. One is black on both sides, one is white on both sides and one is black on one side and white on the other. Given that one draws a card that is black on one side, what is the probability that the other side is black?

Let S be the claim that a silver coin is drawn.

Let D_{2S} be the claim that the drawer with two silver coins is chosen.

Let D_{1S} be the claim that the drawer with one silver coin and one gold coin is chosen.

Bayes' Theorem yields:

$$\begin{aligned} P(D_{2S}/S) &= P(S/D_{2S}) \times P(D_{2S})/P(S) = 1 \times (1/3)/1/2 = 2/3 \\ P(D_{1S}/S) &= P(S/D_{1S}) \times P(D_{1S})/P(S) = 1/2 \times (1/3)/1/2 = 1/3 \end{aligned}$$

Drawing a silver coin is thus evidence that the drawer with two silver coins was selected, relative to the basic assumptions of the problem, namely that each drawer initially has an equal chance of being selected and each coin in the chosen drawer has an equal chance of being selected. Thus, the reasoning in answer 1 that proceeds from taking the alternatives to be that one drew from either the drawer with one silver coin or from the drawer with two silver coins to the conclusion that each is equally likely is mistaken. A simple application of Bayes' theorem, which is a consequence of the basic axioms of probability, together with the idea that E is evidence for H iff $P(H/E) > P(H)$, resolves the conflict between the two answers by revealing the faulty reasoning in answer one. This example involves receiving evidence when a silver coin is drawn, and thus some principle of evidential relevance is required. Of course, the idea that E is evidence for H iff $P(H/E) > P(H)$ is a cornerstone of Bayesianism, and is applied to what may be considered prior probabilities. However, there is no controversy in this case about what these are, nor any reason to balk here at the criterion of evidence, and so we have a satisfactory resolution of the problem.

There remains the issue that the faulty answer seems to stem from an application of the principle of indifference. However, rather than repudiating the principle, Bertrand's box shows that it must be read as requiring not that equal probability be assigned when one recognizes no difference between two options, but where there is no available evidence that distinguishes between them. While those who accept answer 1 might recognize no difference between the alternatives that one has drawn from the drawer with one gold and one silver coin, and the drawer with two silver coins, the basic assumptions of the problem show via Bayes' theorem that drawing a silver coin should be taken as evidence of having drawn from the drawer with two silver coins.

The resolution for Bertrand's box problem does not work in other of the so-called Bertrand cases where the principle of indifference seems to lead to conflicting answers, but in which no evidence is received. Consider drawing balls from an urn whose contents are unknown, and the probability of drawing a red ball.² The alternatives can be taken to be that one draws a red ball or that one draws a ball that is not red. It seems that there is no evidence to favor one over the other, thus the principle of indifference, as clarified above, gives a probability for drawing a red ball of one half. But, if we apply the same reasoning to drawing a blue ball

²This example is discussed in [Kyburg, 1970].

and then to drawing a green ball, we get $Pr(\text{red}) = Pr(\text{blue}) = Pr(\text{green}) = 1/2$, in violation of the probability axioms, which require that the probability of all the possibilities together sum to one. It may be thought however that there are more ways in which a ball can fail to be red, i.e. it can be blue, or green, or yellow, etc., than it can be red.³ This provides a reason not to treat drawing a red ball as equipossible with all such other possibilities together, although it would be a stretch to suggest that this background provides evidence for thinking it is more likely that the ball is not red, than that it is. Rather this case is one in which we lack evidence altogether.

Instead of appealing to some objective evidence to distinguish between possibilities, the suggestion has been made that the conflicting answers delivered by the principle of indifference can be avoided by insisting that consideration of the possibilities fully respect the symmetries in the problem. When the possibility of drawing a red ball is compared just with that of drawing a non-red ball, the symmetry with drawing some other colored ball is ignored. Here invoking symmetry considerations can be used to block the applications of the principle of indifference leading to incoherence. However, unlike in the case of Bertrand's box, we are not left with a definitive probability for drawing a red ball. Here it may be tempting to suppose, as subjective Bayesianism would have it, that one may rationally adopt any one of a variety of priors for the probability of selecting a red ball. But in the face of little real basis for assigning any particular value, perhaps one should not assign any definite probability or even a non-trivial range of probabilities. Without some assumed background that delimits the possibilities, the question itself seems ill posed. The incoherence that arises from applying the principle of indifference first to the possibility of drawing a red ball, and then to drawing blue and green balls, can be seen as stemming from an increasing set of possibilities. Fixing a probability for drawing a red ball requires specifying a set of alternatives, which may just be that the ball is non-red, but once the space of alternatives is divided into more specific possibilities, we are asking for the probability of drawing a red ball in a different scenario. When the background conditions delimit the alternatives to drawing a blue or green ball, symmetry requires that the alternative to a red ball is not that it is non-red, but that it may be green or blue. With the alternatives fixed in this way, the principle of indifference does not lead to incoherence, and it seems reasonable enough to hold in accordance with it that $Pr(\text{red}) = Pr(\text{blue}) = Pr(\text{green}) = 1/3$. While the paradox proper dissolves, we may still wonder about the legitimacy of supposing that the colors are evenly distributed.

There are other examples in which symmetry considerations can be invoked to avoid the conflicting answers of the unqualified principle of indifference, two of which will be noted here. One is Van Fraassen's cube factory in which a factory produces cubes with side-length between 0 and 1 foot.⁴ The first question asks for the probability that a cube chosen at random has a side length between 0 and 1/2 square feet? Following the principle of indifference, the answer appears to be 1/2.

³[Weatherford, 1982]

⁴See [van Fraassen, 1989; Hájek, 2007].

If we ask instead for the probability that a randomly chosen cube has face area between 0 and $\frac{1}{4}$ square feet, thinking of the cubes as constructed in a way that is uniformly distributed over face area, the answer would seem to be $\frac{1}{4}$. But, if the question is posed in terms of the probability of selecting a randomly chosen cube with volume between 0 and $\frac{1}{8}$ cubic feet, we get an answer of $\frac{1}{8}$. Here applying the principle of indifference to the different descriptions yields different probabilities for one and the same state, since the cubes with side between 0 and $\frac{1}{2}$ feet, and those with area between 0 and $\frac{1}{4}$ square feet and those with volume $\frac{1}{8}$ cu feet are the same cubes.

Perhaps the most famous of Bertrand's problems concerns the probability given a circle that a randomly chosen chord is longer than the side of an inscribed equilateral triangle. There are three variables that seem uniformly distributed, which lead by the principle of indifference to three different answers:

Answer 1: By rotating any such triangle, it may be assumed that one end of the chord lies on the vertex of the triangle. Any chord whose other end lies between the other vertices of the triangle will be longer than the side of the triangle. The endpoint of any such chord lies on an arc with $\frac{1}{3}$ the circumference of the circle, so the probability would seem to be $\frac{1}{3}$.

Answer 2: By rotating the triangle it may be assumed that the chord is perpendicular to one side of the triangle. Next consider a radius of the circle that is perpendicular to the side of the triangle to which the chord is parallel. The midpoint of the chord will then lie on that radius. The chord is longer than the side of the triangle, just in case it is closer to the center of the circle than is the side of the triangle. The side of the triangle must bisect the radius and so the probability is $\frac{1}{2}$.

Answer 3: Given a point within the circle, consider a chord that has the point as a midpoint. The chord will be longer than a side of the inscribed triangle provided that the point falls within an inner circle of radius $\frac{1}{2}$ that of the original circle. The inner circle has area $\frac{1}{4}$ that of the original circle and so the probability is $\frac{1}{4}$.

Poincare and E.T. Jaynes both argued that geometric symmetries dictate the second answer [Jaynes, 1973]. Van Fraassen applies the idea to the cube problem, showing that under measures that are dilation invariant [that is measures m such that for positive k , $m(a, b) = m(ka, kb)$], the principle of indifference yields the same probability if we consider cube length and cube area. However, as Jaynes and van Fraassen note, there remain Bertrand-like cases that are apparently immune to such maneuvers. In an example due to Von Mises, we consider a glass with 10oz of wine and water. Suppose at least 1 oz is water and 1oz is wine. What is the probability that the glass contains at least 5oz of water? If we consider the proportion of water to the total and then the proportion of wine to water, we find, as in the case of the cubes, that a uniform distribution on the former leads to a non-uniform distribution on the later. However, in this case, there appears to be nothing that fixes an invariant measure that will yield a unique answer.

2 PUZZLES AND PARADOXES

We have seen that the puzzle created by Bertrand's box is easily resolved. There are ways of avoiding inconsistent answers using the principle of indifference in the case of the colored balls, the cube factory, and Bertrand's chord paradox, although these solutions do still leave some questions. However, the water and wine case remains a problem for the principle of indifference. An option that is often suggested is that the principle of indifference should simply be abandoned. Still, the principle retains some appeal even though it is not quite clear how exactly it could be reformed so as to avoid all of the difficulties that its application generates. It is the seeming plausibility of the principle that makes these cases seem paradoxical. But the fact that the problem in each case may be resisted by limiting or giving up the principle of indifference suggests that these are not to be regarded as cases of genuine paradoxes.⁵ Settling this requires considering the question of what constitutes a paradox.

Sainsbury [1988] characterizes a paradox as

an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises.

In the case of a paradox, appearances deceive as either the premises, the reasoning, or the claim that the conclusion is unacceptable must be defective. Often paradoxes are characterized as involving a contradiction reasoned to from intuitively plausible premises. This is essentially just a variation on Sainsbury's characterization, because where the conclusion is unacceptable, its denial can be offered as a plausible premise, yielding a contradiction. Cases such as Bertrand's box in which two or more conflicting answers are offered can be seen as fitting this pattern insofar as we are willing to take the assumptions of each as premises of a single argument. However, in many instances, this one included, those who are tempted by one answer typically reject the others and their premises, so that they cannot be assimilated to Sainsbury's definition. Instead, we would do better to identify a subclass of paradoxes in which there are two or more conflicting answers to a problem, each involving seemingly acceptable reasoning from plausible premises, where it is at least initially unclear which premises and/or which reasoning should be discarded. This subclass can be covered by maintaining that a paradox

leads to contradictory conclusions from assumptions that have a claim to reasonableness by way of seemingly correct reasoning.

⁵Adding to this are doubts that we should generally expect uniform distributions in cases where we have no reason to suppose that a random process is at work in producing that distribution. This concern applies in the case of drawing balls out of an urn where it is unclear as to the proportion of red balls. Thus, there are two distinct reasons to think that the principle of indifference is flawed. The fact that it can be applied so as to lead to multiple answers is clearly a defect. It must be discarded if it cannot be delimited in all cases so as to avoid this. Beyond this though, it seems that it is not compulsory in cases where there is no reason to assume that the distribution is the result of a random process.

Some will see this characterization as too broad, as it counts as paradoxes cases in which the apparent unacceptability of the premises, conclusions and reasoning is easily resolved, as in the case of Bertrand's box. Rather it might be held that a genuine paradox is one in which an unacceptable conclusion is derived from not just assumptions that are initially plausible, but which are so intuitive or seemingly well supported that they are strongly resistant to revision. Examples of non-probabilistic paradoxes that have this character include at least some versions of the liar and sorites paradoxes. Whether any of the probability puzzles that have been discussed in the literature are genuine paradoxes in this sense remains an open question. While there are lots of puzzles that have been discussed that are quite easily resolved, such as Bertrand's box paradox, and some others that are more complex, but nonetheless seem rightly classified as less than fully paradoxical, including each of the other Bertrand paradoxes, there are a number of puzzles for which there remains substantial disagreement, and thus might rise to the level of full paradox. However, it may be that the disagreements arise simply because there are various aspects of probability that remain poorly understood, and that even those puzzles that have spawned considerable research in recent years will eventually lose their paradoxical status in a way that the liar and sorites paradoxes have continued to resist.

Despite the fact that the above definition will count as paradoxical those puzzles of probability that are easily resolved, and thus perhaps unworthy of the designation, it is nevertheless worthwhile to it adopt here. On this definition, a paradox of probability can be understood as

leading to contradictory claims from assumptions concerning probability that have a claim to reasonableness by way of seemingly correct reasoning, where the reasoning may itself be probabilistic.

It is a virtue that this counts as paradoxical those puzzles involving probability that are often referred to as paradoxes of probability, which include those that are in fact rather easily resolved. Nevertheless, even many such simple problems of probability have the capacity to perplex, and so it seems acceptable to count them as paradoxes. More importantly, this characterization gives rise to a natural way of categorizing the probabilistic paradoxes in terms of the content of the premises and the type of probabilistic reasoning involved in generating the paradox.

A number of paradoxes are easily resolved by consideration of simple deductions from the basic axioms of probability, sometimes in conjunction with Bayes' rule. We have already seen one example of this in Bertrand's box, and others will be noted below. Of course, people frequently violate the rules of probability in ordinary reasoning, as Tversky and Kahneman [1974] have shown, but this does not suggest that the probability axioms fail to capture correct probabilistic reasoning. However, the fact that such reasoning violating the probability axioms seems correct to many supports the broad definition of paradox that would include those puzzles that arise from simple errors of probabilistic reasoning.

A second class of paradoxes involves more substantive probabilistic assumptions

and reasoning. These paradoxes cover those involving infinite spaces, partitioning issues, and principles of evidence. Among these are the doomsday paradox, Simpson's paradox and the Sleeping Beauty problem. The Bertrand paradoxes that do not turn on a failure to account for the evidence, in which symmetry issues arise, belong in this category.

A third class of paradoxes involves not only probabilistic assumptions and/or reasoning, but also the assignment of utilities and the computation of expected value. These include, the St. Petersburg and related paradoxes, the two envelope problem, and Newcomb's problem. In each of these cases, it is the computation of expected value that is at least sometimes fingered as the source of the problem.

3 SIMPLE PUZZLES OF PROBABILITY

3.1 *The Monty Hall Problem*

One of the most famous puzzles involving probability derives from the classic game show in which the host, Monty Hall, presents a contestant with 3 doors and announces that there is a prize behind only one of the doors. The contestant chooses a door, say door 1, after which Monty offers to show what is behind one of the doors that the contestant did not select. Monty then reveals that there is no prize behind one of the other doors, say door 2, and offers the contestant the chance to switch to door 3. The question that is usually posed is whether the contestant should switch, but we may also ask what the contestant's probability for winning should be after Monty reveals that there is no prize behind door 2.⁶ It is assumed here that there is an equal chance that the prize was placed behind each of the three doors. This can be taken to follow from the principle of indifference or else can be built into the problem, so as to avoid any reliance on the principle. It is to be assumed that Monty knows where the prize is and that he will always act on this information to show a door that does not contain a prize. The contestant is taken to be aware of this. The first reaction that many have is that the probability that the prize is behind door 1 should remain the same as the probability that it is behind door 3. Accordingly many people think that the probability is $\frac{1}{2}$ for the two remaining doors and so there is no point to switching. It is reasoned that since you know that Monty will reveal what is behind one of the doors not selected, learning that there is no prize behind door 2 (as opposed to learning this about 3) provides no relevant information about whether the prize is behind door 1. But this is mistaken. Given the assumptions above, the probability that the prize is behind door 1 ($D1$) is indeterminate after Monty shows that there is no prize behind door 2. However, if it is assumed that if $D1$ is true, Monty uses, say, the toss of a fair coin to determine whether he reveals what is behind door 2 or

⁶A structurally identical problem involves three prisoners, in which two are to be executed and the other set free, and the warden gives one of the prisoners the name of another prisoner who is to be executed.

door 3, then Bayes' rule can be applied to calculate the posterior probability for $D1$:

Let $D1, D2, D3$, be the propositions that the prize is behind doors 1, 2 and 3 respectively, and let $S2, S3$ be the propositions that Monty reveals the contents behind 2 and 3. Let P_o be the probability before Monty reveals the contents of one of the doors and P_1 the probability afterwards.

$$\begin{aligned} P_o(D1) &= P_o(D2) = P_o(D3) \\ P_1(D1) &= P_o(D1/S2) = P_o(D1 \& S2)/P_o(S2) = \frac{1/6}{1/2} = 1/3 \\ P_1(D3) &= P_o(D3/S2) = P_o(D3 \& S2)/P_o(S2) = \frac{1/3}{1/2} = 2/3 \end{aligned}$$

With the assumption that Monty chooses randomly between doors 2 and 3, when he is unconstrained in his choice of those two by the location of the prize, it follows that being shown that there is no prize behind door 2 is positively relevant to the prize being behind door 3. This establishes that it is wrong to think that because it is known that Monty will reveal a door, that revealing a particular door cannot provide evidence as to which door the prize is behind. However, if Monty does not choose randomly in the case where he is unconstrained by the location of the choice, the evidential relevance of his revealing no prize behind door 2 can differ. For example, suppose it is known that Monty will always reveal door 2 except where he is constrained not to because the prize is behind door 2. In this case the posteriors are as follows:

$$\begin{aligned} P_1(D1) &= P_o(D1/S2) = P_o(D1 \& S2)/P_o(S2) = \frac{1/3}{2/3} = 1/2 \\ P_1(D3) &= P_o(D3/S2) = P_o(D3 \& S2)/P_o(S2) = \frac{1/3}{2/3} = 1/2 \end{aligned}$$

Since Monty will show door 2 here except where he is constrained not to, doing so does not offer evidence to distinguish between $D1$ and $D3$. This scenario establishes an upper bound on the probability that may reasonably be attributed to $D1$ after Monty reveals one of the other doors. Of course, Monty could adopt a rule of never showing door 2 unless he was constrained to do so. Assuming that the contestant knows this, then it would be known that the prize is behind 3 once 2 is revealed. It might be specified that this possibility is to be excluded by the rules of the game, but short of a procedure that would make the location of the prize certain once one of the doors not chosen is revealed, Monty could still adopt a rule that would make it virtually certain that the prize is behind door 3 in the event that he reveals 2. In any case, where Monty's procedure for choosing a door to reveal is unspecified, $0 < P_1(D1) \leq 1/2$. Since the value of $P_1(D1)$ never exceeds a half and may be less, even where Monty's rule is unknown, the contestant should always switch, unless he knows that Monty will always reveal door 2 except when constrained not to, and even in that case there is no harm in switching.

The Monty Hall problem has been much discussed, in part because of the strong initial temptation to think that Monty does not provide evidence that distinguishes

between the two remaining doors when he opens one of the doors not chosen. There have even been quite a few mathematicians and statisticians who have insisted that the probability of the prize being behind the contestant's chosen door moves from $1/3$ to $1/2$. Despite this, the puzzle belongs in the category of those that are easily resolved using the probability axioms and the concept of evidential relevance. As in the structurally similar case of Bertrand's box, an application of Bayes' rule, using only the basic assumptions of the problem, yields a definitive answer.

3.2 *The Birthday Paradox*

A problem with an answer that is surprising to many involves the chance that at least two people in a group have the same birthday. Given a group of more than 22 people, there is a greater than 50-50 chance that two have the same birthday. In a group of 10, there is a greater than 10% chance of this. These facts can be shown via a straightforward calculation involving the possible combinations:

Let n be the # of people in the group. Ignoring the possibility of a leap year birthday, the probability is given by:

$$1 - 365!/365^n(365 - n)!$$

Although many imagine that the chance is actually much less than 50-50 of two people in a group of 23 having the same birthday, the answer itself is entirely uncontroversial. The complexity of the formula for determining the probability suggests that people are not relying on anything like the formula in estimating the probability, but rather a much simpler conceptualization, which in effect discounts many of the possible combinations. Perhaps the reasoning proceeds something like this: It is highly improbable that any two people chosen at random have the same birthday, so it is imagined that any two such people have different birthdays. Of course, assuming that those two people have different birthdays, a third person chosen at random is highly unlikely to share either of their birthdays. Continuing on in this way, the possibility that two in 23 have the same birthday still looks unlikely. But such reasoning involves discounting the very small probability that two people chosen at random have the same birthday, and the cumulative effect of small probabilities as the number of people increases.

This suggestion as to the source of the error in the birthday paradox is somewhat similar to one of the assumptions that generates the lottery and preface paradoxes. In the lottery paradox, it is assumed that a ticket is purchased from a large number of tickets, one of which is assured of winning. Given the large number of tickets, it seems reasonable to accept of any given ticket that it will not win. From this it appears to follow that none of the tickets will win, contradicting the assumption that one ticket will win.⁷ This paradox has generated considerable discussion. Many would resolve it by questioning the assumption that any given ticket will

⁷The preface paradox is just a variation of the lottery paradox that stems from the practice of taking responsibility for the errors in one's book in the preface. The paradox arises because

not win, arguing that each should only be taken to have a very low probability of winning, and that the assumption that a given ticket will not win does not follow from this. It is something like this rejected assumption that appears to be at work in the birthday paradox. The lottery paradox is nevertheless quite different from the birthday paradox, particularly in that there is no general consensus as to its resolution. Rather than rejecting the assumption that any given ticket will not win, some would block the paradox by rejecting the idea that it follows from the acceptance of the claim that say ticket #1 will lose and acceptance of the claim that ticket #2 will lose, that it should be accepted that both will lose. Thus the lottery paradox does not belong in the category of the easily resolved. Perhaps more importantly, it differs from other paradoxes involving probability, in that it is not principles of probability or probabilistic relevance that are at stake, but rather the relationship between degrees of belief, as measured by probabilities, and acceptance.

3.3 *The Sibling Mystery*

There are several questions involving the chances that a person's sibling is a girl that are easily confused. These are often collectively referred to as the sibling mystery or mysteries, (or the boy-girl paradoxes). The first question asks, given a two-child family in which the older child is a boy, for the probability that his younger sibling is a girl. It is to be assumed that the chance of any given child being a girl is 50%; the facts that the ratio of boys to girls in the population is not exactly 50-50, that the sex of a given child is not entirely independent of that of his or her older siblings, and that some children are intersex, are ignored here. Given these assumptions, the answer to the first question is $1/2$. The mystery comes with the second question, which asks for the probability that a boy from a two-child family has a sister. This looks like the first question, but the answer is different.

The background assumption of both problems is that a boy comes from a two-child family. The possible orderings of boys (*B*) and girls (*G*) in such families are

$$BB, BG, GB, GG$$

By assumption, each is equally probable. The information that the boy is the older of two children rules out the second two possibilities. Since the first two remain equally likely, the probability that the younger sibling is a girl is $1/2$. But learning that a boy is from a two-child family rules out only the fourth case. Since in two of the three remaining cases, the boy has a sister, the answer to the second question is $2/3$. One reason this may seem puzzling at first is that if such a boy is an older sibling, the probability that he has a sister is $1/2$, and if such a boy is a younger sibling, the probability that he has a sister is $1/2$, but if he is either the younger or the older of two siblings, the probability that he has a sister is $2/3$.

the author accepts each of the claims in the book, yet in acknowledging the likelihood that the book contains errors, appears committed to accepting that all the claims in the book are true, and also that they are not.

Things change again if it is assumed that one randomly encounters a boy who has one sibling. It may seem that the probability that he has a sister is the same as in the second example above, but this is incorrect. Given the information that the boy is from a two-child family, we know that he is either the older brother of his sister, the older brother of his brother, the younger brother of his sister or the younger brother of his brother. Each case is equally probable, so the probability that he has a sister is $1/2$. The key here is that in a random draw of a boy from two-child families, one is twice as likely to get a boy from a family of type BB than from type BG or GB .

Such cases are not truly paradoxical. They merely illustrate that care must be taken in specifying the group from which a random selection is made. In the first two cases the draw is over family types, and then some evidence as to which type we have is given. This is perhaps clearer in variants of the first two questions in which a woman is asked about her children. She says she has two, which makes it appropriate to consider random draws from the set $\{BB, BG, GB, GG\}$. This contrasts with the last case in which the draw is from individuals rather than families. Sometimes the problem begins with a boy saying that he has one sibling, and then the probability that he has a sister is asked for. It is unsurprising when posed in this way that both answers are given, for it is unclear whether the problem should be treated as a random encounter of individuals from two-child families. It is also now not so perplexing that the probability that a boy whose one younger sibling is a girl is $1/2$, whereas the probability that a boy from a two-child family has a sister is $2/3$. A random draw from $\{BB, BG, GB\}$ is not the same as a draw from either $\{BB, BG\}$ or a draw from $\{BB, GB\}$.

4 MORE COMPLEX PROBLEMS OF PROBABILITY

4.1 *Simpson's paradox*

A perplexing fact about the association of variables in a population, termed Simpson's paradox, concerns the ability to reverse the association within a sub-population. In a much discussed example involving applicants to graduate school at the University of California, Berkeley, it was observed that women had a higher rejection rate than men overall, while having a lower rejection rate department by department.⁸ Although the overall higher rejection rate suggests the possibility of gender discrimination, this is countered by the fact that admission is determined by evaluation at the departmental level, and in each department the rejection rate for women was actually equal to or lower than that for men. There is a simple explanation that accounts for the overall higher rejection rate, despite the lower rate department by department that is supported by the data in this case, namely that women tended to apply to departments with higher rejection rates.

The interesting fact is that given an association, it is always possible to reverse the correlation in some sub-population. This is a consequence of the fact that for

⁸See [Bickel *et al.*, 1977].

whole numbers it is possible to have:

$$\begin{aligned} a/b < a'/b', c/d < c'/d', \text{ but} \\ a + c/b + d > a' + c'/b' + d' \end{aligned}$$

For discussion see [Malinas and Bigelow, 2004]. While this is perhaps surprising, once it is understood that this simple mathematical fact lies at the heart of the reversed associations, it is clear that these are not truly paradoxical. However, the possibility of reversing an association of variables by repartitioning raises questions about interpreting correlations. Cartwright [1983] called attention to the paradox, arguing that it indicates that one cannot simply use such correlations to represent causal relations, and that some basic causal relationships are needed to interpret the significance of correlations.

4.2 *The Doomsday and Shooting Room Paradoxes*

The shooting room paradox [Leslie, 1996] involves a room in which each person or group of people entering encounter an executioner who rolls a pair of dice. If he rolls double sixes, everyone in the group is killed. If not, they are free to go. It is also the case that at least 9/10 of those entering the room die, which is made possible by the fact that people enter the room in rounds, in increasing numbers, until double sixes are rolled. With probability one, this will occur eventually, and at that time at least ninety percent who have entered in one of the rounds will be in the room during the final, fatal, round. If the executioner rolls a double six on the first round, 100% of those entering the room die, but otherwise the number entering the room increases so that 90% of those entering the room are killed. The paradox arises because assuming that an individual enters the room after the first round, it appears that there are two distinct probabilities that may be attached to the proposition that he will die, namely 1/36, which reflects the chance of double-sixes, and .9, which reflects the percentage of people entering the room who die. However, given the increasing numbers entering the room, the two numbers are compatible. Since it is the role of a double six that determines whether a given individual in the room dies, the chance for an individual of dying is 1/36.

A related paradox draws the conclusion that doomsday will likely come soon, from the premise that I am now alive. Leslie [1990] presents the argument roughly as follows:

If doom is far in the future, then I have very early birth order among all humans, which is unlikely. However, if doom is sooner, given that there is a fairly high percentage of humans that have ever lived living now, it follows that I am a more typical human. Thus, the evidence that I am alive now favors doom sooner over doom later.

The shooting room paradox, in which the largest group of people to enter the room occurs in the final round, is really a kind of special case of the Doomsday

paradox. However, the more general doomsday paradox lacks the clear mechanism that allows for a definitive resolution. In a discussion of Leslie's version of the doomsday argument, Sober [2003] takes it as an argument about likelihoods. The idea is that an observation O favors H_1 over H_2 just in case O is more probable given H_1 than given H_2 , i.e. $P(O/H_1) > P(O/H_2)$. Assuming H_1 and H_2 are the hypotheses of doom sooner and later respectively, the claim is that my being alive now favors early doom. However, Sober objects to Leslie's assumption that $P(O/H_1) > P(O/H_2)$ because it is based on a suspect, a priori, argument that involves treating my place in the temporal birth order as a lottery. Sober thinks this is inappropriate because there is no underlying chance process that can be taken to be at work in producing birth order, but more importantly he takes Leslie's probabilistic claims to be empirically disconfirmed.

Bartha and Hitchcock accept that birth order can be modeled as a lottery, however they use Bayes' theorem to calculate posterior probabilities for H_1 and H_2 [Bartha and Hitchcock, 1999]. They take there to be an asymmetry in the application of Bayes' theorem that leads to the favoring of doom sooner. However, they note that the lottery that Leslie appeals to over possible birth orders presupposes that I exist, which makes me special. They argue that in calculating the posterior probability for H_1 and H_2 , we must first conditionalize on the assumption that I exist at all, and this favors doom later, which they claim cancels out the evidence of doom sooner by dint of my likely place in the birth order. There appear to be various ways of blocking the implausible claim that my existence now makes the demise of the human race more likely to occur sooner rather than later.

Although none of the problems here involve assumptions that are so intuitive that we face a particularly gripping paradox, they each raise questions about the appropriate application of probabilistic principles. They highlight the importance of choosing the right reference class in assessing probabilistic relevance and both the shooting room and Simpson's paradox show the importance of looking to correlations that reflect causally relevant factors. More generally Sober's analysis of the doomsday paradox highlights the idea that the correlations we employ must reflect the sort of process at work that bears upon the outcome of the event at issue. It is a general problem to formulate principles that determine the relevant probabilities. As the next puzzle indicates, there is considerable controversy over such principles.

4.3 *Sleeping Beauty*

The Sleeping Beauty problem centers around the following scenario: Beauty will participate in an experiment in which she will be put into a deep sleep on Sunday night. She will be awakened during the experiment either once or twice depending on the toss of a fair coin. In either case she will be awakened on Monday. She will be unaware upon awakening that it is Monday, but subsequently will be told that it is Monday and then given drugs to put her back to sleep. These drugs will erase her memory of having been awake. For definiteness, it is assumed that the coin is

tossed on Monday night after Beauty is put back to sleep. If the coin lands heads, Beauty will then sleep until Wednesday, at which point the experiment ends. If the coin lands tails, she will be awakened briefly again on Tuesday, unaware of the Monday awakening, and then put back into the deep sleep for the remainder of the experiment. Beauty knows all of this, and will be able to distinguish waking up within the experiment from other awakenings, so that when she awakens during the experiment she will know that it is either Monday or Tuesday.

Before being put to sleep, it is presumably reasonable for Beauty to assign a probability of $1/2$ to the proposition that the coin lands heads. Now, let us assume that the experiment has begun and Beauty has just awakened from a deep sleep. What should her probability be now for heads? There are two plausible answers, each of which has been defended by multiple authors who offer various arguments for their answers. The disagreement centers over whether waking up within the experiment should be seen as evidentially relevant to heads. Elga, who introduced the problem into the philosophical literature, argues that it is, and that Beauty's probability for heads upon awakening should be $1/3$ [Elga, 2000].⁹

Elga starts with the assumption that upon awakening, Beauty is in one of the following scenarios:

H_1 : Heads and it is Monday.

T_1 : Tails and it is Monday.

T_2 : Tails and it is Tuesday.

Let P_s be Beauty's probability function before being put to sleep on Sunday, P_a her probability function upon awakening, and P_m her probability function after being told that it is Monday. Elga argues that being in T_1 is qualitatively like being in T_2 , and hence that $P_a(T_1) = P_a(T_2)$. Elga claims that upon learning that it is Monday, Beauty should update her probabilities by conditionalizing on this information, which can be represented as $H_1 \vee T_1$, and reasons that once Beauty is told that it is Monday, she should attach a probability of $1/2$ to Heads (after all the coin has not yet been tossed). Thus, $P_m(\text{heads}) = P_a(\text{heads}/(H_1 \vee T_1)) = P_a(H_1/(H_1 \vee T_1)) = 1/2$. From this it follows that $P_a(H_1) = P_a(T_1)$, and hence that $P_a(H_1) = P_a(T_1) = P_a(T_2) = 1/3$, and so $P_a(\text{heads}) = 1/3$.

Lewis [2001] argues against Elga that since Beauty knew in advance that she would be awakened, waking up within the experiment conveys no new information that is relevant to the probability of heads and so it should remain at $1/2$. However, he accepts Elga's framework and assumptions except for the claims that $H_1 \vee T_1 \vee T_2$ is relevant to heads and that $P_m(\text{heads}) = 1/2$. From the premise that upon awakening Beauty learns nothing relevant to the probability of heads, Lewis argues that $P_a(\text{heads}) = 1/2$, and paradoxically that $P_m(\text{heads}) = 2/3$. The latter follows from the former claim that $P_a(\text{heads}) = 1/2$, together with the shared assumptions that

$$P_a(\text{heads}/(H_1 \vee T_1)) = P_a(H_1/(H_1 \vee T_1)) = P_a(H_1)/(P_a(H_1) + P_a(T_1)),$$

⁹There is a similar problem in which a driver (or passenger) travels a circular route forgetting whether he has passed through a given intersection [Aumann *et al.*, 1997].

and that

$$P_a(\text{tails}) = P_a(T_1) + P_a(T_2).$$

Dorr [2002] points out that Lewis and Elga's common assumptions that

1. $P_m(\text{heads}) = P_a(\text{heads}/(H_1 \vee T_1))$
2. $P_a(\text{heads}/-(H_1 \vee T_1)) = 0$
3. $0 < P_a(H_1 \vee T_1) < 1$

together with

4. $P_m(H_1) = 1/2$
5. $P_a(H_1) = 1/2$

are probabilistically inconsistent. Elga avoids the inconsistency by rejecting (5), whereas Lewis does so by rejecting (4). But this still leaves Lewis with the un-intuitive consequence that $P_m(\text{heads}) = 2/3$. Instead, one might give up (1), the assumption that one should update by conditionalization upon learning that it is Monday, as Joel Pust [2008] has suggested. This allows one to hold both that no information relevant to heads is gained when Beauty awakens, but also that learning that it is Monday is not relevant to heads either, which is highly plausible. To be sure, Beauty's belief shifts over the course of the experiment cannot be characterized exclusively in terms of conditionalization, since she changes her belief from zero to one in certain propositions, such as that it is Sunday. The inadequacy of conditionalization in dealing with beliefs that change due to loss of information and the acquisition of self-locating beliefs demands a more comprehensive treatment of belief change, and some recent attempts to grapple with this have consequences for the Sleeping Beauty problem. Titelbaum [2008] presents a framework that delivers the answer of $1/3$, whereas Halpern [2005] presents a system in which Beauty's degree of belief upon awakening is $1/2$. Such work shows that the conflicting answers to the Sleeping Beauty problem are not merely an artifact of the Elga and Lewis' representations, and may provide some basis for adjudicating the intuitions that have guided a variety of arguments concerning the problem. One of the key intuitions behind the answer of $1/3$ concerns the fact that if the experiment were to be repeated, in the long run there be would twice as many tails awakenings as heads awakenings, and this suggests that the answer should be $1/3$. However, this answer brings with it various difficulties as well.¹⁰

5 PARADOXES OF PROBABILITY AND DECISION

The probability puzzles considered thus far involve applications of probability in various situations, but do not involve utility. Of course the Monty Hall problem

¹⁰See [White, 2006].

concerns a decision about whether one should switch doors, but this turns upon just the credences that one should have about where the prize is. Utilities, and the issue of how they combine with credences in making rational decisions, do figure prominently in some problems, including the two envelopes and St. Petersburg paradoxes.

5.1 Newcomb's Problem

In this problem, an extraordinarily good predictor of behavior offers you a choice between selecting one opaque box and selecting both the opaque box and a transparent box, which you can see contains one thousand dollars. If he predicted that you would select only the opaque box, then he placed one million dollars in the box. If he predicted that you would select both boxes, then he left the opaque box empty. The situation is summed up in the following decision matrix:

	\$1million in opaque box	\$0 in opaque box
Take both boxes	\$1,001,000	\$1,000
Take opaque box only	\$1,000,000	0

There is some controversy as to the rational choice, but most think that you should take both boxes; the amount of money in the opaque box is fixed, and regardless of what it contains, you will be \$1000 richer if you take both boxes. In addition to the question as to the right choice, we can ask about what rational choice theory prescribes. The rational act is usually taken to be the one that maximizes expected value. While the calculation is generally unproblematic, in this case different versions of decision theory offer different answers. According to Jeffrey's evidential decision theory, one should compute expected value in terms of the conditional probabilities of the outcomes given the acts [Jeffrey, 1983]. Let T be the proposition that both boxes are taken, O that just the opaque box is taken, and M that there is a million dollars in the opaque box. The expected values as computed by Jeffrey's theory are as follows:

$$\begin{aligned}\text{EXP}(T) &= 1,001,000 \times P(M/T) + 1000P(-M/T) \\ \text{EXP}(O) &= 1,000,000 \times P(M/O) + 0 \times (-M/O)\end{aligned}$$

With $P(M/T) \ll P(M/O)$ as given, we have $\text{EXP}(O) > \text{EXP}(T)$. However, this presents a difficulty for Jeffrey's evidential decision theory, since this involves choosing a dominated act. Most decision theorists (including Jeffrey) accept that the right decision is to take both boxes, and accordingly some have suggested

various alternative ways of computing the expected value of an act, which prescribe taking both boxes in the Newcomb problem. Indeed, the problem has proven to be not so much of a paradox, but a spur that has driven the development of causal decision theory. The basic idea stems from observing that the probability of a state occurring may well be high conditional on some act without the act having causal relevance to whether the state occurs. In this case, the probability of ending up with nothing in the opaque box is very high conditional on choosing both boxes, although the choice is causally irrelevant to whether one million dollars is placed in the box. Causal decision theory solves the problem by replacing the conditional probabilities with causal conditional probabilities in calculating expected value, that is, the conditional probability of a state given an act is replaced by the probability that the act will result in the state in question.¹¹

Considered as a problem of decision theory, Newcomb's problem does not seem paradoxical given the development of causal decision theory. However, we can regenerate something of a paradox, by reflecting on the one box solution. One can regard this solution as motivated by the idea that choosing one box goes along, at least in the long run, with being the sort of person whom the predictor will recognize as a one boxer and for whom he will place one million in the opaque box. It seems at least somewhat paradoxical that one should choose an act (taking both boxes) that one regards as not being in accordance with the act, which at least in the long run, can be expected to produce greater expected gains. As the two-boxers correctly point out, the fact that it is in one's interest to be the sort of person who would choose one box does not entail that this is the right decision once the amount of money in the opaque box has been fixed. However, there still appears to be some tension between the general prescription to take both boxes and being the sort of person who picks one box. Of course, those who advocate taking both boxes in the Newcomb problem need not be committed to adopting any sort of general strategy of always taking two boxes, where this would have a causal influence on future plays of the game. Indeed, it is an assumption of the two-box solution that the problem concerns a one-play game, or at any rate that taking both boxes will not have a causal effect on future plays. Even if there is a conflict with assuming that what one does in each play of a Newcomb game is causally irrelevant to subsequent plays, there is seemingly no problem in supposing that taking both boxes in a given play is causally independent of how much money is in the opaque box, and this suffices for the prescription that one should take both boxes.

5.2 *The Cable Guy*

Another apparent problem for decision theory is discussed by Hájek [2005]. The cable guy will install your new cable between 8am and 4pm. You are offered the chance to bet that he will come either during the morning interval from (8 to 12]

¹¹Versions of causal decision theory and the relationships between them are discussed in [Joyce, 1999].

or during the afternoon from (12 to 4). The morning interval includes 12 noon, and so contains an extra moment, but we can assume that the probability that the cable guy comes right at noon is zero, and so we can take the two intervals to be of the same duration. It seems that you should be indifferent between betting on the morning or afternoon. Assuming that the payoff is the same for each bet, your expected value for the two bets should be equal. However, if you bet on the morning interval, there is certain to be some time at which you will regret having chosen to bet on the morning, since the cable guy is to arrive at some point after 8 am. But then, despite the fact that it seems one should be indifferent between betting on the two intervals, one should bet on the afternoon, since betting on the morning interval violates the following plausible principle of rational decision:

Avoid Certain Frustration Principle—Given a choice between two options you should not choose an option for which you are certain that a rational future self will prefer that you had chosen the other, unless both options have this property. [Hájek, 2005]

As stated, the paradox turns on the acceptability of this principle. Although it seems reasonable, Hájek takes the cable guy paradox as showing that it is mistaken.¹² While the principle may often apply, it does seem that it might be overridden in this case. Closer examination of the principle itself suggests another reason for thinking that it is false. Notice that when there is some time t , however fleeting, for which a future act A will certainly be rationally preferred to B , yet many times at which B will almost certainly be preferred to A , the principle requires choosing A , regardless of the likelihood of generally preferring B to A .

Although it seems reasonable to reject the principle to avoid certain frustration, one could also reject the idea that it applies here by suggesting that the certain regret upon which it depends is simply not rational. The cable guy problem is not particularly paradoxical, because its assumptions turn out not to be so plausible after all, although there may still be some question about which of these should be rejected.

5.3 *The St. Petersburg Paradox*

The St. Petersburg paradox arises from a game in which a fair coin is flipped until it comes up tails. The payoff is 2^n , where n is the number of flips. Take a run of the game to be a sequence of $n-1$ heads followed by tails. Since there are an infinite number of possible runs of the game, one for each natural number n , each having positive probability, the expected value of playing the game is infinite. So, it appears rational to pay any finite sum to play the game. The paradox arises from the fact, that a very low payoff is highly probable. It seems reasonable to be unwilling to pay more than a relatively small amount to play a game with a high probability of yielding a low payout.

¹²His response is seconded and further discussed in [Kierland *et al.*, 2008].

One response to the paradox claims that unwillingness to pay more than a small sum to play can be accounted for by appealing to risk aversion.¹³ One trouble with this is that risk aversion is far from universal. If risk aversion is present, then at least to some extent it can be compensated for by offering higher payoffs. In particular, if a person is only willing to pay a very small sum to enter the game, then classical decision theory assumes that this amount can be increased by a corresponding adjustment to the payoffs [Martin, 2004]. Another response is simply to stipulate that there is an upper bound on the value of the consequences of any act. Jeffrey endorses this noting: “Our rebuttal of the St. Petersburg paradox consists in the remark that anyone who offers to let the agent play the St. Petersburg game is a liar, for he is pretending to have an indefinitely large bank” [Jeffrey, 1983, p. 154].

A related puzzle involves the so-called “Pasadena Game” [Nover and Hájek, 2004]. As in the St. Petersburg paradox, a fair coin is tossed until it lands heads, but instead the payoffs grow as $2/n$ and alternate between positive and negative values. This game appears to present a new problem for decision theory. The trouble is that the game’s expectation is a conditionally convergent series and so can be reordered so as to sum to any real number, or to diverge to either positive or negative infinity. For discussion see [Nover and Hájek, 2004].

5.4 Two Envelopes

In some respects the most paradoxical of the decision theoretic puzzles is the two envelopes (or exchange) problem [Jackson *et al.*, 1994].¹⁴ The basic set-up involves two envelopes, with one known to contain twice as much money as the other. You can select one envelope and keep its contents. Suppose it (the envelope on the right) contains $\$X$ dollars. Assuming each envelope has an equal chance of containing the greater amount, the expected value of switching envelopes Y is

$$\frac{1}{2}(X/2) + \frac{1}{2}2X = 5/4X$$

Thus, the basic decision theoretic principle that one should act so as to maximize expected value seems to require switching. Indeed, it suggests that one should be willing to pay a premium to do so. This is paradoxical, since the two envelopes are completely symmetric and so one has no reason to favor one envelope as having the greater amount over the other. Moreover, it seems that one could apply the same reasoning again, so as to justify switching back and forth indefinitely. Additionally, we can imagine a variation in which there are two players, where each is assigned an envelope at random. Each player’s expected value for switching apparently exceeds that of sticking, and yet this is a zero-sum game, and so they cannot both have an advantage in doing so.

¹³See [Weirich, 1984].

¹⁴[Broome, 1995; McGrew *et al.*, 1997; Rawling, 1997; Norton, 1998; Horgan, 2000; Fallis, 2008].

Suppose that after choosing you are permitted to look in the contents of your envelope, and that you find \$100. The expected value of switching is \$125, and so you should switch. There is nothing paradoxical about this. Of course, one might not want to switch due to the decreasing marginal value of money or risk aversion, but putting such considerations aside, switching is the appropriate choice. Here there is no argument for switching back to the first envelope, since sticking now has an expected value of \$125, whereas switching back has an expected value of \$100.

It is now apparent that there is really no opportunity to turn players in the two-envelope game into money pumps. When a player looks into his envelope, he acquires a reason to switch. However, there is no expected gain in switching back, so he will be unwilling to pay to do so. But, if he does not look in his envelope, and it could contain any amount of money, he will not know what amount it would be reasonable to pay to switch. Still, there remains something of a paradox because it seems that one should switch in advance of looking in the envelope, and having done so to switch back. As various authors have pointed out, this requires the assumption that there is no upper bound on the amount of money in the envelopes [Jackson *et al.*, 1994; Broome, 1995; Norton, 1998]. If one had a large sum in one's envelope, then it would be reasonable to think it very likely the other envelope contains the smaller amount. Thus for particular values of X , the probability of the other envelope containing $2X$ will not be $1/2$.

There are clearly many prior distributions for the sum of money in the envelopes such that looking in the right envelope would result in something other than a probability of $1/2$ that the other (left) envelope contains twice (half) of what is in the right envelope. Given such a distribution, the paradox dissolves, and it would be reasonable to have just such a distribution in a typical situation in which one is presented with two envelopes and reliably told that one contains twice the other. However, as Broome points out [Broome, 1995], not all prior distributions are like this, i.e. there are some distributions that determine probabilities $y = 2x$ and $y = x/2$, such that the expected value of y on x (switching) is greater than x . This means that we can regenerate the paradox by building in assumptions guaranteeing that the expected value of y on x (switching) is greater than x . There is then a strengthened two envelope paradox in which it is a given that upon opening one of the envelopes you would think that the other is equally likely to contain half or twice as much money [Fallis, 2008].

The stipulation that it is equally likely that the other envelope contains half or twice as much money as the one selected brings with it certain complications. This requires that there is no maximum amount of money that can be placed in the envelopes, but also a uniform distribution over infinitely many states, which cannot be captured by a standard probability function. John Norton [1998] develops a modified version of the paradox that is more mathematically tractable in which the probability distribution fails to be uniform over infinitely many states, but which satisfies the key assumption that there is a positive expectation of gain in switching for any definite amount that is placed in the envelopes. However, as

Norton observes, where the amount in the envelope is unknown, the infinite sum of the expectation fails to exist. Thus, for each particular amount, the expectation is finite and so one should switch, but there is no recommendation to switch overall, and thus there is a failure of dominance [Norton, 1998; Rawling, 1997]. The failure of the expectation to converge in the strengthened two envelope paradox shows that it is related to the St. Petersburg and Pasadena games [Broome, 1995; Norton, 1998].

5.5 *Other Problems of Decision*

There are many other paradoxes involving probability that arise in decision theory. Among the mostly widely discussed are the Prisoner's dilemma and the Allais paradox. The prisoner's dilemma involves a two-person game in which the best strategy for each individual is such that when both players act upon it, a nonoptimal outcome results. The Allais paradox involves preferences for options that many people apparently endorse, but that violate the Independence axiom, which requires that one's preferences for acts depend only on those states in which their consequences differ. While it is called a paradox, it differs markedly from the others considered here. In particular, the conflict does not derive from intuitively plausible assumptions alone, but rather from a clash between these and some empirical findings about what some seemingly rational people prefer. As such it is not a traditional paradox, despite its name.

In the prisoner's dilemma and Allais paradoxes there seems to be little controversy about the probabilistic assumptions involved. Instead these paradoxes raise complex philosophical issues concerning the axioms of rational preference and the principle of utility maximization. Of course, each of the paradoxes in this section involves the concept of expected value, and many of them suggest that some qualification, or refinement, is needed in order to avoid paradox. Many agree though that the remedies involve adjustments, rather than radical changes to decision theory. As noted, one response to the St. Petersburg paradox is to insist that there must be an upper bound on the value of any game. For any actual case that one might encounter, this restriction is unimportant. However, problems such as the prisoner's dilemma and the Allais paradox have lead some to reject the fundamental principles of rational preference and rational choice. As such they are not merely small puzzles, but the sort of problems that like Russell's paradox, which led to a rethinking of the foundations of set theory, have prompted the development of diverse theories of rationality.

ACKNOWLEDGEMENTS

Thanks to an anonymous referee for helpful comments on an earlier draft of this paper.

BIBLIOGRAPHY

- [Aumann *et al.*, 1997] R. J. Aumann, S. Hart, and M. Perry. The Forgetful Passenger, *Games and Economic Behavior* 20:117-120, 1997.
- [Bartha and Hitchcock, 1999] P. Bartha and C. Hitchcock. No One Knows the Date or the Hour: An Unorthodox Application of Rev. Bayes's Theorem, *Philosophy of Science* 66 (3):S339-S353, 1999.
- [Bickel *et al.*, 1977] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex Bias in Graduate Admissions: Data from Berkeley, in William B. Fairley and Frederick Mosteller (eds.), *Statistics and Public Policy*, Reading: Addison-Wesley, 1977.
- [Broome, 1995] J. Broome. The Two-envelope Paradox, *Analysis* 55 (1):6-11, 1995.
- [Cartwright, 1983] N. Cartwright. *How the Laws of Physics Lie*. Oxford: Oxford University Press, 1983.
- [Clark, 2002] M. Clark. *Paradoxes from A to Z*. New York: Routledge, 2002.
- [Dorr, 2002] C. Dorr. Sleeping Beauty: in defense of Elga, *Analysis* 62 (4):292-296, 2002.
- [Elga, 2000] A. Elga. Self-locating Belief and the Sleeping Beauty Problem, *Analysis* 60 (2):143-147, 2000.
- [Fallis, 2008] D. Fallis. Resolving the Strengthened Two Envelope Paradox, presented at the *American Philosophical Association, Central Division Meeting*, Chicago, 2008.
- [Hájek, 2005] A. Hájek. The Cable Guy Paradox, *Analysis* 65:112-119, 2005.
- [Hájek, 2007] A. Hájek. *Interpretations of Probability* (Winter 2007). <http://plato.stanford.edu/archives/win2007/entries/probability-interpret/>
- [Halpern, 2005] J. Y. Halpern. Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems, in T. S. Gendler and J. Hawthorne (eds.), *Oxford Studies in Epistemology* Oxford University Press, 111-142, 2005.
- [Horgan, 2000] T. Horgan. The Two-Envelope Paradox, Nonstandard Expected Utility, and the Intensionality of Probability, *Nous* 34:578-603, 2000.
- [Hunter and Madachy, 1975] H. A. H. Hunter and J. H. Madachy. *Mathematical Diversions*. Toronto: Dover, 1975.
- [Jackson *et al.*, 1994] F. Jackson, P. Menzies, and G. Oppy. The Two Envelope Paradox, *Analysis* 54:43-45, 1994.
- [Jaynes, 1973] E. T. Jaynes. The Well-Posed Problem, *Foundations of Physics* 3:477-492, 1973.
- [Jeffrey, 1983] R. C. Jeffrey. *The Logic of Decision*. Second ed. Chicago: The University of Chicago Press, 1983.
- [Joyce, 1999] J. M. Joyce. *The Foundations of Causal Decision Theory*, *Cambridge Studies in Probability, Induction and Decision Theory*. Cambridge: Cambridge University Press, 1999.
- [Kierland *et al.*, 2008] B. Kierland, B. Monton, and S. Ruhmkorff. Avoiding Certain Frustration, Reflection, and the Cable Guy Paradox, *Philosophical Studies* 138:317-333, 2008.
- [Kyburg, 1970] H. E. Kyburg, Jr. *Probability and Inductive Logic*. Toronto: The Macmillan Company, 1970.
- [Leslie, 1990] J. Leslie. Is the End of the World Nigh, *The Philosophical Quarterly* 40 (158):65-72, 1990.
- [Leslie, 1996] J. Leslie. *The End of the World*. London: Routledge, 1996.
- [Lewis, 2001] D. Lewis. Sleeping Beauty: reply to Elga, *Analysis* 61 (3):171-176, 2001.
- [Malinas and Bigelow, 2004] G. Malinas and J. Bigelow. *Simpson's Paradox* (Spring 2004). <http://plato.stanford.edu/archives/spr2004/entries/paradox-simpson/>
- [Martin, 2004] R. Martin. *The St. Petersburg Paradox* (2004), E. Zalta, ed. <http://plato.stanford.edu/archives/fall2004/entries/paradox-stpetersburg/>
- [McGrew *et al.*, 1997] T. J. McGrew, D. Shier, and H. S. Silverstein. The Two-Envelope Paradox Resolved, *Analysis* 57 (1):28-33, 1997.
- [Norton, 1998] J. D. Norton. When the Sum of Our Expectations Fails Us: The Exchange Paradox, *Pacific Philosophical Quarterly* 79:34-58, 1998.
- [Nover and Hájek, 2004] H. Nover and A. Hájek. Vexing Expectations, *Mind* 113:237-249, 2004.
- [Pust, 2008] J. Pust. Sleeping Beauty, Conditionalization, and Knowledge De Praesenti, presented at the *American Philosophical Association, Pacific Division Meeting*, Pasadena, CA, 2008.
- [Rawling, 1997] P. Rawling. Perspectives on a Pair of Envelopes, *Theory and Decision* 43:253-277, 1997.

- [Sainsbury, 1988] R. M. Sainsbury. *Paradoxes*. Cambridge: Cambridge University Press, 1988.
- [Sober, 2003] E. Sober. An Empirical Critique of Two Versions of the Doomsday Argument-Gott's Line and Leslie's Wedge, *Synthese* 135:415-430, 2003.
- [Titelbaum, 2008] M. G. Titelbaum. The Relevance of Self-Locating Beliefs, *Philosophical Review*, 2008.
- [Tversky and Kahneman, 1974] A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases, 185:1124-1131, 1974.
- [van Fraassen, 1989] B. C. van Fraassen. *Laws and Symmetry*. New York: Oxford University Press, 1989.
- [Weatherford, 1982] R. Weatherford. *Philosophical Foundations of Probability Theory*. London: Routledge and Kegan Paul, 1982.
- [Weirich, 1984] P. Weirich. The St. Petersburg Gamble and Risk, *Theory and Decision* 17:193-202, 1984.
- [White, 2006] R. White. The Generalized Sleeping Beauty Problem: A Challenge for Thirders, *Analysis* 66:114-119, 2006.

STATISTICAL PARADOXES: TAKE IT TO THE LIMIT

C. Andy Tsao

1 INTRODUCTION

There are two kinds of paradoxes. The first kind arises from confusion. People think there is paradox but actually there is none once the confusion is clarified. The second kind arises from a conflict among principles we find compelling. There is a fundamental tension in our thinking brought out by these paradoxes. We can not simply resolve them by deepening our understanding. Rather, we have to recognize an uncomfortable truth. Naturally, the paradoxes of the second kind are more interesting. Nonetheless, the taxonomy of paradoxes are usually as difficult as the paradoxes themselves.

In this article, we will discuss two statistical paradoxes: Lindley's paradox and the Fieller-Creasy Problem — see [Lindley, 1957] and [Fieller, 1954] respectively. Both arose in the early 60's yet they are not fully resolved so far as we know. They are paradoxes of limits. We think it is worthwhile to note that: mathematics may not be as rigorous as we hope it to be — it can be quite tricky, especially when we push it to the limit.

Our statistical paradoxes can be illustrated by the following example. Consider the length of Cleopatra's nose as θ_C (cm). We are interested in knowing:

- What is θ_C ?
- What is the possible range of θ_C ?

These can be answered by the statistical theory of point estimation and confidence intervals. Moreover, we might also like to know

1. Is $\theta_C = 15(cm)$?
2. Is θ_C equal to θ_E , the nose length of Elizabeth Taylor? Formally
 - Is $D = \theta_E - \theta_C = 0$?
 - Is $\rho = \theta_E/\theta_C = 1$?

These are typical problems of hypothesis testing. The first is a question for the (population) mean of one population while the second is one for the comparison of means from two populations. Note that there are at least two ways of representing the equality, the difference of zero or the ratio of 1.

In classical (or frequentist) statistics, θ_C is considered as a parameter, a fixed yet unknown quantity. We do not know what the true value of θ_C is; hence resort to the observation. Suppose we can measure Cleopatra's nose n times in an independent manner. The commonly assumed model is that $X_1, X_2, \dots, X_n$ are independently identically distributed as a normal random variable with mean θ_C and variance $\sigma^2 > 0$, abbreviated as $X_1, \dots, X_n \sim_{iid} N(\theta_C, \sigma^2)$. Similarly, we assume $Y_1, \dots, Y_m \sim_{iid} N(\theta_E, \sigma_E^2)$, observations of the length of nose of Ms. Taylor.

Lindley's paradox occurs in hypothesis testing problems where the null hypothesis is just one value (point null hypothesis). For example, for testing $\theta = 15$ versus $\theta \neq 15$, Lindley's paradox suggests that the p-value is more liberal than its Bayesian contenders. In other words, based on the same observation, the classical statistical evidence that the nose length is 15 cm differs from those Bayesian evidences (posterior probabilities with respect to some priors). Moreover, it contradicts the common belief that the classical statistical procedures are objective and more conservative than the Bayesian procedures. In the context of hypothesis testing, conservative tests are harder to reject the null hypothesis than their other contenders. The Fieller-Creasy problem highlights the problems about formulation and how a sound classical/frequentist statistical procedure can have problematic properties. First, the problem formulation of noses of equal lengths affects choices of the suitable statistical procedures: equality represented by $D = 0$ or by $\rho = 1$ will lead to different tests and, consequently, different conclusions. In addition, the confidence intervals for the ratio are quite problematic. [Gleser and Hwang, 1987] shows that for a 95% confidence interval for ρ , the ratio of nose lengths, has a positive probability being the entire parameter space $(-\infty, \infty)$ or sets excluding a neighborhood of 0.

Lindley's paradox and the Fieller-Creasy problem have been widely discussed ever since their discovery. Our paper is not intended to be an exhaustive review. Instead, we hope to highlight some interesting perspectives on the paradoxes and, hopefully, remind philosophers, statisticians and users of statistics of these uncomfortable facets of statistics.

2 FREQUENTIST VS. BAYESIAN

Lindley's paradox and the Fieller-Creasy problem are important illustrations of the Frequentist-Bayesian discrepancy. The discrepancy starts with the different interpretations of probability. Explicitly, frequentists and Bayesians treat the uncertainty about the parameter and the randomness of the sampling data in different ways. These differences in turn affect the choices of statistical procedures and the guarantee of these procedures.

For easy exposition, we will assume the normality. Let $X_1, \dots, X_n \sim_{iid} N(\theta, \sigma^2)$, random sample from a normal probability density function (pdf), $f(x|\theta)$, with unknown parameter $\theta \in \Theta = (-\infty, \infty)$ and known $\sigma^2 > 0$. Typical statistical problem formulations are point estimation (estimate what θ is), confidence interval (an interval estimate of θ) and hypotheses testing (which hypothesis about θ is more likely).

The frequentist considers the parameter θ a unknown yet fixed quantity and $X_1, \dots, X_n$ as one random sample (of size n) from one of the possible infinite experiments. The commonly recommended estimator of θ

$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n). \quad (1)$$

A $(1 - \alpha)$ confidence interval of θ is

$$C(X) = [\bar{X} - z_{\alpha/2}\sigma_n, \bar{X} + z_{\alpha/2}\sigma_n] \quad (2)$$

where z_α such that $P(Z < z_\alpha) = 1 - \alpha$, $\sigma_n^2 = \sigma^2/n$ and Z is a $N(0, 1)$ random variable.

For testing

$$H_0 : \theta = \theta_0 \quad vs. \quad H_0 : \theta \neq \theta_0 \quad (3)$$

The recommended α level test rejects H_0 when

$$\frac{|\bar{X} - \theta_0|}{\sigma_n} > z_{\alpha/2}. \quad (4)$$

In addition for testing (3), the p-value is defined as

$$p(x) = P\left(\frac{|\bar{X} - \theta_0|}{\sigma_n} > \frac{|\bar{x} - \theta_0|}{\sigma_n}\right) = P\left(|Z + \Delta| > \frac{|\bar{x} - \theta_0|}{\sigma_n}\right) \quad (5)$$

where $\Delta = \frac{\theta - \theta_0}{\sigma_n}$.

These frequentist statistical procedures are “recommended” because they have good frequentist properties. To assess whether (1) estimate θ well, (2) covers θ or (4) makes accurate decision, one needs to know the true value of θ . However, because the parameter is unknown, there is no way to verify how these procedures (for point estimation, etc) perform with respect to the current observed data $x_1, \dots, x_n$. The notational difference is worth noting: X 's refer to the random observation while x 's refer to the observed value of X 's in the given experiment. To discuss the guarantees or properties, frequentists resort to the long-run average performances, for example, unbiasedness in point estimation, confidence coefficient (minimum coverage probability) and significance level (maximum of probability of Type I error). All of them interpret probability as long-run frequency via Law of Large Number (LLN)

$$E(X) = \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K X^{(k)} \quad (6)$$

where $X^{(k)}$, an iid sample of $f(x|\theta)$ from the k -th experiment.

The probability of event A can be viewed as the expectation of $1_{[X \in A]}$, the indicator function of A , hence

$$P(A) = E1_{[X \in A]} = \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K 1_{[X^{(k)} \in A]}. \quad (7)$$

Similarly

$$E(\bar{X}) = \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K \bar{X}^{(k)} \quad (8)$$

where $\bar{X}^{(k)} = \frac{1}{n}(X_1^{(k)} + \cdots + X_n^{(k)})$, an iid sample of size n from the k -th experiment. It guarantees that the estimator $\bar{X}$ will hit the true θ when averaging over applications in the long-run. The confidence interval (2) has $(1 - \alpha)$ *confidence coefficient*, that is,

$$\min_{\theta \in \Theta} P_{\theta}(\theta \in C(X)) = 1 - \alpha. \quad (9)$$

This indicates that $C(X)$ will cover the true θ about $(1 - \alpha)100\%$ in the long-run applications in repeated experiments. The α level test (4) guarantees that

$$\max_{\theta \in \Theta_0} P_{\theta}(\text{Reject } H_0) = \alpha. \quad (10)$$

Define Θ as the parameter space (collection of all possible values of θ) and Θ_0 as the null space (the collection of all θ such that the null hypothesis is true). Recall that all of the probabilities and the expectation are interpreted as long run frequencies in the frequentist inference framework. Moreover, we would like to point out that the minimum and maximum (or, more rigorously, infimum and supremum) in the definitions of (9) and (10) reflect the guarantee under the worst scenario. In this sense, classical/frequentist statistical procedures are built to be conservative.

Bayesians formulate the uncertainty of the parameter θ through a prior $\pi(\theta)$. Under our normal setting, one of the mathematically convenient priors is the conjugate prior $\pi(\theta) \sim N(\mu, \tau^2)$ where μ is a real number and $\tau^2 > 0$. Upon observing $x = (x_1, \dots, x_n)'$, the posterior pdf of θ can be derived by Bayes Theorem

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_{\theta \in \Theta} \pi(\theta)f(x|\theta)d\theta} \sim N(\mu_x, \rho^{-1})$$

where

$$\mu_x = \frac{1}{\rho} \left(\frac{\mu}{\tau^2} + \frac{\bar{x}}{\sigma_n^2} \right) = \frac{\sigma_n^2 \mu + \tau^2 \bar{x}}{\sigma_n^2 + \tau^2} \text{ and} \quad (11)$$

$$\rho = \frac{1}{\tau^2} + \frac{1}{\sigma_n^2} = \frac{\sigma_n^2 + \tau^2}{\sigma_n^2 \tau^2}. \quad (12)$$

where $\Theta = (-\infty, \infty)$, the parameter space.

For most Bayesians, the posterior distribution contains all the information needed for statistical inferences. Under the squared error loss, the Bayes estimate of θ with respect to $\pi(\theta)$, is the posterior mean $E_{\pi(\theta|x)}\theta = \mu_x$. The expression of μ_x in (11) can be viewed as a weighted average of sample mean $\bar{x}$ and prior mean μ . In addition, for very large sample size n , μ_x is essentially $\bar{x}$. On the other hand, if τ^2 goes to infinity, μ_x becomes $\bar{x}$; τ^2 goes to zero μ_x becomes μ . In terms of estimating θ under the current normal-normal setting, the Bayes point estimate is μ_x and the frequentist point estimate is $\bar{x}$. This is a perfect illustration of widely held intuition/belief: as the (prior) information diffuses or a “non-informative” prior is used, the Bayes inference coincides with the frequentist inference; Bayesian inference is similar to the frequentist one when sample size is very large.

The Bayes interval estimate analogous to the frequentist confidence interval for θ is the Bayes credible set. We say $C_\pi(x)$ is a $(1 - \alpha)$ credible set for θ if

$$P_{\pi(\theta|x)}(\theta \in C_\pi(x)) = 1 - \alpha. \quad (13)$$

To some Bayesians, it is of little point to consider hypothesis testing problems. They argue that the posterior distribution provides better and complete picture regarding θ . Nonetheless, others develop tests in the Bayesian framework. Roughly speaking, Bayesian tests are based on posterior probability of Θ_0 and Θ_1 ,

$$P_{\pi(\theta|x)}(\theta \in \Theta_0) \text{ and } P_{\pi(\theta|x)}(\theta \in \Theta_1) \quad (14)$$

through their ratio (posterior odd) or Bayes factor. Typically, if $P_{\pi(\theta|x)}(\theta \in \Theta_1)$ is (much) larger than $P_{\pi(\theta|x)}(\theta \in \Theta_0)$, Bayesians will reject H_0 . In the simple versus simple hypothesis testing problem, Bayesian testing procedures again are similar to frequentist testing procedures: Bayesian tests depend on the a weighted likelihood ratio and the frequentist one depends on likelihood ratio.

3 LINDLEY'S PARADOX

Hypotheses testing is a well-established problem formulation of statistics. It is widely used in many applied fields and p-values are implemented in almost all statistical software as standard outputs. Casual users of statistics might think of the hypothesis tests as synonymous with p-values. This seems quite natural for them to have this impression. Standard statistical textbooks claim that the p-value is the data-dependent evidence against the null hypothesis. Statistical

packages include p-values rather than specifying the significance level. Academic journal papers of most fields lavishly quote p-values as statistical support for their scientific discovery. It is common practice that comparing p-values with 0.05 for “significant” scientific results.

Lindley’s paradox can be viewed as a paradox about p-values. It casts a shadow on the practice of hypotheses testing. Namely,

- P-value is NOT conservative statistical evidence (against the null hypothesis) compared with its Bayes contenders (with respect to some reasonable family of priors)
- Frequentist and Bayesian statistical measures do not necessarily coincide with each other even for diffuse priors

Specifically, let $X_1, \dots, X_n \sim_{iid} N(\theta, \sigma^2)$ with unknown $\theta \in \Theta = \mathcal{R}$ and known $\sigma^2 > 0$. Consider the hypothesis testing problems:

$$H_0 : \theta = 0 \quad vs. \quad H_1 : \theta \neq 0. \quad (15)$$

Note that so far as two-sided hypotheses testing is concerned, without loss of generality, any point null hypothesis $\theta = \theta_0$ can be translated to $\theta = 0$. It is well recognized that this formulation might be unrealistic but nonetheless provides great simplification. In addition, the assignment of prior probability on 0 can be problematic for Bayesians. The more appropriate *imprecise hypothesis testing* problem has been proposed.

$$H_0^\epsilon : |\theta| \leq \epsilon \quad vs. \quad H_1^\epsilon : |\theta| > \epsilon. \quad (16)$$

for some $\epsilon > 0$. One rationale is that $|\theta| < \epsilon$ are practically zero in applications for suitable ϵ . Some Bayesian approaches have been taken for (16), for example, [Verdinelli and Wasserman, 1996], [Delampady, 1989] and [Gómez-Villegas and Sánchez-Manzano, 1992]. Nonetheless, (15) has been suggested as a reasonable approximation to (16) for practical ϵ choices. See, for example, [Berger and Delampady, 1987] for arguments and elaboration on this point.

For testing (15), the recommended and practiced test is the uniformly most powerful unbiased α test: At level α

$$\text{Rejecting } H_0 \text{ if } |\bar{X}|/\sigma_n > z_{\alpha/2} \quad (17)$$

where $Z \sim N(0, 1)$ and $z_{\alpha/2}$ is the upper $\alpha/2$ cutoff point of standard normal such that $P(Z > z_{\alpha/2}) = \alpha/2$ and $\sigma_n^2 = \sigma^2/n$. And the corresponded p-value is

$$p(\bar{x}) = P_0(|\bar{X}| > |\bar{x}|) = 2P(|Z| > \frac{|\bar{x}|}{\sigma_n}) \quad (18)$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and its observed (realized) sample version $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Lindley’s paradox states that for large n , there exists moderate $\bar{x}$ such that

- The p-value is less than a small α , say 0.05.
- The posterior probability of $\theta = 0$ can be greater than $1 - \alpha$.

With suitable choice of prior which assigns $\pi_0 = P_\pi(\theta = 0)$ and distributes the rest of the prior probability over the alternative space through diffuse density. For example, when the diffuse prior is chosen as a normal pdf with mean μ and large variance τ^2 ,

$$\pi(\theta) = \begin{cases} \pi_0 & \text{if } \theta = 0, \\ \frac{(1-\pi_0)}{\sqrt{2\pi\tau^2}} \exp[-(\theta - \mu)^2/\tau^2] & \text{otherwise.} \end{cases} \quad (19)$$

A natural Bayes estimate of $\theta = 0$ is

$$P_{\pi(\theta|\bar{x})}(\theta = 0) = E_{\pi(\theta|\bar{x})}1_{[\theta=0]} \quad (20)$$

where $1_{[\theta=0]}$ is the indicator function which equals 1 if $\theta = 0$, otherwise equals 0. Note that (20) is the posterior expectation of $1_{[\theta=0]}$ and the Bayes estimate with respect to squared error loss.

For large n , moderate $|\bar{x}|$ will render large (greater than $1 - \alpha$) $P_{\pi(\theta|\bar{x})}(\theta = 0)$ and small (less than α) p-value. This is the paradox. A paradox epitomizes the conflict between Bayesian and frequentist evidences of assessing whether $\theta = 0$. This paradox has deep implications and causes concern about using hypothesis testing. See [Lindley, 1957], [Schafer, 1982] and references therein. Also [Berger and Delampady, 1987], [Berger and Sellke, 1987], [Casella and Berger, 1987], and more recently, [Tsao, 2006a] and [Tsao, 2006b] are studies motivated by this perplexity.

Is Lindley's paradox a paradox of confusion or a paradox of conflict? Here we focus on some *assumptions* that underpin Lindley's paradox

1. The point null hypothesis is a reasonable approximation to the more realistic interval null hypothesis.
2. Singularity of the indicator function, $1_{[\theta=0]}$, does not create artificial irregularity which in turn yields the paradox.
3. It is justifiable to compare $E_{\pi(\theta|\bar{x})}1_{[\theta=0]}$ with the p-value as Bayesian and frequentist evidences respectively for assessing possibility of $\theta = 0$.
4. Assignment of $\pi_0 = P_\pi(\theta = 0)$ does not seriously affect the results.

The mathematical reduction through limit at 0 might create some irregularities. Note that these forementioned *assumptions* are not mathematical conditions such that the Lindley's paradox will hold, they are underlying the arguments and proofs. For example, while the point null formulation (15) simplifies the (frequentist) computation of the distribution of the test statistic under the null space, it makes the Bayesian computation much harder. Inevitably, it induces the following questions:

What is a reasonable π_0 to be assigned? Robustness of this assignment? Estimation of an indicator function takes only value 0 except one point is another difficult task. Will the paradox still hold if we replace it by a smoother version of the indicator function? Recently, [Tsao, 2006a] uses

$$\beta(\theta) = 2\Phi\left(\frac{-|\theta|}{\sigma/\sqrt{m}}\right), \quad (21)$$

with m being a positive integer, as a smooth version of the indicator function $1_{[\theta=0]}$ and note that the p-value as a maximum likelihood estimate for (21). Under this smooth null approach, it is shown that although the maximum likelihood estimate is more extreme than their Bayes contenders, the discrepancy is much less serious than that in the Lindley's paradox. It suggests that although the maximum likelihood estimate tends to be more extreme than Bayes estimates with respect to some "reasonable" priors, the marked difference between them might be due to the singularity of the problem. The singularity, $\theta = 0$, is artificially chosen for mathematical convenience. There are certainly other formulations for near zero θ values from application perspectives. In light of this observation, we think p-value or a maximum likelihood estimate do have different behaviours but the discrepancy noted in Lindley's paradox might be mild if we replace the limiting singularity by some smooth version of formulation. For more discussion on these assumptions, please refer to [Tsao, 2006a].

Lindley's paradox is often quoted as an example of irreconcilability between Classical/Frequentism and Bayesianism. However, p-value itself is controversial statistical evidence. Its validity is also under active research. A line of research, for example [Berger and Delampady, 1987], [Berger and Sellke, 1987], [Casella and Berger, 1987] and [Hwang *et al.*, 1992], studies the validity of p-values under a robust Bayesian and a formal decision theoretical framework. Unfortunately, so far as we know, there seems to be no consensus. Therefore, we think Lindley's paradox is better understood as a warning about the p-value, the common practice, as the evidence against null hypothesis since the p-value tends to be more liberal (i.e. smaller) than Bayes estimates when the data is moderately significant.

Before closing this section, we would like to point out that a confidence interval can be a very informative supplement to the p-value in hypothesis testing. It helps us to see how the statistical significance corresponds to practical significance in the given context. Suppose that the nose length of Cleopatra's is measured as 15.0001 cm based on the average of many independent random sample. Can we say the true nose length $\theta_C = 15$? Practically? Yes; statistically? Maybe not. The difference 0.0001 might not be practically different but can be statistical significant when the sample size is large or the variance is small. P-value captures the statistical significance but it should be carefully examined in the context of the problem. A remedy to this is to look at the (two-sided) confidence interval of θ at the same time. By the duality of test and confidence set, see for example Theorem 9.2.1 of [Casella and Berger, 1990], the confidence interval can also be

used for hypotheses testing. Specifically, in our context,

$$\text{Rejecting } H_0 \text{ if } |\bar{X}|/\sigma_n > z_{\alpha/2} \quad \text{if and only if } 0 \text{ does not belong to } C(X). \quad (22)$$

where $C(X) = \bar{X} \pm z_{\alpha/2}\sigma_n$, a $(1 - \alpha)$ confidence interval for θ . The test statistic $(\bar{X} - 0)/\sigma_n$ is a standardized difference of $(\bar{X} - 0)$. It measures the difference in the standard normal scale in the sense that the test statistic is distributed as a standard normal random variable. In our problem, this scale is σ_n . If the scale is too small for practical consideration, it will be immediately noticed through the confidence interval. For example, a 95% confidence interval of $[15.00005, 15.00015]$ will lead to a very small p-value yet this difference makes no practical importance — the nose length is essentially 15 cm.

4 FIELLER-CREASY PROBLEM

When we say Cleopatra and Elizabeth Taylor have the same nose length, we mean $D = \theta_E - \theta_C = 0$ or $\rho = \theta_E/\theta_C = 1$. Mathematically, they are equivalent statements. We tend to think either problem formulation is fine and the statistical procedures will provide consistent answers. Unfortunately, it is not so. The difference formulation is widely applied. The statistical procedures, point estimation, confidence intervals and hypotheses testings, are similar to the one-population problems and well-studied in statistical theory. On the other hand, the ratio formulation generates important differences in the corresponding statistical procedures such as the confidence intervals. It is noted that the 95% Fieller's confidence interval, the popular confidence interval for the ratio ρ , has a positive probability of having infinity length.

In the simplest setting, let $X \sim N(\theta_1, \sigma^2)$ and $Y \sim N(\theta_2, \sigma^2)$, the Fieller's confidence set for $\rho = \theta_1/\theta_2$, where $\theta_2 \neq 0$, is

$$(23) \quad C_F(\mathbf{X}) = \left\{ \rho : \frac{|X - Y\rho|}{\sqrt{1 + \rho^2}} < z_{\alpha/2} \right\}$$

where $\mathbf{X} = (X, Y)'$.

The ratio parameters are of importance in the field of biology and bioassay. See, for example, [Finney, 1978], [Govindarajulu, 1988] and [Chow and Liu, 1992]. Proposed by [Fieller, 1954], the Fieller's confidence set, denoted by $C_F(X)$, is a popular confidence set for ratios, denoted by ρ . Despite its popularity, it is one of the most well-known embarrassments to frequentist confidence theory; namely, with a positive probability the set $C_F(\mathbf{X})$, as a $1 - \alpha$ confidence set, can be the whole parameter space. Nonetheless, it is not a trivially bad set estimator. The justification is well-founded from pivotal quantity, likelihood ratio test and profile likelihood arguments.

Furthermore, the possibility of infinite length of $C_F(\mathbf{X})$ can not be avoided. In fact, it was proved in [Gleser and Hwang, 1987] that any confidence interval

for some errors-in-variables models having almost surely finite length has zero confidence coefficient, the infimum of the coverage probability over the parameter space. These models include the normal model we assumed with the ratio as the parameter. A remedy is to provide a data-dependent confidence reported by estimated confidence approach. This is, however, beyond the scope of current paper. The interested readers are referred to [Kiefer, 1977] and [Berger, 1988].

[Fu, 1995] suggests that the problem with the ratio is due to the singularity induced by taking the limit of the denominator parameter to 0. The ratio shoots to positive or negative infinity. This is yet another example when mathematics takes to the limit, there emerges some irregularity. However, there is something more fundamental: a statistical procedure may not have good pre-data (frequentist) and post-data (Bayesian) performances at the same time. Here we refer to pre-data performance as the long-run frequentist properties such as confidence coefficient (infimum of the coverage probability), accuracy and expected length; the post-data performance refers to the properties such as the posterior coverage probability, the (realized) length, where the data is considered given.

We think the paradox of Fieller-Creasy problem arises from confusion and conflict about the Bayes credible set and the frequentist confidence interval. While confidence intervals are taught as standard set procedures in statistics textbooks, users of statistics often pay less attention in the interpretation of confidence intervals. For example, the confidence coefficient 95% for a 95% confidence interval refers to the minimum coverage probability of the confidence interval as a random procedure and the probability is interpreted through the long run frequency. Moreover, these realized confidence intervals sometimes are mistaken as 95% Bayes credible sets. Since many classical confidence intervals have constant coverage probability over the parameter space, they seldom correspond to Bayes credible set with respect to proper priors. In light of this observation, we think it is natural or at least anticipated that the confidence interval for the ratio to have strange post-data behaviours. The interested readers are referred to, for example, [Tsao, 1998] for poor properties for conditional implementation of Fieller's confidence sets.

5 CONCLUDING REMARK

Lindley's paradox and the Fieller-Creasy problem are often quoted as the contradiction or inconsistency between frequentist and the Bayesian inferences. Frequentist approach focuses on the pre-data aspects of the statistical inference while Bayesian approach concentrate on post-data perspectives. We acknowledge the difference between these two paradigms of statistics. However, the discrepancy is not necessarily marked in practice. For example, [Mukhopadhyay and DasGupta, 1997] shows that it is possible to construct $(1 - \alpha)$ HPD Bayes credible sets with minimum coverage (frequentist) coverage probability uniformly close to $(1 - \alpha)$ as well. The drastic discrepancy, in our opinion, is induced by the singularity or irregularity in the formulation or the setting of the approach. In the Lindley's

paradox, the estimation of the indicator function of a point null; in the Fieller-Creasy problem, the ratio parameter when the denominator approaches to 0.

Recall the “belief/intuition” we mentioned in Section 2,

when a diffuse prior or a “non-informative” prior is used, the Bayes inference coincides with the frequentist inference; the Bayesian inference is similar to the frequentist one when sample size is very large.

Note that both diffuse prior or large sample size correspond to the cases when the prior variance or the sample size go to infinity. They are the limiting situations. We think there are fundamental differences between frequentist and Bayesian approaches. They do not necessarily coincided or are reconcilable with each other. Within one domain, frequentist or Bayesian alike, the problem of limits remains but the interpretations and guarantees are clear. However, if we pass from one domain into the other, every problem needs to be proceeded with caution.

In mathematics, the existence of a solution of an equation often depends on what the space/set in which we are looking for the solution. For example, for $x^2 + 1 = 0$ there exists no real root but there are roots in the complex space. As in the fable of the blind and the elephant, maybe we are not wise enough to see the whole picture and the paradox remains unresolved. Yet if we can humbly assume there are some animals unknown to us, it is hopeful that sometime in the future our knowledge domain will be broaden to lift these contradictions.

Are these statistical paradoxes of confusion or of confliction? Well, we will leave the taxonomy of these paradoxes to our wise readers.

ACKNOWLEDGMENTS

The author would like to thank two anonymous referees for their careful reading and helpful comments. The research is supported by NSC 97-2118-M-259-003-MY2, Taiwan.

BIBLIOGRAPHY

- [Berger, 1985] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Second Edition, Springer-Verlag, New York, 1985.
- [Berger, 1988] J. O. Berger. An alternative: The estimated confidence approach. In *Statistical Decision Theory and Related Topics IV*, Vol. 1 (S. S. Gupta and J. O. Berger, eds.). 85–90. Springer-Verlag, New York, 1988.
- [Berger and Delampady, 1987] J. O. Berger and M. Delampady. Testing precise hypothesis (with discussion). *Statistical Science* **2**, 317–352, 1987.
- [Berger and Sellke, 1987] J. O. Berger and T. Sellke. Testing a point null hypothesis: the irreconcilability of p-values and evidence. *Journal of American Statistical Association* **82**, 112–122, 1987.
- [Casella and Berger, 1987] G. Casella and R. L. Berger. Reconciling Bayesian and Frequentist evidence in the one-sided testing problem. *Journal of American Statistical Association* **82**, 106–111. 1987 .
- [Casella and Berger, 1990] G. Casella and R. L. Berger. *Statistical Inference*. Brooks/Cole, Thomson Information/Publishing Group, 1990.

- [Chow and Liu, 1992] S. C. Chow and J. P. Liu. Design and analysis of bioavailability and bioequivalence studies. Marvel Decker, Inc. New York, 1992.
- [Delampady, 1989] M. Delampady. Lower bounds on Bayes factors for interval null hypotheses. *Journal of American Statistical Association* **84**, 120–124, 1989.
- [Fieller, 1954] E. C. Fieller. Some problems connected in interval estimation. *Journal of Royal Statistical Society Series B* **16**, 175–183, 1954.
- [Finney, 1978] D. J. Finney. Statistical methods in biological assay, 3rd ed., Charles Griffin and Company Ltd, 1978.
- [Fu, 1995] J. Fu. Personal communication, 1995.
- [Govindarajulu, 1988] Z. Govindarajulu. Statistical Techniques in Bioassay. Karger, Switzerland, 1988.
- [Gleser and Hwang, 1987] L. Gleser and J. T. Hwang. The nonexistence of $100(1-\alpha)\%$ confidence set of finite expected diameter in error in variable and related models. *Annals of Statistics* **15**, 1351–1362, 1987.
- [Gómez-Villegas and Sánchez-Manzano, 1992] M. A. Gómez-Villegas and G. E. Sánchez-Manzano. Bayes factors in testing precise hypotheses. *Communications in Statistics: Theory and Methods* **21**, 1707–1715, 1992.
- [Hwang et al., 1992] J. T. G. Hwang, G. Casella, C. Robert, M. Wells and R. Farrel. Estimation of accuracy in testing. *Annals of Statistics* **20**, 490–509, 1992.
- [Kiefer, 1977] J. Kiefer. Conditional confidence and confidence estimators. (with discussion), *Journal of American Statistical Association* **72**, 789–827, 1977.
- [Lehmann, 1986] E. L. Lehmann. Testing Statistical Hypotheses, 2nd Ed. Wiley, New York, 1986.
- [Lindley, 1957] D. V. Lindley. A statistical paradox. *Biometrika* **44**, 187–192, 1957.
- [Mukhopadhyay and DasGupta, 1997] S. Mukhopadhyay and A. DasGupta. Uniform Approximation of Bayes Solutions and Posteriors : Frequentistly Valid Bayes Inference *Statistics and Decisions* **15**, 51–73, 1997.
- [Schafer, 1982] G. Schafer. Lindley's paradox. *Journal of American Statistical Association* **77**, 325–334, 1982.
- [Tsao, 2006a] C. A. Tsao. A note on Lindley's paradox. *Test* **15**, 125–139, 2006.
- [Tsao, 2006b] C. A. Tsao. Assessing post-data weight of evidence. *Journal of Statistical Planning and Inference* **136**, 4012–4025, 2006.
- [Tsao, 1998] C. A. Tsao. Conditional coverage probability of confidence intervals of errors-in-variables and related models. *Statistics and Probability Letters* **40**, 165–170, 1998.
- [Verdinelli and Wasserman, 1996] I. Verdinelli and L. Wasserman. Bayes factors, nuisance parameters and imprecise tests. *Bayesian Statistics 5* (Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M. eds.), 765–771. Oxford University Press, 1996.

Part VIII

Statistics and Inductive Inference

This page intentionally left blank

STATISTICS AS INDUCTIVE INFERENCE

Jan-Willem Romeijn

1 STATISTICAL PROCEDURES AS INDUCTIVE LOGICS

An inductive logic is a system of inference that describes the relation between propositions on data, and propositions that extend beyond the data, such as predictions over future data, and general conclusions on all possible data. Statistics, on the other hand, is a mathematical discipline that describes procedures for deriving results about a population from sample data. These results include predictions on future samples, decisions on rejecting or accepting a hypothesis about the population, the determination of probability assignments over such hypotheses, the selection of a statistical model for studying the population, and so on. Both inductive logic and statistics are calculi for getting from the given data to propositions or results that transcend the data.

This suggests that there is a strong parallel between statistics and inductive logic. In fact, it does not take much imagination to view statistical procedures as inferences: the input components, primarily the data, are the premises, and the result of the procedure is the conclusion. In this rough and ready way, statistical procedures can be understood as defining particular inductive logics. However, the two disciplines have evolved more or less separately. In part this is because there are objections to viewing classical statistics as inferential, although this is not true for all statistical procedures. For another part, it may be because inductive logic has been dominated by the Carnapian programme. Perhaps statisticians have not recognised inductive logic as a discipline that is much like their own.

However this may be, I think it is time for a rapprochement. There are, to my mind, good reasons for investigating the parallel between inductive logic and statistics along the lines suggested above. First, framing the statistical procedures as inferences in a logic may clarify the presuppositions of these procedures. Second, by relating statistics to inductive logic, techniques and insights from inductive logic may be used to enrich statistics. And finally, showing the parallels between inductive logic and statistics may show the relevance, also to inductive logicians themselves, of their discipline to the sciences, and thereby direct further research in this field.

With this aim in mind, I consider a number of statistical procedures in this chapter, and I investigate whether they can be seen as part of an inductive logic, or otherwise whether they can, at least partly, be translated into such a logic. I start by describing induction in formal terms, and I introduce a general notion of

probabilistic inductive inference. This provides a setting in which both statistical procedures and inductive logics may be captured. I subsequently discuss a number of statistical procedures, and show how they can, and cannot, be captured by certain inductive logics.

The first statistical procedure is Neyman-Pearson hypotheses testing. This procedure was introduced as explicitly non-inferential, and so it should strictly speaking not be captured by an inductive logic. On the other hand, power and significance are often interpreted inferentially. At the end of the chapter I devise an inductive logic that may be used to warrant such an interpretation. The second statistical procedure is parameter estimation. I briefly discuss Fisher's theory of maximum likelihood estimators, and I show that there is a certain relation with the inductive logic developed by Carnap. A third statistical procedure is Bayesian statistics. I show that it can be captured in a probabilistic inductive logic that relates to Carnapian inductive logic via the representation theorem of de Finetti. This leads to a discussion of Bayesian statistics in relation to Bayesian inductive logic.

Given the nature of the chapter, the discussion of statistical procedures is relatively short. Many procedures cannot be dealt with. Similarly, I cannot discuss in detail the many inductive logics devised within Carnapian inductive logic. For the former, the reader may consult other chapters in this volume, in particular the chapter by Festa. For the latter, I refer to [Hartmann *et al.*, 2009], specifically the discussions of inductive logic contained therein.

2 OBSERVATIONAL DATA

As indicated, inductive inference starts from propositions on data, and ends in propositions that extend beyond the data. An example of an inductive inference is that, from the proposition that up until now all observed pears were green, we conclude that the next few pears will be green as well. Another example is that from the green pears we have seen we conclude that all pears are green, period. The key characteristic is that the conclusion says more than what is classically entailed by the premises.

Let me straighten these inferences out a bit. First, I restrict attention to propositions on empirical facts, thus leaving aside such propositions as that pears are healthy, or that God made them. Second, I focus on the results of observations of particular kinds of empirical fact. For example, the empirical fact at issue is the colour of pears, and the results of the observations are therefore colours of individual pears. There can in principle be an infinity of such observation results, but what I call data is always a finite sequence of them. Third, the result of an observation is always one from a designated partition of properties, usually finite but always countable. In the pear case, it may be {red, green, yellow}. I leave aside observations that cannot be classified in terms of a mutually exclusive set of properties.

I now make these ideas on what counts as data a bit more formal. The concept

I want to get across is that of a sample space, in which single observations and sequences of observations can be represented as sets, called events. After introducing the observations in terms of a language, I define sample space. All the probabilities in this chapter will be defined over sample space, because probability is axiomatized as a measure function over sets. However, the expressions may be taken as sentences from a logical language just as well.

We denote the observation of individual i by Q_i . This is a propositional variable, and we denote assignments or valuations of this variable by q_i^k , which represents the sentence that the result of observing individual i is the property k . A sequence of such results of length t , starting at 1, is denoted with the propositional variable S_t , and its assignment with $s^{k_1 \dots k_t}$, often abbreviated as s_t . In order to simplify notation, I denote properties with natural numbers, so $k \in K = \{0, 1, \dots, n-1\}$. For example, if the observations are the aforementioned colours of pears, then $n = 3$. I write red as 0, green as 1, and yellow as 2, so s^{012} says that the first three pairs were red, green, and yellow respectively. Note further that there are logical relations among the sentences, like $s^{012} \rightarrow q_2^1$. Together, the expressions s_t and q_i^k form the observation language.

Now we develop a set-theoretical representation of the observations, a so-called sample space, otherwise known as an observation algebra. To this aim, consider the set of all infinitely long sequences K^ω , that is, all sequences like 012002010211112..., each encoding the observations of infinitely many pears. Denote such sequences with u , and write $u(i)$ for the i -th element in the sequence u . Every sentence q_i^k can then be associated with a particular set of such sequences, namely the set of u whose i -th element is k :

$$q_i^k = \{u \in K^\omega : u(i) = k\}.$$

Clearly, we can build up all finite sequences of results $s^{k_1 \dots k_t}$ as intersections of such sets:

$$s^{k_1 \dots k_t} = \bigcap_{i=1}^t q_i^{k_i}.$$

Note that entailments in the language now come out as set inclusions: we have $s^{012} \subset q_2^1$. Instead of a language with sentences q_i^k and logical relations among such sentences, I will in the following employ the algebra $\mathcal{Q}$, built up by the sets q_i^k and their conjunctions and intersections.

I want to emphasise that the notion of a sample space introduced here is really quite general. It excludes a continuum of individuals and a continuum of properties, but apart from that, any data recording that involves individuals and that ranges over a set of properties can serve as input. For example, instead of pears having colours we may think of subjects having test scores. Or of companies having certain stock prices. The sample space used in this chapter follows the basic structure of most applications in statistics, and of almost all applications in inductive logic.

3 INDUCTIVE INFERENCE

Now that I have made the notion of data more precise, let me turn to inductive inference. Consider the case in which I have observed three green pears: s^{111} . What can I conclude about the next pear? Or about pears in general? From the structure of the data itself, it seems that we can conclude depressingly little. We might say that the next pear is green, q_4^1 . But as it stands, each of the sets $s^{111k} = s^{111} \cap q_4^k$, for $k = 0, 1, 2$, is a member of the sample space, or in terms of the logical language, we cannot derive any sentence q_4^k from the sentence s^{111} . The event of observing three green pears is consistent with any colour for the next pear. Purely on the basis of the classical relations among observations, as captured by the language and the sample space, we cannot draw any inductive conclusion.

Perhaps we can say that given three green pears, the next pear being green is more probable? This is where we enter the domain of probabilistic inductive logic. We can describe the complete population of pears by a probability function over the observational facts,

$$P : \mathcal{Q} \mapsto [0, 1].$$

Every possible pear q_{t+1}^k , and also every sequence of such pears $s^{k_1 \dots k_t}$, receives a distinct probability. The probability of the next pear being of a certain colour, conditional on a given sequence, is expressed as $P(q_{t+1}^k | s^{k_1 \dots k_t})$. Similarly, we may wonder about the probability that all pears are green, which is again determined by the probability assignment, in this case $P(\{\forall i : q_i^1\})$. All such probabilistic inductive inferences are determined by the full probability function P .

The central question of any probabilistic inductive inference or procedure is therefore how to determine the function P , relative to the data that we already have. What must the probability of the next observation be, given a sequence of observations gone before? And what is the right, or otherwise the preferable, distribution over all observations given the sequence? Both statistics and inductive logic aim to provide an answer to these questions, but they do so in different ways.

In order to facilitate the view that the statistical procedures are logical inferences, it will be convenient to keep in mind a particular understanding of probability assignments P over the sample space, or observation algebra, $\mathcal{Q}$. Recall that in classical two-valued logic, a model of the premises is a complete truth valuation over the language, subject to the rules of logic. Because of the correspondence between language and algebra, the model is also a complete function over the algebra, taking the values $\{0, 1\}$. Accordingly, the premises of some deductive logical argument are represented as a set of models over the algebra. By analogy, we may consider a probability function over an observation algebra as a model too. Just like the truth value assignment, the probability function is a function over an algebra, only it takes values in the interval $[0, 1]$, and it is subject to the axioms of probability.

Probabilistic inductive logics use probability models for the purpose of inductive inference. In particular, the premises of a probabilistic inductive argument can be represented as a set, possibly a singleton, of probability assignments. But

there are widely different ways of understanding the inferential step, i.e., the step running from the premises to the conclusion. The most straightforward of these, and the one that is closest to classical statistical practice, is to associate a probability function P , or otherwise a set of such functions, with each sample s_t . The inferential step then runs from the data s_t and a large set of probability functions P , possibly all conceivable functions, towards a more restricted set, or even towards a single P . The resulting inductive logic is called *ampliative*, because the restriction on the set of probability functions that is effected by the data, i.e. the conclusion, is often stronger than what follows from the data and the initial set of probability functions, i.e. the premises, by deduction.

We can also make the inferential step precise by analogy to a more classical, *non-ampliative* notion of entailment. As will become apparent, this kind of inferential step is more naturally associated with what is traditionally called inductive logic. It is also associated with a basic kind of probabilistic logic, as elaborated in [Hailperin, 1996] and more recently in [Haenni *et al.*, 2009], especially section 2. Finally, this kind of inference is strongly related to Bayesian logic, as advocated by [Howson, 2003]. It is the kind of inductive logic favored in this chapter.

An argument is said to be classically valid if and only if the set of models satisfying the premises is contained in the set of models satisfying the conclusion. The same idea of classical entailment may now be applied to the probabilistic models over sample space. In that case, the inferential step is from one set of probability assignments, characterised by a number of restrictions associated with premises, towards another set of probability assignments, characterised by a different restriction that is associated with a conclusion. The inductive inference is called valid if the former is contained in the latter, i.e., if every model satisfying the premises is also a model satisfying the conclusions. In such a valid inferential step, the conclusion does not amplify the premises.

As an example, say that we fix $P(q_1^0) = \frac{1}{2}$ and $P(q_1^1) = \frac{1}{3}$. Both these probability assignments can be taken as premises in a logical argument, and the models of these premises are simply all probability functions P over $\mathcal{Q}$ for which these two valuations hold. By the axioms of probability, we can derive that any such function P will also satisfy $P(q_1^2) = \frac{1}{6}$, and hence also that $P(q_1^2) < \frac{1}{4}$. On its own, the latter expression amounts to a set of probability functions over the sample space $\mathcal{Q}$ in which the probability functions that satisfy both premises are included. In other words, the latter assignment is classically entailed by the two premises. Along exactly the same lines, we may derive a probability assignment for a statistical hypothesis h conditional on the data s_t , written as $P(h|s_t)$, from the input probabilities $P(h)$, $P(s_t)$, and $P(s_t|h)$, using the theorem of Bayes.

The classical, non-ampliative understanding of entailment may thus be used to reason inductively, towards predictions and statistical hypotheses that themselves determine a probability assignment over data. In the following I will focus primarily on such non-ampliative inductive logical inferences to investigate statistical procedures. I first discuss Neyman-Pearson hypothesis testing and Fisher's maximum likelihood estimation in their own terms, showing that they are best understood

as ampliative inductive inferences. Then I discuss Carnapian inductive logic and show that it can be viewed as a non-ampliative version of parameter estimation. This leads to a discussion of Bayesian statistical inference, which is subsequently related to a generalisation of Carnapian inductive logic, Bayesian inductive logic. The chapter ends with an application of this logic to Neyman-Pearson hypothesis testing.

As indicated, Carnapian inductive logic is most easily related to non-ampliative logic. So, viewing statistical procedures in this perspective makes the latter more amenable to inductive logical analysis. But I do not want to claim that I thereby lay bare the real nature of the statistical procedures. Rather, I hope to show that the investigation of statistics along these specific logical lines clarifies and enriches statistical procedures. Furthermore, as indicated, I hope to stimulate research in inductive logic that is directed at problems in statistics.

4 NEYMAN-PEARSON TESTING

The first statistical application concerns the choice between two statistical hypotheses, that is, two fully specified probability functions over sample space. In the above vocabulary, it concerns the choice between two probabilistic models, but we must be careful with our words here, because in statistics, models often refer to sets of statistical hypotheses. In the following, I will therefore refer to complete probability functions over the algebra as *hypotheses*.

Let $\mathcal{H} = \{h_0, h_1\}$ be the set of hypotheses, and let $\mathcal{Q}$ be the sample space, that is, the observation algebra introduced earlier on. We can compare the hypotheses h_0 and h_1 by means of a Neyman-Pearson test function. See [Barnett, 1999] and [Neyman and Pearson, 1967] for the details.

DEFINITION 1 Neyman-Pearson Hypothesis Test. Let F be a function over the sample space $\mathcal{Q}$,

$$F(s_t) = \begin{cases} 1 & \text{if } \frac{P_{h_1}(s_t)}{P_{h_0}(s_t)} > r, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where P_{h_j} is the probability over the sample space determined by the statistical hypothesis h_j . If $F = 1$ we decide to reject the null hypothesis h_0 , else we accept h_0 for the time being.

Note that, in this simplified setting, the test function is defined for each set of sequences s_t separately. For each sample plan, and associated sample size t , we must define a separate test function.

The decision to accept or reject a hypothesis is associated with the so-called

significance and power of the test:

$$\text{Significance}_F = \alpha = \int_{\mathcal{Q}} F(s_t) P_{h_0}(s_t) ds_t,$$

$$\text{Power}_F = 1 - \beta = \int_{\mathcal{Q}} F(s_t) P_{h_1}(s_t) ds_t.$$

The significance is the probability, according to the hypothesis h_0 , of obtaining data that leads us to reject the hypothesis h_0 , or in short, the type-I error of falsely rejecting the null hypothesis, denoted α . Similarly, the power is the probability, according to h_1 , of obtaining data that leads us to reject the hypothesis h_0 , or in short, the probability under h_1 of correctly rejecting the null hypothesis, so that $\beta = 1 - \text{Power}$ is the type-II error of falsely accepting the null hypothesis. An optimal test is one that minimizes the significance level, and maximizes the power. Neyman and Pearson prove that the decision has optimal significance and power for, and only for, likelihood-ratio test functions F . That is, an optimal test depends only on a threshold for the ratio $\frac{P_{h_1}(s_t)}{P_{h_0}(s_t)}$.

Let me illustrate the idea of Neyman-Pearson tests. Say that we have a pear whose colour is described by q^k , and we want to know from what farm it originates, from farmer Maria (h_0) or Lisa (h_1). We know that the colour composition of the pears from the two farms are different:

Hypothesis \ Data	q^0	q^1	q^2
h_0	0.00	0.05	0.95
h_1	0.40	0.30	0.30

If we want to decide between the two hypotheses, we need to fix a test function. Say that we choose

$$F(q^k) = \begin{cases} 0 & \text{if } k = 2, \\ 1 & \text{else.} \end{cases}$$

In the definition above, which uses a threshold for the likelihood ratio, this comes down to choosing a value for r somewhere between $\frac{6}{19}$ and 14, for example $r = 1$. The significance level is $P_{h_0}(q^0 \cup q^1) = 0.05$, and the power is $P_{h_1}(q^0 \cup q^1) = 0.70$. Now say that the pear we have is green, so $F = 1$ and we reject the null hypothesis, concluding that Maria did not grow the pear with the aforementioned power and significance.

Note that from the perspective of ampliative inductive logic, it is not too far-fetched to read an inferential step into the Neyman-Pearson procedure. The test function F brings us from a sample s_t and two probability functions, P_{h_j} for $j = 0, 1$, to a single probability function P_{h_1} , or P_{h_0} , over the sample space $\mathcal{Q}$. So we might say that the test function is the procedural analogue of an inductive

inferential step, as discussed in Section 3. This step is ampliative because both probability functions P_{h_j} are consistent with the data. Ruling out one of them cannot be done deductively.¹

Neyman-Pearson hypothesis testing is sometimes criticised because its results generally depend on the probability function over the entire sample space, and not just on the probability of those elements in sample space corresponding to the actual events, the observed sample for short. That is, the decision to accept or reject the null hypothesis against some alternative hypothesis depends not just on the probability of what has actually been observed, but also on the probability assignment over everything that could have been observed. A well-known illustration of this problem concerns so-called optional stopping. But here I want to illustrate the same point with an example that can be traced back to [Jeffreys, 1931] p. 357, and of which a variant is discussed in [Hacking, 1965].²

Instead of the hypotheses h_0 and h_1 above, say that we compare the hypotheses h_0^* and h_1 .

Hypothesis \ Data	q^0	q^1	q^2
h_0^*	0.05	0.05	0.90
h_1	0.40	0.30	0.30

We determine the test function $F(q^k) = 1$ iff $k = 0$, by requiring the same significance level, $P_{h_0^*}(q^0) = 0.05$, resulting in the power $P_{h_1}(q^0) = 0.40$. Now imagine that we observe q^1 again. Then we accept h_0^* . But this is a bit odd, because the hypotheses h_0 and h_0^* have the same probability for q^1 ! So how can the two test procedures react differently to this observation? It seems that, in contrast to h_0 , the hypothesis h_0^* escapes rejection because it allocates some probability to s^0 , an event that does not occur. This causes a shift in the area within sample space on which the hypothesis h_0^* is rejected. This phenomenon gave rise to the famed complaint of Jeffreys that “a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred: if indeed h_0 is true, it will be rejected after observing q_1 , simply because it gives q_0 a zero probability.”

This illustrates how the results of a Neyman-Pearson procedure depends on the whole probability function that a hypothesis defines over the sample space, and

¹There are attempts to make these ampliative inferences more precise, by means of a form of default reasoning, or a reasoning that employs a preferential ordering over probability models. Specifically, so-called evidential probability, proposed by [Kyburg, 1974] and more recently discussed by [Wheeler, 2006], is concerned with inferences that combine statistical hypotheses, which are each accepted with certain significance levels. However, in this chapter I will not investigate these logics. They are not concerned with inferences from the data to predictions or to hypotheses, but rather with inferences from hypotheses to other hypotheses, and from hypotheses to predictions.

²I would like to thank Jos Uffink for bringing this example to my attention. As far as I can see, the exact formulation of the example is his.

not just on the probability defined for the actual observation. From the perspective of an inductive logician, it may therefore seem “a remarkable procedure”, to cite Jeffreys again. But it must be emphasised that Neyman-Pearson statistics was never intended as an inference in disguise. It is a procedure that allows us to decide between two hypotheses on the basis of data, generating error rates associated with that decision. Neyman and Pearson themselves were very explicit that the procedure must not be interpreted inferentially. Rather than inquiring into the truth and falsity of a hypothesis, they were interested in the probability of mistakenly deciding to reject or accept a hypothesis. The significance and power concern the probability over data given a hypothesis, not the probability of hypotheses given the data.³

5 FISHER’S PARAMETER ESTIMATION

Let me turn to another important classical statistical procedure, so-called parameter estimation. I focus in particular on an estimation procedure first devised by Fisher, estimation by maximum likelihood.

The maximum likelihood estimator determines the best among a much larger, possibly infinite, set of hypotheses. Again it depends entirely on the probability that the hypotheses assign to points in the sample space. See [Barnett, 1999] and [Fisher, 1956] for more detail.

DEFINITION 2 Maximum Likelihood Estimation. Let $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$ be a set of hypotheses, labeled by the parameter θ , and let $\mathcal{Q}$ be the sample space. Then the maximum likelihood estimator of θ ,

$$\hat{\theta}(s_t) = \{\theta : \forall h_{\theta'} (P_{h_{\theta'}}(s_t) \leq P_{h_\theta}(s_t))\}, \quad (2)$$

$\hat{\theta}$ for short, is a function over the elements s_t in the sample space.

So the estimator is a set, typically a singleton, of those values of θ for which the likelihood of h_θ on the data s_t is maximal. The associated best hypothesis we denote with $h_{\hat{\theta}}$. Note that this estimator is a function over the sample space, associating each s_t with a hypothesis, or a set of them.

Often the estimation is coupled to a so-called confidence interval. Restricting the parameter space to $\Theta = [0, 1]$ for convenience, and assuming that the true value is θ , we can define a region in sample space within which the estimator function is not too far off the mark. Specifically, we might set the region in such a way that it covers $1 - \epsilon$ of the probability P_{h_θ} :

$$\int_{\theta-\Delta}^{\theta+\Delta} P_{h_\theta}(\hat{\theta}) d\hat{\theta} = 1 - \epsilon.$$

³This is not to say that Neyman-Pearson statistics cannot be viewed from an inferential angle. See Section 9 for an inferential account.

We can provide an unproblematic frequentist interpretation of the interval $\hat{\theta} \in [\theta - \Delta, \theta + \Delta]$: in a series of estimations, the fraction of times in which the estimator $\hat{\theta}$ is further off the mark than Δ will tend to ϵ . The smaller the region, the more reliable the estimate. Note, however, that this interval is defined in terms of the unknown true value θ .

Some applications allow for the derivation of a region of parameter values within which the true value θ can be expected to lie.⁴ The general idea is to define a set of parameter values R within which the data are not too unlikely, $R(s_t) = \{\theta : P_{h_\theta}(s_t) > \epsilon\}$ for some small value $\epsilon > 0$. Now in terms of the integral above, we can swap the roles of θ and $\hat{\theta}$ and define the so-called central confidence interval:

$$\text{Conf}_{1-\epsilon}(\hat{\theta}) = \left\{ \theta : |\theta - \hat{\theta}| < \Delta \text{ and } \int_{\hat{\theta}-\Delta}^{\hat{\theta}+\Delta} P_{h_\theta}(\hat{\theta}) d\theta = 1 - \epsilon \right\}.$$

Via the function $\hat{\theta}(s_t)$, every element of the sample space s_t is assigned a region $\text{Conf}_{1-\epsilon}$ of parameter values, interpreted as the region within which we may expect to find the true value θ . Note, however, that swapping the roles of θ and $\hat{\theta}$ in the integral is not unproblematic. We can only interpret the integral as a probability if $P_{h_\theta}(\hat{\theta} + \delta) = P_{h_{\theta-\delta}}(\hat{\theta})$ for all values of δ , or in other words, if for fixed $\hat{\theta}$ the function $P_{h_\theta}(\hat{\theta})$ is indeed a probability density over θ . In other cases, the interval cannot be taken as expressing the expected accuracy of the estimate, or at least not without further critical reflection.

Let me illustrate parameter estimation in a simple example on pears again. Say that we are interested in the colour composition of pears from Emma's farm, and that her pears are red, q_i^0 , or green, q_i^1 . Any ratio between these two kinds of pears is possible, so we have a set of hypotheses h_θ , called multinomial hypotheses, for which

$$P_{h_\theta}(q_i^1 | s_{t-1}) = \theta, \quad P_{h_\theta}(q_i^0 | s_{t-1}) = 1 - \theta \quad (3)$$

with $\theta \in [0, 1]$. The hypothesis h_θ fixes the portion of green pears at θ , and therefore, independently of what pears we saw before, on the assumption of the hypothesis h_θ the probability that a randomly drawn pear from Emma's farm is green is θ . The type of distribution over $\mathcal{Q}$ that is induced by these hypotheses is called a Bernoulli distribution, or a multinomial distribution.

⁴The determination of such regions is similar in nature to the determination of so-called fiducial probability. [Fisher, 1930; Fisher, 1935; Fisher, 1956] developed the notion of fiducial probability as a way of capturing parameter estimation in terms of a non-ampliative entailment relation, basically deriving a probability assignment over hypotheses without assuming a prior probability over statistical hypotheses at the outset. The fiducial argument is controversial, however, and its applicability is limited to particular statistical problems. [Seidenfeld, 1979] provides a detailed discussion of the restricted applicability of the argument in cases with multiple parameters. [Dawid and Stone, 1982] argue that in order to run the fiducial argument, one has to assume that the statistical problem can be captured in a functional model that is smoothly invertible. In this chapter, I will not discuss the fiducial argument and explore another non-ampliative representation instead.

The idea of Fisher's maximum likelihood estimation is that we choose the value of θ for which the probability that the hypotheses h_θ gives to the data $s^{k_1 \dots k_t}$ is maximal. Say that we have observed a sequence of pears s^{000101} . The probability of these data given the hypothesis h_θ is

$$P_{h_\theta}(s^{000101}) = \prod_{i=1}^t P_{h_\theta}(q_i^{k_i} | s_{i-1}) = \theta^2(1-\theta)^4. \quad (4)$$

Note that the probability of the data only depends on the number of 0's and the number of 1's in the sequence. Now the above likelihood function is maximal at $\theta = \frac{1}{3}$, so $\hat{\theta} = \frac{1}{3}$. More generally, defining t_1 as the number of 1's in the sequence s_t , the maximum likelihood estimator is

$$\hat{\theta}(s_t) = \frac{t_1}{t}. \quad (5)$$

Note finally that for a true value θ , the probability of finding the estimate in the interval $\frac{t_1}{t} \in [\theta - \Delta, \theta + \Delta]$ increases for larger data sequences. Fixing the probability at $1 - \epsilon$, the size of the interval will therefore decrease with increasing sample size.

This completes the introduction into parameter estimation. The thing to note is that the statistical procedure can be taken as the procedural analogue of an ampliative logical inference, running from the data to a probability assignment over the sample space. We have $\mathcal{H}$ as the set of probability models from which the inference starts, and by means of the data we then choose a single $h_{\hat{\theta}}$ of these, or a set C_{95} , as our conclusion. In the following I aim to investigate whether there is a non-ampliative logical representation of this inductive inference.

6 CARNAPIAN LOGICS

A straightforward way of capturing parameter estimation in a logic is by relating it to the logic of induction developed by [Carnap, 1950; Carnap, 1952]. Historically, Carnapian inductive logic can lay most claim to the title of inductive logic proper. It was the first systematic study into probabilistic predictions on the basis of data.

The central concept in Carnapian inductive logic is logical probability. Recall that the sample space $\mathcal{Q}$, also called the observation algebra, corresponds to an observation language, comprising of sentences such as "the second pear is green", or formally, q_2^1 . The original idea of Carnap was to derive a probability assignment over the language on the basis of symmetries within the language. In the example, we have three mutually exclusive properties for each pear, and in the absence of any further knowledge, there is no reason to think of any of these properties as special or as more, or less, appropriate than the other two. The symmetry inherent to the language suggests that each of the sentences q_i^k for $k = 0, 1, 2$ should get equal probability:

$$P(q_i^0) = P(q_i^1) = P(q_i^2) = \frac{1}{3}.$$

The idea of logical probability is to fix a unique probability function over the observation language, or otherwise a strongly restricted set of such functions, on the basis of symmetries.

Next to symmetries, the set of probability functions can also be restricted by certain predictive properties. As an example, we may feel that yellow pears are more akin to green pears, so that finding a yellow pear decreases the probability for red pears considerably, while it decreases the probability for green pears much less dramatically. That is,

$$\frac{P(q_{t+1}^1 | s_{t-1} \cap q_t^2)}{P(q_{t+1}^0 | s_{t-1} \cap q_t^2)} > \frac{P(q_{t+1}^1 | s_{t-1})}{P(q_{t+1}^0 | s_{t-1})}.$$

How such relations among properties may play a part in determining the probability assignment P is described in the literature on analogy reasoning. See [Festa, 1996; Maher, 2000; Romeijn, 2006]. Interesting recent work on relations between predictive properties in the context of analogical predictions can also be found in [Paris and Waterhouse, 2008].

Any Carnapian inductive logic is defined by a number of symmetry principles and predictive properties, determining a probability function, or a set of such functions. One very well-known inductive logic, discussed at length in [Carnap, 1952], employs a probability assignment characterised by the following symmetries,

$$\begin{aligned} P(q_i^k) &= P(q_i^{k'}), \\ P(s^{k_1 \dots k_i \dots k_t}) &= P(s^{k_i \dots k_1 \dots k_t}), \end{aligned} \quad (6)$$

for all values of i , t , k , and k' , and for all values k_i with $1 \leq i \leq t$. The latter of these is known as the exchangeability of observations: the order in the observations does not matter to their probability. The inductive logic at issue employs a particular version of exchangeability, known as the requirement of restricted relevance,

$$P(q_{t+1}^k | s_t) = f(t_k, t), \quad (7)$$

where t_k is the number of earlier instances q_i^k in the sequence s_t and t the total number of observations. Together these symmetries and predictive properties determine a particular set of probability assignments P , for which we can derive the following consequence:

$$P(q_{t+1}^k | s_t) = \frac{t_k + \frac{\lambda}{n}}{t + \lambda}, \quad (8)$$

where n the number of values for k . The parameter $0 \leq \lambda \leq \infty$ can be chosen at will. Predictive probability assignments of this form are called Carnapian λ -rules.

The probability distribution in Equation (8) has some striking features. Most importantly, for any of the probability functions P satisfying the aforementioned symmetries, we have that

$$P(q_{t+1}^k | s_{t-1} \cap q_t^k) > P(q_{t+1}^k | s_{t-1}).$$

This predictive property is called *instantial relevance*: the occurrence of q_t^k increases the probability for q_{t+1}^k . It was a success for Carnap that this typically inductive effect is derivable from the symmetries alone. By providing an independent justification for these symmetries, Carnap effectively provided a justification for induction, thereby answering the age-old challenge of Hume.⁵

Note that the outlook of Carnapian logic is very different from the outlook of classical statistical procedures, like Fisher's parameter estimation or Neyman-Pearson testing. Classical statistics starts with statistical hypotheses, each associated with a probability functions over a sample space, and then chooses the best fitting one on the basis of the data. By contrast, Carnapian logic starts with a sample space and a number of symmetry principles and predictive properties, that together fix a set of probability functions over the sample space. Just like the truth tables restrict the possible truth valuations, so do these principles restrict the logical probability functions, albeit not to a singleton, as λ can still be chosen freely. But from the point of view of statistics, Carnap is thereby motivating, from logical principles, the choice for a particular set of hypotheses.

Recall that classical statistics was naturally associated with ampliative inductive inference. By contrast, if we ignore the notion of logical probability and concentrate on the inferential step, Carnapian inductive logics fall very neatly within the template for non-ampliative inductive logic that I laid down at the beginning. By means of a number of symmetry principles and predictive properties, we fix a set of probability assignments over the sample space. The conclusions are then reached by working out specific consequences for probability functions within this set, using the axioms of probability. In particular, Carnapian inductive logic looks at the probability assignments conditional on various samples s_t , deriving that they all satisfy *instantial relevance*, for example. Importantly, the symmetries in the language appear as premises in the inductive logical inference. They restrict the set of probability assignments that is considered in the inference.

Despite these differences in outlook, ampliative against non-ampliative, we can identify a strong similarity between parameter estimation, as discussed in Section 5, and the predictive systems of Carnapian logic. To see this, note that the procedure of parameter estimation can be used to determine the probability of the next piece of data. In the example on pears, once we have observed s^{000101} and thus chosen $h_{\frac{1}{3}}$, we may on the basis of that predict that the next pear has a probability of $\frac{1}{3}$ to be green. In other words, the function $\hat{\theta}$ is a predictive system, much like any other Carnapian inductive logic. We can write

$$P(q_{t+1}^k | s_t) = P_{h_{\hat{\theta}(s_t)}}(q_{t+1}^k).$$

The estimation function $\hat{\theta}$ by Fisher is thus captured in a single probability function

⁵As recounted in [Zabell, 1982], earlier work that connects exchangeability to the predictive properties of probability functions was done by [Johnson, 1932] and [de Finetti, 1937]. But the specific relation with Hume's problem noted here is due to Carnap: he motivated predictive properties such as Equation (8) independently, by the definition of logical probability, whereas for the subjectivist de Finetti these properties did not have any objective grounding.

P . So we can present the latter as a probability assignment over sample space, from which estimations can be derived by a non-ampliative inference.

Let me make this concrete by means of the example on red and green pears. In the Carnapian prediction rule of Equation (8), choosing $\lambda = 0$ will yield the observed relative frequency as predictions. And according to Equation (5) these relative frequencies are also the maximum likelihood estimators. Thus, for each set of possible observations, $\{s^{k_1 \dots k_t} : k_i = 0, 1\}$, the Carnapian rule with $\lambda = 0$ predicts according to the Fisherian estimate.⁶

Unfortunately the alignment of Fisher estimation and Carnapian inductive logic is rather problematic. Already for estimations for multinomial hypotheses, it is not immediate how we can define the corresponding probability assignment over sample space, and whether we thereby define a coherent probability function at all. For more complicated sets of hypotheses, and the more complicated estimators associated with it, the corresponding probability assignment P may be even less natural, or possibly incoherent. Moreover, the principles and predictive properties that may motivate the choice of that probability function will be very hard to come by. In the following I will therefore not discuss the further intricacies of capturing Fisher's estimation functions by Carnapian prediction rules. However, Carnapian rules will make a reappearance in the next two sections, because in a much more straightforward sense, they are the predictive counterpart to Bayesian statistics. indexFisher, R. A.

7 BAYESIAN STATISTICS

The defining characteristic of Bayesian statistics is that probability assignments do not just range over data, but that they can also take statistical hypotheses as arguments. As will be seen in the following, Bayesian inference is naturally represented in terms of a non-ampliative inductive logic, and it also relates very naturally to Carnapian inductive logic.

Let $\mathcal{H}$ be the space of statistical hypotheses h_θ , and let $\mathcal{Q}$ be the sample space as before. The functions P are probability assignments over the entire space $\mathcal{H} \times \mathcal{Q}$. Since h_θ is a member of the combined algebra, it makes sense to write $P(s_t|h_\theta)$ instead of the $P_{h_\theta}(s_t)$ written in the context of classical statistics. We can define Bayesian statistics as follows.

DEFINITION 3 Bayesian Statistical Inference. Assume the prior probability

⁶Note that the probability function P that describes the estimations is a rather unusual one. After three green pears for example, s^{111} , the probability for the next pear to be red will be 0, so that $P(s^{1110}) = 0$. By the standard axiomatisation and definitions of probability, the probability of any observation q_5^0 conditional on s^{1110} is not defined. But if the probability function P is supposed to follow the Fisherian estimations, then we must have $P(q_5^0|s^{1110}) = \frac{1}{4}$. To accommodate the probability function imposed by Fisher's estimations, we may change the axiomatisation of probability. In particular, we may adopt an axiomatisation in which conditional probability is primitive, as described in [Rényi, 1970]. Alternatively, we can restrict ourselves to estimations based on the observation of more than one property.

$P(h_\theta)$ assigned to hypotheses $h_\theta \in \mathcal{H}$, with $\theta \in \Theta$, the space of parameter values. Further assume $P(s_t|h_\theta)$, the probability assigned to the data s_t conditional on the hypotheses, called the likelihoods. Bayes' theorem determines that

$$P(h_\theta|s_t) = P(h_\theta) \frac{P(s_t|h_\theta)}{P(s_t)}. \quad (9)$$

Bayesian statistics outputs the posterior probability assignment, $P(h_\theta|s_t)$.

See [Barnett, 1999] and [Press, 2003] for a more detailed discussion. The further results form a Bayesian inference, such as estimations and measures for the accuracy of the estimations, can all be derived from the posterior distribution over the statistical hypotheses.

In this definition the probability of the data $P(s_t)$ is not presupposed, because it can be computed from the prior and the likelihoods by the law of total probability,

$$P(s_t) = \int_{\Theta} P(h_\theta) P(s_t|h_\theta) d\theta.$$

The result of a Bayesian statistical inference is not always a posterior probability. Often the interest is only in comparing the ratio of the posteriors of two hypotheses. By Bayes' theorem we have

$$\frac{P(h_\theta|s_t)}{P(h_{\theta'}|s_t)} = \frac{P(h_\theta)P(s_t|h_\theta)}{P(h_{\theta'})P(s_t|h_{\theta'})},$$

and if we assume equal priors $P(h_\theta) = P(h_{\theta'})$, we can use the ratio of the likelihoods of the hypotheses, the so-called Bayes factor, to compare the hypotheses.

Let me give an example of a Bayesian procedure. Consider the hypotheses of Equation (3), concerning the fraction of green pears in Emma's orchard. Instead of choosing among them on the basis of the data, assign a so-called Beta-distribution over the range of hypotheses,

$$P(h_\theta) \propto \theta^{\lambda/2-1} (1-\theta)^{\lambda/2-1} \quad (10)$$

with $\theta \in \Theta = [0, 1]$. For $\lambda = 2$, this function is uniform over the domain. Now say that we obtain a certain sequence of pears, s^{000101} . By the likelihood of the hypotheses as given in Equation (4), we can derive

$$P(h_\theta|s^{000101}) = \theta^{\lambda/2+1} (1-\theta)^{\lambda/2+3}.$$

More generally, the likelihood function for the data s_t with numbers t_k of earlier instances q_i^k is $\theta^{t_1} (1-\theta)^{t_0}$, so that

$$P(h_\theta|s_t) \propto \theta^{\lambda/2-1+t_1} (1-\theta)^{\lambda/2-1+t_0}. \quad (11)$$

is the posterior distribution over the hypotheses. This posterior is derived by the axioms of probability theory alone, specifically by Bayes' theorem.

As said, capturing this statistical procedure in a non-ampliative inference is relatively straightforward. The premises are the prior over the hypotheses, $P(h_\theta)$ for $\theta \in \Theta$, and the likelihood functions, $P(s_t|h_\theta)$ over the algebras $\mathcal{Q}$, which are determined for each hypothesis h_θ separately. These premises are such that only a single probability assignment over the space $\mathcal{H} \times \mathcal{Q}$ remains. In other words, the premises have a unique probability model. Moreover, all the conclusions are straightforward consequences of this probability assignment. They can be derived from the assignment by applying theorems of probability theory, primarily Bayes' theorem.

Before turning to the relation of Bayesian inference with Carnapian logic, let me compare it to the classical procedures sketched in the foregoing. In all cases, we consider a set of statistical hypotheses, and in all cases our choice among these is informed by the probability of the data according to the hypotheses. The difference is that in the two classical procedures, this choice is absolute: acceptance, rejection, and the appointment of a best estimate. In the Bayesian procedure, by contrast, all this is expressed in a posterior probability assignment over the set of hypotheses.

Note that this posterior over hypotheses can be used to generate the kind of choices between hypotheses that classical statistics provides. Consider Fisherian parameter estimation. We can use the posterior to derive an expectation for the parameter θ , as follows:

$$E[\theta] = \int_{\Theta} \theta P(h_\theta|s_t) d\theta. \quad (12)$$

Clearly, $E[\theta]$ is a function that brings us from the hypotheses h_θ and the data s_t to a preferred value for the parameter. The function depends on the prior probability over the hypotheses, but it is in a sense analogous to the maximum likelihood estimator. In analogy to the confidence interval, we can also define a so-called credal interval from the posterior probability distribution:

$$\text{Cred}_{1-\epsilon} = \left\{ \theta : |\theta - E[\theta]| < d \text{ and } \int_{E[\theta]-d}^{E[\theta]+d} P(h_\theta|s_t) d\theta = 1 - \epsilon \right\}.$$

This set of values for θ is such that the posterior probability of the corresponding h_θ jointly add up to $1 - \epsilon$ of the total posterior probability.

Most of the controversy over the Bayesian method concerns the determination and interpretation of the probability assignment over hypotheses. As for interpretation, classical statistics objects to the whole idea of assigning probabilities to hypotheses. The data have a well-defined probability, because they consist of repeatable events, and so we can interpret the probabilities as frequencies, or as some other kind of objective probability. But the probability assigned to a hypothesis cannot be understood in this way, and instead expresses an epistemic state of uncertainty. One of the distinctive features of classical statistics is that it rejects such epistemic probability assignments, and that it restricts itself to a straightforward interpretation of probability as relative frequency.

Even if we buy into this interpretation of probability as epistemic uncertainty, how do we determine a prior probability? At the outset we do not have any idea of which hypothesis is right, or even which hypothesis is a good candidate. So how are we supposed to assign a prior probability to the hypotheses? The literature proposes several objective criteria for filling in the priors, for instance by maximum entropy or by other versions of the principle of indifference, but something of the subjectivity of the starting point remains. The strength of the classical statistical procedures is that they do not need any such subjective prior probability.

8 BAYESIAN INDUCTIVE LOGIC

While Bayesian statistics differs strongly from classical statistics, it is much more closely related to the inductive logic of Carnap. In this section I will elaborate on this relation, and indicate how Bayesian statistical inference and inductive logic may have a fruitful common future.

To see how Bayesian statistics and Carnapian inductive logic hang together, note first that the result of a Bayesian statistical inference, namely a posterior, is naturally translated into the result of a Carnapian inductive logic, namely a prediction,

$$P(q_{t+1}^1 | s_t) = \int_0^1 P(q_{t+1}^1 | h_\theta \cap s_t) P(h_\theta | s_t) d\theta, \quad (13)$$

by the law of total probability. Furthermore, consider the posterior probability over multinomial hypotheses. Recall that the parameter θ is the probability for the next pear to be green, as defined in Equation (3). By Equations (12) and (13) we have

$$\begin{aligned} E[\theta] &= \int_{\Theta} \theta P(h_\theta | s_t) d\theta \\ &= \int_0^1 P(q_{t+1}^1 | h_\theta \cap s_t) P(h_\theta | s_t) d\theta \\ &= P(q_{t+1}^1 | s_t), \end{aligned}$$

This shows that in the case of multinomial statistical hypotheses, the expectation value for the parameter is the same as a predictive probability.

The correspondence between Bayesian statistics and Carnapian inductive logic is in fact even more striking. We can work out the integral of Equation line (13), using Equation (10) as the prior and the multinomial hypotheses defined in Equation 3, to obtain

$$P(q_{t+1}^1 | s_t) = \frac{t_1 + \frac{\lambda}{2}}{t + \lambda}. \quad (14)$$

This means that there is a specific correspondence between certain kinds of predictive probabilities, as described by the Carnapian λ -rules, and certain kinds of Bayesian statistical inferences, namely with multinomial hypotheses and priors

from the family of Dirichlet distributions, which generalise the Beta-distributions used in the foregoing.

On top of this, the equivalence between Carnapian inductive logic and Bayesian statistical inference is more general than is shown in the foregoing. Instead of the well-behaved priors just considered, we might consider any functional form as a prior over the hypotheses h_θ , and then wonder what the resulting predictive probability is. As de Finetti showed in his representation theorem, the resulting predictive probability will always comply to a predictive property known as exchangeability, which was given in Equation (6). Conversely, and more surprisingly, any predictive probability complying to the property of exchangeability can be written down in terms of a Bayesian statistical inference with multinomial hypotheses and some prior over these hypotheses. In sum, de Finetti showed that there is a one-to-one correspondence between the predictive property of exchangeability on the one hand, and Bayesian statistical inferences using multinomial hypotheses on the other.

It is insightful to make this result by de Finetti explicit in terms of the non-ampliative inductive logic discussed in the foregoing. Recall that a Bayesian statistical inference takes a prior and likelihoods as premises, leading to a single probability assignment over the space $\mathcal{H} \times \mathcal{Q}$ as the only assignment satisfying the premises. We infer probabilistic consequences, such as the posterior and the predictions, from this probability assignment. Similarly, a Carnapian inductive logic is characterised by a single probability assignment, defined over the space $\mathcal{Q}$, from which the predictions can be derived. The representation theorem by de Finetti effectively shows an equivalence between these two probability assignments: when it comes to predictions, we can reduce the probability assignment over $\mathcal{H} \times \mathcal{Q}$ to an assignment over $\mathcal{Q}$ only.

For de Finetti, this equivalence was very welcome. He had a strictly subjectivist interpretation of probability, believing that probability expresses uncertain belief only. Moreover, he was eager to rid science of its metaphysical excess baggage to which, in his view, the notion of objective chance belonged. So de Finetti applied his representation theorem to argue against the use of multinomial hypotheses, and thereby against the use of statistical hypotheses more generally. Why refer to these obscure chances if we can achieve the very same statistical ends by employing the unproblematic notion of exchangeability? The latter is a predictive property, and it can hence be interpreted as an empirical and as a subjective notion.

The fact is that statistics, as it is used in the sciences, is persistent in its use of statistical hypotheses. Therefore I want to invite the reader to consider the inverse application of de Finetti's theorem. Why does science use these obscure objective chances? As I argue extensively in [Romeijn, 2005], the reason is that statistical hypotheses provide invaluable help by, indirectly, pinning down the probability assignments over $\mathcal{Q}$ that have the required predictive properties. Rather than reducing the Bayesian inferences over statistical hypotheses to inductive predictions over observations, we can use the representation theorem to capture relations between observations in an insightful way, namely by citing the statistical

hypotheses that may be true of the data. As further illustrated in [Romeijn, 2004; Romeijn, 2006], enriching inductive logic in this way improves the control that we have over predictive properties.

Finally, it may be noted that this view on inductive logic is comparable to the “presupposition view” in [Festa, 1993], which takes a similar line with regards to the choice of λ in Carnapian inductive logic. It is also strongly related to the views expressed by Hintikka in [Auxier and Hahn, 2006], and I want to highlight certain aspects of this latter view in particular. In response to Kuipers’ overview of inductive logic, Hintikka writes that “Inductive inference, including rules of probabilistic induction, depends on tacit assumptions concerning the nature of the world. Once these assumptions are spelled out, inductive inference becomes in principle a species of deductive inference.” Now the symmetry principles and predictive properties used in Carnapian inductive logic are exactly the tacit assumptions Hintikka speaks about. As explained in the foregoing, the use of particular statistical hypotheses in a Bayesian inference comes down to the very same set of assumptions, but now these assumptions are not tacit anymore: they have been made explicit as the choice for a particular set of statistical hypotheses. Therefore, the use of statistical hypotheses that I have advertised above may help us to get closer to the ideal of inductive logic envisaged by Hintikka.

9 NEYMAN-PEARSON TEST AS AN INFERENCE

In this final section, I investigate whether we can turn the Neyman-Pearson procedure of Section 4 into an inference within Bayesian inductive logic. This might come across as a pointless exercise in statistical yoga, trying to make Neyman and Pearson relax in a position that is far from natural. However, the exercise will nicely illustrate the use of Bayesian inductive logic. Moreover, I think that it will bring Neyman-Pearson testing and inductive logic closer together, and thereby stimulate research on the intersection of inductive logic and statistics in the sciences.

An additional reason for investigating Neyman-Pearson hypothesis testing in this framework is that in many practical applications, scientists are tempted to read the probability statements about the hypotheses inversely: the significance is often taken as the probability that the null hypothesis is true. Although emphatically wrong, this inferential reading has a strong intuitive appeal to users. The following will make explicit that in this reading, the Neyman-Pearson procedure is effectively taken as a kind of non-ampliative inductive inference.

First, we construct the space $\mathcal{H} \times \mathcal{Q}$, and define the probability functions P_{h_j} over the sample spaces $\langle h_j, \mathcal{Q} \rangle$. For the prior probability assignment over the two hypotheses, we take $P(h_0) \in (l, u)$, meaning that $l < P(h_0) < u$. Finally, we adopt the restriction that $P(h_0) + P(h_1) = 1$. This defines a set of probability functions over the entire space, serving as a starting point of the inference.

Next we include the data in the probability assignments. Crucially, we coarse-

grain the observations to the simple observation f^j , with

$$f^j = \{s_t : F(s_t) = j\},$$

so that the observation simply encodes the value of the test function. It follows from this coarse-graining that we obtain the type-I and type-II errors as the likelihoods of the observations,

$$\begin{aligned} P(f^1|h_0) &= \alpha, \\ P(f^0|h_1) &= \beta. \end{aligned}$$

Finally we use Bayes' theorem to derive a set of posterior probability distributions over the hypotheses, according to

$$\frac{P(h_1|f^j)}{P(h_0|f^j)} = \frac{P(f^j|h_1)P(h_1)}{P(f^j|h_0)P(h_0)}.$$

Note that the quality of the test, in terms of size and power, will be reflected in the posteriors. If, for example, we find an observation s_t that allows us to reject the null hypothesis, so f^1 , then as long as $\alpha < 1 - \beta$, meaning that the significance is smaller than the power, we find that $\underline{P}(h_0|f^1) < \underline{P}(h_0)$ and $\overline{P}(h_0|f^1) < \overline{P}(h_0)$. The larger the difference between significance and power, the larger the difference between posteriors and priors.

Note, however, that we have not yet decided on a fully specified prior probability over the statistical hypotheses. This echoes the fact that classical statistics does not make use of a prior probability. However, it is only by restricting the prior probability over hypotheses in some way or other that we can make the Bayesian rendering of the results of Neyman and Pearson work. In particular, if we choose $(l, u) = (0, 1)$ for the prior, then we find $(l', u') = (0, 1)$ for the posterior as well. However, if we choose

$$l \geq \frac{\beta}{\beta + 1 - \alpha}, \quad u \leq \frac{1 - \beta}{1 - \beta + \alpha},$$

we find for all $P(h^0) \in (l, u)$ that $\overline{P}(h_0|f^1) < \frac{1}{2} < \underline{P}(h_1|f^1)$. Similarly, we find $\underline{P}(h_0|f^0) > \frac{1}{2} > \overline{P}(h_1|f^0)$. So with this interval prior, an observation s_t for which $F(s_t) = 1$ tilts the balance towards h_1 for all the probability functions P in the interval, and vice versa.

Let me illustrate the Bayesian inference by means of the above example on pears. We set up the sample space and hypotheses as before, and we then coarse-grain the observations to f^j , corresponding to the value of the test function, $f^1 = q^0 \cup q^1$ and $f^0 = q^2$. We obtain

$$\begin{aligned} P(f^1|h_0) &= P(q^0 \cup q^1|h_0) = \alpha = 0.05 \\ P(f^0|h_1) &= P(q^0 \cup q^1|h_1) = \beta = 0.30 \end{aligned}$$

Choosing $P(h_0) \in (0.24, 0.93)$, this results in $P(h_0|f^0) = (0.50, 0.98)$, and $P(h_0|f^1) = (0.02, 0.50)$.

Depending on the choice of prior, one can argue that the resulting Bayesian inference replicates the Neyman-Pearson procedure: if the probability over hypotheses expresses our preference over them, then indeed f^0 makes us prefer h_0 and f^1 makes us prefer h_1 . Importantly, the inference fits the entailment relation mentioned earlier: we have a set of probabilistic models on the side of the premises, namely the set of priors over $\mathcal{H}$, coupled to the full probability assignments over $\langle h_j, \mathcal{Q} \rangle$ for $j = 0, 1$. And we have a set of models on the conclusion side, namely the set of posteriors over $\mathcal{H}$. Because the latter is computed from the former by the axioms of probability, the two sets include the same probability functions. Therefore the conclusion is classically entailed by the premises.

The above example shows that we can imitate the workings of a Neyman-Pearson test in Bayesian inductive logic, and thus in terms of a non-ampliative inductive inference. But the imitation is far from perfect. For one, the results of a Bayesian inference will always be a probability function. By contrast, Neyman-Pearson statistics ends in a decision to accept or reject, which is a binary decision instead of some sort of weak or inconclusive preference. Of course, there are many attempts to weld a binary decision onto the probabilistic end result of a Bayesian inference, for example in [Levi, 1980] and in the discussion on rational acceptance, e.g., [Douven, 2002]. In particular, we might supplement the probabilistic results of a Bayesian inference with rules for translating the probability assignments into decisions, e.g., we choose h_0 if we have $\underline{P}(h_0|s_t) > \frac{1}{2}$, and similarly for h_1 . However, the bivalence of Neyman-Pearson statistics cannot be replicated in a Bayesian inference itself. It will have to result from a decision-theoretic add-on to the inferential part of Bayesian statistics.

More generally, the representation in probabilistic logic will probably not appeal to advocates of classical statistics. Quite apart from the issue of binary acceptance, the whole idea of assuming a prior probability, however unspecific, may be objected to on the principled ground that probability functions express long-term frequencies, and that hypotheses cannot have such frequencies.

There is one attractive feature, at least to my mind, of the above rendering, that may be of interest in its own right. With the representation in place, we can ask again how to understand the example by Jeffreys, as considered in Section 4. Following [Edwards, 1972], it illustrates that Neyman and Pearson tests do not respect the likelihood principle, because they depend on the probability assignment over the entire sample space and not just on the probability of the observed sample. However, in the Bayesian representation we do respect the likelihood principle, but in addition we condition on f^j , not on q^k . In fact the whole example hinges on how the samples are grouped into regions of acceptance and rejection. Instead of adopting the diagnosis by Hacking concerning the likelihood principle, we could therefore say that the approach of Neyman and Pearson takes the observations in terms of a rather coarse-grained partition of information. In other words, rather than saying that Neyman-Pearson procedures violate the likelihood principle, we

can also say that the procedures crucially depend on how the observed sample is framed, and thus violate the principle of total evidence.

10 IN CONCLUSION

In the foregoing I have discussed three statistical procedures, to wit, Neyman-Pearson hypotheses testing, Fisher's maximum likelihood estimation, and Bayesian statistical inference. These three procedures were seen to relate to inductive logic in a variety of ways.

The two classical approaches were connected most naturally to ampliative inductive inference, running from a set of probability functions and the data to a restricted set of such functions. However, I have also related both procedures to non-ampliative inferences. First I connected parameter estimation to Carnapian inductive logic. Then I related this logic to Bayesian statistical inference, which was seen to be non-ampliative already. Further, I have indicated how Carnapian inductive logic can be extended to Bayesian inductive logic, which accommodates the use of statistical hypotheses and thus captures Bayesian statistics. Finally, I have illustrated the latter logic by giving a non-ampliative account of Neyman-Pearson hypothesis testing.

I hope that portraying statistical procedures in the setting of inductive logic has been illuminating. In particular, I hope that the relation between Carnapian inductive logic and Bayesian statistics stimulates research on the intersection of the two. Certainly, some research in this area has already been conducted; see for example [Skyrms, 1991; Skyrms, 1993; Skyrms, 1996] and [Festa, 1993]. Following these contributions, [Romeijn, 2005] argues that an inductive logic that includes statistical hypotheses in its language is closely related to Bayesian statistical inference, and some of these views have been reiterated in this chapter. However, I believe that there is much room for improvement. Research on the intersection of inductive logic and statistical inference can certainly enhance the relevance of inductive logical systems to scientific method and the philosophy of science. In parallel, I believe that insights from inductive logic may help to clarify the foundations of statistics.

ACKNOWLEDGEMENTS

I want to thank Roberto Festa for comments on an earlier draft of this chapter, and Jos Uffink for drawing my attention to Jeffreys' amusing quote on the "remarkable procedure". I also thank the Spanish Ministry of Science and Innovation (Research project FFI2008-1169) for financial support. This research was carried out as part of a project funded by the Dutch Organization of Scientific Research (NWO VENI-grant nr. 275-20-013).

BIBLIOGRAPHY

- [Auxier and Hahn, 2006] R.E. Auxier and L.E. Hahn, editors. *The Philosophy of Jaako Hintikka*. Open Court, Chicago, 2006.
- [Barnett, 1999] V. Barnett. *Comparative Statistical Inference*. John Wiley, New York, 1999.
- [Carnap, 1950] R. Carnap. *Logical Foundations of Probability*. University of Chicago Press, 1950.
- [Carnap, 1952] Rudolf Carnap. *The Continuum of Inductive Methods*. University of Chicago Press, Chicago, 1952.
- [Dawid and Stone, 1982] A. P. Dawid and M. Stone. The functional-model basis of fiducial inference (with discussion). *Annals of Statistics*, 10(4):1054–1074, 1982.
- [de Finetti, 1937] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1):1–68, 1937.
- [Douven, 2002] I. Douven. A new solution to the paradoxes of rational acceptability. *The British Journal for the Philosophy of Science*, 53:391–410, 2002.
- [Edwards, 1972] A.W.F. Edwards. *Likelihood*. Cambridge University Press, 1972.
- [Festa, 1993] R. Festa. *Optimum Inductive Methods*. Dordrecht: Kluwer, 1993.
- [Festa, 1996] R. Festa. Analogy and exchangeability in predictive inferences. *Erkenntnis*, 45:89–112, 1996.
- [Fisher, 1930] Ronald A. Fisher. Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535, 1930.
- [Fisher, 1935] Ronald A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6:317–324, 1935.
- [Fisher, 1956] Ronald A. Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, 1956.
- [Hacking, 1965] I. Hacking. *The Logic of Statistical Inference*. Cambridge University Press, Cambridge, 1965.
- [Haenni et al., 2009] Rolf Haenni, Jan-Willem Romeijn, Greg Wheeler, and Jon Williamson. *Probabilistic Logics and Probabilistic Networks*. Springer, 2009.
- [Hailperin, 1996] T. Hailperin. *Sentential Probability Logic*. Lehigh University Press, 1996.
- [Hartmann et al., 2009] S. Hartmann, D. Gabbay, and J. Woods, editors. *Handbook for the History of Logic: Inductive Logic (Volume 10)*. College Publications, 2009.
- [Howson, 2003] C. Howson. Probability and logic. *Journal of Applied Logic*, 1(3–4):151–165, 2003.
- [Jeffreys, 1931] H. Jeffreys. *Scientific Inference*. Cambridge University Press, , Cambridge, 1931.
- [Johnson, 1932] W. Johnson. Probability: the deductive and inductive problems. *Mind*, 49:409–423, 1932.
- [Kyburg, 1974] Henry E. Kyburg, Jr. *The Logical Foundations of Statistical Inference*. D. Reidel, Dordrecht, 1974.
- [Levi, 1980] Isaac Levi. *The enterprise of knowledge: an essay on knowledge, credal probability, and chance*. MIT Press, Cambridge MA, 1980.
- [Maher, 2000] P. Maher. Probabilities for two properties. *Erkenntnis*, 52:63–81, 2000.
- [Neyman and Pearson, 1967] J. Neyman and E. Pearson. *Joint Statistical Papers*. University of California Press, Berkeley, 1967.
- [Paris and Waterhouse, 2008] J. Paris and P. Waterhouse. Atom exchangeability and instantial relevance. *unpublished manuscript*, 2008.
- [Press, 2003] J. Press. *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. John Wiley, New York, 2003.
- [Rényi, 1970] A. Rényi. *Probability Theory*. North Holland, Amsterdam, 1970.
- [Romeijn, 2004] J.W. Romeijn. Hypotheses and inductive predictions. *Synthese*, 141(3):333–64, 2004.
- [Romeijn, 2005] J.W. Romeijn. *Bayesian Inductive Logic*. PhD dissertation, University of Groningen, 2005.
- [Romeijn, 2006] J.W. Romeijn. Analogical predictions for explicit similarity. *Erkenntnis*, 64:253–280, 2006.
- [Seidenfeld, 1979] T. Seidenfeld. *Philosophical Problems of Statistical Inference: Learning from R.A. Fisher*. Reidel, Dordrecht, 1979.

- [Skyrms, 1991] B. Skyrms. Carnapian inductive logic for markov chains. *Erkenntnis*, 35:35–53, 1991.
- [Skyrms, 1993] B. Skyrms. Analogy by similarity in hyper-Carnapian inductive logic. In J. Earman, A. I. Janis, G. Massey, and N. Rescher, editors, *Philosophical Problems of the Internal and External Worlds*, pages 273–282. University of Pittsburgh Press, Pittsburgh, 1993.
- [Skyrms, 1996] B. Skyrms. *Statistics, Probability, and Game*, chapter Carnapian Inductive Logic and Bayesian Statistics, pages 321–336. IMS Lecture Notes, 1996.
- [Wheeler, 2006] Gregory Wheeler. Rational acceptance and conjunctive/disjunctive absorption. *Journal of Logic, Language and Information*, 15(1-2):49–63, 2006.
- [Zabell, 1982] S. Zabell. W. E. Johnson’s “sufficientness” postulate. *Annals of Statistics*, 10:1091–99, 1982.

Part IX

Various Issues about Causal Inference

This page intentionally left blank

COMMON CAUSE IN CAUSAL INFERENCE

Peter Spirtes

1 INTRODUCTION

One of the major impediments to reliably inferring both qualitative and quantitative causal relations from non-experimental data is the possibility that there may be unobserved common causes of observed variables. Suppose, for example, the correlation between two observed variables is measured, e.g. *Barometer Reading* and *Rainfall*, and that there is no domain background knowledge about the causal relations between these variables. A qualitative causal question is of the form “Will manipulating the barometer reading (e.g. by rotating a dial showing the barometer reading) affect subsequent rainfall?” A quantitative causal questions is of the form “How much does manipulating the barometer reading to m affect subsequent rainfall?”¹

The possibility of unobserved common causes presents no problems for causal inference when simple randomized experiments are possible, beyond the ordinary statistical problems of inferring a population distribution from a sample.² For example, if it is possible to randomly select a number and set the value of the barometer reading according to the outcome of the randomizing device, it is clearly possible to answer both the qualitative and quantitative causal questions simply by observing the subsequent rainfall, given a large enough sample size. However, in many instances randomized experiments cannot be performed due to ethical considerations, or practical considerations such as the amount of time and money required to perform experiments. This article will consider cases where randomized experiments are not possible.

For example, the *direct effect* of *Barometer Reading* on *Rainfall* relative to the set of variables $\mathbf{S} = \{\textit{Rainfall}, \textit{Atmospheric Pressure}, \textit{Barometer Reading}\}$ can ideally be experimentally determined by performing the following two manipulations of variables in $\mathbf{S}$. In both manipulations *Atmospheric Pressure* is set to the same

¹This example will be revisited in section 3.2.3 in more detail. Although qualitative features of the causal relationships in this example are widely known, it is highly idealized in a number of ways, including not being specific about precisely how the variables are measured. As in many social science models, the example does not specify time indices for the variables, which raises a number of interesting questions about the interpretations of these models [Fisher, 1970]. However, these issues are not particular problems for models with unobserved common causes and will not be discussed in detail here.

²On the other hand, in sequential randomized trials with unobserved common causes, the estimation of causal effects can be quite intricate [Robins, 1986].

constant c . In the first manipulation *Barometer Reading* is set to the constant d , and in the second manipulation, *Barometer Reading* is set to $d + 1$. The average difference in *Rainfall* resulting from these two different manipulations is the direct effect of *Barometer Reading* on *Rainfall* relative to the set of variables $\mathbf{S}$.

Suppose however that the experiments cannot be performed. Causal modelers in a variety of social sciences have an easy way to answer quantitative causal questions, given (often unrealistically) strong background assumptions. For example, suppose it is assumed what the time order of events is, that the causal relationships are linear, and that there are no unobserved common causes of any pair of variables in $\mathbf{S}$. The direct effect of *Barometer Reading* on *Rainfall* relative to $\mathbf{S}$ is equal to the regression coefficient relating *Barometer Reading* to *Rainfall* when *Rainfall* is regressed on both *Barometer Reading* and *Atmospheric Pressure* (all of the variables in $\mathbf{S}$ that temporally precede *Rainfall*) in the unmanipulated population (i.e. the population where the barometer dial is not manually set). The coefficient of *Barometer Reading* in this regression will be zero, which is the correct answer about the direct effect of *Barometer Reading* on *Rainfall* relative to $\mathbf{S}$. (See section 3.2.3 for more details on regression.)

But, without the assumption that there are no unmeasured common causes, the regression method is not reliable as a method for calculating the direct effect. For example, if *Atmospheric Pressure* is not observed, and $\mathbf{S}'$ is the set of observed variables $\{\textit{Rainfall}, \textit{Barometer Reading}\}$, then *Atmospheric Pressure* is an unmeasured common cause of *Barometer Reading* and *Rainfall*. The direct effect of *Barometer Reading* on *Rainfall* relative to $\mathbf{S}'$ is still zero. Regressing *Rainfall* on all of the *observed* variables (i.e. variables in $\mathbf{S}'$) that precede it in time results in a regression of *Rainfall* on *Barometer Reading*. Because of the non-zero correlation between *Barometer Reading* and *Rainfall*, the regression coefficient when *Barometer Reading* is regressed on *Rainfall* is also non-zero. In this case, the regression coefficient is not equal to the direct effect of *Barometer Setting* on *Rainfall* relative to $\mathbf{S}'$.

One partial solution to this problem that is commonly employed in many social sciences is to attempt to observe as many potential common causes (i.e. variables occurring prior to the effect) as possible. However, in many cases there is no reliable way to tell if all of the common causes have been observed. Without background knowledge, there is no way to tell whether the regression coefficient of *Barometer Reading* when *Rainfall* is regressed on *Barometer Reading* is non-zero because *Barometer Reading* causes *Rainfall*, or because it does not cause *Rainfall* but some unobserved third variable, e.g. *Atmospheric Pressure*, causes both *Rainfall* and *Barometer Reading*. Even worse, it is not generally recognized that including *more* potential common causes in a regression can make the difference between the regression coefficient and the direct effect being estimated *larger* [Spirtes *et al.*, 1998]. See section 3.2.3 for more details and examples.

There is no reliable method of causal inference of the effect of manipulating *Barometer Reading* on *Rainfall* from the *correlation* between *Barometer Reading* and *Rainfall* alone. This is the truth in the maxim “Correlation is not causation.”

However, with some additional assumptions in many cases reliable causal inference is possible even when there may be unrecorded causes unknown prior to the analysis of the data, and time order is not known. The following sections describe a variety of assumptions and algorithms for reliable causal inference, even when there may be unobserved common causes.

There are many types of statistical/causal models that postulate unobserved common causes. These include principal components models, factor analytic models (see section 4.2.4), item response models, some structural equation models, Rasch models, and finite mixture models. (See e.g. [Bartholomew & Knott, 1999; Harman, 1976]). They differ in the families of distributions they represent, and the kinds of constraints they entail on marginal distributions over observed variables—the constraints that help to make possible inferences to the existence and causal roles of unobserved variables. It should be noted that the statistical terminology is ambiguous. For example, “factor analysis,” denotes a family of methods (of dubious reliability) for generating models with latent variables; a “factor analysis” model is any model that might be produced by such a procedure — the very same model, if produced by hand, might be termed a “multiple indicator model”. One and the same statistical model might, or might not, be given a causal interpretation. This article will not describe all of these models in detail, but will use structural equation models (SEMs, section 2) as representative of a large category of causal models with unobserved common causes (including principal components and factor analysis models as a special case).³

At least in publications, common methods of causal inference in the social sciences either assume at the outset that an appropriate causal explanation will involve unobserved common causes, or assume that any common causes are themselves among the measured variables. Such procedures beg important questions: are there unobserved common causes at work? If so, how can they be found, and how can their causal relations be found? What causal relations among the measured variables can be found in spite of unobserved common causes? These questions are the focus of this survey.

Section 2 describes structural equation models. Section 3 introduces the concept of a manipulation, in order to clearly distinguish several different kinds of causal inference from statistical inference. Section 4 discusses problems for reliable inference of qualitative causal relations from observational data and background assumptions, and some approaches that have been taken to solving these problems. Section 5 describes some important open problems. Section 6 is an appendix that defines some of the technical terms used in the text.

³In the epidemiological literature, the problem of inferring the effects of manipulations is described as the problem of inferring potential outcomes. See [Rubin, 1974].

2 STRUCTURAL EQUATION MODELS

The set of variables in a structural equation model (SEM) can be divided into two subsets, the “error variables” or “error terms,” and the substantive variables (for which there is no standard terminology in the literature). The substantive variables are the variables of interest, but they are not necessarily all observed. In SEM *C*, shown in Figure 1, the substantive variables are the variables $\{X, Y, Z, W\}$. (The various parts of Figure 1 are explained in more detail below). For example, in a simplified model of money supply in equilibrium [Wyatt, 2004], Y could represent investment, Z could represent money supply, X could represent *GDP*, and W could represent monetary base. Typically, for each substantive variable X there is a corresponding error term for X that represents all of the causes of X that are not substantive variables.⁴ Each substantive variable such as X occurs on the left hand side of one equation that relates the value of X to the direct causes of X plus the error term ε_X on the right hand side of the equation [Bollen, 1989].

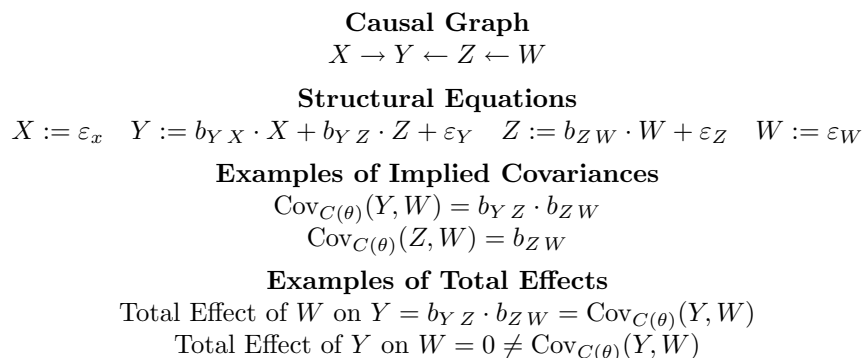


Figure 1. SEM *C*

For purposes of exposition, the error terms will be assumed to be Gaussian and the structural equations will assumed to be linear, unless explicitly stated otherwise. When the error terms are Gaussian, and the structural equations are linear, all of the substantive variables will also be Gaussian. Without loss of generality, the substantive variables can be transformed to have distribution $N(0, 1)$ (where N indicates the distribution is Gaussian, the first number represents the mean, and the second number represents the standard deviation) and the error terms can be transformed to have zero mean.

For example, in SEM *C* the structural equation for Y is “ $Y := b_{YX} \cdot X + b_{YZ} \cdot Z + \varepsilon_Y$ ”, where b_{YX} ⁵ and b_{YZ} are linear structural coefficients, and ε_Y is the error

⁴Error terms under a causal interpretation should be distinguished from the residuals from a regression (i.e. the difference between the predicted value from a regression and the actual value of a variable), although under certain conditions they are equal. See section 3.2.3.

⁵In general, if Y has a non-zero coefficient in the structural equation for X , it is denoted by

term for Y . Using Lauritzen's notation (Lauritzen, 2001) the structural equation uses an assignment operator, “ $:=$ ”, to relate the left hand side and the right hand side. This indicates that the variables on the right hand side are to be interpreted as causes of the left hand side and that the equation can be used not only to describe the relationships between variables in the existing population, but also to predict the effects of manipulating some of the variables (see section 3 for details.)

SEMs have two forms — a free parameter form, and a fixed parameter form. In the free parameter form, the linear coefficients in the structural equations (e.g. b_{YZ}), and the covariance matrix of the error terms (e.g. $\text{var}(\varepsilon_Y)$) are variables. In the fixed parameter form, the linear coefficients in the structural equations (e.g. b_{YZ}), and the covariance matrix among the error terms (e.g. $\text{var}(\varepsilon_Y)$) are constants. The context will make it clear whether b_{YZ} refers to a variable (in a SEM with free parameters) or a constant (in a SEM with fixed parameters.)

If SEM C has free parameters, $C(\theta)$ represents the fixed parameter SEM where the free parameter have been assigned fixed values according to the assignment θ (e.g., $\theta = \{b_{YZ} = b_{YX} = b_{ZW} = 0.4, \text{var}(\varepsilon_X) = \text{var}(\varepsilon_W) = 1, \text{var}(\varepsilon_Y) = 0.68, \text{var}(\varepsilon_Z) = 0.84\}$, where in this example the error terms are assumed to be uncorrelated).

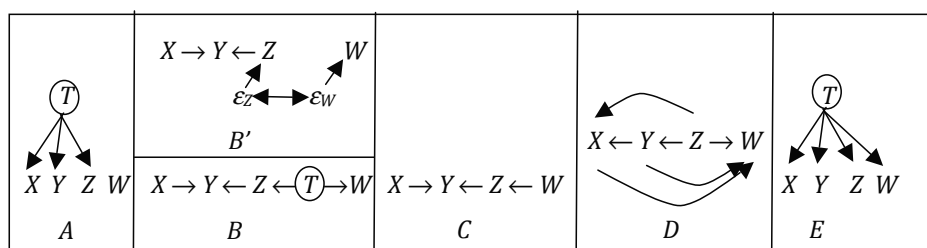


Figure 2. Alternatives to SEM C

For example, in SEM B of Figure 2, if the set of substantive variables is $\{X, Y, Z, T, W\}$ then the causal graph is $X \rightarrow Y \leftarrow Z \leftarrow T \rightarrow W$. However, if the substantive variables are $\{X, Y, Z, W\}$ the corresponding causal graph is shown in B' of Figure 2. If T is left out of the substantive variables, then the “other non-substantive causes” of Z and W both contain T , which correlates the error terms of Z and W . The $\varepsilon_Z \leftrightarrow \varepsilon_W$ edge is typically taken to imply that there is some unobserved common cause of Z and W , but it is less specific than the $Z \leftarrow T \rightarrow W$ representation of a common cause (because the former does not specify how many common causes there are or their relationships to each other, while the latter does.)

b_{XY} .

Let $\mathbf{X}$ be the set of all substantive variables in SEM C .⁶ The structural equations together with the distribution of the error terms in $C(\theta)$ entail a probability distribution $f_{C(\theta)}(\mathbf{X})$ over the substantive variables $\mathbf{X}$. If the joint distribution is multivariate Gaussian, the distribution $f_{C(\theta)}(\mathbf{X})$ is completely characterized by the covariance matrix $\Sigma_{C(\theta)}$ over the substantive variables, and the means μ of the substantive variables. $f_{C(\theta)}(\mathbf{X}) \sim N(\Sigma_{C(\theta)}, \mu)$ is a function of the fixed parameter values, e.g. $\text{cov}_{C(\theta)}(Y, W) = b_{YZ} \cdot b_{ZW}$.⁷ If, for a given covariance matrix Σ here exists an assignment of parameter values θ of SEM C so that the entailed covariance matrix is Σ (i.e. $\Sigma_{C(\theta)} = \Sigma$) say that SEM C *represents* Σ . If, for each Σ that can be represented by SEM C , there is a unique assignment of values to the free parameters θ such that $\Sigma_{C(\theta)} = \Sigma$, then the free parameters of SEM C are *identifiable*. If the free parameters of SEM C are identifiable, then given a covariance matrix Σ that is represented by SEM C , the values of the free parameters are uniquely determined by Σ .

For purposes of illustration, it will be assumed that the graphs are acyclic unless explicitly stated otherwise. The assumptions of Gaussian error terms, linear structural equations, and acyclic graphs simplify the examples while still illustrating the basic problems for causal inference and the basic strategies for solving the problems. Many of the algorithms for causal inference that succeed on this case can be extended to work on more general cases (which will be pointed out in the corresponding discussion.)

The following rule explains how to use the graph of a SEM C to calculate $\text{Cov}_{C(\theta)}(Y, W)$. The rule for calculating covariances from the graph in this way will be helpful in explaining why causal inference with unobserved common causes introduces problems that do not occur when there are no unobserved common causes.

A *path* between X and W is a sequence of adjacent edges starting with X and ending with W , or starting with W and ending with X , where each edge occurs at most once. (A path between X and W is also a path between W and X .) For example, $X \rightarrow Y \leftarrow Z \leftarrow W$ is a path in the graph of SEM C between X and W , and also a path between W and X . A *directed path* from W to Y is a special kind of path in which all of the edges on the path point towards Y . $Y \leftarrow Z \leftarrow W$ is a directed path from W to Y , but not a directed path from Y to W (because the arrows point towards Y , not W). $X \rightarrow Y \leftarrow Z \leftarrow W$ is not a directed path from W to X because not all of the arrows point towards X ; it is also not a directed path from X to W , because not all of the arrows point towards W . A *trek* between Y and W is either a directed path from Y to W , or a directed path from W to Y , or a pair of directed paths from some variable Z to W and Y that intersect only at Z . (A trek between Y and W is also a trek between W and Y .) There are no

⁶Individual variables are in italics, and sets of variables are in boldface.

⁷In matrix form, the structural equations are $\mathbf{X} = \mathbf{B}\mathbf{X} + \varepsilon$, where ε is the set of all error terms, and $\mathbf{B}$ is the structural coefficient matrix. If the covariance matrix among the ε is Σ_ε , $\Sigma_{S(\theta)} = (\mathbf{I} - \mathbf{B})^{-1} \Sigma_\varepsilon (\mathbf{I} - \mathbf{B})^{-1T}$, where $\mathbf{I}$ is the identity matrix, $(\mathbf{I} - \mathbf{B})^{-1}$ is the inverse of $(\mathbf{I} - \mathbf{B})$, and $(\mathbf{I} - \mathbf{B})^{-1T}$ is the transpose of the inverse of $(\mathbf{I} - \mathbf{B})$.

treks between X and W in SEM C because the only path between X and W is not a directed path from X to W , or a directed path from W to X , and there is no pair of directed paths to X and W from some third variable. However, there is a trek between Y and W , namely $Y \leftarrow Z \leftarrow W$ (which is also a trek between W and Y).

A *trek product* is the product of the linear structural coefficients associated with each edge on the trek.⁸ For example, the trek product of $Y \leftarrow Z \leftarrow W$ is $b_{YZ} \cdot b_{ZW}$. It can be shown that $\text{Cov}_{C(\theta)}(Y, W)$ is the *trek sum*, i.e. the sum of all trek products between Y and W [Spirtes *et al.*, 2001]. SEM C has only one trek between Y and W , so $\text{cov}_{C(\theta)}(Y, W) = b_{YZ} \cdot b_{ZW}$. $\text{cov}_{C(\theta)}(X, W) = 0$ because there is no trek between X and W in SEM C .

3 CONDITIONING VERSUS MANIPULATING

The fundamental difference between a statistical but non-causal model and a statistical/causal model is that the latter can be used to both represent a family of probability distributions and calculate the effects of manipulating variables (roughly performing randomized experiments on the variables), while the former can be used to represent a family of distributions but cannot be used to calculate the effects of manipulating variables. This section elaborates on this distinction.

3.1 Conditioning

The probability density function⁹ of Y *conditional* on $W = m$ (denoted $f(Y|W = m)$) represents the density of Y in the *subpopulation* where $W = m$, and is defined from the joint distribution as $f(Y|W = m) = f(Y, W)/f(W = m)$ (where $f(Y, W)$ is the joint density of Y and W , and $f(W = m) \neq 0$). The conditional density depends only upon the joint density (assuming the values of the variables conditioned on do not have density 0), and does not depend upon the causal relationships. When a variable W is conditioned on, this intuitively represents *seeing* the value of W .

For simplicity, instead of using joint density to illustrate various concepts about conditioning and manipulating, the means of variables will be used instead. The mean of Y conditional on $W = m$ will be denoted $E(Y|W = m)$. In SEM $C(\theta)$, it can easily be shown that $E_{C(\theta)}(Y|W = m) = m \cdot \text{cov}_{C(\theta)}(Y, W) = m \cdot b_{YZ} \cdot b_{ZW}$ (where b_{YZ} and b_{ZW} are the structural coefficients in $C(\theta)$.)

Conditional densities and conditional probabilities are useful for problems (e.g. diagnosis) in which the value of the variable of interest is expensive or difficult

⁸This assumes that the variance of each substantive variable is equal to 1. Otherwise the product of structural coefficient on the trek should also be multiplied by the variance of the unique vertex on the trek that has no edges pointing into it (the *source* of the trek.)

⁹A probability density function represents a probability distribution for continuous variables in terms of integrals. In the case of Gaussian variables, it is completely specified by the covariance matrix and the means of the variables.

to measure, but other related variables that have information about the variable of interest are more easily observed. For example, the fact that the probability of measles conditional on spots is much higher than the probability of measles conditional on no spots is useful information in diagnosing whether someone has measles, since directly measuring the presence of measles is much more difficult and expensive than observing that they have spots.

3.2 *Manipulating*

In contrast to conditioning, a manipulated probability density is not a density in a subpopulation of the population, but is a density in a (possibly) hypothetical population formed by manipulating the value of a variable or variables in a causal system.

For example, in an experiment in which W is randomized (which might be impossible to do in practice), whatever causes of W existed in the existing population are replaced by the outcome of a randomization device as the sole cause of W . In contrast to conditioning, which corresponds to *seeing*, manipulating corresponds to *doing* [Pearl, 2000]. Intuitively, the density of rainfall after seeing that barometer reading is m (conditioning on the barometer reading) is different than the density of rainfall after manipulating the barometer reading to m . (The assumption is made that the manipulation is ideal. The only direct effect of an ideal manipulation is on the variable being manipulated — any other effects are due to the change in the variable being manipulated.)

If there is a structural equation model that correctly represents the causal structure of the general population, then manipulating the value of a variable W is represented by a new structural equation model. The new structural equation model replaces the old structural equation for W by a new structural equation that relates W to the outcome of the randomizing device; in addition, all of the other structural equations are left the same [Strotz and Wold, 1960; Spirtes *et al.*, 2001; Pearl, 2000].

For example, if SEM C is true, then the effect of randomizing Y to some new distribution can modeled by replacing “ $Y := b_{YX} \cdot X + b_{YZ} \cdot Z + \varepsilon_Y$ ” with “ $Y := \varepsilon'_Y$ ”, where ε'_Y represents the output of the randomizing device.

When Y is manipulated this way, a new model $C(\theta')$ is created out of $C(\theta)$, where $\theta' = \theta$, except that in θ' , $\text{var}(\varepsilon_Y)$ is replaced by $\text{var}(\varepsilon'_Y)$, and b_{YW} and b_{YX} are set to 0. $C(\theta')$ has a new set of structural equations, a new entailed distribution over the substantive variables, a new entailed covariance matrix, and a new causal graph that represents the new structural equations. The causal graph for $C(\theta')$ is the same as the original causal graph for $C(\theta)$, except that all of the edges coming into the manipulated variable Y are removed because those variables have coefficients fixed at zero in the new set of structural equations. (In addition to the kinds of manipulations discussed here, there are more general kinds of manipulations. See e.g. [Pearl, 2000; Spirtes *et al.*, 2001].)

Since by design the outcome of the randomizing device is not caused by any of

the other variables in the system, and is a direct cause only of Y , the randomizing device can be treated as an exogenous error term, and Y is equal to the error term. In the degenerate case, a manipulation can set every member of a population to have the same level of Y , in which case the structural equation that describes Y in the experimental population sets the level of Y to a constant (i.e. it sets the variance of the error term to zero.) The result of modifying the set of structural equations in this way can lead to a density in the randomized population that is not necessarily the same as the density in any subpopulation of the population. (For more details see [Spirtes *et al.*, 2001; Pearl, 2000]).

The *direct effect* of Y on X relative to the set of variables $\mathbf{S}$ represents the following experimental quantity. Perform two separate manipulations of all of the variables in $\mathbf{S}$ except for X . In both manipulations all of the variables in $\mathbf{S}$ except for X and Y are set to the same constant c . In the first manipulation Y is set to the constant d , and in the second manipulation, to $d+1$. The average difference in X resulting from these two different manipulations is the direct effect of Y on X relative to the set of variables $\mathbf{S}$. The direct effect of Y on X relative to $\mathbf{S} = \{X, Y\}$ (that is only Y is manipulated) is the *total effect* of Y on X .

If the *direct effect* of Y on X relative to $\mathbf{S}$ is non-zero, then Y is a *direct cause* of X relative to $\mathbf{S}$. If SEM C is true, W is not a direct cause of Y relative to $\{X, Y, Z, W\}$, because two manipulations that disagree on the value assigned to W but agree on the values assigned to X and Z produce the same average value of Y . On the other hand, relative to $\{X, Y, W\}$, W is a direct cause of Y , because two manipulations that disagree on the values assigned to W but agree on the values assigned to X , produce different average values of Y . (For a discussion of total and direct effects see [Bollen, 1989]). If the total effect of Y on X is non-zero, then Y is a *cause* of X .¹⁰

A set $\mathbf{S}$ of variables is *causally sufficient* if every variable that is a direct cause (relative to $\mathbf{S}$) of any pair of variables in $\mathbf{S}$ is also in $\mathbf{S}$. Intuitively, a set of variables is causally sufficient if no common causes of pairs of variables in the set have been left out of the set. If SEM C is true, for $\mathbf{S} = \{X, Y, Z, W\}$, $\mathbf{S}$ and every subset of $\mathbf{S}$ is causally sufficient. In contrast, suppose SEM B (shown in Figure 2) with causal graph $X \rightarrow Y \leftarrow Z \leftarrow T \rightarrow W$ is true. In that case $\mathbf{S}$ is not causally sufficient because T is a direct cause of W and Z relative to $\mathbf{S}$, and both W and Z are in $\mathbf{S}$. As pointed out in section 2, given the non-causally sufficient set of variables $\mathbf{S}$, a SEM such as B' of Figure 2 that correctly represents both the population covariance matrix and the causal relations requires the introduction of a correlated error between ε_Z and ε_W .

¹⁰This terminology has the counter-intuitive consequence that Y can fail to be a *cause* of X , but can still be a *direct cause* of X relative to a particular set of variables $\mathbf{S}$. There is an edge $Y \rightarrow X$ in the causal graph of a SEM when Y is a direct cause of X relative to the variables in the SEM even if Y is not a cause of X .

3.2.1 Total Effects in SEM C

Adapting Lauritzen's notation [Lauritzen, 2001] the mean of Y when W is manipulated to the constant value m will be written as $E(Y||W = m)$, where the double bar distinguishes this operation from the single bar used to denote conditioning. $E_{C(\theta)}(Y||W = m)$ represents the mean of Y that is entailed by $C(\theta)$ when W is manipulated to have the value m , by replacing the structural equation for W in $S(\theta)$ with the new structural equation $W = m$.

In the case of linear models, there is an easy way to calculate the total effect of Y on W as long as the causal graph is acyclic. The total effect of Y on W *along a given directed path* from W to Y is simply the product of the edge coefficients along the directed path. For example, in SEM C , the effect of W on Y along the directed path $Y \leftarrow Z \leftarrow W$ is $b_{YZ} \cdot b_{ZW}$. The total effect of W on Y is the sum over all directed paths from W to Y of the effect of W on Y along each directed path (the *directed path sum*). Since in SEM C there is only one directed path from W to Z , the total effect of W on Y is also $b_{YZ} \cdot b_{ZW}$. (The manipulated mean of Y when W is manipulated to m and the pre-manipulation mean of Y is zero, is then equal to m times the total effect, or $m \cdot b_{YZ} \cdot b_{ZW}$.) One intuitive consequence of this method of calculating manipulated means is that if Y is not a cause of W (i.e. there is no directed path from Y to W in the causal graph), then the manipulated mean and the unmanipulated mean of W are the same.

In SEM C , $E_{C(\theta)}(Y||W = m) = m \cdot b_{YZ} \cdot b_{ZW} = E_{C(\theta)}(Y|W = m)$. This follows from the fact that $\text{cov}_{C(\theta)}(Y, W)$ (used to calculate the conditional mean) is equal to the trek sum, and the total effect (used to calculate a manipulated mean) is equal to a directed path sum. However, in SEM C , the treks and the directed paths are the same, i.e. $Y \leftarrow Z \leftarrow W$. Hence the covariance and the total effect are the same, and the conditional and manipulated means are the same.

In contrast, by assumption $E_{C(\theta)}(Y) = 0$; so in this case, both the manipulated mean ($E_{C(\theta)}(Y||W = m)$) and the conditional mean ($E_{C(\theta)}(Y|W = m)$) are different from $E_{C(\theta)}(Y)$. This is intuitively correct, since according to $C(\theta)$, manipulating the value of W (*doing*) changes the value of Y because it is a cause of Y , and conditioning on (*seeing*) W contains information about the value of Y .

3.2.2 Total Effects in SEM B

Suppose that SEM B is the alternative SEM of Figure 2. Let Σ be the covariance matrix over the observed variables X, Y, Z, W . It can be shown that for every θ of SEM C , there is a θ' of SEM B such that $\Sigma_{B(\theta')} = \Sigma_{C(\theta)}$, i.e. for every assignment θ of values to the free parameters of SEM C there is an assignment θ' of values to the free parameters of SEM B such that the covariance matrix entailed by $B(\theta')$ is the same as the covariance matrix entailed by $C(\theta)$ over the observed variables. Suppose then that θ' is chosen so that $\Sigma_{B(\theta')} = \Sigma_{C(\theta)}$.

Intuitively, the common cause T of Y and W contributes to the covariance between Y and W , but there is no causal influence of W on Y because there is no directed path from W to Y . The reason that in SEM B the conditional mean

is not equal to the manipulated mean is that in SEM B there is a trek between W and Y ($Y \leftarrow Z \leftarrow T \rightarrow W$) that is not a directed path. So the trek product over $Y \leftarrow Z \leftarrow T \rightarrow W$ contributes to the covariance, and hence the conditional mean, but it does not contribute to the total effect of W on Y , and hence does not contribute to the manipulated mean.

Slightly more formally, $E_{B(\theta')}(Y|W = m) = m \cdot \text{cov}_{B(\theta')}(Y, W)$, which is equal to m times the sum of trek products between W and Y . $Y \leftarrow Z \leftarrow T \rightarrow W$ is a trek in B , and so $E_{B(\theta')}(Y|W = m) = m \cdot b_{YZ} \cdot b_{ZT} \cdot b_{WT}$. In contrast, the total effect of W on Y according to B is the sum of directed path products from W to Y . However, there are no directed paths from W to Y in B , so the total effect is zero. In summary, $E_{B(\theta')}(Y||W = m) = E_{B(\theta')}(Y) = 0 \neq E_{B(\theta')}(Y|W = m)$.

One lesson to be drawn is that reliably estimating the quantitative effect of one variable on another (e.g. W on Y) requires first knowing the qualitative causal relationships of the variables represented by the graph (e.g. whether W causes Y as in SEM C , or W does not cause Y as in SEM B).

3.2.3 Regression and Structural Coefficients

One way to predict the value of Y from a set of variables $\mathbf{O}$ is to regress Y on $\mathbf{O}$. For example, suppose that given a covariance matrix Σ the goal is to predict the value of Y after having observed values of $\mathbf{O} = \{X, Z, W\}$. Regression is a procedure that takes Σ as input and outputs a linear equation “ $\hat{Y} = r_{YZ.\mathbf{O}} \cdot Z + r_{YX.\mathbf{O}} \cdot X + r_{YW.\mathbf{O}} \cdot W$ ”, where $\hat{Y}$ is the predicted (estimated) value of Y from the values of X, Z , and W , and $r_{YZ.\mathbf{O}}$ is a real number that is the partial regression coefficient of Z when Y is regressed on $\mathbf{O}$. It can be shown that if $X = x, Z = z$, and $W = w$ then $E(\hat{Y}) = r_{YZ.\mathbf{O}} \cdot z + r_{YX.\mathbf{O}} \cdot x + r_{YW.\mathbf{O}} \cdot w = E(Y|X = x, Z = z, W = w)$. So regression gives an unbiased estimate of the conditional mean.¹¹ The output of the regression depends only on Σ , and is unaffected by how the variables are causally related.

Under certain assumptions about the causal relations between X, Y, Z , and W , the partial regression coefficients are also equal to linear structural coefficients associated with edges in a causal graph of a SEM [Spirtes *et al.*, 1998b]. In those cases, the regression coefficients can be used not only to estimate the conditional mean, they can also be used to estimate manipulated means. In a directed graph of SEM C , Z is a *descendant* of Y (and Y is an *ancestor* of Z) if there is a directed path from Y to Z . $\mathbf{ND}(C, Y)$ is the set of all non-descendants of Y in the graph of SEM C , other than Y itself. For example, $\mathbf{ND}(C, Y) = \{X, W, Z\}$ and $\mathbf{ND}(C, X) = \{W, Z\}$. In a SEM C over a causally sufficient set of variables (without correlated errors and a graph that is acyclic) if X is a non-descendant of Y in the causal graph (i.e. a non-effect of Y) then the structural coefficient b_{YX} (i.e. the coefficient of X in the structural equation for Y) is equal to $r_{YX.\mathbf{ND}(C,Y)}$,

¹¹An unbiased estimator is defined in the Appendix. The regression equation gives the best unbiased linear predictor of Y from X, Z , and W .

that is the coefficient of X when Y is regressed on $\mathbf{ND}(C, Y)$.¹² For example, to find b_{YX} and b_{YZ} in $C(\theta)$, regress Y on X, Z , and W using as input $\Sigma_{C(\theta)}$. It follows that $r_{YZ.\mathbf{ND}(Y)} = b_{YZ}$, $r_{YX.\mathbf{ND}(Y)} = b_{YX}$, and $r_{YW.\mathbf{ND}(Y)} = b_{YW}$. (In SEM C , b_{YW} is zero because W is not a *parent* of Y , i.e. there is no edge $W \rightarrow Y$ in the graph of C .) The use of regression coefficients to predict Y from conditioning (seeing) values of other variables does not require knowledge of causal relations. In contrast, the use of regression coefficients to estimate structural coefficients (and hence to predict the effects of manipulations) does require knowledge of causal relations between the variables; i.e. it requires knowing which variables are not effects of Y , and that the set of variables is causally sufficient. Suppose first that J

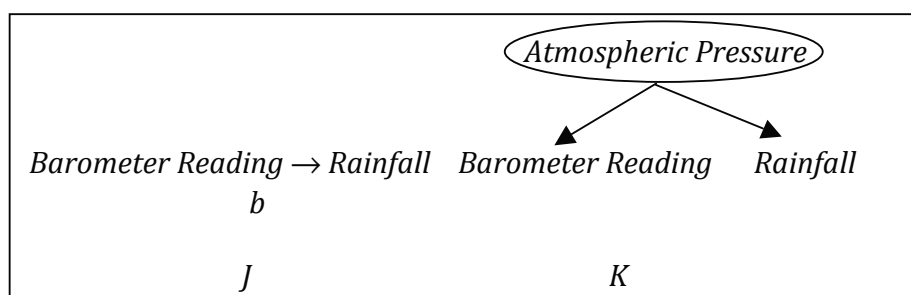


Figure 3. SEMs J and K

in Figure 3 is the true SEM. According to J , the total effect of *Barometer Reading* on *Rainfall* is the directed path sum and is equal to the structural coefficient b , which is the coefficient resulting from regressing *Rainfall* on *Barometer Reading*. If SEM K is true, then the total effect of *Barometer Reading* on *Rainfall* is equal to the coefficient of *Barometer Reading* in the structural equation for *Rainfall* (which is zero) and is equal to the coefficient resulting from regressing *Rainfall* on *Barometer Reading* and *Atmospheric Pressure*. Without knowing which of J or K is true, it is not possible to determine which regression gives the true direct (and in this case total) effect. Unfortunately, given just the covariance between *Barometer Reading* and *Rainfall*, there is no way to determine which of J or K is true. And if K is true, and *Atmospheric Pressure* is unmeasured, then there is no way to perform the regression of *Rainfall* on *Barometer Reading* and *Atmospheric Pressure*. This problem would not arise if it were known that there are no unobserved common causes of *Rainfall* and *Barometer Reading*; in that case

¹²It is only necessary to regress Y on its parents in SEM C for the regression coefficients to equal the structural coefficients. However, in many circumstances, it is not known which variables are parents of Y , but it is known which variables are non-descendants of Y (e.g. all variables that occur earlier than Y). When the regression coefficients are equal to the structural coefficients, the regression residuals (the difference between the predicted and actual values of Y) can be interpreted as due to unmeasured causes of Y , and hence are equal to error terms under a causal interpretation.

the only SEM compatible with background assumptions (about time order) is J , and the total effect of *Barometer Reading* on *Rainfall* can be reliably estimated.¹³

One practice that causal modelers in the social sciences use to try to partially solve this problem is to observe as many potential common causes of *Rainfall* and *Barometer Reading* as possible (e.g. variables that occur prior to *Rainfall* and *Barometer Reading*), and regress on all of them. It is well known that regressing on many variables can create various statistical problems at small sample sizes, but at large enough sample sizes, and assuming that there are no deterministic relationships among the measured variables, these statistical problems can be overcome. Unfortunately, it is generally not possible to reliably test when all of the common causes of *Rainfall* and *Barometer Reading* have been observed (although with some additional assumptions, in some cases the FCI algorithm described in section 4.2.3 can reliably detect test this). In addition, regressing on additional potential common causes can actually make the difference between the structural coefficient and the regression coefficient larger [Spirtes *et al.*, 1998b]. For example, in Figure 4, if Z is temporally prior to Y , simple calculations show $b_{YX} = r_{YX} \neq r_{YX.Z}$ even though Z is a potential common cause. This is because $\{X, Y\}$ is causally sufficient, but $\{X, Y, Z\}$ is not.

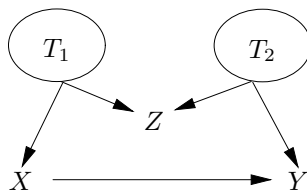


Figure 4. Adding more variables to regression can make estimates more biased

4 ESTIMATING MANIPULATED MEANS

One goal of causal inference is to obtain a qualitative understanding of the causal relationships between variables, and how a population distribution is generated. In that case, inference of the causal graph, or as much about the causal graph that can be reliably discovered is the goal. A second goal is to obtain quantitative information about causal relations between variables — e.g. if W is manipulated to m , what will the mean of Y become? Estimates of quantitative causal relations typically require first inferring qualitative causal relationships, so both issues will be discussed together.

¹³In general, if a set of variables in a SEM G is causally sufficient, the structural coefficient b_{YX} is equal to the partial regression coefficient $r_{YX.\mathbf{O}}$, for any set $\mathbf{O}$ that contains X , as long as $\mathbf{O}$ does not contain any descendant of Y , and X is d-separated from Y conditional on $\mathbf{O}$ in the graph of G with the $X \rightarrow Y$ edge removed [Spirtes *et al.*, 1998b]. D-separation is explained in section 4.1.1.

The following discussion concentrates on estimating¹⁴ manipulated means, but similar strategies can be used to estimate other causal quantities. As discussed in section 3.2.1., $E_{C(\theta)}(Y|W = m) = m \cdot b_{YZ} \cdot b_{ZW}$, while $E_{B(\theta)}(Y|W = m) = 0$. So the mean of Y when W is manipulated to m is a function of which SEM is true, and of the values of the parameters in the SEM. One way to estimate the manipulated mean, which will be discussed in detail in this section, is to (i) use the sample to find the correct SEM (e.g. SEM C), (ii) find a formula for $E_{C(\theta)}(Y|W = m)$ in terms of the free parameters of the correct SEM (e.g. $m \cdot b_{YZ} \cdot b_{ZW}$), (iii) estimate the values of the free parameters of the SEM from the data (e.g. $\hat{b}_{YZ}$ and $\hat{b}_{ZW}$) and (iv) substitute the estimated values into the formula for the manipulated mean (e.g. $m \cdot \hat{b}_{YZ} \cdot \hat{b}_{ZW}$). The actual process is a little more complicated because without a known time order, and with the possibility of unmeasured common causes, typically it is not possible to reliably find the true SEM; rather is only possible to reliably find a set of SEMs that contains the true SEM.

The principles of estimating the effects of manipulations can be illustrated with the 5 causal graphs of Figure 2. The first step is to search among the SEMs and select the best SEMs or set of SEMs. In order to use sample data to select SEMs, it is necessary to make some assumptions that link SEMs (and in particular the graphs of SEMs) to probability distributions. The Causal Markov Assumption is implicit in much of the practice of structural equation modeling. (For extended discussions of the Causal Markov Assumption, see [Spirtes *et al.*, 2001; Cartwright, 1994; Glymour, 1999]).

4.1.1 Causal Markov Assumption

$\mathbf{X}$ and $\mathbf{Y}$ are *independent* in a probability density function f , denoted $I_f(\mathbf{X}, \mathbf{Y})$ when $f(\mathbf{X}|\mathbf{Y}) = f(\mathbf{X})$. (The f subscript will be left out of I_f when it does not introduce any ambiguity.) Intuitively, this means that the values of the variables in $\mathbf{Y}$ contain no information about the values of the variables in $\mathbf{X}$. In the case of Gaussian variables, $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if each variable $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, $\text{cov}(X, Y) = 0$. Similarly, $\mathbf{X}$ is *independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$* (denoted $I_f(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$) when $f(\mathbf{X}|\mathbf{Y}, \mathbf{Z}) = f(\mathbf{X}|\mathbf{Z})$. Intuitively, this means that the values of the variables in $\mathbf{Y}$ contain no information about the values of the variables in $\mathbf{X}$, once the values of the variables in $\mathbf{Z}$ are already known. In the case of Gaussian distribution f , $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if each variable $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, $\text{cov}_f(X, Y|\mathbf{Z}) = 0$ (where $\text{cov}_f(X, Y|\mathbf{Z})$ is the covariance between X and Y conditional on $\mathbf{Z}$ in distribution f). For Gaussian distributions, the conditional covariance does not depend upon the particular value of $\mathbf{Z}$ conditioned on (although the conditional mean does).

$I_{S(\theta)}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ denotes that $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in the density over the substantive variables entailed by $S(\theta)$. $I_S(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ denotes that a SEM S (with free parameters) entails $I_{S(\theta)}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ for all assignment of values θ to its free parameters; in other words $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in

¹⁴See section 5 for a technical appendix defining estimators and their properties.

every distribution represented by S . If $I_S(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ then S is said to *entail* that $\mathbf{X}$ and $\mathbf{Y}$ are independent conditional on $\mathbf{Z}$. Since $I_S(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ does not depend upon the parameter values of S , it is possible to determine whether $I_S(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ from the graph of S ; in this case the graph of S is also said to *entail* that $\mathbf{X}$ and $\mathbf{Y}$ are independent conditional on $\mathbf{Z}$.¹⁵ (In cases where it does not create confusion, the graph of S will simply be referred to as S). It is possible that $I_{(\theta)}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ for some, but not all θ (see an example in section 4.1.3). In general the fact that it is not the case that $I_S(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ does not mean that there is *no* assignment of values θ to the free parameters for which $I_{S(\theta)}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ — it entails only that it is not the case that $I_{S(\theta)}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ for *all* θ .

The following assumption is used to relate causal relations to probability distributions.

Weak Causal Markov Assumption: For a causally sufficient set of variables $\mathbf{V}$ in a population N , if no variable in $\mathbf{X}$ causes any variable in $\mathbf{Y}$, and no variable in $\mathbf{Y}$ causes any variable in $\mathbf{X}$, then $\mathbf{X}$ and $\mathbf{Y}$ are independent (i.e. in the Gaussian case, members of $\mathbf{X}$ and members of $\mathbf{Y}$ are pairwise uncorrelated.)

The Weak Causal Markov Assumption has the consequence that the error terms for causally sufficient sets of variables are independent. Simon’s famous analysis of “spurious correlation” [Simon, 1985] is precisely an application of the Weak Causal Markov Assumption to explain correlated errors as the result of unobserved common causes. The examples that Bollen gives of why an error term for a variable X might be correlated with one of the causes of X other than sampling problems are all due to causal relations between the error term and other causes of X , and hence an application of the Weak Causal Markov Assumption [Bollen, 1989]. (For a discussion of the Causal Markov Assumption, and conditions under which it should not be assumed, see [Spirtes *et al.*, 2001]).

For deterministic causal models such as SEMs, the Weak Causal Markov Assumption also entails another version of the Causal Markov Assumption, i.e. that for causally sufficient sets of variables, all variables are independent of the their non-effects (non-descendants in the causal graph) conditional on their direct causes (parents in the causal graph) [Spirtes *et al.*, 2001].¹⁶

Some of the causal discovery algorithms described later that use the Causal Markov Assumption merely assume that causally sufficient sets of variables *exist*, not that the causally sufficient sets of variables are all *observed*, i.e. there can be unobserved common causes. The SEMs in Figure 2 illustrate the use of this

¹⁵In acyclic graphs with no double-headed arrows, $\mathbf{X}$ is *d-separated* from $\mathbf{Y}$ conditional on $\mathbf{Z}$ is a purely graphical relationship among three disjoint sets of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ such that $\mathbf{X}$ is d-separated from $\mathbf{Y}$ conditional on $\mathbf{Z}$ in DAG G if and only if G entails that $\mathbf{X}$ and $\mathbf{Y}$ are independent conditional on $\mathbf{Z}$. A variable A is a collider on a path if the path contains edges $B \rightarrow A \leftarrow C$. For disjoint sets $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, $\mathbf{X}$ is *d-separated* from $\mathbf{Y}$ conditional on $\mathbf{Z}$ in G if and only if there is no path between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ such that every collider on the path is in $\mathbf{Z}$ or has a descendant in $\mathbf{Z}$, and no non-collider on the path is in $\mathbf{Z}$. (See [Pearl, 1988] for details). *d*-separation can also be extended to graphs with cycles and double-headed arrows.

¹⁶For non-deterministic causal models, the alternative Causal Markov Assumption is usually made directly. (For different alternative versions of Markov relations, see [Lauritzen *et al.*, 1990]).

consequence of the Weak Causal Markov Assumption in causal inference. Suppose that SEM C of Figure 2 has the true causal graph. In SEM C , $\text{Cov}_C(W, Z) = b_{ZW}$, so $\sim I(W, Z)$ as long as $b_{ZW} \neq 0$. In contrast, $I_A(W, Z)$, so A cannot represent the population distribution at all, regardless of what parameter values it is assigned. So, even for the assignment θ of parameters to A that most closely matches the population distribution, a statistical test will reject $A(\theta)$ at a large enough sample size.

In contrast, it is not possible to eliminate SEM D from consideration using just the Causal Markov Assumption as long as T is not observed. If SEM D is true, the Causal Markov Assumption does not *entail* any zero covariance or conditional covariances among X , Y , W , and Z . However, just because there are no zero conditional covariance among X , Y , W , and Z that hold for all values of the free parameters of D does not imply that there are no zero conditional covariance relations among X , Y , W , and Z that hold for some values of the free parameters of D . Indeed, for any assignment of parameter values θ to C there exists an assignment of parameter values θ' to D that represents the same distribution over the observed variables X , Y , Z , and W .¹⁷

Hence, at least for some parametric families (e.g. Gaussian, multinomial) in order to reliably draw conclusions about effects of manipulations from observational data, some additional assumptions must be made. (This is shown more formally in [Spirtes *et al.*, 2001; Robins *et al.*, 2003]). The next section will describe one such assumption.

4.1.2 Causal Faithfulness Assumption

In any SEM over a causally sufficient set of variables, there is an adjacency between two variables X and Y if and only if the SEM does not entail that X and Y are independent conditional on any subset $\mathbf{Z}$ of the other variables [Spirtes *et al.*, 2001]. Even though SEM C entails that X and Z are independent, and SEM D does not (because SEM D contains an edge from Z to X), nevertheless there could be an assignment of parameter values to SEM D that has the same distribution as $C(\theta)$. That is, it could be that $I_{D(\theta')}(X, Z)$ for some θ' (i.e. $\text{cov}_{D(\theta')}(X, Z) = 0$) even though it is not the case that $I_D(X, Z)$ (i.e. it is not the case that D entails $\text{cov}(X, Z) = 0$ for all values of the free parameters). $\text{Cov}_{D(\theta')}(X, Z) = \beta_{XZ} + (\beta_{YZ} \cdot \beta_{XY})$, so if $\text{cov}_{D(\theta')}(X, Z) = 0$, then $\beta_{XZ} = -(\beta_{YZ} \cdot \beta_{XY})$. Although C and D agree on the total effect of Z on X (zero according to both), they disagree on the total effect of Y on X (zero according to SEM C , β_{XY} according to SEM D).

So if it is observed that $I(X, Z)$ there are at least two possible explanations, SEMs C and D . There are several arguments why, in the absence of evidence to the contrary, C should be the preferred explanation.

¹⁷However, if at most one error term is non-Gaussian, but the structural equations are linear, then SEM D cannot represent any distribution represented by SEM C , and hence can be eliminated from consideration without any further assumptions [Shimizu *et al.*, 2006].

SEM C explains the independence of X and Z as a consequence of no causal connection between the variables. In contrast SEM D explains the independence as a consequence of a large direct effect of Z on X cancelled *exactly* by a large indirect effect of Z on X (via the effect on Y). But sciences typically assumes that, unless there is evidence to the contrary, an improbable and unstable cancellation of parameters (as in SEM D) does not hide real causal influences. When a theory cannot explain an empirical regularity save by invoking a special parameterization, most scientists are uneasy with the theory and look for an alternative [Glymour, 1980].

Second, this cancellation is improbable (in the sense that if a zero conditional correlation is not entailed, the measure of the set of free parameter values for any DAG that lead to such cancellations is zero for any “smooth” prior probability distribution¹⁸ e.g. Normal, exponential, etc., over the free parameters).

Finally, SEM C is simpler than SEM D . SEM C has fewer free parameters than SEM D . Because SEM D imposes a proper subset of the conditional independence constraints imposed by SEM C , SEM D represents a proper superset of the distributions that can be represented by SEM C . There is even a well defined sense in which the set of distributions represented by SEM D is of higher dimension than the set of distributions represented by SEM C [Geiger *et al.*, 2002]. So SEM D is more complex than SEM C in a precisely defined way.

The assumption that a causal influence is not hidden by coincidental cancellations can be expressed as following for SEMs. A probability density function f is *faithful* to the graph G of a SEM if and only if every conditional independence relation true in f is entailed by G .

Causal Faithfulness Assumption: For a causally sufficient set of variables $\mathbf{V}$ in a population N , the population distribution is faithful to the causal graph for N .¹⁹

The Causal Faithfulness Assumption requires preferring SEM C to SEM D , because parameter values θ' for which $I_{D(\theta')}(X, Z)$ would violate the Causal Faithfulness Assumption. The Causal Faithfulness Assumption limits the SEMs considered to those SEMs in which population conditional independence constraints are entailed by causal structure, rather than by particular values of the parameters. There can be sometimes be good reasons to believe that the more complicated model is true, and the reason that a conditional independence relation holds is not due to the structure of the graph, e.g. when there are deterministic relationships among the substantive variables, equality constraints upon free parameters [Spirtes *et al.*, 2001], or cases where policy makers are intentionally optimizing

¹⁸A smooth measure is absolutely continuous with Lebesgue measure.

¹⁹The assumption can be considerably weakened and still permit reliable causal inference, but the weaker assumption requires more complicated algorithms with more complex and somewhat less informative output. The weakened assumption is closer to the informal intuition. The Adjacency Causal Faithfulness Assumption states: For a causally sufficient set of variables $\mathbf{S}$, if X is a direct cause of Y (so X and Y are adjacent in the causal graph for population N), then for any $\mathbf{Z} \subseteq \mathbf{S}$, $\sim I(X, Y | \mathbf{Z})$ in population N [Ramsey *et al.*, 2006].

the values of variables [Hoover, 2001]. In those cases, the Causal Faithfulness Assumption should not be made.

4.1.3 Markov Equivalence-Over- $\mathbf{O}$ Classes

According to SEM B , $I_B(Z, W|T)$. However it is not the case that $I_C(Z, W|T)$ because SEM C does not contain the unobserved variable T . This conditional independence is not directly testable, because T is not observed. However, SEMs B and C entail the same set of conditional independence relations among just the observed variables $\mathbf{O} = \{X, Y, Z, W\}$. SEMs B and C are *Markov equivalent-over- $\mathbf{O}$* if they entail the same set of conditional independence relations among the variables in $\mathbf{O}$. Hence when $\mathbf{O} = \{X, Y, Z, W\}$ the Markov equivalence-over- $\mathbf{O}$ class contains SEM B as well as SEM C . It is easy to see that the Markov equivalence-over- $\mathbf{O}$ class contains an infinite number of SEMs, because adding more and more unobserved common causes of X and Y does not change the conditional independence relations entailed over $\mathbf{O}$. (If two SEMs both have the same set of substantive variables $\mathbf{V}$ (observed or unobserved), and they are Markov equivalent-over- $\mathbf{V}$, they are said to be Markov-equivalent.)

If SEM C is true, SEM A can be eliminated from consideration by the Causal Markov Assumption, and SEM D can be eliminated from consideration by the Causal Faithfulness Assumption. However, neither of these assumptions eliminates SEM B from consideration.

Since SEM B and SEM C entail the same set of conditional independence relations over $\mathbf{O}$, it is not possible to eliminate SEM B from consideration without either adding more assumptions or background knowledge, or using features of the probability distribution that are not conditional independence relations. Without a principled way to choose between SEMs B and C , the best a search algorithm can reliably do is to return both SEM B and SEM C , as well as the other graphs Markov equivalent to SEMs B and C , as possibilities.

SEMs B and C are *distribution equivalent-over- $\mathbf{O}$* if and only if for any assignment of parameter values θ to C there exists an assignment of parameter values θ' to B that entails the same marginal distribution over $\mathbf{O}$, and vice versa. (If two SEMs have the same set of substantive variables $\mathbf{V}$ (observed or unobserved) and they are distribution equivalent-over- $\mathbf{V}$, they are said to be *distribution-equivalent*.)²⁰ If all of the error terms are Gaussian, then SEMs B and C are distribution equivalent-over- $\mathbf{O}$ as well as Markov equivalent-over- $\mathbf{O}$. In such cases, the best that a reliable search algorithm can do is to return the entire Markov equivalence-over- $\mathbf{O}$ class, regardless of what features of the marginal distribution that it uses.

²⁰Strictly speaking, distribution equivalence over- $\mathbf{O}$ is relative to a particular parametric family. So two SEMs that are distribution-equivalent-over- $\mathbf{O}$ for a parametric family that allows only Gaussian error terms are not distribution-equivalent-over- $\mathbf{O}$ for a parametric family that allows non-Gaussian error terms.

In contrast, SEMs D and E are Markov equivalent-over- $\mathbf{O}$, but they are not distribution equivalent-over- $\mathbf{O}$. SEM E , unlike SEM D , entails that $\text{Cov}_E(X, Z) \cdot \text{Cov}_E(Y, W) = \text{Cov}_E(X, W) \cdot \text{Cov}_E(Y, Z)$ and $\text{Cov}_E(Y, Z)$ cannot all be negative (this follows easily from the trek sum rule for calculating covariances.) This is a constraint on the marginal distribution over $\mathbf{O}$ that is not a conditional independence constraint. So when unobserved common causes are allowed, SEMs that are Markov equivalent-over- $\mathbf{O}$ are not necessarily distribution equivalent-over- $\mathbf{O}$ even when all of the error terms are Gaussian and the structural equations are linear.

When Markov equivalence-over- $\mathbf{O}$ fails to entail distribution equivalence-over- $\mathbf{O}$, then using conditional independence relations alone for causal inference is still correct, but it is not as informative as theoretically possible. For example, assuming causal sufficiency and non-Gaussian errors [Shimizu *et al.*, 2006], conditional independence tests can at best reliably determine the correct Markov equivalence class, while using other features of the sample distribution (higher order moments of the distribution) can be used to reliably determine a unique graph. The situation is analogous to linear SEMs and Gaussian errors, but without causal sufficiency. Using conditional independence tests it is possible to reliably infer a Markov equivalence-over- $\mathbf{O}$ class of DAGs; however, because there are other features of the distribution over $\mathbf{O}$ that are not being used, this set is theoretically not as small as possible. For example, if $\text{Cov}_E(X, Z) \cdot \text{Cov}_E(Y, W) \neq \text{Cov}_E(X, W) \cdot \text{Cov}_E(Y, Z)$ are all negative then SEM E can be eliminated from consideration, even if all of the conditional independence relations entailed by E (none in this case) are true in the population. The difficulty is that in many cases it is not known how to use the extra information contained in the density to reliably narrow down the set of DAGs output. However, in some cases it is known how to use the extra information (see section 4.2.4).

There are algorithms for testing when two SEMs are Markov equivalent-over- $\mathbf{O}$ [Spirtes and Richardson, 1996], although they are computationally much more intensive than comparable algorithms for testing when two SEMs without unobserved common causes are Markov equivalent. While there are recently developed algorithms for determining when two SEMs are distribution equivalent-over- $\mathbf{O}$, they are so computationally intensive they are only practical for SEMs with a few variables [Geiger and Meek, 1999].

4.1.4 Partial Ancestral Graphs — Representing Markov Equivalence-over- $\mathbf{O}$ Classes

A partial ancestral graph (PAG) [Spirtes *et al.*, 1995; Ali *et al.*, 2005; Zhang, 2007] represents a Markov equivalence-over- $\mathbf{O}$ class of DAGs. X and Y are adjacent in a PAG that represents a Markov equivalence-over- $\mathbf{O}$ class if X and Y are not entailed to be independent conditional on any subset of $\mathbf{O}$. The edge contains a “o” endpoint on the X end if for some DAGs in the Markov equivalence-over- $\mathbf{O}$ class, X is an ancestor of Y and in other DAGs it is not an ancestor of Y ; the edge contains a tail at X if and only if X is an ancestor of Y in every DAG in the Markov equivalence-over- $\mathbf{O}$ class; and the edge contains an arrowhead at X if and

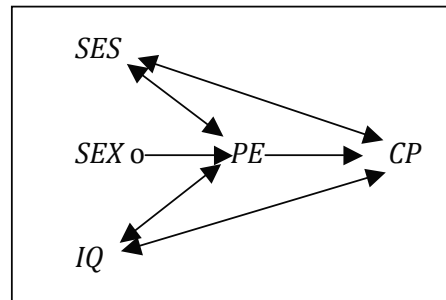


Figure 5. College Plans PAG

only if X is not an ancestor of Y in every DAG in the Markov equivalence-over- $\mathbf{O}$ class.

An example of a PAG that was constructed from data measured on the college plans of high school students is shown in Figure 5.²¹

SES is socio-economic status, PE is parental encouragement, CP is college plans, and IQ is intelligence quotient [Spirtes *et al.*, 2001; Sewell and Shah, 1968]. The double-headed arrow between CP and SES indicates that there is an unobserved common cause of SES and CP . The $PE \rightarrow CP$ arrow indicates that PE is a direct cause (relative to the observed variables) of CP , and that there are no unobserved common causes of PE and CP . The edge $SEX \circ \rightarrow PE$ indicates that PE is not a cause of SEX , but the statistical information is compatible with SEX being a cause of PE , or there being an unobserved common cause of SEX and PE .

4.1.5 Scoring SEMs

Together, the Causal Markov and Causal Faithfulness Assumptions entail a number of conditional independence facts about the population distribution. However, even if the conditional independence relations hold in the population, they typically will not hold in the sample. For a sample, what is a good measure of how compatible a set of conditional independence relations that are entailed by a SEM are with the sample data? A similar question can be asked of those parametric families that entail constraints that are not conditional independence constraints.

Under the assumption of no unmeasured common causes, there are a number of different scores of fit that are consistent in the following sense: under the Causal Markov and Causal Faithfulness Assumptions, in the large sample limit, the set of SEMs with the highest score is guaranteed to contain the true SEM with probability 1. For example, the Bayes Information Criterion is a penalized likelihood score that rewards a SEM for assigning a high likelihood to the sample

²¹This model created is not a SEM because the variables were discretized, but the causal relations can be represented by a PAG and the PAG can be created with the same FCI algorithm (section 4.2.3) using the appropriate tests for conditional independence for discrete variables.

for the maximum likelihood estimate of the parameters, and penalizes a SEM that represents a set of probability distributions that has high dimension [Haughton, 1988; Chickering, 2003]. In the case of a multi-variate Gaussian SEM M , for a given sample of size n

$$\text{BIC}(M, \text{sample}) = -2L(\Sigma_{M(\hat{\theta})}, \text{sample}) + \ln(n) \cdot df_M,$$

where

- $\hat{\theta}$ is the maximum likelihood estimate of the parameters for model M from sample;
- $L(\Sigma_{M(\hat{\theta})}, \text{sample})$ is the likelihood of $\Sigma_{M(\hat{\theta})}$; and
- df_M is the degrees of freedom (dimensionality) of the SEM M .

While the use of penalized likelihood scores, such as BIC, for SEMs with no unobserved common causes is not problematic, there are major statistical problems in scoring models using penalized likelihood scores for SEMs with unobserved common causes. In order to calculate a BIC score it is necessary to calculate a maximum likelihood estimate of the SEM parameters, and the dimensionality of the set of marginal distributions over the observed variables represented by the SEM.

However, with the exception of a few distributions such as Gaussian, or multinomial, even when the joint distribution falls into a family of distributions that is well understood, the marginal distributions will not. In cases where the marginal distribution is a member of a well understood family of distributions, the parameters of the SEM may not be identifiable at all, and hence it is not possible to find a maximum likelihood estimate of the parameters from the data over the observed marginal. Furthermore, even in those cases where the maximum likelihood estimates can be calculated (such as SEM E in Figure 2) the actual calculations typically involve an iterative hill-climbing algorithm that is much more computationally expensive than regression and can get stuck in local maxima.

In addition, there are both theoretical and practical difficulties in calculating the dimensionality of the set of marginal probability distributions that are represented by a SEM with unobserved common causes. The dimensionality is not well defined for some values of the parameters, and is difficult to calculate even when it is well defined [Geiger *et al.*, 2002].

This problem is caused by the fact that while the unobserved conditional independence relations (those that involve the unobserved common cause) cannot be tested directly, they can nevertheless entail constraints on the marginal distribution that are not conditional independence relations. These non-conditional independence constraints present both a problem and an opportunity. On the one hand, if it is known how to use them to eliminate some graphs from consideration (as in certain special cases) then they strengthen the causal inferences that can be made. For example, if $\text{Cov}(X, Z) \cdot \text{Cov}(Y, W) \neq \text{Cov}(X, W) \cdot \text{Cov}(Y, Z)$ are

all negative, then SEM E cannot be the correct SEM, even if it is compatible with all of the conditional independence relations over $\mathbf{O}$. On the other hand, the non-conditional independence constraints are the reason that the maximum likelihood estimate can get stuck in local maxima and that the dimensionality of the marginal distributions represented by a SEM with unobserved common causes is sometimes undefined or difficult to calculate.

4.2 Search

One of the major reasons that causal inference is harder than purely statistical inferences (estimating a probability distribution or features of a distribution from a sample) is that reliable statistical inference requires finding only one simple model that fits the data well, whereas reliable causal inference requires finding all simple models that fit the data well. Suppose, for example, that SEM C is the true causal SEM. Let G be the a SEM that has a DAG that is the same as the DAG of SEM C , but reverses the direction of the $Z \leftarrow W$ edge to $Z \rightarrow W$, i.e. the DAG for SEM G is $X \leftarrow Y \leftarrow Z \rightarrow W$. SEMs G and C are distribution equivalent (relative to a linear Gaussian parameterization). Because C and G represent the same set of distributions, maximum likelihood estimates of the parameters that characterize the joint distribution are the same. If a search over SEMs found that G was a simple model that fit the data well, and the goal is to use the SEM only for statistical inference, the search can reasonably stop. The fact that some other model such as C is another simple model that fits the data well does not prevent G from providing reliable inference about the joint distribution.

In contrast, if a SEM is to be used for causal inference, the fact that there is another SEM C that is as simple as SEM G and fits the data equally well does affect whether G can be reliably used for causal inference. The difference is that SEMs G and C make different predictions about the effects of some (but not all) manipulations. Hence, if a search of SEMs yielded that SEM G was simple and fit the data well, it is not possible to draw that conclusion that SEM G could be used for reliable causal inference. Finding that another SEM C is equally simple and fits the data equally well, but makes different predictions about the effects of some manipulations indicates that it is not possible to draw reliable conclusions about whether SEMs G or C is correct about the effects of the manipulations. On the other hand, if all of the SEMs that are equally simple and fit the data equally well agree about the effect of a manipulation, then the effect of the manipulation can be reliably inferred. This implies that searching for SEMs that are to be used for causal inference requires not just locating one good SEM, but all good SEMs.

4.2.1 Informal Search

In practice, search for causal models with unobserved common causes is often informal, and based on a combination of background assumptions together with statistical tests of the causal models. If a model is rejected by a statistical test, the researcher looks for a modification of the original model that will pass a statistical

test. The search typically halts when a model that is compatible with background assumptions does not fail a statistical test. Often, only the final model is presented, and the search itself is not described. Searches of this kind are only as reliable as the reliability of the background assumptions, and the extent to which the space of alternatives compatible with the background assumptions was searched. (For an example of a case where a search is described, see [Rodgers and Maranto, 1989]). Rodgers and Maranto show that different disciplines often start from very different causal models, and have different background assumptions, even when investigating the same question. Furthermore, unless the background assumptions are very extensive, or the number of variables is tiny, it is not feasible to estimate and test all of the models compatible with background assumptions. This is further complicated by the fact that for causal inference it is not sufficient to find one model that passes a statistical test; instead it is necessary to find all such models.

4.2.2 *Score Based Search*

For SEMs that have no unobserved common causes, the Greedy Equivalence Search [Chickering, 2003] returns a Markov equivalence class of SEMs that has the highest score (using one of several possible scores, including BIC). The output Markov equivalence class contains the true SEM with probability 1 in the large sample limit (under the Causal Markov and Faithfulness Assumptions, and assuming either a linear Gaussian model or a multinomial model with an acyclic causal graph). Although in the worst case the running time of the algorithm is exponential in the number of variables, in practice it can be used on graphs with relatively few edges but hundreds of variables.

However, a score based search over SEMs with unobserved common causes is much more difficult for two reasons. There are an infinite number of SEMs to search, and it is not clear how to structure the search efficiently. Scoring a SEM with an unobserved common cause faces both theoretical and practical difficulties. No score-based search that is both feasible and reliable in the large sample limit is known at this time for SEMs with unobserved common causes.

The LISREL and EQS programs have automated score-based searches that are versions of the informal search method described in section 4.2.1. However, these searches suffer from the difficulties one would expect from score-based searches over SEMs with unmeasured common causes. The scores depend upon being able to perform maximum likelihood estimates of the parameters, and the iterative algorithms necessary to do this often fail to converge or converge to the wrong values when the initial model is far from the truth. Even in cases in which the parameter estimates do not go awry, the searches have low reliability [Spirtes *et al.*, 1990].

4.2.3 *Constraint Based Search*

In contrast to a score based search, a constraint based search does not require the estimation of parameters, or the calculation of the dimensionality of the marginal

distribution — it requires only being able to perform the appropriate tests of the constraints (e.g. conditional independence constraints) used in the search. Furthermore, although the number of DAGs with unobserved common causes is infinite, the number of Markov equivalence-over- $\mathbf{O}$ classes is finite. The Fast Causal Inference (FCI) algorithm [Spirtes *et al.*, 2001; Spirtes *et al.*, 1995; Zhang, 2007] performs a series of conditional independence tests and constructs a PAG on the basis of those tests. In the large sample limit, it returns a PAG that contains the true SEM with probability 1 under the Causal Markov and Causal Faithfulness Assumptions (for linear Gaussian densities or multinomial distributions). The major differences between using constraint based searches under the assumption of no unobserved common causes, and constraint based searches without that assumption are that in the latter case the algorithms are computationally more complex, and the output is less informative. The PAG in Figure 5 was constructed by the FCI algorithm [Spirtes *et al.*, 2001] from data on 10,318 high school students [Sewell and Shah, 1968].

4.2.4 Search Using Vanishing Tetrad Constraints

Searches using conditional independence constraints are correct, but completely uninformative for some common kinds of data sets. Consider the SEM in Figure 6.

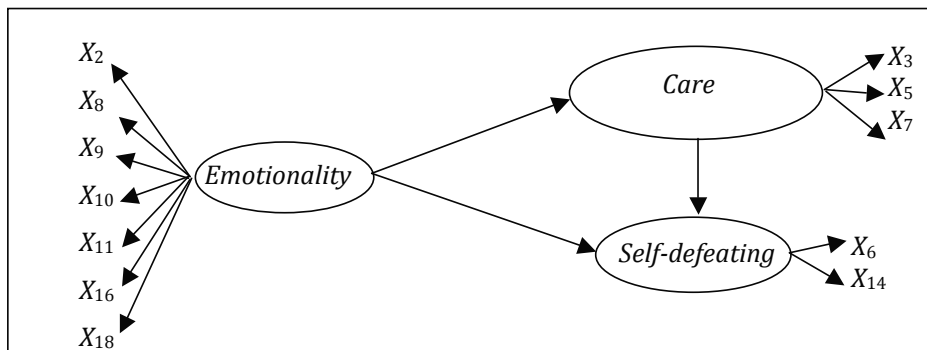


Figure 6. SEM S

The data comes from a survey of test anxiety indicators administered to 335 grade 12 male students in British Columbia [Gierl and Todd, 1996]. The survey contains 20 measures of symptoms of anxiety under test conditions. Each question is about a symptom of anxiety. For example, question 8 is about how often one feels “jittery when taking tests”. The answer is observed on a four-point scale (almost never, sometimes, often, or almost always) ; it will be assumed that the variables are approximately Gaussian. Each X variable represents an answer to a question on the survey. For reasons to be explained later, not all of the questions

on the test have been included in the model. There are three unobserved common causes in the model: *Emotionality*, *Care about achieving* (which will henceforth be referred to as *Care*) and *Self-defeating*. The test questions are of little interest in themselves; of more interest is what information they reveal about some unobserved psychological traits. If S is correct, there are no conditional independence relations among the $\mathbf{X}$ variables alone — the only entailed conditional independencies require conditioning on an unobserved common cause. Hence the FCI algorithm would return a completely unoriented PAG in which every pair of variables in $\mathbf{X}$ is adjacent. Such a PAG makes no predictions at all about the effects of manipulations of the observed variables. Furthermore, in this case, the effects of manipulating the observed variables (answers to test questions) is of no interest — the interesting questions are about the effects of manipulating the unobserved variables and the qualitative causal relationships between them.

Models such as S are *multiple indicator models*, and can be divided into two parts: the measurement model, which contains the edges between the unobserved variables and the observed variables (e.g. $\text{Emotionality} \rightarrow X_2$), and the structural model, which contains the edges between the unobserved variables (e.g. $\text{Emotionality} \rightarrow \text{Care}$).

The various $\mathbf{X}$ variables in $S(\{X_2, X_3, X_5, X_7, X_8, X_9, X_{10}, X_{11}, X_{14}, X_{16}, X_{18}\})$ were chosen with the idea that they indirectly measure some psychological trait that cannot be directly observed. Ideally, the $\mathbf{X}$ variables can be broken into clusters, where each variable in the cluster is caused by exactly one unobserved cause common to the members of the cluster, and a unique error term uncorrelated with every other error term, and nothing else. From the values of the variables in the cluster, it is then easy to make inferences about the value of the unobserved common cause. Such a measurement model is called *pure*. The measurement model of S is pure.

If the measurement model is impure (i.e. there are multiple common causes of a pair of variables in $\mathbf{X}$, or some of the $\mathbf{X}$ variables cause each other) then drawing inferences about the values of the common causes is much more difficult. Consider the set $\mathbf{X}' = \mathbf{X} \cup \{X_{15}\}$. If X_{10} caused X_{15} , the measurement model over the expanded set of variables would not be pure. If a measurement model for a set $\mathbf{X}'$ of variables is not pure, it is nevertheless possible that some subset of $\mathbf{X}'$, such as $\mathbf{X}$, has a pure measurement model. If the only reason that the measurement model is impure is that X_{10} causes X_{15} then $\mathbf{X} = \mathbf{X}' \setminus \{X_{15}\}$ does have a pure measurement model, because all the “impurities” have been removed. S does not contain all of the questions on the survey precisely because various tests described below indicated that they some of them needed to be excluded in order to have a pure measurement model.

The task of searching for a multiple indicator SEM model can then be broken into two parts: first finding clusters of variables so that the measurement model is pure; second, use the pure measurement model to make inferences about the structural model.

Factor analysis is often used to determine the number of unmeasured common

causes in a multiple indicator model, but there are important theoretical and practical problems in using factor analysis in this way. Factor analysis constructs models with unobserved common causes (factors) of the observed $\mathbf{X}$ variables. However, factor analysis models typically connect each unobserved common cause (factor) to each $\mathbf{X}$ variable, so the measurement model is not pure. A major difficulty with giving a causal interpretation to factor analytic models is that the observed distribution does not determine the covariance matrix among the unobserved factors. Hence, a number of different factor analytic models are compatible with the same observed data [Harman, 1976]. In order to reduce the underdetermination of the factor analysis model by the data, it is often assumed that the unobserved factors are independent of each other; however, this is clearly not an appropriate assumption for unobserved factors that are supposed to represent actual causes that may causally interact with each other. In addition, simulation studies indicate that factor analysis is not a reliable tool for estimating the correct number of unobserved common causes [Glymour, 1998].

On this data set, factor analysis indicates that there are 2 unobserved direct common causes, rather than 3 unobserved direct common causes [Bartholomew *et al.*, 2002]. If a pure measurement model is constructed from the factor analytic model by associating each observed in $\mathbf{X}$ only with the factor that it is most strongly associated with, the resulting SEM fails a statistical test (has a p-value of zero) [Silva *et al.*, 2006].

A search for pure measurement models that depends upon testing vanishing tetrad constraints (described below) is an alternative to factor analysis. Conceptually, the task of building a pure measurement model from the observed variables can be broken into 3 separate tasks:

1. Select a subset of the observed variables that form a pure measurement model.
2. Determine the number of clusters (i.e. the number of unobserved common causes) that the observed variables measure.
3. Cluster the observed variables into the proper groups (so each group has exactly one unobserved direct common cause.)

It is possible to construct pure measurement models using vanishing tetrad constraints as a guide [Silva *et al.*, 2006]. A *vanishing tetrad constraint* holds when $\text{Cov}(X, Y) \cdot \text{Cov}(Z, W) - \text{Cov}(X, Z) \cdot \text{Cov}(Y, W) = 0$.

A pure measurement model entails that each X_i variable is independent of every other X_j variable conditional on its unobserved parent, e.g. S entails X_2 is independent of X_j conditional on *Emotionality*. These conditional independence relations cannot be directly tested, because *Emotionality* is not observed. However, together with the other conditional independence relations involving unobserved variables entailed by S , they imply vanishing tetrad constraints on the observed variables that reveal information about the measurement model that does not

depend upon the structural model among the unobserved common causes. The basic idea extends back to Spearman's attempts to use vanishing tetrad constraints to show that there was a single unobserved factor of intelligence that explained a variety of observed competencies [Spearman, 1904].

Because X_2 and X_8 have one unobserved direct common cause (*Emotionality*), and X_3 and X_5 have a different unobserved direct common cause (*Care*), SEM S entails $\text{Cov}_S(X_2, X_3) \cdot \text{Cov}_S(X_5, X_8) = \text{Cov}_S(X_2, X_5) \cdot \text{Cov}_S(X_3, X_8) \neq \text{Cov}_S(X_2, X_8) \cdot \text{Cov}_S(X_3, X_5)$ for all values of the free parameters.²² (This is easy to see using the trek sum rule for calculating covariances.) On the other hand, because X_2 , X_8 , X_9 , and X_{10} all have one unobserved common cause (*Emotionality*) as a direct common cause, the following vanishing tetrad constraints are entailed by SEM S : $\text{Cov}_S(X_2, X_8) \cdot \text{Cov}_S(X_9, X_{10}) = \text{Cov}_S(X_2, X_9) \cdot \text{Cov}_S(X_8, X_{10}) = \text{Cov}_S(X_2, X_{10}) \cdot \text{Cov}_S(X_8, X_9)$ [Spirtes *et al.*, 2001]. The BuildPureClusters algorithm uses the vanishing tetrad constraints as a guide to the construction of pure measurement models, and in the large sample limit reliably succeeds if a pure measurement model among a subset of the observed variables exists [Silva *et al.*, 2006] and each unmeasured common cause has at least three pure indicators.

Once a pure measurement model has been constructed, it is possible to estimate the covariances among the unobserved common causes. These covariances can then be given as input to an algorithm such as FCI to return a PAG among the unobserved common causes. In this example, the PAG returned contains an undirected edge between every pair of unobserved common causes. (SEM S is an example that is compatible with the PAG, but any other orientation of the edges among the three unobserved common causes that does not create a cycle is also compatible with the pattern.) The resulting SEM (or set of SEMs) passes a statistical test with a p-value of 0.47.

4.3 Calculating the Effects of Interventions from SEMs with Unobserved Common Causes

Given a SEM without unobserved common causes, there is a simple algorithm for estimating the total effect of a manipulation. First, write the total effect of the manipulation as a function of the free parameters using the path sum rule; second, estimate the values of the free parameters; third, substitute the estimated values of the free parameters into the equation for the effect of manipulation.

This algorithm is applicable to some SEMs with unobserved common causes as well, e.g. SEM S of Figure 6. There are algorithms that can consistently estimate the free parameters of SEM S from just the observed $\mathbf{X}$ variables (even though it is not possible to estimate the structural coefficients by simply regressing each variable on its parents, because the parents are unobserved).

²²The inequality is based on an extension of the Causal Faithfulness Assumption that states that vanishing tetrad constraints that are not entailed for all values of the free parameters by the true causal graph are assumed not to hold.

One commonly used method for estimating causal effects when there are unmeasured common causes is through the use of instrumental variables [Staiger and Stock, 1997]. The simplest kind of instrumental variable model has the causal graph shown in Figure 7 where *GAS PRICE* is the instrument. There are no regressions on the observed variables that are unbiased estimates of the linear coefficient relating *OPERATING COSTS* to *MILES* (because of the unobserved common cause *T*). Moreover, the structural coefficients relating *T* to *MILES* and *OPERATING COSTS* are not identifiable from the observed covariance matrix. However, the structural coefficient relating *OPERATING COSTS* to *MILES* is equal to $\text{Cov}(\text{GAS PRICE}, \text{MILES}) / \text{Cov}(\text{GAS PRICE}, \text{OPERATING COSTS})$, and hence is still identifiable.

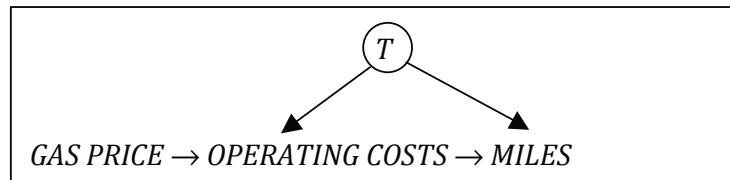


Figure 7. Instrumental Variable Model

In order to perform instrumental variable estimation, it is necessary to find some variable *X* to serve as an instrument; ideally for accurate estimations, *X* is strongly correlated with the cause, and uncorrelated with the error term of the effect [Staiger and Stock, 1997]. Unfortunately, in simple SEMs such as Figure 7 it is not possible to test whether *X* is an instrument (e.g. that *X* does not directly cause *Z*); this must be concluded from background assumptions about the domain. Recent algorithms have generalized instrumental variable estimation to a wider class of models.

Another approach to estimating the effects of manipulations in some models with unobserved common causes is based on a causal calculus that takes as input a causal graph and the marginal probability distribution over the observed variables [Spirtes *et al.*, 2001; Pearl, 2000]. The causal calculus has been incorporated into algorithms that search for an expression that relates the effect of a manipulation directly to the marginal probability distribution over the observed variables (instead of relating the manipulation to a function of the free parameters). The recently constructed algorithms have been shown to be complete (they either returns an answer or says “can’t be estimated”) [Pearl, 1995; Huang and Valtorta, 2006; Shpitser and Pearl, 2006a; Shpitser and Pearl, 2006b]. However, the algorithm requires knowing the true SEM, which in general cannot be determined by the observed marginal distribution alone.

In cases where a SEM is linear, but at most one of the errors is Gaussian, recently developed algorithms calculate the effect of one manipulated observed variable on other observed variables from the observed marginal distribution [Hoyer *et al.*,

2006; Hoyer *et al.*, in press].

4.3.1 *Calculating the Effects of Manipulation from a Markov Equivalence-over-O Class of SEMs with Unobserved Common Causes*

Suppose that instead of knowing the true SEM, it is known that the true SEM lies in a Markov equivalence-over-**O** class represented by a PAG (which is the case unless background assumptions are strong), e.g. the PAG for SEM G . It is not known whether the true SEM is SEM B , SEM C , SEM G , or some other SEM in the Markov equivalence-over-**O** class that has unobserved common causes. There is an algorithm that takes a PAG as input, and in some cases where all of the SEMs that are represented by the PAG agree on the effects of a particular manipulation, return a formula for the effect of the manipulation in terms of the marginal distribution over **O**, and otherwise returns “Can’t tell”.

Because a Markov equivalence-over-**O** class is a proper superset of a Markov equivalence class over SEMs without unobserved common causes, stronger inferences can be drawn when the assumption of no unobserved common causes is made. For example, because the only variables without unobserved common causes in the Markov equivalence-over-**O** class are SEMs C and G , if it is known that there are no unobserved common causes then it is known that X causes Y . In contrast, if there is a possibility that there are unobserved common causes, then the true SEM may have an unobserved common cause of X and Y . So given the assumption of no unobserved common causes and hence the edge $X \rightarrow Y$ it is possible to calculate the effect of X on Y ; but given the PAG (i.e. allowing the possibility of unobserved common causes and the edge $X \rightarrow Y$) it is not possible to calculate the effect of X on Y . However, even allowing for the possibility of unobserved common causes, it is possible to calculate that the effect of Y on X is zero from the PAG (from the arrowhead at $X \rightarrow Y$). There are also examples where non-zero effects can be calculated from PAGs. For example, it is possible to calculate the effect of parental encouragement on college plans from the PAG of Figure 5.

Introducing the possibility of unobserved common causes reduces the number of manipulated quantities that can be calculated, and increases the complexity of calculating the manipulated quantities that can still be calculated.

4.4 *Summary*

In summary, the following steps constitute a reliable method of estimating the effects of manipulations from sample data and background assumptions (e.g. acyclicity) when there are no unobserved common causes.

1. Perform a search to find the best SEM or set of SEMs. For some parametric families (e.g. linear Gaussian or discrete) the best that can be done is to find the best Markov equivalence-over-**O** class of SEMs. For other parametric families it is possible to find a single SEM (e.g. linear but non-Gaussian error terms.)

2. Determine if all of the SEMs in the best set of SEMs agree on the effect of a manipulation. (If this is a single SEM, this is trivial. If it is a Markov equivalence class of SEMs, there are algorithms for doing this. For intermediate cases, further research is needed.) If they do agree, output a formula that relates the effect of the manipulation to the marginal distribution over the observed variables entailed by the SEMs in the best class. If they don't, output "Can't tell" and halt.
3. For one of the SEMs in the best set of SEMs, estimate the entailed marginal density over the observed variables.
4. Substitute the estimated marginal density over the observed variables into the formula relating the effect of the manipulation to the entailed observed marginal density of the model.²³

5 OPEN PROBLEMS²⁴

The questions listed below are areas of active research that have produced some answers, but many open questions remain. The citations given are just a sample from larger bodies of research.

5.1 Models

There are a wide variety of causal models that have been employed in different disciplines. These include Bayesian Networks, Chain Graphs, Partial Ancestral Graphs, Markov Decision Processes, Structural Equation Models, Propensity Scoring, Information Theory, and Granger Causality. The relative advantages and disadvantages of these models and the relationships between these models are partially, but not completely understood. What new models are appropriate for different domains, e.g. feedback or reversible systems [Spirtes *et al.*, 1993; Spirtes *et al.*, 2001]? What new models are appropriate for different combinations of kinds of data, e.g. experimental and observational [Eberhardt and Glymour, 2006; Eberhardt *et al.*, 2005; Yoo and Cooper, 2004; Yoo *et al.*, 2006; Danks, 2002; Cooper and Yoo, 1999]? What new models are appropriate for different kinds of background assumptions, and different families of densities?

5.2 Model Scores

What kind of scores can be used to best evaluate causal models from various kinds of data? While some scores, such as BIC, have good large sample properties, they are difficult to compute or cannot be applied to some causal models, and may not

²³This works for simple cases, but in some cases the actual algorithm is slightly more complicated.

²⁴This section benefited greatly from suggestions made by Constantin Aliferis.

have good small sample properties. In a related vein, what are good families of prior densities that capture various kinds of background assumptions?

5.3 *Search Algorithms*

How can search algorithms be improved to incorporate different kinds of background assumptions, search over different classes of causal models, run faster, handle more variables and larger sample sizes, be more reliable at small sample sizes, and produce output that is as informative as possible?

5.4 *Properties of Search Algorithms*

For existing and novel causal search algorithms, what are their semantic and syntactic properties (e.g. soundness, consistency, maximum informativeness)? What are their statistical properties (pointwise consistency, uniform consistency, sample efficiency)? What are their computational properties (computational complexity)?

5.5 *Assumptions*

What plausible alternatives are there to the Causal Markov and Faithfulness Assumptions? Are there other assumptions might be weaker and hold in more domains and applications without much loss about what can be reliably inferred? Are there stronger assumptions that are plausible for some domains that might allow for stronger causal inferences? How often are these assumption violated, and how much do violations of these assumptions lead to incorrect inferences?

There are special assumptions, such as linearity, which can improve the strength of causal conclusions that can be reliably inferred, and the speed and sample efficiency of algorithms that draw the conclusions. What other density families or stronger assumptions about a domain are there that are plausible for some domains and how can they be used to improve causal inference?

Can various statistical assumptions be relaxed? For example, what if the sample selection process is not i.i.d., but may be causally affected by variables of interest [Richardson and Spirtes, 2002; Spirtes *et al.*, 1995; Cooper, 1995; Cox and Wermuth, 1996; Cooper, 2000]?

5.6 *Deriving Consequences of Causal Models*

Shpitser and Pearl have given complete algorithms for deriving the consequences of various causal models with hidden common causes in terms of the unmanipulated density and the given manipulation [Shpitser and Pearl, forthcoming]. Partial extensions of these results to deriving consequences from sets of causal models have been given [Zhang, 2008]; are there further extensions to derivations from sets of causal models?

It is often useful to quickly derive constraints (e.g. vanishing tetrad constraints) on marginal densities from causal models with hidden common causes, in order to

guide search. Are there other constraints on densities that can be derived from causal models, and how can they be incorporated into search algorithms?

5.7 Evaluation

What are the most appropriate performance measures for causal inference algorithms? What benchmarks can be established? What is the best research design for testing causal inference algorithms?

5.8 Interconnections

Many different domains have studied causal discovery including Artificial Intelligence, Econometrics, Markov Decision Processes, Operations Research, Control Theory, Experimental Design, and Statistics. What are the formal connections between the different models, assumptions, and algorithms used in each of these domains? What can each of these domains learn from the others?

6 APPENDIX

This appendix contains some definitions of the properties of estimators that are mentioned in the text.

An estimator ϕ_n of a model parameter μ (such as the mean of a variable) is a function from samples O^n of size n to a real number - that is, for each sample, the estimator outputs an estimate of the quantity μ . The quality of an estimator ϕ_n of μ can be measured by its *mean squared error*: that is, the mean (over all randomly independently selected samples of size n) of $(\phi_n(O^n) - \mu)^2$, the square of the difference between the real number output by the estimator and μ . Let $\bar{\phi}_n$ be the average output of $\phi_n(O^n)$ (with respect to the sampling distribution of the population density). The mean squared error is the sum of two terms: the bias and the variance of the estimator. The *bias* of the estimator ($\bar{\phi}_n - \mu$), i.e. the difference between the mean output of the estimator and the true value μ . The *variance* of the estimator is the mean of $(\phi_n(O^n) - \bar{\phi}_n)^2$, i.e. mean of the squared difference between the output of the estimator and the average output of the estimator). A pointwise consistent estimator is one in which the mean squared error approaches zero as the sample size approaches infinity. The quality of an estimator at a finite sample size depends upon the two factors of the bias of an estimator and the variance of the estimator. (There are other desirable properties that estimators can have, such as being uniformly consistent — roughly that it is possible to put probabilistic bounds on the size of the error at a given sample size — that will not be discussed here. (See [Robins *et al.*, 2003])).

One kind of estimator that is commonly employed is a maximum likelihood estimator, which under mild regularity conditions has a number of desirable properties such as pointwise consistency. Suppose that SEM C is given and the goal is to estimate the values of the free parameters. For assignment θ of values to the free

parameters of SEM C , there is an implied covariance matrix. Given the implied covariance matrix, it is possible to determine the density of drawing a sample that has the observed sample covariance matrix; this is the likelihood of the data for θ . A maximum likelihood estimator selects the assignment θ of values to the free parameters that makes the sample data most likely. In the case of a SEM over a causally sufficient set of variables, the maximum likelihood estimate of the linear coefficient b_{YZ} (denoted by $\hat{b}_{YZ}$) is the regression coefficient of Z when Y is regressed on its non-descendants in the causal graph (e.g. in the case of SEM C , the regression coefficient of Z when Y is regressed on Z , X , and W). In SEM C , the formula for the total effect of W on Y is $m \cdot b_{YZ} \cdot b_{ZW}$. Substituting the maximum likelihood estimates of $\hat{b}_{YZ}$ and $\hat{b}_{ZW}$ into the formula yields $m \cdot \hat{b}_{YZ} \cdot \hat{b}_{ZW}$, which is a maximum likelihood estimate of the total effect of W on Y .

ACKNOWLEDGEMENTS

I would like to thank an anonymous reviewer for many helpful comments, and Constantin Aliferis for detailed suggestions about the description of open problems.

BIBLIOGRAPHY

- [Ali *et al.*, 2005] R. A. Ali, T. S. Richardson, P. Spirtes, and J. Zhang. Towards Characterizing Markov Equivalence Classes for Directed Acyclic Graphs with Latent Variables. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, 2005.
- [Bartholomew and Knott, 1999] D. J. Bartholomew and M. Knott. *Latent Variable Models and Factor Analysis (Kendall's Library of Statistics, 7)*. A Hodder Arnold Publication, 1999.
- [Bartholomew *et al.*, 2002] D. J. Bartholomew, F. Steele, I. Moustaki, and J. I. Galbraith. *The Analysis and Interpretation of Multivariate Data for Social Scientists (Texts in Statistical Science Series)*. Chapman and Hall/CRC, 2002.
- [Bollen, 1989] K. A. Bollen. *Structural Equations with Latent Variables*. Wiley-Interscience, 1989.
- [Cartwright, 1994] N. Cartwright. *Nature's Capacities and Their Measurements (Clarendon Paperbacks)*. Oxford University Press, USA, 1994.
- [Chickering, 2003] D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3, 507-554, 2003.
- [Cooper, 1995] G. F. Cooper. *Causal Discovery from data in the presence of selection bias*. Paper presented at the Fifth International Workshop on AI and Statistics, 1995.
- [Cooper, 2000] G. F. Cooper. *Causal Modeling and Discovery Under Selection*. Paper presented at the Sixteenth Uncertainty In Artificial Intelligence Conference, San Francisco, CA, 2000.
- [Cooper and Yoo, 1999] G. F. Cooper and C. Yoo. *Causal Discovery from a Mixture of Data*. Paper presented at the Fifteenth Uncertainty In Artificial Intelligence Conference, San Francisco, CA, 1999.
- [Cox and Wermuth, 1996] D. R. Cox and N. Wermuth. *Multivariate Dependencies: Models, Analysis and Interpretation (Monographs on Statistics and Applied Probability)*. Chapman and Hall/CRC, 1996.
- [Danks, 2002] D. Danks. Learning the causal structure of overlapping variable sets. *Lect Notes Comput Sc*, 2534, 178-191, 2002.
- [Eberhardt, 2009] F. Eberhardt and C. Glymour. 4 N-1 Experiments Suffice to Determine the Causal Relations Among N Variables. *Innovations in Machine Learning: Theory And Applications*, 2009.

- [Eberhardt *et al.*, 2005] F. Eberhardt, C. Glymour, and R. Scheines. *On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables*. Paper presented at the 21st Conference on Uncertainty in Artificial Intelligence, 2005.
- [Fisher, 1970] F. Fisher. A Correspondence Principle for Simultaneous Equation Models. *Econometrica*, 38, 73-92, 1970.
- [Geiger *et al.*, 2002] D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified Exponential Families: Graphical Models and Model Selection. *Ann. Statist.*, 30, 1412-1440, 2002.
- [Geiger and Meek, 1999] D. Geiger and C. Meek. *Quantifier Elimination for Statistical Problems*. Paper presented at the 15th Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, 1999.
- [Gierl and Todd, 1996] M. J. Gierl and R. W. Todd. A Confirmatory Factor Analysis of the Test Anxiety Inventory Using Canadian High School Students. *Educational and Psychological Measurement*, 56(2), 315-324, 1996.
- [Glymour, 1998] C. Glymour. What Went Wrong? Reflections on Science by Observation and the Bell Curve. *Philosophy of Science*, 65(1), 1-32, 1998.
- [Glymour, 1999] C. Glymour. Rabbit Hunting. *Synthese: An International Journal for Epistemology, Methodology and Philosophy of Science*, 121, 55-78, 1999.
- [Glymour, 1980] C. Glymour. *Theory and evidence*. Princeton, N.J: Princeton University Press, 1980.
- [Harman, 1976] H. H. Harman. *Modern Factor Analysis*. University Of Chicago Press, 1976.
- [Haughton, 1988] D. M. A. Haughton. On the Choice of a Model to Fit Data from an Exponential Family. *The Annals of Statistics*, 16(1), 342-355, 1988.
- [Hoover, 2001] K. D. Hoover. *Causality in Macroeconomics*. Cambridge University Press, 2001.
- [Hoyer *et al.*, 2006] P. O. Hoyer, S. Shimizu, and A. Kerminen. *Estimation of linear, non-gaussian causal models in the presence of confounding latent variables*. Paper presented at the Third European Workshop on Probabilistic Graphical Models, 2006.
- [Hoyer *et al.*, in press] P. O. Hoyer, S. Shimizu, A. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, in press.
- [Huang, 2006] Y. Huang and M. Valtorta. *Identifiability in Causal Bayesian Networks: A Sound and Complete Algorithm*. Paper presented at the Twenty-First National Conference on Artificial Intelligence, Edinboro, Scotland, 2006.
- [Lauritzen, 2001] S. L. Lauritzen. Causal Inference from Graphical Models. In D. Barndorff-Nielsen and C. Kluppelberg (pp. 63-107). London/Baton Rouge: Chapman and Hall, 2001.
- [Lauritzen *et al.*, 1990] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H. G. Leimer. Independence properties of directed Markov fields. *Networks*, 20, 491-505, 1990.
- [Pearl, 1995] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4), 669-688, 1995.
- [Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
- [Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
- [Ramsey *et al.*, 2006] J. Ramsey, P. Spirtes, and J. Zhang. *Adjacency-Faithfulness and Conservative Causal Inference*. Paper presented at the 22nd Conference on Uncertainty in Artificial Intelligence, 2006.
- [Richardson and Spirtes, 2002] T. Richardson and P. Spirtes. Ancestral graph Markov models. *Ann Stat*, 30(4), 962-1030, 2002.
- [Robins, 1986] J. M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods-application to the healthy worker survivor effect. *Mathematical Modelling*, 7, 1393-1512 1986.
- [Robins *et al.*, 2003] J. M. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3), 491-515, 2003.
- [Rodgers and Maranto, 1989] R. C. Rodgers and C. L. Maranto. Causal-Models Of Publishing Productivity In Psychology. *J Appl Psychol*, 74(4), 636-649, 1989.
- [Rubin, 1974] D. B. Rubin. Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies. *Journal of Educational Psychology*, 66(5), 688-701, 1974.
- [Sewell and Shah, 1968] W. H. Sewell and V. P. Shah. Social Class, Parental Encouragement, and Educational Aspirations. *Am J Sociol*, 73(5), 559-572, 1968.

- [Shimizu *et al.*, 2006] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7, 2003-2030, 2006.
- [Shpitser and Pearl, 2006a] I. Shpitser and J. Pearl. *Identification of Conditional Intervention Distributions*. Paper presented at the 22nd Annual Conference on Uncertainty in Artificial Intelligence, Arlington, VA, 2006.
- [Shpitser and Pearl, 2006b] I. Shpitser and J. Pearl. *Identification of Joint interventional Distributions in Recursive Semi-Markovian Causal Models*. Paper presented at the Twenty-First National Conference on Artificial Intelligence, Menlo Park, California, 2006.
- [Shpitser and Pearl, forthcoming] I. Shpitser and J. Pearl. Complete Identification Methods for the Causal Hierarchy. *Journal of Machine Learning Research*, forthcoming.
- [Silva *et al.*, 2006] R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *J Mach Learn Res*, 7, 191-246, 2006.
- [Simon, 1985] H. A. Simon. Spurious Correlation: A Causal Interpretation. In H. M. Blalock (Ed.), *Causal Models in the Social Sciences*. (pp. 7-22). MacMillan, 1985.
- [Spearman, 1904] C. Spearman. General Intelligence objectively determined and measured. *American Journal of Psychology*, 15, 201-293, 1904.
- [Spirtes *et al.*, 1995] P. Spirtes, C. Meek, and T. S. Richardson. *Causal inference in the presence of latent variables and selection bias*. Paper presented at the Eleventh Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, 1995.
- [Spirtes *et al.*, 1998a] P. Spirtes, T. Richardson, C. Meek, R. Scheines, and C. Glymour. Using path diagrams as a structural equation modeling tool. *Sociol Method Res*, 27(2), 182-225, 1998.
- [Spirtes and Richardson, 1996] P. Spirtes and T. S. Richardson. *A Polynomial Time Algorithm For Determining DAG Equivalence in the Presence of Latent Variables and Selection Bias*. Paper presented at the Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics, 1996.
- [Spirtes *et al.*, 1998b] P. Spirtes, T. S. Richardson, C. Meek, and R. Scheines. Using path diagrams as a structural equation modeling tool. *Sociological Methods and Research*, 27, 182-225, 1998.
- [Spirtes *et al.*, 1990] P. Spirtes, R. Scheines, and C. Glymour. Simulation Studies of the Reliability of Computer-Aided Model Specification Using the TETRAD II, EQS and LISREL VI Programs. *Sociological Methods and Research*, 3-66, 1990.
- [Spirtes *et al.*, 1993] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag Lectures in Statistics, 1993.
- [Spirtes *et al.*, 2001] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search, Second Edition (Adaptive Computation and Machine Learning)*. The MIT Press, 2001.
- [Staiger and Stock, 1997] D. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3), 557-586 1997.
- [Strotz and Wold, 1960] R. H. Strotz and H. O. A. Wold. Recursive VS Nonrecursive Systems- An Attempt At Synthesis. *Econometrica*, 28(2), 417-427, 1960.
- [Wyatt, 2004] G. Wyatt. *Macroeconomic Models in a Causal Framework*. Exempla Books, 2004.
- [Yoo and Cooper, 2004] C. Yoo and G. F. Cooper. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. *Artif Intell Med*, 31(2), 169-182, 2004.
- [Yoo *et al.*, 2006] C. Yoo, G. F. Cooper, and M. Schmidt. A control study to evaluate a computer-based microarray experiment design recommendation system for gene-regulation pathways discovery. *J Biomed Inform*, 39(2), 126-146, 2006.
- [Zhang, 2007] J. Zhang. *A Characterization of Markov Equivalence Classes for Directed Acyclic Graphs with Latent Variables*. Paper presented at the 23rd Conference on Uncertainty in Artificial Intelligence, Vancouver, BC, 2007.
- [Zhang, 2008] J. Zhang. Causal Reasoning with Ancestral Graphs. *Journal of Machine Learning Research*, 9, 1437-1474, 2008.

This page intentionally left blank

THE LOGIC AND PHILOSOPHY OF CAUSAL INFERENCE: A STATISTICAL PERSPECTIVE

Sander Greenland

The topic of causality is a vast one in science and philosophy, so vast that even a limited review would require a book. Yet most theories have not found favor among empirical researchers – by whom I mean those whose primary job is to collect and analyze data, as opposed to philosophers or theoreticians. This chapter will thus concern only the few statistical theories for causation and causal inference that have produced methods now widespread in practice.

In an attempt to avoid the confusion that often accompanies narrative descriptions of causation and causal inference (especially in applied sciences), this chapter uses rather stark and purely logical descriptions, and will assume the reader has at least some familiarity with probability and statistics. Detailed illustrations and applications, along with philosophical discussions, can be found in the references. Special emphasis will be given to issues of causal inference from uncontrolled observations (observational studies), in which the effect under study becomes difficult to separate from other, extraneous phenomena; the latter are often called “bias sources” or “systematic errors”.

Before embarking on these descriptions, I will touch briefly on the relevance (or possible lack thereof) for statistics of philosophies of causation.

DO WE NEED PHILOSOPHY OF CAUSATION FOR A STATISTICAL THEORY OF CAUSAL INFERENCE?

It is possible to distinguish two kinds of inference: Inference to causal models from observations, and inference from causal models to the effects of manipulations. Inference to causal models may be viewed as trying to construct a general set of laws from existing observations that can be tested with and applied to new observations. In statistics this problem is subsumed under the topic of model specification or model building. Inference from causal models may be viewed as deducing tests and making decisions based on proposed or accepted laws, which in statistics is subsumed under topics of testing, estimation, and decision theory.

In applied statistics, the feedback between these two directions of inference is often summarized as a cycle of model proposal → model test → model revision → model test that continues until available tests cease to have practical impact on the model [Box, 1980]. There are familiar controversies about whether cycles of this form lead toward “truth” or simply toward effective tools for prediction and

manipulation (e.g., [Kuhn, 1970a; 1970b]), and whether the philosophical debate surrounding causal inference stems from the fact that the word “causation” evokes some notion of a deeper truth about the world hidden from current view.

Of interest then is that the most successful statistical model of causation, the potential-outcomes model discussed below, has attracted theoretical criticisms precisely because it contains counterfactual elements hidden from randomized-experimental test (e.g., [Dawid, 2000]). These criticisms have been dismissed by applied statisticians (see the discussion following [Dawid, 2000]), who understand that the manipulative account inherent in potential-outcomes models fits well with the more instrumentalist or predictive view of causation than critics admit. Indeed, these models can be and have been used to great success with no worry about whether their hidden elements need to be taken seriously [Greenland, 2004], just as the celestial cogs and wheels once used to display the Ptolemaic model of celestial motions were no obstacle to its considerable predictive success.

Given this instrumentalist view, it might seem that causal inference maybe distinguished from other inferences only due to its emphasis on manipulation rather than prediction. From a statistical viewpoint, the distinction between prediction and causal inference is semantic, not philosophical: Causal inference is merely special case of prediction in which we are concerned with predicting outcomes under alternative manipulations. Because only one of the alternatives can be carried out, only one of the outcomes can be observed, resulting in nonidentification. But the solution to this problem is no different than in problems of pure prediction: We simply assume some limited form of isotropy, in which predictive regularities (whether labeled “predictive” or “causal”) persist over the space and time spans of interest, at least enough to justify generalizations across the spans. Whether a deeper analysis is warranted for practice remains to be seen.

POTENTIAL OUTCOMES AND STRUCTURAL EQUATIONS

Manipulative accounts of causation, including those with counterfactuals, have deep roots in the history of modern science. Informal outlines for causal inference may be traced as far back as the development of experimental science. After all, typical definitions of “experiment” include an element of experimenter “control” of conditions, implying that such control will affect the outcome. Early in these developments, however, Hume [1739; 1748, p. 115] recognized that the definition of “cause” implicit in much usage carried the seeds of intractable underdetermination or, as known in statistics, *nonidentification*; that is, observation alone could not determine or identify whether one condition caused another.

To formalize the notion of causation and delineate the identification problem, consider the following model (introduced by Neyman [1923]) which became established in the experimental literature in the mid-20th century, was later extended and popularized for observational research [Rubin, 1990], and is now standard in much of statistics where it is called the *potential-outcomes* or “counterfactual” model. Suppose we observe a subject i to have a particular outcome after being

given a treatment. (“Subject” is here merely a term for observational unit; it may be a plot of land, a laboratory animal, or a population, as opposed to a person.) Let X be the variable ranging over the treatment possibilities, and let x and x^* be two distinct treatments (that is, values for X with $x^* \neq x$). Let Y be the variable ranging over the outcome (response) possibilities, and let y and y^* be two distinct outcomes (that is, values for Y with $y^* \neq y$). As an example, X might encode a range of treatment options for women with perimenopausal complaints, such as unopposed estrogen therapy, opposed estrogen therapy, placebo treatment, and no treatment, while Y could indicate survival over the decade following treatment initiation ($Y = 1$ if the woman survives, 0 if not), or Y could be the survival time (lifespan) following treatment initiation.

A common notion of cause and effect is then captured as follows: Receiving treatment x (i.e., having $X = x$) *caused* the outcome Y to be y for the subject relative to having instead $X = x^*$, if the actual outcome and treatment were y and x , but would have been y^* had x^* been administered instead. The two outcomes y and y^* are then called the *potential outcomes* corresponding to treatments x and x^* for the subject, and the difference $y - y^*$ is called a measure of the *effect* on the subject of giving $X = x$ instead of (or relative to) $X = x^*$. In the example, $y - y^*$ could be the difference of survival time with unopposed estrogen therapy (x) versus placebo (x^*).

The nonidentification problem reflects that the subject’s response y^* to treatment x^* is not observed if the received treatment x is not equal to x^* , and therefore the effect $y - y^*$ cannot be computed from the observations. For example, we cannot observe how long a woman would survive under placebo therapy if in reality she receives unopposed estrogen, and so we cannot compute the effect that receiving unopposed estrogen rather than placebo had on this woman. This problem in causal statements is often highlighted by noting that the premise of the conditional “If X had been x^* instead of x , Y would have been y^* ” is counterfactual (contrary to fact) [Lewis, 1973ab]. Any statistical inference about the effect must therefore invoke physical assumptions that give precise meaning to the counterfactual conditional. It must also invoke “identification” assumptions that allow construction of estimates and tests of the average or expected effect $E(Y - Y^*)$ from the data actually observed.

For further details and citations on the model see [Greenland *et al.*, 1999; Greenland, 2004], or some of the many other reviews (e.g., [Morgan and Winship, 2007; Pearl, 2009]). The reader is warned however that (notwithstanding the enormous contributions by Rubin to the model), much of the sociologic literature misattributes the model to Rubin, some going so far as to call the model the “Rubin Causal Model” (e.g., [Holland, 1986]), and thus misses large segments of the literature on the model in the experimental, econometric, and health sciences.

Causal Laws and Structural Equations

Basic science often provides an empirical pattern or a more formal physical theory that predicts the subject's outcome Y as a function of the treatment X . This function might be subject-specific, tailored to specifics of the subject's characteristics. For example, suppose the subject is an ordinary ceramic dish, the treatment is dropping it flat on a concrete floor from height x , and the outcome is breakage ($Y = 1$) or not ($Y = 0$). Ordinary experience provides us a rough theory that says dropping the dish two meters will cause breakage whereas dropping it one millimeter will not; that is, $X = 2000\text{mm}$ will cause $Y = 1$ relative to $X = 1\text{mm}$. A different theory would apply to a steel dish. Especially in physics, such experience may eventually give rise to a mathematical "law" or model $f(x)$ relating Y to X for each of a broad class of subjects. Examples include laws governing behavior of charged particles in response to an electric field of a given strength, or more limited laws governing the size of predator populations in response to prey abundance.

Such theoretical mechanisms or laws shift the uncertainty about the counterfactuals (which are unobserved potential outcomes) to uncertainty about the mechanisms or laws connecting the outcome variable Y to the antecedent variable X . Indeed, that shift is often promoted as a major force for progress in experimental science, as follows: Suppose one proposes a general physical theory that says or implies that $Y = f(x)$ whenever $X = x$ for each subject in a given class. Observing these predictions fail — that is, observing enough subjects in the class who have $Y \neq f(x)$ — can then be grounds for discarding or modifying the theory.

A functional relation $Y = f_i(x)$ supplying the outcome of subject i under different possible treatments is often called a *structural equation* for the subject. It is important to note that the equation gives the variation in the outcome Y as X varies *within a single subject*; that is, it shows how Y varies as X is varied while i is held constant. This single-subject feature captures the counterfactual nature of causal laws.

Although it is not the defining feature of a structural equation, the within-subject property distinguishes the equation from ordinary "regression" functions of statistics, which describe *associations*. An association is the variation in Y as one moves *across* subjects with different X ; that is, both X and i are varying in a regression. Nonetheless, because analytic statistical methodology is heavily invested in estimating associations, and because associations are all that are statistically identified in nonexperimental settings, much of statistical theory for causal inference comprises delineation of assumptions that allow deduction of regression equations from structural equations [Berk, 2003; Pearl, 2009].

Causal Null Hypotheses

A large portion of standard statistical theory concerns testing of associational "null hypotheses," which assert absence of association in a population or distribution from which the observations are supposed to have arisen. In causal inference these

nulls become no-effect hypotheses or “causal nulls,” which in their strong or strict form state that the structural equation is constant for each subject; that is, $f_i(x) = c_i$ for each subject i . This hypothesis need not correspond to the hypothesis of no association, which asserts that Y does not change across different subjects with different X values. But much of statistical theory for causal inference comprises delineation of assumptions that allow deduction of associational null hypotheses from causal null hypotheses. When such an associational null is identifiable, its rejection by a statistical test implies that the causal null hypotheses should be rejected as well.

The Causal Identification Problem

As just described, there is one element that makes a theory causal and which demands more than mere observation for inference: The multiple possible values for Y for each subject i , given different values of X . Let x_i be the observed value of X for subject i , and suppose we have a theory that predicts Y as a function $f_i(x_i)$ of X . No matter how many subjects yield the predicted outcome $f_i(x_i)$ upon passive observation, and so are in accord with the causal theory that $Y = f_i(x)$, that theory cannot be deemed more than a good *description* of how Y and X will be associated as one looks *across* subjects. In terms of passive observation, it merely predicts how plots the pairs x_i, y_i will look as i varies across subjects.

The theory’s predictions may be borne out among the treatment assignments (X distribution) we observe, but we cannot be sure the theory would have succeeded under other, counterfactual treatment assignments. In algebraic terms, we might see $Y_i = f_i(x_i)$ for all the subjects, even though Y_i might not have equaled $f_i(x^*)$ for some $x^* \neq x_i$. In sum, a theory may be a very good description of what we observe but a poor predictor of what we *would* observe upon intervening to change treatment assignment X . In other words, no matter how well it predicts associations across subjects, it need not tell us how the outcome Y would change upon changing X *within* subjects.

Confounding and Randomization

Whether by the investigator, nature, or another party, the treatment x_i given subject i might have been determined in a way that is associated with Y across subjects, apart from any causal (within-subject) link from X to Y . The term *confounding* is often used to refer to this condition [Greenland *et al.*, 1999], although it is also known as “nonignorability of the treatment-assignment mechanism” [Rubin, 1991]. Another way to describe this condition is that there is between-subject variation in Y that is not due to within-subject variation of Y with X ; in other words, given a fixed value x for X , it is variation in the potential outcome Y_x with X across subjects. Earlier, informal discussions of causal inference described this condition as “extraneous variation,” or that portion of between-subject association of X and Y that is not due to an effect of X on Y . They recognized that such

covariation of X and Y would remain present even if the causal null hypothesis were correct, and thus would distort causal inferences or tests of that hypothesis [Mill, 1843].

Sources of extraneous variation are sometimes called “confounders,” although the term *confounder* is often defined in more strict terms [Greenland and Pearl, 2007]; see the section on causal diagrams below). Recognizing the impossibility of eliminating or even knowing all confounders in biological work, R.A. Fisher [1932, 1935] developed an elegant theory of *randomized experiments* to allow statistical “control for” confounding. This control is accomplished by enforcing a known distribution for confounding under sufficiently strict causal hypotheses (such as the null hypothesis).

In its basic form, randomization theory replaces vague ignorance about the degree of confounding with a fully specified probability distribution for the observed outcomes under the sharp null hypothesis of no effect $Y_i = f_i(x) = c_i$ (Y constant across X within i). Consider classical permutation inference (e.g., [Cox and Hinkley, 1974, Ch.6]) in an experiment that assigns values of X among N subjects, and let R be the treatment-assignment variable. R ranges over possible values for X ; thus, when subject i is assigned to have $X = x$, the subject’s value r_i for R is equal to x . Note however that the subject may deviate from assigned treatment, resulting in the actual value x_a of X not being equal to the assigned value x (in which case $X \neq R$).

If assignment R has no effect on Y , the observed outcomes $y_1, \dots, y_N$ should be the same regardless of the assignments $r_1, \dots, r_N$; in other words, they should be the c_i in the null model $Y_i = f_i(x) = c_i$. Thus, given this causal null hypothesis, we may regard the outcome list $(y_1, \dots, y_N)$ as if it were fixed from the start of the experiment and thus independent of a subsequent random treatment allocation. A known randomization scheme then allows one to compute exact probabilities for any allocation of these fixed outcomes among the treatment levels, including the allocation observed.

From these probabilities, we can compute null distributions of test statistics (such as the sample sum of cross products $\sum_i x_i y_i$ used for trend tests) and thus compute P -values (“observed significance levels”) for testing the causal null hypothesis. If the possible treatment allocations are merely permutations of the actual treatment-assignment list $r_1, \dots, r_N$, the resulting test is known as a *permutation test* for the effect. Test of means may also be used, noting that the total-sample mean of Y remains fixed under the null hypothesis, regardless of allocation.

As a special case of the foregoing methods, consider binary X and Y with perfect random allocation of N_1 subjects to $X = 1$ and N_0 subjects to $X = 0$, and let $N = N_1 + N_0$. Then $\sum_i x_i y_i$ is the number of subjects in the $X = 1, Y = 1$ cell of the 2 by 2 table of X and Y . Under the causal null hypothesis, $(y_1, \dots, y_N)$ is a fixed vector of potential outcomes, and so the $Y = 1$ marginal total in the experiment is fixed at $y_+ = \sum_i y_i$ (so is the sample mean of Y , y_+/N , which is the proportion of subjects having $Y = 1$). Under the null, randomization becomes nothing more

than randomly allocating the fixed outcomes $(y_1, \dots, y_N)$ to the $X = 1$ and $X = 0$ categories, which induces a hypergeometric distribution for $\sum_i x_i y_i$, as derived by Fisher for his exact test [Cox and Hinkley, 1974, Ch. 5].

Causality and Conditionality

The conditionality problem illustrates how the introduction of a causal component into a statistical model can resolve previous ambiguities in choice of a statistical procedure. This resolution comes from explicitly modeling the otherwise hidden within-subject dimension underlying causal questions, and shows how statistical questions can arise even when no ordinary sample-to-population inference problem exists.

There has been a long-running controversy in statistics concerning whether the use of permutation tests is justifiable when the sample distribution of Y (comprising the observed values $y_1, \dots, y_N$) is not fixed in advance by the investigator. Some of the arguments for these tests appeal to rather abstruse and somewhat controversial principles of ancillarity or conditionality [Little, 1989] while others are based on favorable repeated-sampling (frequency) properties when $y_1, \dots, y_N$ represent a sample from a distribution for Y .

If one is concerned only with the observed N units, the question of association of X and Y among those sampled is purely empirical rather than inferential, in that it is answered by simply plotting or cross-tabulating the observed pairs (x_i, y_i) . In this regard, it is no different a question than asking the heights of the N tallest mountains on earth: Accept the reported measurements and you have your answer.

The absence of a statistical inference problem in this descriptive question has led many to automatically identify the fixed-margin controversy and even causal inference with the problem of inference to a larger population. Add however the causal dimension and we have a problem of inference from observed properties of the sample (the observed distribution of the pairs x_i, y_i) to unobserved properties of the same sample, namely the unobserved potential outcomes of the sampled subjects. As described above, the fixed Y margin can be deduced directly from the causal null hypothesis for the observed sample of units ($i = 1, \dots, N$), with no reference to sampling from a larger population. In particular, the fixed Y margin is just a physical property of the sample under the causal null hypothesis among those sampled [Greenland, 1991].

There is however a connection to inference about a population from which the observed units are sampled. First, note that rejection of the null for the sample implies rejection for the population: The sample is part of the population, and thus finding an effect in the sample implies that there is an effect in the population (namely, in the part that composes the sample). The converse condition is that failure to reject the null for the sample should correspond to failure to reject for the population. This converse is not logically necessary, but violating it would amount to asserting that causation exists in a population even though we would not assert causation exists in the portion we observed. At the very least such an

assertion would seem paradoxical [Greenland, 1991].

The General Causal Inference Problem as a Missing-Data Problem

To extend statistical reasoning about causation beyond the null hypothesis, it is essential to add a model for the distributions of the potential outcomes. In the general form of this model, the univariate outcome variable Y is replaced by a fixed, baseline covariate vector $\mathbf{Y}$ with components Y_x indexed by values x of the treatment X ; these Y_x are the potential-outcome variables, one for each possible value x of X . (More generally, as described later, X may be a vector $\mathbf{X}$ that indexes the potential-outcome vector $\mathbf{Y}$.)

The treatment-allocation variable becomes a vector $\mathbf{R}$ of indicators with components R_x , where $R_x = 1$ if and only if component Y_x of $\mathbf{Y}$ is observed. The observed value $\mathbf{r}_i$ of $\mathbf{R}$ for subject i thus displays which component of $\mathbf{Y}$ was observed (if any), and is a vector of zeros except possibly a single 1 at the component corresponding to the actual treatment. As a consequence, $r_{+i} = \sum_x r_{xi}$ is either 0 or 1 (no or one component of $\mathbf{Y}$ observed). This conceptual framework allows one to view causal-inference problems as special types of missing-data problems, in which only one component of $\mathbf{Y}$ can be observed (i.e., no more than one component of $\mathbf{R}$ can be 1), leaving the rest of $\mathbf{Y}$ as “missing data” [Rubin, 1991].

Because only one component of $\mathbf{Y}$ can be observed, the joint distribution $\Pr(\mathbf{Y}=\mathbf{y})$ of the components of $\mathbf{Y}$ is not statistically identified — that is, distinct distributions for $\mathbf{Y}$ can lead to exactly the same distribution for the observations. Assuming that every subject has an observed outcome, those observations are the subject-specific dot products $\mathbf{R}'\mathbf{Y} = \sum_x R_x Y_x$, which equal the observed outcomes (the Y_x for which $R_x = 1$), and $\mathbf{R}$, which shows the treatment that was received (i.e., the R_x that equals 1). Without further assumptions or experimental control, all we may identify is $\Pr(\mathbf{R}'\mathbf{Y} = \mathbf{r}'\mathbf{y}, \mathbf{R}=\mathbf{r})$ and functions of it, including the conditional distributions $\Pr(Y_x = y | R_x = 1)$. In the nonexperimental settings, it may be prudent to search for the weakest identification assumptions consonant with our practical goals.

Standardization and Inference on Marginal Effects

The goal of most statistical causal inference is to compare the marginal distributions $\Pr(Y_x = y)$ of the Y_x across X . These comparisons are identified under the independence assumption that for all x , $\Pr(Y_x = y) = \Pr(Y_x = y | R_x = 1)$, sometimes called “weak ignorability” (the stronger but inessential condition that $\mathbf{R}$ is independent of $\mathbf{Y}$ is then called “strong ignorability”), and which corresponds to absence of confounding. The paradigmatic example arises when $\mathbf{R}$ is randomized by the investigator, for then $\mathbf{R}$ is independent of everything, including any potential-outcome vector $\mathbf{Y}$.

In nonrandomized studies, ignorability conditions are usually unacceptable assumptions. One strategy for this situation is to pretend instead that $\mathbf{R}$ and $\mathbf{Y}$ are

independent conditional on a vector of fully observed covariates $\mathbf{Z}$, called *strong ignorability given $\mathbf{Z}$* . A weaker condition sufficient for practical applications is the analogous set of X -specific relations,

$$(1) \quad \Pr(Y_x = y | R_x = 1, \mathbf{Z} = \mathbf{z}) = \Pr(Y_x = y | \mathbf{Z} = \mathbf{z})$$

for all x , called *weak ignorability given $\mathbf{Z}$* .

Now consider the following “standardization” formula, which biostatisticians, demographers, and epidemiologists may recognize as a modern version of the classical formula for “direct standardization” to the covariate distribution $\Pr(\mathbf{Z} = \mathbf{z})$:

$$(2) \quad \Pr(Y_x = y) = \sum_z \Pr(Y_x = y | \mathbf{Z} = \mathbf{z}) \Pr(\mathbf{Z} = \mathbf{z})$$

This equation is just a basic probability relation displaying $\Pr(Y_x = y)$ as a covariate-probability weighted average of the covariate-specific potential-outcome probabilities $\Pr(Y_x = y | \mathbf{Z} = \mathbf{z})$. Applying assumption (1) to equation (2) yields

$$(3) \quad \Pr(Y_x = y) = \sum_z \Pr(Y_x = y | R_x = 1, \mathbf{Z} = \mathbf{z}) \Pr(\mathbf{Z} = \mathbf{z}).$$

This equation provides the desired marginals $\Pr(Y_x = y)$ in terms of the observable distribution $\Pr(Y_x = y | R_x = 1, \mathbf{Z} = \mathbf{z})$, assuming the standardization is adequate to remove confounding. Equation (3) is thus sometimes termed “no confounding of the marginal effects given $\mathbf{Z}$ ”. It is a weaker condition than assumption (1) (weak ignorability given $\mathbf{Z}$) because deviations from (1) may average to zero over $\mathbf{Z}$ and thus preserve equation (3), or at least leave it an acceptable approximation. In other words, in theory, stratification by $\mathbf{Z}$ need not be sufficient for estimating conditional effects in order to be sufficient for estimating marginal effects.

Using the fact that $\Pr(R_x = 1 | \mathbf{Z} = \mathbf{z}) = \Pr(R_x = 1, \mathbf{Z} = \mathbf{z}) / \Pr(\mathbf{Z} = \mathbf{z})$, equation (3) may be rewritten in an equivalent form

$$(4) \quad \Pr(Y_x = y) = \sum_z \Pr(Y_x = y, R_x = 1, \mathbf{Z} = \mathbf{z}) / \Pr(R_x = 1 | \mathbf{Z} = \mathbf{z}).$$

This equation displays $\Pr(Y_x = y)$ as an *inverse-probability weighted* (IPW) average of the unconditional observation probabilities $\Pr(Y_x = y, R_x = 1, \mathbf{Z} = \mathbf{z})$. The allocation probabilities $\Pr(R_x = 1 | \mathbf{Z} = \mathbf{z})$ that form the inverse weights are sometimes called “propensity scores” and can be generalized to continuous and time-dependent treatment processes [Robins, 1999ab]. Equation (4) shows how these scores, if known or at least identifiable, can be used to estimate a marginal potential-outcome distribution $\Pr(Y_x = y)$ under the $\mathbf{Z}$ -conditional weak ignorability assumption, which licenses the derivation of equation (3) and hence (4) from equation (2).

Summary

Classical permutation arguments are based on enforcing a distribution for the observation-indicator vector $\mathbf{R}$ (e.g., by treatment randomization) and then using

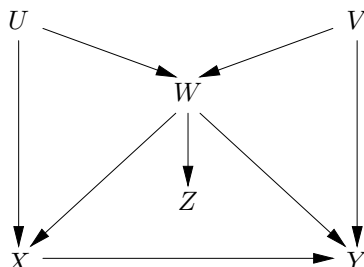


Figure 1. Example of a DAG

this distribution as the source of subsequent probability statements about the data. These arguments are not heavily emphasized in most statistical training, and yet are the ones most directly linked to potential-outcome models of causation. They should be contrasted with regression statistics, which base their probability statement on assumed distributions for the observed outcome $\mathbf{R}'\mathbf{Y}$ given observed covariates $\mathbf{Z}$.

The advantage of the treatment-based approach to causal inference is clear when indeed the distribution of $\mathbf{R}$ is known or at least identifiable, as in experiments: It seems far better to use an identified distribution than one that is merely assumed. Nonetheless, in nonexperimental (observational) research, all distributions become mere assumptions when not based on known mechanisms. The choice between approaches then comes down to judgments about which assumptions are more plausible or at least more palatable. Current research on statistical methods for causal inference includes development of “multiply robust” procedures, which retain their statistical validity under broader conditions than either treatment-based or outcome-based modeling [Kang and Shafer, 2007].

CAUSAL SYSTEMS AND CAUSAL DIAGRAMS

Suppose now we have a time-sequenced set of structural equations, or *causal system*, in which (for example) the output w of a function $f(u, v)$ may become an input to a later function $g(u, w)$ with output x . (Equivalently, suppose that for a variable W with corresponding potential-outcome vector $\mathbf{W}$, the observed value $W = \mathbf{R}'_W \mathbf{W}$ may be part of the index vectors for the potential outcomes of a subsequent variable.) We may then illustrate the system using a *directed acyclic graph* (DAG) in which arrows connect input variables to output variables [Pearl, 2009; Glymour and Greenland, 2008; Spirtes *et al.*, 2001]. Such graph is a *causal diagram* if (as here) the arrows are interpreted as links in causal chains.

Figure 1 provides an example, illustrating a system of structural equations

$$\begin{aligned} U &= f_U(\varepsilon_U), & V &= f_V(\varepsilon_V), & W &= f_W(u, v, \varepsilon_W), \\ X &= f_X(u, w, \varepsilon_X), & Y &= f_Y(v, w, x, \varepsilon_Y), & Z &= f_Z(w, \varepsilon_Z), \end{aligned}$$

where the inputs $\varepsilon_U, \varepsilon_V, \varepsilon_W, \varepsilon_X, \varepsilon_Y, \varepsilon_Z$ are “purely random disturbances,” that is, inputs that are independent random variables (but not necessarily identically distributed). Traditionally, such disturbances are left implicit, i.e., understood to be present but not shown.

The entire system may be viewed as a multivariate model for the graphed variables, with the graph encoding various constraints on the joint distribution of these variables [Lauritzen, 1996; Spirtes *et al.*, 2001; Pearl, 2009]. In particular, the distribution of the disturbances induces a joint distribution of the graphed variables which obeys the *Markov decomposition*. That is, the joint distribution of the graphed variables decomposes into factors, one for each graphed variable, that give the probability of each variable given its graphical/functional inputs (“parents”). For Figure 1 the decomposition is (in an abbreviated notation)

$$\Pr(u, v, w, x, y, z) = \Pr(u)\Pr(v)\Pr(w|u, v)\Pr(x|u, w)\Pr(y|v, w, x)\Pr(z|w).$$

Because U and V have no inputs within the system (they are “exogenous”), their factors $\Pr(u)$ and $\Pr(v)$ are unconditional.

The decomposition provides not only the original joint distribution, but also a formula for the effect on that distribution of shifting the functional relations or distributions of any subset of the variables. For example, *randomization* of W usually refers to an intervention that replaces $W = f_W(u, w, \varepsilon_W)$ by $W = f_W(\varepsilon_W^*)$ where the distribution of ε_W^* (and hence W) is determined by the investigator and hence is known. The resulting joint distribution of the variables is then

$$\Pr(u, v, w, x, y, z) = \Pr(u)\Pr(v)\Pr(w)\Pr(x|u, w)\Pr(y|v, w, x)\Pr(z|w),$$

the factor $\Pr(w|u, v)$ being replaced by the new randomization distribution for $W, \Pr(w)$. The corresponding graph lacks arrows into W .

Suppose that instead of randomizing W we intervene to force all values of W to a particular value w_0 , without altering any other functional relation (if this can be done). We then replace the equation $W = f_W(u, w, \varepsilon_W)$ by the equation $W = w_0$. The resulting joint distribution of the variables is then 0 except at $W = w_0$, where it is

$$\Pr(u, v, w_0, x, y, z) = \Pr(u)\Pr(v)\Pr(x|u, w_0)\Pr(y|v, w_0, x)\Pr(z|w_0),$$

the factor $\Pr(w|u, v)$ being replaced by the new, forced distribution $\Pr(w_0) = 1$. The corresponding graph lacks arrows into W , and has $W = w_0$ in place of W . Repeating this exercise for other values of W shows how the system responds to various interventions that fix or set W to particular values, again presuming this is can be done without disturbing other systemic relations [Pearl, 2009].

Some Useful Elements of Graph Theory

To describe further consequences of the Markov decomposition relevant for causal inference we need several concepts from graphical probability theory. Two variables in a graph are *adjacent* if they are connected by an arrow. Consider a

“path” in a graph, a nonrepeating sequence of variables in which successive sequence members are adjacent. Effects of one variable on another are transmitted by causal sequences or *causal pathways*, which in a causal diagram are *causal* or *directed* paths in which each arrow points to the tail of the next arrow in the path. More precisely, given a causal diagram, the existence of a directed path from one variable to another is a necessary but not sufficient condition for an effect to occur. Thus, in Figure 1, $U \rightarrow W \rightarrow Y$ means U *can* affect Y via its effect on W , but *not* that U does affect Y . The graph is *acyclic* if (as assumed here) no variable in the graph affects itself (meaning there is no feedback loop in the graph).

A central problem of causal inference is distinguishing causation from mere probabilistic dependence, or “association” as it is often called. Graph theory provides a quick visual distinction, via the following concepts. A variable is a *collider* on a path if it is at a meeting of two arrowheads along the path; otherwise it is a *noncollider* on the path. In Figure 1, W is a collider on the path UWV but a noncollider on the paths UWY and XWY . A path is said to be *closed* or *blocked* at a collider, and *open* or *unblocked* at a noncollider. The entire path is open if it has no collider, otherwise it is closed. Open paths include but are not limited to causal pathways, a fact which (as discussed below) reflects classic problems in causal inference.

It is often helpful to think of associations as signals flowing through the graph. Given a graph, associations can flow through or be transmitted by open paths. Open paths themselves are merely conduits for the transmission, however. More precisely, given a graph, the existence of an open path between two variables is a necessary but not sufficient condition for an association between them. (It should be noted however that the presence of an open path will in practice almost always correspond to the presence of an association, although more likely a miniscule one if the path is long [Greenland, 2003].)

Conversely, a sufficient (but not necessary) condition for two variables to be unassociated (independent) is that there is no open path between (that is, any path between them is closed). Thus, we can immediately spot the independencies that must hold in the graphed distribution by just seeing whether two variables have no open path between them. If a statistical test of these independencies rejects them (that is, detects associations where none should be, according to the graph), that result may be taken as evidence against the posited causal system that gave rise to the graph.

Biases and Confounding

Now suppose we are interested in an effect of one variable (the target antecedent) on another variable (the target outcome). Any open path between these target variables that is not part of this effect is a *biasing path*, because it provides a pathway for association between the target variables that is not due to the target effect.

To illustrate, suppose our interest is in the net (total) effect of the antecedent W

on the outcome Y in Figure 1. This effect equals the net association transmitted through all the causal paths from W to Y , which are WY and WXY . There are, however, two other open paths from W to Y : $WUXY$ and WVY . Thus, the association we observe will be the net transmission through all four paths, which may be far from the net transmission through the two target paths WY and WXY ; that is, the signal of interest may be seriously corrupted by unwanted signals through $WUXY$ and WVY . These unwanted signals (transmissions along biasing paths) are examples of *biases*, although the concept of “bias” subsumes other phenomena as well (such as distortions due to measurement error).

Unconditionally, a biasing path for a net effect in a DAG must pass through a shared (“common”) cause of the target variables. What is more, it must consist of two segments, one being a causal path from the shared cause to the target antecedent, and the other a causal path from the shared cause to the target outcome that does not include the target antecedent. For example, suppose again that our inferential target is the net effect of W on Y in Figure 1, which has biasing paths $WUXY$ and WVY . $WUXY$ can be decomposed into a causal path UW from U to W and a causal path UXY from U to Y , joined at the shared cause U of W and Y . Similarly, WVY can be decomposed into a causal path VW from V to W and a causal path VY from V to Y , joined at the shared cause V of W and Y .

Any bias that is transmitted via a common cause of the target variables is an example of confounding, in that it contributes to the association of the target antecedent variable with the potential target outcomes (that is, it contributes to nonignorability). More generally, confounding arises from association that is transmitted along biasing paths that terminate with an effect on the target outcome. Thus, as described earlier, confounding is association due to “extraneous” effects on the target outcome. Those effects are said to “confound” the target effect. Correspondingly, a *confounding path* is a biasing path that terminates with an arrow into the outcome of interest [Greenland and Pearl, 2007]. In Figure 1, both of the biasing paths for the effect of W on Y ($WUXY$ and WVY) are confounding paths, because both terminate with an arrow into (effect on) the target outcome Y .

If confounding occurs, variables within the responsible confounding paths are often called *confounders*. Because the entire confounding path is open, any confounder must be linked by open paths to both target variables, and must have associations with both target variables. The converse is not correct, however: A variable associated with both target variables need not be a confounder. For example, when examining the net effect of U on Y in Figure 1, X could be associated with U and Y but would not be a confounder because it does not lie on a confounding path.

Given a DAG with no conditioning, it can be shown that all biasing paths are confounding paths, and vice-versa. Conditioning, however, may open biasing paths, some of which may not be confounding paths. We thus now turn to the concept of conditioning in graphs.

Conditioning and Control

Let G and C be disjoint subsets of variables in the graph, with g and c being sets of values for G and C . Independencies in the conditional distribution $\Pr(g|c)$ implied by the graph may then be seen using just a few more concepts. One key notion is that the open/closed status of a variable along a path is reversed by conditioning (stratifying) on the variable: A collider becomes open and noncollider becomes closed. As a consequence, the status of paths may reverse. For example, if we condition on W in Figure 1, the closed path $XUWVY$ becomes open and now can transmit associations; it thus becomes a confounding path for the X effect on Y . At the same time, the open paths $XUWY$, $XWVY$, and XWY become closed and can no longer transmit associations, and thus are no longer confounding paths.

It should be noted that, in accord with ordinary language, experimental sciences use the term “control” to refer to a physical alteration of a system to remove sources of bias, such as randomization. In observational sciences, however, “control of a variable” is often used more broadly to include conditioning on a variable, whether it removes bias or creates bias. Thus, conditioning on W in Figure 1 will “control” (remove) any confounding of the X effect on Y that was present in the original system, but at the same time may introduce new confounding by opening paths that were previously closed.

Not all biasing paths opened by conditioning are confounding paths, however. For example, suppose our target effect is the net effect of U on V in Figure 1. Because there is no causal pathway from U to V , this effect is zero, or “null.” But conditioning on W will open the path UWV , allowing unwanted association to flow from U to V . In other words, conditioning on W transforms UWV into a biasing path which is not a confounding path, because it does not terminate with an arrow into V . Bias that results from such conditioning on a shared effect of the target variables is often called “Berksonian,” in honor of the discoverer of this type of bias, Joseph Berkson [Glymour and Greenland, 2008].

Conditioning can also produce bias due to closing target paths. For example, suppose our target effect is the net effect of U on Y in Figure 1 (associations transmitted via UXY , UWY , and $UWXY$). This target effect is equal to the unconditional association of U and Y , because there is no unconditional biasing path. Conditioning on W will open the path $UWVY$, allowing unwanted association to flow from U to Y . In other words, conditioning on W transforms $UWVY$ into a biasing path (which is a confounding path, since it terminates in an arrow into Y). But conditioning on W will also close the target paths UWY and $UWXY$, blocking part of the effect (signal) of interest. The association of U and Y conditional on W may thus bear little resemblance to the target effect.

Suppose instead that our interest is only in that part of the effect of U on Y not mediated by W (UXY in Figure 1). A standard strategy in the social-science literature is to then condition on W in order to block the effects mediated by W (UWY and $UWXY$ in Figure 1). Unfortunately, this literature almost always overlooks the fact that the same strategy can introduce bias via the pathways

opened by conditioning on the intermediate W ($UWVY$ in Figure 1)

Conditioning on effects of a variable can also partially reverse its status on paths. For example, conditioning on a variable affected by a collider on a path from X to Y can partially open the path and hence can result in new bias if the rest of the path is open after the conditioning. For example, if our target is the net effect of X on Y in Figure 1, conditioning on Z (affected by W) can partially close the confounding paths passing through W ($XUWY$, XWY , and $XWVY$) yet partially open the path $XUWVY$, which becomes a confounding path.

Collider Bias, Response Bias, and Selection Bias

Any biasing path opened by conditioning (whether full or partial) must pass through at least one collider; hence any bias that results from a newly opened path may be called *collider bias* [Greenland, 2003]. Of particular interest are those instances in which collider bias results from the process that determines how subjects come to be included in the statistical analysis of a target effect or association. The process is usually described in terms of subject response (to requests for participation) or subject selection (whether selection by the researcher or self-selection by the subject). Any bias that results from the process is thus often called “response bias” or “selection bias.” As mentioned above, “Berksonian bias” usually refers to situations in which both the target variables affect inclusion.

A more general and accurate term for all these biases is *inclusion bias*. To describe their shared structural and graphical representation, suppose Z is an indicator variable for inclusion in the analysis, with $Z = 1$ if a subject is included and $Z = 0$ if not. All associations observed must then be conditional on $Z = 1$. To say inclusion is random (“random sampling” for analysis) means that the structural equation for Z is $Z = \varepsilon_Z$, where ε_Z is a random indicator independent of all other random disturbances; Z will then have no causes in the corresponding graph (it will be exogenous in the graph). But if inclusion is affected by more than one causal pathway, Z will appear in the graph as a collider or as a variable affected by a collider, and observed associations may suffer considerable bias from the forced conditioning on $Z = 1$.

To illustrate, consider again Figure 1, with Z the inclusion indicator, and again with the target being the net effect of U on Y . Here, W may influence whether a subject is included or not. As a consequence, the forced conditioning on $Z = 1$ can partially open the path $UWVY$, which becomes a confounding path, and partially close the target paths UWY and $UWXY$. The net bias from these changes may be considerable, if Z is strongly affected by W , or minor if Z is only weakly affected by W .

DISCUSSION

The above models can be extended to deal with another major source of bias in observational research, measurement error. Because the topic is quite involved and brings in many elements not related to causality, it has not been included here. Notably however, measurement error expands the nonidentification problem to noncausal associations, thus further weakening identification of causal effects from observed associations; see [Greenland, 2005a; 2009; 2010] for examples and discussion.

The above models also extend to consideration of longitudinal (time-varying) treatments such as medical regimens, but again many technical elements arise in applications [Robins, 1999ab]. For a discussion of relations of the models to the sufficient-component cause model common in epidemiology (which is equivalent to the INUS model of Mackie [1965]) see [Greenland and Brumback, 2002; VanderWeele and Robins, 2008].

As may be apparent from the above presentation, statistical and structural representations of causation bypass most of the philosophic subtleties associated with the complex topic of cause and effect. This bypass has facilitated applications and may reflect the task-oriented attitude of most scientists. Nonetheless, it should not lead one to overlook some serious practical problems that are usually ignored.

Perhaps the largest problem is the possible ambiguity in what it means to intervene on a variable or to “change” its level. This ambiguity can render ambiguous the concept of a potential-outcome vector $\mathbf{Y}$ for X . After all, if Y_x is a counterfactual component of $\mathbf{Y}$ (as all but one component must be), its value may depend in a dramatic fashion on exactly how X would come to be x if x is counterfactual [Greenland, 2005b; Hernán, 2005]. In causal diagrams the same problem is utterly invisible. These are not fatal objections to the models, for the models have proven useful whenever the meaning of interventions and outcomes is unambiguous (for example when X is measles vaccination and Y is the subsequent occurrence of measles). But they are quite disconcerting when the models are used to make claims about the impact of (for example) “eradication of childhood disease”: The effect of such an ambiguous action depends dramatically on exactly how it is carried out (e.g., by vaccinating, by curing, or by killing children).

A related problem of less practical concern, but nonetheless disconcerting, is that potential-outcome models and their structural-equation generalizations seem to use an informal notion of causation to define actual effects. In particular, “setting” or forcing a treatment variable X to a particular level x (whether that is done in response to a random-number generator or in response to extraneous factors) is a causal command left undefined in the account. Here again the accompanying causal-diagram theory is silent, taking the causal interpretation of its arrows as a primitive.

Regardless of any objections and problems, the statistics and observational science literature employing potential-outcome models has exploded over the past few decades, while causal diagrams have spread rapidly in the wake. It thus seems

important that those interested in issues of causality become familiar with these formal yet practical tools for causal inference.

ACKNOWLEDGEMENTS

I am grateful to Katherine Hoggatt and a referee for helpful comments on the initial draft of this paper.

BIBLIOGRAPHY

- [Berk, 2004] R. A. Berk. *Regression analysis: a constructive critique*. Newbury Park, CA: Sage, 2004.
- [Box, 1980] G. E. P. Box. Sampling and Bayes inference in scientific modeling and robustness. *J R Stat Soc Ser A*; 143:383–430, 1980.
- [Cox and Hinkley, 1974] D. R. Cox and D. V. Hinkley. *Theoretical statistics*. New York: Chapman and Hall, 1974.
- [Dawid, 2000] A. P. Dawid. Causal inference without counterfactuals (with comments). *J Am Stat Assoc* 95: 407–428, 2000.
- [Fisher, 1932] R. A. Fisher. *Statistical methods for research workers*, 4th ed. London: 1932.
- [Fisher, 1935] R. A. Fisher. *The design of experiments*. Oliver and Boyd, Edinburgh, 1935.
- [Glymour and Greenland, 2008] M. M. Glymour and S. Greenland. Causal diagrams. In K. J. Rothman, S. Greenland and T. L. Lash. *Modern Epidemiology*, 3rd edition. Philadelphia, PA: Lippincott Williams & Wilkins, 2008.
- [Greenland, 1991] S. Greenland. On the logical justification of conditional tests for two-by-two contingency tables. *Am Statist*, 45:248–251, 1991.
- [Greenland, 2003] S. Greenland. Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, 14:300–306, 2003.
- [Greenland, 2004] S. Greenland. An overview of methods for causal inference from observational studies. In A. Gelman and X. L. Meng, eds. *Applied Bayesian modeling and causal inference from an incomplete-data perspective*. New York: Wiley, 2004.
- [Greenland, 2005a] S. Greenland. Multiple-bias modeling for analysis of observational data (with discussion). *J R Stat Soc series A*, 168:267–308, 2005a.
- [Greenland, 2005b] S. Greenland. Epidemiologic measures and policy formulation: Lessons from potential outcomes (with discussion). *Emerg Themes Epidemiol*, 2:1–4, 2005b. (Originally published as “Causality theory for policy uses of epidemiologic measures”. Chapter 6.2 in: C. J. L. Murray, J. A. Salomon, C. D. Mathers, and A. D. Lopez, eds. *Summary Measures of Population Health*. Cambridge, MA: Harvard University Press/WHO, 291–302.)
- [Greenland, 2009] S. Greenland. Relaxation penalties and priors for plausible modeling of non-identified bias sources. *Statistical Science*, 24:195–210, 2009.
- [Greenland, 2010] S. Greenland. Overthrowing the tyranny of null hypotheses hidden in causal diagrams. Ch. 22 in R. Dechter, H. Geffner, and J. Y. Halpern, eds. *Heuristics, Probabilities, and Causality: A Tribute to Judea Pearl*, pp. 365–382. London: College Publications, 2010.
- [Greenland and Brumback, 2002] S. Greenland and B. A. Brumback. An overview of relations among causal modeling methods. *Int J Epidemiol*, 31:1030–1037, 2002.
- [Greenland and Pearl, 2007] S. Greenland and J. Pearl. Causal diagrams. In S. Boslaugh, ed. *Encyclopedia of epidemiology*. Thousand Oaks, CA: Sage Publications, 2007: 149–156.
- [Greenland et al., 1999] S. Greenland, J. M. Robins and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14:29–46, 1999.
- [Hernán, 2005] M. A. Hernán. Hypothetical interventions to define causal effects—afterthought or prerequisite? *Am J Epidemiol*, 162:618–620, 2005.
- [Holland, 1986] P. W. Holland. Statistics and causal inference (with discussion). *J Am Stat Assoc*, 81:945–970, 1986.
- [Hume, 1978] D. Hume. *A treatise of human nature*. Oxford: Oxford University Press, 1888; 2nd ed, 1978. (Original publication, 1739.)

- [Hume, 1988] D. Hume. *An Enquiry Concerning Human Understanding*. Open Court Publishing Company, Chicago, 1988, p. 115. (Original publication 1748.)
- [Kang and Schafer, 2007] J. D. Y. Kang and J. L. Schafer. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:477-580, 2007.
- [Kuhn, 1970a] T. S. Kuhn. *Reflections on my critics*. In: Lakatos I, Musgrave A, eds. *Criticism and the growth of knowledge*. Cambridge: Cambridge University Press, 1970a.
- [Kuhn, 1970b] T. S. Kuhn. *The structure of scientific revolutions*, 2nd ed. Chicago: University of Chicago Press, 1970b, Chapter XIII.
- [Lauritzen, 1996] S. Lauritzen. *Graphical Models*. New York: Oxford, 1996.
- [Lewis, 1973a] D. Lewis. Causation. *J Philos*, 70:556-567, 1973a. (Reprinted with postscript in: Lewis D. Philosophical papers. New York: Oxford University Press, 1986.)
- [Lewis, 1973b] D. Lewis. *Counterfactuals*. Blackwell, Oxford, 1973b.
- [Little, 1989] R. J. A. Little. On testing equality of two independent binomial proportions. *Am Statist*, 43:283-288, 1989.
- [Mackie, 1965] J. L. Mackie. Causes and conditions. *Am Philos Q*, 2:245-255, 1965. Reprinted in Sosa E, Tooley M, eds. *Causation*. New York: Oxford, 1993, 33-55.
- [Mill, 1956] J. S. Mill. *A System of Logic, Ratiocinative and Inductive*. Longmans Green, London 1956, Chapter X. (Original publication 1843)
- [Morgan and Winship, 2007] S. L. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press, 2007.
- [Neyman, 1990] J. Neyman. On the application of probability theory to agricultural experiments: Essay on principles, Section 9. 1923; Partial translation from the original French in *Statistical Science*, 5, 465-480, 1990.
- [Pearl, 2009] J. Pearl. *Causality*, 2nd ed. New York: Cambridge, 2009.
- [Robins, 1999a] J. M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, M.E. Halloran and D. Berry. eds., IMA Volume 116, pp 95-134. New York: Springer-Verlag, 1999a.
- [Robins, 1999b] J. M. Robins. Association, causation, and marginal structural models. *Synthese*, 121: 151-179, 1999b.
- [Rubin, 1923] D. B. Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5:472-480, 1990.
- [Rubin, 1991] D. B. Rubin. Practical implications of modes of statistical inference for causal effects, and the critical role of the assignment mechanism. *Biometrics*, 47:1213-1234, 1991.
- [Spirtes et al., 2001] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. Cambridge, MA, MIT Press, 2001.
- [VanderWeele and Robins, 2008] T. J. VanderWeele and J. M. Robins. Empirical and counterfactual conditions for sufficient cause interactions. *Biometrika*, 95:49-61, 2008.

Part X

Some Philosophical Issues
Concerning Statistical
Learning Theory

This page intentionally left blank

STATISTICAL LEARNING THEORY AS A FRAMEWORK FOR THE PHILOSOPHY OF INDUCTION

Gilbert Harman and Sanjeev Kulkarni

Statistical Learning Theory (e.g., [Hastie *et al.*, 2001; Vapnik, 1998; 2000; 2006; Devroye *et al.*, 1996]) is the basic theory behind contemporary machine learning and pattern recognition. We suggest that the theory provides an excellent framework for the philosophy of induction (see also [Harman and Kulkarni, 2007]).

Inductive reasons are often compared with deductive reasons. Deductive reasons for a conclusion guarantee the conclusion in the sense that the truth of the reasons guarantees the truth of the conclusion. Not so for inductive reasons, which typically do not provide the same sort of guarantee. One part of the philosophy of induction is concerned with saying what guarantees there are for various inductive methods.

There are various paradigmatic approaches to specifying the problem of induction.

For example, Reichenbach [1949] argued, roughly, that induction works in the long run if anything works in the long run. His proposal has been followed up in interesting ways in the learning in the limit literature (e.g. [Putnam, 1963; Osherson *et al.*, 1982; Kelly, 1996; Schulte, 2002]). The paradigm here is to envision a potentially infinite data stream of labeled items, a question Q about that stream, and a method M that proposes an answer to Q given each finite initial sequence of the data. If the data sequence consists of a series of letters of the alphabet, one question might be whether every “A” in the sequence is followed by a “B”; then the issue is whether there is a method for answering that question after each datum, a method that will eventually give the correct answer from that point on.

A second paradigm assumes one has an initial *known subjective* probability distribution satisfying certain more or less weak conditions along with a method for updating one’s probabilities, e.g. by conditionalization, and proves theorems about the results of such a method [Savage, 1954; Jeffrey, 2004].

Statistical learning theory, which is our topic, represents a third paradigm which assumes there is an *unknown objective* probability distribution that characterizes the data and the new cases about which inferences are to be made, the goal being to do as well as possible in characterizing the new cases in terms of that unknown objective probability distribution. The basic theory attempts to specify what can be proved about various methods for using data to reach conclusions about new cases.

We will be concerned with what we take to be basic statistical learning theory, which is concerned with what can be proved about various inductive methods, given minimal assumptions about the background probability distribution. We are interested in results that hold no matter what that background probability distribution is (as long as the minimal assumptions are satisfied). In other words, we are interested in “worst case” results: even in the worst case, such and such holds.

We begin by sketching certain aspects of basic statistical learning theory, then comment briefly on philosophical implications for reliability theories of justification, Popper’s [1979; 2002] appeal to falsifiability, simplicity as relevant to inductive inference, and whether inductive inference from observed cases to a conclusion about a new case presupposes an inference to a generalization covering old and new cases.

PATTERN RECOGNITION

One basic problem discussed in statistical learning theory is the *pattern recognition problem*: “How can data be used to find good rules for classifying new cases on the basis of the values of certain features of those cases?” As we have indicated, the simplest version of the problem presupposes that there is an unknown statistical probability distribution that specifies probabilistic relations between feature values of each possible case and its correct classification and also specifies how likely various cases are to come up either as data or as new cases to be classified. In this simplest version, these probabilities are assumed to be identically distributed and are independent of each other, but no other assumption about the probability distribution is made.

For example, the Post Office wants to use machines to sort envelopes on the basis of handwritten zip codes. The data are samples of actual handwritten zip codes as they have appeared on envelopes that have been received by the Post Office, samples that have been classified by human operators as representing one or another zip code (or as not representing a zip code). In a standard version of the problem the handwritten cases are digitized and presented as an N by M grid of light intensities, so there are $N \times M$ features, the value of each of which is specified as an intensity of light in the corresponding pixel of the grid. A rule for classifying new cases maps each possible value of the $N \times M$ features into a particular zip code (or into a decision that the features shown do not represent a zip code).

The feature values of a possible case can be represented as a vector $\bar{x} = (x_1, x_2, \dots, x_D)$ or equivalently as a point in a D -dimensional feature space whose coordinates are specified by $\bar{x}$. The i th co-ordinate of a point in the feature space corresponds to the value of the i th feature for an item represented by that point. A pattern recognition rule can be thought of as a rule for assigning labels to each point in the feature space, where the label assigned to the point represents the rule’s verdict for a case with those features.

In order to evaluate pattern recognition rules, an assumption is needed about the value of getting the right answer and the cost of getting the wrong answer. In many applications, the value of any right answer is taken to be zero, and the cost of any wrong answer is set at one. In other cases, different values and costs might be assigned. (For example, in some medical diagnostic contexts, a false negative verdict might be assigned a greater cost than a false positive verdict.)

Pattern recognition rules are then better or worse depending on the expected value of using them, where expectations are determined by the values and costs of right and wrong answers, the probabilities of the various cases, and the probabilities that the answer given by the rules are correct for each of those cases.

If the value of right answers is set at zero, the best pattern recognition rules have the least expected costs. If the value of right answers is set at zero and the cost of wrong answers is set at 1, then the expected cost corresponds to the probability of error. In this case, the best pattern recognition rules have the smallest probability of error.

The Post Office may want to assign different values and costs to correct or incorrect decisions about various cases, but on the other hand may be satisfied (at least at the beginning) to try to find rules with the least probability of error.

The best pattern recognition rules for a given problem are those rules with the least expected cost. Any rule of this sort is called a *Bayes rule*. There is always at least one such rule. (There is more than one Bayes rule only if for at least one possible case the expected value of two different decisions about the case are tied.)

Without loss of generality, we might assume that we are only concerned with the probability of error of a rule. So, in the rest of our discussion we suppose we are interested in minimizing the probability of error of decisions about new cases.

If one is concerned with *yes/no* issues, so that each item is either an instance of a certain category or not, we can represent a *yes* classification as 1 and a *no* classification as 0. In this case, a pattern recognition rule is equivalent to a specification of the set of points in the feature space that the rule classifies as *yes* or 1. In the post office example, one might be concerned with whether a given written figure should be classified as a “9” (yes=1) or not (no=0).

BAYES ERROR RATE R^*

To review: We suppose each item has D features $x_1, x_2, \dots, x_D$, where each feature x_i takes real values. (In the Post Office example, there are $N \times M$ features, each of which might take any of as many as 256 intensity values.) We represent the values of the features of a given item with the feature vector $\bar{x} = (x_1, x_2, \dots, x_D)$. We can also represent the values of the features of an item as a point in a D -dimensional *feature space*.

In our simplified case, a pattern recognition rule maps each feature vector $\bar{x}$ into a *yes/no* decision. We use 1 for *yes* and 0 for *no*. The rule maps points in the feature space into 1 and 0. We can specify the rule by specifying the set of points that it maps into 1.

We suppose there is a statistical probability distribution P specifying the actual statistical probability relations among feature values and the correct classification of items whose features have those values and also specifying how likely items with those feature values are to come up.

A rule with minimal probability of error, i.e. a Bayes rule, maps $\bar{x}$ into 1, if $P(1|\bar{x}) > P(0|\bar{x})$, and maps $\bar{x}$ into 0, if $P(1|\bar{x}) < P(0|\bar{x})$. (It does not matter what a Bayes rule maps $\bar{x}$ into when these conditional probabilities are equal.)

Applying this rule to a fixed observed $\bar{x}$, the probability of error is the smaller of $P(0|\bar{x})$ and $P(1|\bar{x})$, i.e., $\min\{P(0|\bar{x}), P(1|\bar{x})\}$. Therefore, the overall probability of error of a Bayes decision rule, the *Bayes error rate*, denoted by R^* is

$$R^* = \sum_{\bar{x}} P(\bar{x}) \min\{P(0|\bar{x}), P(1|\bar{x})\}$$

In cases involving probability densities, as specified by a probability density function p , the Bayes error rate is given by

$$R^* = \int \min\{P(0|\bar{x}), P(1|\bar{x})\} p(\bar{x}) d\bar{x}$$

USING DATA TO LEARN THE STATISTICAL PROBABILITY DISTRIBUTION?

The best pattern recognition rule is determined by the statistical probability distribution P . We assume that this distribution is initially unknown. In this case then, we need to use data in order to come up with a good rule.

Suppose that various cases arise with known values of various observable features and that an “expert” tells us the correct classification of these cases. We want to be able to use the data, once we have enough of it, to find a rule that will do well on new cases (as assessed by the same “expert”) given their feature values with performance close to that of the Bayes rule.

An idea that doesn’t usually work is to use such data to learn the probability distribution P , which is then used to find the Bayes rule. The idea would be to appeal to the probabilistic “law of large numbers,” from which we can infer that in the long run the observed frequency with which a feature vector $\bar{x}$ is classified as 1 will converge to the statistical probability of its being classified as 1. In other words, with probability approaching 1, for any ϵ however small, there will be a t such that after t occurrences of the feature vector $\bar{x}$, the frequency with which the label associated with $\bar{x}$ is 1 will be within ϵ of the statistical probability of the label’s being 1.

If there are finitely many points in the feature space, each representing an event with the features $\bar{x}$, then we might consider labeled items as they come to our attention and note the frequencies with which each given $\bar{x}$ is a 1. In each case,

the frequency that a given $\bar{x}$ is found to be a 1 will converge to its statistical probability of being 1. Because there are only finitely many such events, there will be *uniform convergence* of these frequencies. In other words, with probability approaching 1, for any ϵ however small, there will be a t such that after t items of data, the frequency of each $\bar{x}$ being 1 will be within ϵ of its probability of being 1.

Unfortunately, this can take a very long time. For example, in Post Office case, suppose there the grid of pixels were only 10 by 100. Then if there were 256 possible intensity values, there would be 256^{1000} different feature vectors. Even if there were only 2 possible intensity values, *on* and *off*, there would still be $2^{1000} > 10^{300}$ different feature vectors. Given that the current age of the universe is less than 10^{20} seconds, there would not be time for even a tiny fraction of those feature vectors to show up in anyone's lifetime!

We need a method that does not require having to learn the whole probability distribution. Statistical learning theory is the theory of such methods.

EMPIRICAL RISK MINIMIZATION

What is a good way to choose a rule of pattern recognition that will maximize expected value, or in our special case, to minimize probability of error? One initially appealing idea might be simply to choose some rule or other that has the least error on cases that are included in the data. But too many rules have that property. For any given rule, there will be other rules making the same decisions on cases in the data but making all possible decisions on cases not part of the data.

Any workable inductive procedure must therefore have some sort of *inductive bias*. It must favor certain rules of pattern recognition over others that make exactly the same decisions for cases in the data.

The crudest inductive bias simply restricts the relevant pattern recognition rules to a selected class of rules C . Then a policy of *empirical risk minimization*, given some data, selects a rule from C with minimal cost on the data. But of course not all choices of the class of rules C are equally good. As we have just observed, if C contains all possible rules, a policy of enumerative induction or empirical risk minimization fails to give any advice at all about new cases (or amounts to choosing classifications of new cases at random).

On the other hand, if C does not contain all rules, it may fail to contain a Bayes rule—a best rule of pattern classification in this instance. Indeed, it may fail to contain anything whose performance is close to a Bayes rule. It might even contain only the worst possible rules!

But let us put this last problem to the side for a moment. We will come back to it shortly.

Consider this question: what has to be true of the set C of rules so that, no matter what the unknown background probability distribution, empirical risk minimization eventually does as well as possible with respect to the rules in C ? More precisely, what has to be true of the set C in order to guarantee that, with

probability approaching 1, *no matter what the unknown probability distribution*, given more and more data, the probability of error for the rules that empirical risk minimization endorses at each stage eventually approaches the minimum value of probability of error of rules in C ?

A fundamental result of statistical learning theory is that the set of rules in C cannot be too rich, where the richness of C is measured by its *VC-dimension*. Let us explain.

Shattering and VC-dimension

Recall that we are restricting attention to rules that map a set of D feature values into a *yes/no* verdict for a particular categorization, where we use 1 for *yes* and 0 for *no*. As we have mentioned, we can represent such a rule in terms of a D dimensional *feature space* in which each point is labeled 1 or 0. Each point in that space represents a possible set of features. The i th co-ordinate of a point in the feature space corresponds to the value of the i th feature for that point. The label attached to the point represents that rule's verdict for a case with those features.

Now consider a set of N points in a given feature space and consider the 2^N possible ways to label those points as *yes* or *no*. If for each possible way of labeling those points there is a rule in C that agrees with that labeling, we say that the rules in C *shatter* those N points.

To say that the rules in C shatter a particular set of points is to say that no possible assignment of verdicts to those points (those possible cases) "falsifies" the claim that some rule in C correctly represents those verdicts, to use terminology from Popper [1979; 2002].

The finite VC-dimension of a set of rules C is the largest finite number N for which some set of N points is shattered by rules in C . If C has no finite VC-dimension, its VC-dimension is infinite.

Fundamental Result

Recall that empirical risk minimization says to select a rule in C with least error on the data. Vapnik and Chervonenkis [1968] show that empirical risk minimization works, no matter what the unknown statistical probability distribution is, if, and only if, C has finite VC dimension. More precisely (subject to mild measurability conditions):

If and only if C has finite VC dimension: with probability approaching 1, then, no matter what the unknown probability distribution, given more and more data, the probability of error for the rules that empirical risk minimization endorses at each stage eventually approaches the minimum value of probability of error of rules in C .

The Vapnik Chervonenkis result also provides information about how much data are needed for empirical risk minimization to produce a good result no matter what

the unknown statistical probability distribution is. If the rules in C have VC dimension V , there will be a function $m(V, \epsilon, \delta)$ indicating the maximum amount of data needed (no matter what the unknown statistical probability distribution) to ensure that the probability is less than δ that enumerative induction will endorse a hypothesis with an probability of error rate that exceeds the minimum probability of error rate for rules in C by more than ϵ . (A smaller ϵ indicates a better approximation to the minimum error error for rules in C and a smaller δ indicates a higher probability that the rules endorsed will be within the desired approximation to that minimum probability of error.)

Where there is such a function $m(V, \epsilon, \delta)$ there is what has come to be called “Probably Approximately Correct” (or “PAC”) learning (terminology due to [Valiant, 1984]).

Example: Perceptron Learning

A *perceptron* is a simple classifier that takes the weighted sum of the D input feature values (along with an additional constant input value) and outputs +1 for *yes* if the result of the weighted sum is greater than some threshold T and outputs 0 for *no* otherwise. Given data, it is easy to find a threshold and weights for such a perceptron that yield the least error (or cost) on that data.

Any classification rule that can be represented by such a perceptron divides the D dimensional feature space into two regions, the *yes* region and the *no* region, where the regions are separated by a line, plane, or hyperplane. In other words, a perceptron classifier *linearly separates* the points in the feature space. The class C in this case is the class of linear separation rules in that feature space.

The VC-dimension of such linear separations of a D dimensional feature space is $D + 1$, which is finite, so the the result mentioned in the previous subsection applies. We can know how many items of data are needed in order probably to approximate the best such separation.

Here we can return to the worry we temporarily put aside, because it applies to perceptron learning. The worry is that the class C of rules used for empirical risk minimization may fail to include the best rule, the Bayes rule and may indeed contain no rule whose performance is even close to that of the Bayes rule.

Obviously, many possible classification rules are not and cannot be approximated by linear separation rules. The XOR rule is a well known example. Suppose there are two features, x_1 and x_2 , each of which takes a real value between -1 and $+1$, and the correct classification rule, which is also the Bayes rule, takes the value 1 if and only if $x_1 \times x_2 < 0$. Here there is a two dimensional feature space in which the points in the upper left quadrant and the lower right quadrant are to be labeled 1 and the points in the upper right quadrant and lower left quadrant are to be labeled 0. Clearly there is no way to draw a line separating the points to be labeled 1 from those to be labeled 0. Indeed, if the statistical probability density is evenly distributed over the feature space, any linear separation has a significant probability of error while the Bayes rule has a probability of error of 0!

Example: Feed-forward Neural Network Learning

This last worry can be somewhat alleviated by using a feed-forward neural network with several layers of perceptrons. Inputs go to perceptrons in the first layer whose outputs are inputs to perceptrons in the second layer, and so on to a final perceptron that outputs 0 or 1 depending on whether the weighted sum of its inputs are above or below a certain threshold. A fixed network with fixed connection strengths between nodes and fixed thresholds can be used to classify inputs as 0 or 1, so such a network therefore represents a classification rule. As before, varying connection strengths and thresholds (while retaining all connections) yields other rules.

There are more or less good methods for finding a rule represented by a given structure of perceptrons that has least error on given data. These learning methods typically involve smoothing the threshold functions of all but the final perceptron and use a kind of “gradient descent” that may or may not become stuck in a “local minimum” that is not a global minimum. In any case there will be one or more setting of connection strengths and thresholds that minimizes error on the data.

It can be shown that, the set of all rules represented by possible ways of varying connection strengths and thresholds for any particular feed forward network has a finite VC-dimension. So, the PAC learning criterion is satisfied. Furthermore, any (nonpathological) classification rule can be approximated arbitrarily closely by feed forward neural networks with enough layers by adding more nodes. So it is possible to ensure that the error rate of the best rule that can be represented by such a network is not far from the error rate of the best possible rule, the Bayes rule.

Of course, adding nodes increases the VC dimension of such a network, which means more data will be needed to guarantee a certain level of performance under the PAC criterion. So, there is a trade-off between the amount of data needed to satisfy the PAC criterion and how closely the error rate of the best rule in C is to the Bayes error rate.

Furthermore, no matter how large the VC dimension of C , as long as it is finite, there is no guarantee that, no matter what the background statistical probability distribution, the error rate of the rule selected via empirical risk minimization will converge to the Bayes error.

DATA COVERAGE BALANCED AGAINST SOMETHING ELSE

Instead of using pure empirical risk minimization, an alternative learning procedure balances empirical error or cost on the data against something else—often called *simplicity*, although (as we will see) that is not always a good name for the relevant consideration—and then allows C to have infinite VC dimension.

There are versions of this strategy which, with probability approaching 1, will eventually come up with rules from C whose probability of error approaches that of the Bayes rule in the limit. However, because the rules in C have infinite VC

dimension, the PAC result does not hold. So “eventually” might be a very long time.

One version of this alternative strategy is concerned with rules that can be specified in a particular notation. Each rule in C is assigned a number, its length as measured by the number of symbols used in its shortest description in that notation. Given data, the procedure is to select a rule for which (for example) the sum of its empirical error (or cost) on the data plus its length is minimal.

Another version identifies C with the union of a nested series of classes each of finite VC-dimension $C_1 \subset C_2 \subset \dots \subset C_n \subset \dots$, where the VC-dimension of C_i is less than the VC-dimension of C_{i+1} . Given data, the procedure is then to select a rule that minimizes the sum of its empirical error (or cost) on the data plus the number i of the smallest class C_i to which the rule belongs. This version is called *structural risk minimization*.

These two kinds of ordering can be quite different. If rules are ordered by description length, linear separations will be scattered throughout the ordering, because the length of a description of a linear rule will depend on the number of symbols needed to specify various constants in those rules. In the ordering of classes of rules by VC-dimension, linear separations can be put ahead of quadratic separations, for example, because linear separations in a given space have a lower VC-dimension than quadratic separations in the same space.

We return later to the question of whether any approach of this sort is best described as balancing *simplicity* against data-coverage.

Example: Support Vector Machines

The final perceptron in a feed forward neural network makes a linear decision in the space represented by its inputs. The earlier parts of the feed forward network can therefore be thought of as mapping the feature space into the space for which the final perceptron makes a linear decision.

This suggests a more general strategy for solving pattern recognition problems. Map the feature space into another space—perhaps a space with many more dimensions—and then make a linear decision in that space that minimizes error on the data. It is true that the more dimensions to the other space, the higher the VC-dimension of the rules represented by linear separations in that space. However, *support vector machines* get around this problem by using *wide margin separations* instead of just any linear separation.

Linear separations are hyperplanes that have no thickness. Wide margin separations are thick hyperplanes, hyperslabs. Vapnik [2006] observes that, if the relevant points in a space are confined to a hypersphere of a given size, then the VC-dimension of wide margin separations of those points are often much less than the VC-dimension of pure linear separations.

Even if the space in question has infinitely many dimensions, the VC-dimension of wide-margin separations of points in that hypersphere is finite and inversely related to the size of the margin, the thickness of the hyperslabs.

Support vector machines first map the feature space to a very large or infinite dimensional space in which images of points in the feature space are confined to a hyper volume of fixed radius. Then a wide margin separation of points in the transformed space is selected by trading off empirical error on the data against the width of margins of the best wide margin separations.

Transduction

The learning methods discussed so far use labeled data to find a rule that is then used to classify new cases as they arise. Furthermore, these methods all involve learning total classifications. Nearest neighbor methods, perceptrons, multilayer feed-forward networks, and standard support vector machines all yield rules that assign a classification to every possible set of features.

We could modify some of these methods to provide only partial classifications. For example, we could modify support vector machines so as not to choose among the various separating hyperplanes internal to the selected separating hyperslab. The points in this between space would be left unclassified. The system would still be an inductive method, since it would classify some, perhaps many, new cases in accordance with a rule derived from labeled data, but the rule would not be a total rule, since it would not characterize all points in the space.

Suppose we are using a method that in this way provides only a partial classification of cases and a case arises to be classified in the intervening space of previously unclassified cases. Vapnik [1998; 2000; 2006] considers certain *transductive* methods for classifying such new cases, methods that use information about what new cases have come up to be classified and then select a subset of separations that (a) correctly classify the data and (b) agree on their classifications of the new cases. In one version, the selected separations also (c) disagree as much as possible on the classifications of other possible cases.

An important related version of transduction uses not only the information that certain new cases have come up to be classified but also the information that there is a certain set U (“universum”) of examples that are hard to classify. In this version, transduction selects the subset of linear separations satisfying (a) and (b) but disagreeing as much as possible on the classification of the hard cases in U . Transduction performs considerably better than other methods in certain difficult real-life situations involving high-dimensional feature spaces where there is relatively little data.

PHILOSOPHICAL IMPLICATIONS

We conclude by briefly mentioning a few ways in which statistical learning theory has philosophical implications. It is relevant to reliability theories of epistemic justification, it helps to situate some of Popper’s often misunderstood appeals to falsifiability, it casts doubt on claims about the importance of simplicity consid-

erations in inductive reasoning, and it illuminates discussion of direct inductive inference.

Reliability

Many philosophical epistemologists (e.g., [Goldman, 1986; 2002; Bishop and Trout, 2005]) argue that epistemology should be concerned with the *reliability* of methods of belief formation (or more generally of belief revision). We believe that the relevant notion of reliability makes sense only in relation to an appeal to something like the sort of background statistical probability distribution that figures in the pattern recognition problem studied in statistical learning theory. For example, the statistical likelihood that a method will yield true answers might provide a measure of the reliability of the method. If so, statistical learning theory provides one sort of foundation for philosophical epistemology.

VC Dimension and Popperian Falsifiability

There is an interesting relation between the role of VC dimension in the PAC result and the emphasis on falsifiability in Karl Popper's writings in the philosophy of science. Popper [1934] famously argues that the difference between scientific hypotheses and metaphysical hypotheses is that scientific hypotheses are "falsifiable" in a way that metaphysical hypotheses are not. To say that a certain hypothesis is falsifiable is to say that there is possible evidence that would not count as consistent with the hypothesis.

According to Popper, evidence cannot establish a scientific hypothesis, it can only "falsify" it. A scientific hypothesis is therefore a falsifiable *conjecture*. A useful scientific hypothesis is a falsifiable hypothesis that has withstood empirical testing.

Recall that enumerative induction requires a choice of a set of rules C . That choice involves a "conjecture" that the relevant rules are the rules in C . If this conjecture is to count as scientific rather than metaphysical, according to Popper, the class of rules C must be appropriately "falsifiable."

Many discussions of Popper treat his notion of falsifiability as an all or nothing matter, not a matter of degree. But in fact Popper does allow for degrees of difficulty of falsifiability [2002, sections 31-40]. For example, he asserts that a linear hypothesis is more falsifiable — easier to falsify — than a quadratic hypothesis. This fits with VC theory, because the collection of linear classification rules has a lower VC dimension than the collection of quadratic classification rules.

However, Popper's measure of degree of difficulty of falsifiability of a class of hypotheses does not quite correspond to VC-dimension (Corfield et al, 2005). Where the VC-dimension of a class C of hypotheses is the largest number N such that *some* set of N points is shattered by rules in C , what we might call the "Popper dimension" of the difficulty of falsifiability of a class is the largest number N such that *every* set of N points is shattered by rules in C . This difference between *some*

and *every* is important and VC-dimension turns out to be the key notion rather than Popper-dimension.

Popper also assumes that the falsifiability of a class of hypotheses is a function of the number of parameters used to pick out instances of the class. This turns out not to be correct either for Popper dimension or VC dimension, as discussed below.

This suggests that Popper's appeal to degree of falsifiability would be improved by adopting VC-dimension as the relevant measure in place of his own measure.

Simplicity

We now want to say something more about Popper's [1972; 2002] discussion of scientific method. Popper argues that there is no justification for any sort of inductive reasoning, but he does think there are justified scientific methods.

In particular, he argues that a version of structural risk minimization best captures actual scientific method (although of course he does not use the term "structural risk minimization"). In his view, scientists accept a certain ordering of classes of hypotheses, an ordering based on the number of *parameters* needing to be specified to be able to pick out a particular member of the class. So, for example, for real value estimation on the basis of one feature, linear hypotheses of the form $y = ax + b$ have two parameters, a and b , quadratic hypotheses of the form $y = ax^2 + bx + c$ have three parameters, a , b , and c , and so forth. So, linear hypotheses are ordered before quadratic hypotheses, and so forth.

Popper takes this ordering to be based on "falsifiability" in the sense at least three data points are needed to "falsify" a claim that the relevant function is linear, at least four are needed to "falsify" the claim that the relevant function is quadratic, and so forth.

In Popper's somewhat misleading terminology, data "falsify" a hypothesis by being inconsistent with it, so that the hypothesis has positive empirical error on the data. He recognizes, however, that actual data do not show that a hypothesis is false, because the data themselves might be noisy and so not strictly speaking correct.

Popper takes the ordering of classes of hypotheses in terms of parameters to be an ordering in terms of "simplicity" in one important sense of that term. So, he takes it that scientists balance data-coverage against simplicity, where simplicity is measured by "falsifiability" [Popper, 2002, section 43].

We can distinguish several claims here.

- (1) Hypothesis choice requires an ordering of nested classes of hypotheses.
- (2) This ordering represents the degree of "falsifiability" of a given class of hypotheses.
- (3) Classes are ordered in accordance with the number of parameters whose values need to be specified in order to pick out specific hypotheses.

- (4) The ordering ranks *simpler* hypotheses before more *complex* hypotheses.

Claim (1) is also part of structural risk minimization. Claim (2) is similar to the appeal to VC dimension in structural risk minimization, except that Popper's degree of falsifiability does not coincide with VC dimension, as noted in above. As we will see in a moment, claim (3) is inadequate and, interpreted as Popper interprets it, it is incompatible with (2) and with structural risk minimization. Claim (4) is at best terminological and may just be wrong.

Claim (3) is inadequate because there can be many ways to specify the same class of hypotheses, using different numbers of parameters. For example, linear hypotheses in the plane might be represented as instances of $abx + cd$, with four parameters instead of two. Alternatively, notice that it is possible to code a pair of real numbers a, b as a single real number c , so that a and b can be recovered from c . That is, there are functions such that $f(a, b) = c$, where $f_1(c) = a$ and $f_2(c) = b$.¹ Given such a coding, we can represent linear hypotheses as $f_1(c)x + f_2(c)$ using only the one parameter c . In fact, for any class of hypotheses that can be represented using P parameters, there is another way to represent the same class of hypotheses using only 1 parameter.

Perhaps Popper means claim (3) to apply to some ordinary or preferred way of representing classes in terms of parameters, so that the representations using the above coding functions do not count. But even if we use ordinary representations, claim (3) conflicts with claim (2) and with structural risk minimization.

To see this, consider the class of sine curves $y = a + \sin(bx)$ that might be used to separate points in a one dimensional feature space, represented by the points on a line between 0 and 1. Almost any n distinct points in this line segment are shattered by curves from that class. So this class of sine curves has infinite "falsifiability" in Popper's sense (and infinite VC-dimension) even though only two parameters have to be specified to determine a particular member of the set, using the sort of representation Popper envisioned. Popper himself did not realize this and explicitly treats the class of sine curves as relatively simple in the relevant respect [1934, Section 44].

The fact that this class of sine curves has infinite VC dimension (as well as infinite falsifiability in Popper's sense) is some evidence that the relevant ordering of hypotheses for scientific hypothesis acceptance is not a simplicity ordering, at least if sine curves count as "simple".

Transduction as Direct Induction

Vapnik [2000, p. 293] says that transduction does not involve first inferring an *inductive generalization* which is then used for classification. Harman [1965; 1967] argues that any such inference should always be treated as a special case of *inference to the best explanation* where the relevant sort of explanation appeals to a generalization. But the apparent conflict here appears to be merely terminological.

¹For example, f might take the decimal representations of a and b and interleave them to get c .

Transduction differs from the other inductive methods we have been discussing in this way: the classification of new cases is not always based on an inductive generalization from labeled data alone. That is because transduction also makes use of the information that certain new cases have come up to be assessed.

On the other hand, transduction does involve the implicit acceptance of a generalization G , corresponding to the selected subset of separations in the transformed higher dimensional space. So, transduction does involve inductive generalization, even if not inductive generalization from the labeled data alone, since it makes use of the extra information that certain new cases have come up to be assessed.

It is true that, although the data include what new cases have come up, the classifications that transduction gives to these new cases are not treated as data. When additional new cases arise, transduction applied to the old plus the new cases can modify the classifications. It might therefore be said that the principle G derived from accepting the new classifications is hostage to the new cases in a way that inductive generalizations from labeled data are not. But transduction treats the fact that certain new cases have come up as data and new data always have the potentiality to change what rule should be accepted.

In other words, there is a sense in which transduction does not involve inductive generalization, because the relevant generalization is not arrived at from the labeled data alone, and there is a sense in which transduction does involve inductive generalization, because it does arrive at a general rule based on labeled data plus information about what new cases have come up.

What is important and not merely terminological is that, under certain conditions, transduction gives considerably better results in practice than those obtained from methods that use labeled data to infer a rule which is then used to classify new cases [Joachims, 1999; Vapnik, 2000; Weston *et al.*, 2003; Goutte *et al.*, 2004].

CONCLUSION

To summarize, we began by sketching certain aspects of statistical learning theory and then commented on philosophical implications for reliability theories of justification, on Popper's appeal to falsifiability, on simplicity as relevant to inductive inference, and on whether direct induction is an instance of inference to the best explanation.

BIBLIOGRAPHY

- [Bishop and Trout, 2005] M. A. Bishop and J. D. Trout. *Epistemology and the Psychology of Human Judgment*. Oxford: Oxford University Press, 2005.
- [Corfield *et al.*, 2005] D. Corfield, B. Schölkopf, and V. Vapnik. Popper, Falsification and the VC-dimension. Max Planck Institute for Biological Cybernetics Technical Report No. 145, 2005.
- [Devroye *et al.*, 1996] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

- [Goldman, 1986] A. Goldman. *Epistemology and Cognition*. Cambridge, MA: Harvard University Press, 1986.
- [Goldman, 2002] A. Goldman. *Pathways to Knowledge: Private and Public*. Oxford: Oxford University Press, 2002.
- [Goutte *et al.*, 2004] C. Goutte, N. Cancedda, E. Gaussier, and H. Dèjean. Generative vs Discriminative Approaches to Entity Extraction from Label Deficient Data. *JADT 2004, 7es Journ'ees internationales d'Analyse statistique des Donn'ees Textuelles*, Louvain-la-Neuve, Belgium, 10-12 mars, 2004.
- [Harman and Kulkarni, 2007] G. Harman and S. Kulkarni. *Reliable Reasoning: Induction and Statistical Learning Theory*, Cambridge, MA: MIT Press, 2007.
- [Hastie *et al.*, 2001] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2001.
- [Jeffrey, 2004] R. Jeffrey. *Subjective Probability (The Real Thing)*. Cambridge, England: Cambridge University Press, 2004.
- [Joachims, 1999] T. Joachims. Transductive Inference for Text Classification Using Support Vector Machines. In I. Bratko and S. Dzeroski, editors, *Proceedings of the 16th International Conference on Machine Learning: 200-9*. San Francisco: Morgan Kaufmann, 1999.
- [Kelley, 1996] K. T. Kelley. *The Logic of Reliable Inquiry*. Oxford: Oxford University Press, 1996.
- [Osherson *et al.*, 1982] D. Osherson, M. Stob, and S. Weinstein. *Systems That Learn*. Cambridge: MIT Press, 1982.
- [Popper, 1979] K. Popper. *Objective Knowledge: An Evolutionary Approach*. Oxford: Clarendon Press, 1979.
- [Popper, 2002] K. Popper. *The Logic of Scientific Discovery* (London: Routledge, 2002), 2002.
- [Putnam, 1963] H. Putnam. Degree of Confirmation and Inductive Logic. In *The Philosophy of Rudolf Carnap*, ed. A. Schillp. LaSalle, Indiana: Open Court, 1963.
- [Reichenbach, 1949] H. Reichenbach. *The Theory of Probability*. Berkeley: University of California Press, 1949.
- [Savage, 1954] L. J. Savage. *The Foundations of Statistics*. New York: Wiley, 1954.
- [Schulte, 2002] O. Schulte. Formal Learning Theory. *Stanford Encyclopedia of Philosophy* Edward N. Zalta, editor, 2002. <http://plato.stanford.edu/>.
- [Valiant, 1984] L. G. Valiant. A Theory of the Learnable, *Communications of the ACM* 27, pp. 1134-1142, 1984.
- [Vapnik and Chervonenkis, 1968] V. Vapnik and A. Ja. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities (in Russian), *Doklady Akademii Nauk USSR* 181, 1968. Translated into English as "On the uniform convergence of relative frequencies of events to their probabilities", *Theory of Probability and Its Applications* 16 (1971), pp. 264-280.
- [Vapnik, 1998] V. Vapnik. *Statistical Learning Theory*. New York: Wiley, 1998.
- [Vapnik, 2000] V. Vapnik. *The Nature of Statistical Learning Theory*, second edition. New York, Springer, 2000.
- [Vapnik, 2006] V. Vapnik. *Estimation of Dependencies Based on Empirical Data*, 2nd edition. New York: Springer, 2006.
- [Weston *et al.*, 2003] J. Weston, F. Pérez-Cruz, O. Bousquet, O. Chapelle, A. Elisseeff, and B. Schölkopf. KDD Cup 2001 Data Analysis: Prediction of Molecular Bioactivity for Drug Design-Binding to Thrombin, 2003.

This page intentionally left blank

TESTABILITY AND STATISTICAL LEARNING THEORY

Daniel Steel

Emerging in the 1960s from attempts to model learning in neural networks and machines, statistical learning theory subsequently expanded into a more general theory of learning from statistical data [Vapnik, 2000]. Statistical learning theory is of philosophical interest in a number of ways, but especially because one of its central concepts, Vapnik-Chervonenkis (VC) dimension, bears more than a passing resemblance to Karl Popper's notion of testable or falsifiable theories [Corfield *et al.*, 2005; Harman and Kulkarni, 2007, 50-52]. In this essay, I explore this connection with an emphasis on the underlying motivations of the two concepts. The concept of VC dimension is involved in two central results in statistical learning theory: the first identifies finite VC dimension as a necessary condition for long run convergence, while the second shows how a preference for lower VC dimension can improve "the rate of convergence" [Vapnik, 2000, 21, 83]. When viewed from this broad perspective, there are striking similarities in motivation between VC dimension in statistical learning theory and Popper's falsificationism. Popper's prohibition on unfalsifiable theories was grounded in the conviction that testability is necessary for science to progress towards closer approximations of the truth, or "verisimilitude" [1963]. Since unfalsifiable theories have infinite VC dimension, there is an obvious link between the first central result of statistical learning theory and one of the central tenets of Popper's philosophy of science. But the similarity between Popper and statistical learning theory does not end here. Popper emphasized that testability comes in degrees, and he thought that a preference for more testable theories would increase the rate of scientific progress. Moreover, the concept of VC dimension is similar though not equivalent to Popper's concept of theory dimension, which was one of his proposals about how to make the notion of degrees of testability more precise [1959, chapter 6], and both are promoted as means for making convergence more efficient.

But despite these similarities, there are some important differences in the technical details and, what is of greater interest here, in the fundamental aims of Popper's falsificationism and statistical learning theory. Popper's falsificationism is a staunchly scientific realist perspective driven by the goal of enhancing the efficiency of scientific progress towards truth [Popper, 1963]. In contrast, statistical learning theory aims to minimize the expected errors of predictions [Vapnik, 2000, 18]. On the face of it, these aims are very different. For example, the aim of minimizing expected predictive error can justify selecting a simple hypothesis

even when it is known that the truth is more complex [Vapnik, 2000, 116]. The similarity of the concepts of VC and Popper dimension, therefore, raises some intriguing questions about the connection between predictive accuracy and efficient convergence to the truth. In the final section of this essay, I explain how an account of Ockham’s razor from the perspective of formal learning theory provides some insights on this issue.

VC DIMENSION

Statistical learning theory aims to provide a general model of what Vapnik informally terms “learning from examples” or more formally a “function estimation model.” This model consists of the following three elements.

1. A generator (G) of random vectors $x \in R^n$, drawn independently from a fixed but unknown probability function $F(x)$.
2. A supervisor (S) who returns an output value y to every input vector x , according to a conditional distribution function $F(y|x)$, also fixed but unknown.
3. A learning machine (LM) capable of implementing a set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, where Λ is a set of parameters. [Vapnik, 2000, 17]

In other words, the learning machine is fed a set of data points x and corresponding values of y generated from the unknown conditional distribution function $F(y|x)$, and then selects a function from the set $f(x, \alpha)$ and uses it to make predictions about future data generated by $F(y|x)$. For example, suppose the problem is to infer gender from heights and weights. In this case, the vector x consists of height-weight pairs, while y indicates gender, which for simplicity we can suppose must be either male or female. Then the “training” data consists of a set of height-weight pairs together with the gender for some number of people. Then the learning machine attempts to select the function that will do the best job of predicting gender by height and weight in subsequent batches of data. This is a simple example of what is known as *pattern recognition*.

The learning machine typically restricts attention to functions that satisfy a certain functional form, for instance, linear functions. As a result, the conditional distribution function $F(y|x)$ cannot be assumed to be one of the alternatives that the learning machine may choose. The aim, then, is not discover $F(y|x)$ but instead to select the parameters so that the function $f(x, \alpha)$ that minimizes expected predictive error. Expected predictive error of a function $f(x, \alpha)$, denoted by $R(\alpha)$, is defined as follows.

$$R(\alpha) = \int L(y, f(x, \alpha)) dF(x, y)$$

The quantity $L(y, f(x, \alpha))$ is called the *loss function*. In pattern recognition, the loss function equals 0 if $y = f(x, \alpha)$ and 1 otherwise. In other types of inference problem, the loss function is defined differently, for example, in regression estimation it is $(y - f(x, \alpha))^2$.

The aim, then, is to select the function $f(x, \alpha)$ that has the smallest value of $R(\alpha)$. However, since $F(x)$ and $F(y|x)$ (and hence $F(x, y)$) are unknown, it is not possible to find the function that minimizes predictive error through a direct application of the above equation. Instead, the training data are the basis for deciding which function $f(x, \alpha)$ to select. The most straightforward way to do this is to choose the function that minimizes error on the training data. The error on the training data for a function $f(x, \alpha)$ is denoted by $R_{emp}(\alpha)$ and defined as follows.

$$R_{emp}(\alpha) = \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i, \alpha))$$

In this equation, n is the number of observations in the training data. Thus, in pattern recognition, the function that minimizes $R_{emp}(\alpha)$ is simply the one that makes the fewest mistakes in the training data.

The expression *empirical risk minimization* (ERM) is used to refer to selecting a function $f(x, \alpha)$ that minimizes error on the training data as just explained. Statistical learning theory, then, focuses on the following three questions [Vapnik, 2000, 21]:

1. Under what conditions does ERM ensure $R_{emp}(\alpha)$ converges to $R(\alpha)$ in the long run (i.e. when is ERM “consistent”)?
2. What is the rate of the convergence and how can it be controlled?
3. How can algorithms be constructed to control the rate of convergence?

The concept of VC dimension is connected to the answers to these questions, especially question number 2. VC dimension is defined in terms of the concept of shattering. The best way to approach these rather intricate concepts is by means of a simple example.

Consider again our example of predicting gender on the basis of height and weight. In this example, height-weight pairs can be represented by points on a plane that are labeled + for female or – for male. Suppose we restrict attention to functions that group data points into +’s and –’s by drawing straight lines between them, which we can call the linear functions. Now suppose that we are given three pairs of heights and weights, represented as three points on the plane, but we are not yet told the genders of the people whose heights and weights we have been given (see figure 1). In this situation, shattering a set of data points means being able to draw a straight line between them so that all the +’s are on one side of the line and all the –’s are on the other, *no matter which points are labeled + and which –*. In other words, for each possible way of assigning the +’s and –’s, there is a straight line that separates them. Thus the intuitive origin of



Figure 1. Three points representing height-weight pairs for three individuals

the term “shatter”: if the set of functions shatters the set of data points, then one is assured of being able to “break down” the data points according to the values of variable one wishes to predict. It is easy to see that the set of three data points in figure 1 is shattered by the linear functions (see figure 2). However, not every configuration of three data points is shattered by the linear functions, for example, they do not shatter a set of three perfectly collinear points (see figure 3). In addition, no set of four data points is shattered by the linear functions. To see this, consider a set of four data points. If the data points are perfectly collinear, then they are not shattered by the set of straight lines for the same reason that three perfectly collinear points are not. Suppose then that the four points are not collinear. But in that case the linear functions do not shatter them either (see figure 4). Thus, in this example, the linear functions shatter some sets of three data points, but shatter no sets of four or more data points. The concept of shattering is used to define VC dimension. A set of functions has a VC dimension h if and only if the set of functions shatters *some* group of h many data points and *no* group of *more* than h data points. Thus, in the example described in the height-weight-gender example the VC dimension of the linear functions is three.

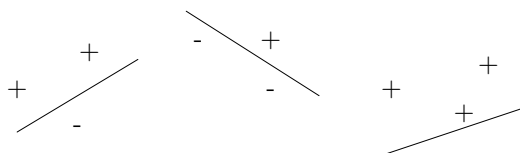


Figure 2. The set of linear functions shatters these three data points

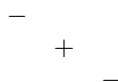


Figure 3. The linear functions do not shatter three perfectly collinear data points



Figure 4. The linear functions do not shatter any set of four data points

Given the above example, the general definitions of shattering and VC dimension can be stated more intelligibly. Consider a set of functions $f(x, \alpha), \alpha \in \Lambda$ and a set of data points $x_1, \dots, x_n$. This set of functions is said to *shatter* $x_1, \dots, x_n$ if, for any possible assignment of values to $y_1, \dots, y_n$, there is an $\alpha \in \Lambda$ such that $f(x, \alpha)$ makes no errors on this set of data. In the example above, $f(x, \alpha)$ is the set of linear functions, so the parameters α would include the slope and intercept. The data points (the x_i 's) are the height-weight pairs, and the y_i 's are the genders represented by the '+'s and '-'s. The VC dimension h of a set of functions $f(x, \alpha), \alpha \in \Lambda$, then, is the maximum number such that some set of h data points can be shattered by $f(x, \alpha), \alpha \in \Lambda$.¹ Notice that the definitions of shattering and VC dimension do not assume that $f(x, \alpha)$ is the set of linear functions — it could be any set of functions. Nor do they assume that data points consist of pairs of measurements (e.g. of height and weight). Each data point might consist of one measurement or three or however many one likes.

The concept of VC dimension is connected to questions 1 through 3 above by way of the some central mathematical results of statistical learning theory. The first is that ERM ensures that $R_{emp}(\alpha)$ converges to $R(\alpha)$ *only when the VC dimension of the set of functions is finite* [Vapnik 2000, 80; Harman and Kulkarni, 2007, 48-49]. In other words, finite VC dimension is a necessary condition for being assured of converging to the most predictively accurate function in the long run. As Vapnik explains, finite VC dimension is closely related to Popper's concept of a falsifiable hypothesis [Vapnik, 2000, 47-55]. That point will be discussed in further detail below. The second connection has to do with controlling the rate of convergence. Although it is good to know that one's methods will hone in on the right answer in the long run, it is also very important, and perhaps even more important, to think about how the method behaves in the meantime. This issue is linked to the classic problem of how to balance maximizing fit with the data versus over fitting [Forster and Sober, 1994].

When fitting a function to a set of data points, it is better other things being equal to prefer functions that classify those points more accurately. In pattern recognition, for instance, ERM would direct one to select the function that makes the fewest errors. But fit with the data is not the only relevant consideration. Because there is randomness in the data, some of the observed variation does not reflect the underlying relationship that one wants to approximate. Thus, selecting

¹This is actually the definition of VC dimension for indicator functions, in which y must be either 1 or 0. See Vapnik [2000, 80] for the generalization of this definition to real functions.

a function that perfectly fits *all* variation in the data is not necessarily a good idea, since some of that variation may not indicate anything about what to expect in the future. Thus, it is a commonplace that curve fitting ought to involve some balance between good fit with the data and some procedure for avoiding *over* fitting. In statistical learning theory, VC dimension is central to reducing the chance of over fitting the data. The higher the VC dimension of a set of functions, the greater its capacity to fit the data. For example, while the linear functions do not shatter the four data points in figure 4, those data points would be shattered by sets of higher order polynomial functions. To make this point more concrete, consider functions that divide the data points into +’s and –’s by drawing parabolas between them, which we can call the parabolic functions. The VC dimension of the parabolic functions in the height-weight-gender example is four. Thus, selecting the function from a set with higher VC dimension increases the chances of having a better fit with the data, but it also increases the chance of over fitting. The challenge, then, is to decide how to balance these two conflicting concerns.

In statistical learning theory, this challenge is addressed by deriving a probabilistic upper limit on the expected predictive error given the error in the training data and VC dimension [Vapnik, 2000, chapter 4]. More specifically, with probability $1 - \eta$, $R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(\ln(2n/h)+1) - \ln(\eta/4)}{n}}$.²

Recall that $R(\alpha)$ indicates the expected predictive error of the function $f(x, \alpha)$, $R_{emp}(\alpha)$ represents the error of that function on the training data, h is the VC dimension of the set from which functions are selected, and finally n is the number of data points in the training data. To see how this result would be applied, consider again the example of predicting gender on the basis of height and weight. In that example, the VC dimension h of the set of linear functions is three. If we chose instead to select a function from the parabolic functions, h would equal four. So, the question is which set of functions to select our function from, for example, just the linear functions or a set including both linear and parabolic functions. Choosing a set with lower VC dimension will result in a smaller value for $\sqrt{\frac{h(\ln(2n/h)+1) - \ln(\eta/4)}{n}}$, but sets with lower VC dimension will also typically have a higher training error $R_{emp}(\alpha)$. Thus, the above inequality allows one to choose the set of functions that yields the lowest probabilistic upper bound on the expected predictive error, $R(\alpha)$.

Testability

In *The Logic of Scientific Discovery*, Popper characterized a testable or falsifiable theory as one that is capable of being “refuted by experience” [1959, 18]. As is well known, Popper insisted that only falsifiable theories were genuinely scientific, but not because they were more likely to be true. To the contrary, Popper emphasized that the more falsifiable a theory is, the more it sticks its neck out,

²Like other results from statistical learning theory, this assumes that past and future data are generated by the same independent and identically distributed probability distribution.

and hence the more *improbable* it is [1963, 217-220]. Popper took this as proof that science does not aim primarily for probable theories, but instead for highly informative ones. The reason for preferring falsifiable theories, according to Popper, is that by so doing we further scientific progress. This point comes out in *The Logic of Scientific Discovery* in connection with what Popper terms “conventionalist stratagems” [1959, 57-61]. Conventionalism treats scientific theories as true by definition, so that if some apparent conflict arises between the theory and observation, that conflict must be resolved by rejecting something other than the theory. Popper admitted that there is no logical contradiction to be found in conventionalism, but he argued that it was nevertheless highly problematic on methodological grounds. In particular, conventionalism would obstruct the advancement of scientific knowledge, and we should therefore firmly commit to rules of scientific method that disallow conventionalist stratagems [1959, 61-62]. This connection between falsifiability and scientific progress is easy to appreciate in light of some of Popper’s favorite examples of theories that failed to abide by his strictures, for instance, Marxism and Freudian psychology. According to Popper, the Marxist and Freudian traditions were case studies of how treating scientific theories as unquestionable truths could lead researchers into a morass of ad hoc explanations that stultified the advancement of knowledge.

Popper set out his view of scientific progress in greater detail in *Conjectures and Refutations* [1963, 231-248]. The central theme of that proposal is quite simple. Scientific progress in Popper’s sense occurs when a scientific theory is refuted and replaced with another that is closer to the truth. As we do not have direct access to the truth, progress would typically be judged by some more indirect means. For example, suppose that one theory T is refuted and replaced by another T^* such that (1) T^* passes all of the severe tests that T passed, (2) T^* passes the tests that T failed, and (3) T^* makes new predictions that turn out to be correct. If this happens, then Popper thought that we have good reason to say that T^* is closer to the truth than T . For example, Popper thought that Einstein’s General Theory of Relativity satisfied these conditions with respect to Newtonian Mechanics. It is obvious that falsifiable theories are a necessary ingredient in this picture of scientific progress. On Popper’s account, the advancement of science is driven by refuting theories and replacing them with better theories that generate new discoveries. Therefore, unfalsifiable theories — or theories we decide to save at all costs by means of “conventionalist stratagems” — halt progress in its tracks. Moreover, Popper believed that this process of conjectures and refutations would, in the long run, lead scientists closer and closer to the truth, although we may never know at any given time how close we are (or aren’t).

Popper’s claimed, then, that testable theories are necessary if science hopes to converge to the truth in the long run. It is easy to see an analogy between Popper’s claim on this score and the result from statistical learning theory that finite VC dimension is a necessary condition for long run convergence to the function that minimizes expected predictive error. A set of functions Φ would be unfalsifiable if, for any possible set of data, there is a function in Φ that can fit that data with zero

error. Recall that the VC dimension h of a set of functions is the maximum number such that some set of h data points can be shattered by that set. An unfalsifiable set of functions, then, would have no such maximum and hence would have infinite VC dimension. Thus, a basic result of statistical learning theory coincides with Popper's intuition that falsifiability is a necessary ingredient for being assured of homing in on the truth in the long run. Vapnik concludes his discussion of the relationship between falsifiability and statistical learning theory by remarking "how amazing Popper's idea was" [2000, 55].

Popper also proposed that the falsifiability or testability of theories could come in degrees. Degrees of testability are clearly important for Popper's vision of scientific progress. For when one theory is refuted, there may be several possible replacements and Popper would presumably recommend that we choose the most testable of the viable alternatives. Moreover, Popper's reasoning naturally suggests that going with the most testable theory would accelerate scientific progress. After all, a barely testable theory might not halt progress altogether, but it could certainly slow it down. Indeed, the idea that degrees of testability are linked to the rate of scientific progress is hinted at in the epigraphs to *Conjectures and Refutations*.

Experience is the name everyone gives to their mistakes.

Oscar Wilde

Our whole problem is to make the mistakes as fast as possible. . .

John Archibald Wheeler

More testable theories rule out more and typically will be refuted faster than less testable ones. So, it is easy to guess Popper's meaning here: the more testable our theories, the faster our mistakes, and the more rapid the advancement of science.

In *The Logic of Scientific Discovery*, Popper suggested two grounds for comparing degrees of testability [1959, chapter 6]. The first was a subclass relation. For instance, the theory that planets move in circles around the sun is a subclass of the theory that they move in ellipses, and hence the former theory is more easily refuted (i.e. more testable) than the latter. However, it is the second of Popper's suggestions for how to compare degrees of testability that is most pertinent to our concerns here. Popper proposed that the dimension of a theory be understood in terms of the number of data points needed to refute it. More specifically, if $d + 1$ is the minimum number of data points needed to refute the theory t , then the *Popper dimension* of t is d [1959, 113-114]. The difference between Popper and VC dimension can be neatly made in terms of shattering. Suppose we think of theories as sets of functions. If the Popper dimension of a theory of functions is d , then no set of only d data points can refute the theory and hence the theory shatters *every* group of d many data points. On the other hand, if the VC dimension of the theory is h , then that set shatters *some but not necessarily all* groups of h many data points. This difference is illustrated by the example of predicting gender by height and weight discussed above. In that example, the linear functions shatter every set of two data points, some but not all sets of three data points, and no sets

of four data points. Consequently, the Popper dimension of the linear functions in this case is two, while the VC dimension is three. Further divergences between Popper and VC dimension occur in cases in which data points consist of more than two measurements. For example, suppose we wanted to predict diabetes on the basis of blood pressure, body mass index, and cholesterol level. In this case, the data points would be spread out in a three dimensional space, and the linear functions would separate data points with flat planes. In this situation, the Popper dimension of the linear functions remains two (since a flat plane cannot separate three perfectly collinear points in a three dimensional space) but the VC dimension of the linear functions in this case would be four.³

However, there are some cases in which Popper and VC dimension coincide. For instance, consider a very simple example in which one wishes to predict the colors of balls drawn from an urn, which may be either blue or red.⁴ In this example, the x_i 's consist of balls drawn from the urn and the y_i 's of an indication of the color (blue or red) of each ball. Suppose that 99 balls have been drawn so far, and all are red. The functions, then, tell us what to predict about the colors of the future balls given this data. What we might call the inductive function directs us to predict that all future balls will be red. Another set of functions directs us to predict that the balls will switch at some future time from red to blue and stay blue from then on. We can call these the anti-inductive functions. Notice that in this example a fixed number of data points cannot be arranged into distinct configurations as in figures 2 and 3, and hence there is no difference between shattering some and shattering all configurations of n many data points. As a result, Popper and VC dimension are equivalent in this case. For example, the inductive function does not shatter the next data point, since it would be refuted if the next ball is blue. Hence, its Popper and VC dimension is zero. In contrast, the anti-inductive functions can shatter the next data point, since the switch from red to blue might begin with the next ball or it might begin later. However, the anti-inductive functions do not shatter the next two data points, since none of them can accommodate a switch to blue followed by an immediate switch back to red. Thus, the Popper and VC dimension of the anti-inductive functions is one.

Let us sum up the similarities and differences between VC dimension and Popper's notion of degrees of testability, beginning with the similarities. The two concepts are similar in spirit, coincide in some simple examples, and track one another in some other examples. In addition, similar claims are made on behalf of both: testability and finite VC dimension are claimed to be necessary for convergence in the long run, and a preference for lower Popper dimension (i.e. greater testability) and lower VC dimension are both said to promote the faster convergence. Moreover, there is a further similarity between Popper and VC dimension that is rarely remarked upon: both of these concepts presuppose without explana-

³To see this, arrange the points in the following way. Put three of the four points on the same plane but not collinear, and place the fourth point on a different plane from the other three: then you can separate any assignment of +’s and -’s with a flat plane.

⁴This example is borrowed from Goodman [1946].

tion some natural or preferred way of dividing data up into points or units. Since different modes of expression might result in different ways of carving up data into units, this means that neither concept is language invariant.⁵ Now let us turn to the differences between Popper and VC dimension. First, there are differences in technical details of the two concepts, as explained above. Furthermore, Popper never provided any precise articulation or proof of his claims about the link between testability and convergence, while Vapnik and others have done this for VC dimension. Thus, statistical learning theory represents a very significant step forward from Popper's work. Finally, there are also some important differences in philosophical motivation that I will discuss in the next section.

Scientific Realism, Empirical Risk Minimization, and Simplicity

The conceptual connection between falsificationism and statistical learning theory is intriguing and suggestive. In one sense, the similarity is understandable given that both are proposed as a means for promoting faster convergence to a correct answer. In another sense, however, the correspondence is surprising because different things are meant by "correct answer" in the two cases. For Popper, convergence meant convergence to the truth. In statistical learning theory, convergence means convergence to the function that has the least expected predictive error in comparison to the other functions in a given set. Moreover, in Popper's scheme a refuted theory should be replaced by another that might be true given what we know, while in statistical learning theory it may be appropriate to select a function from a set that we know does not contain the true one. For example, we might minimize expected predictive error by choosing a linear function even if it is known that the true conditional distribution function is not linear [Kelly and Mayo-Wilson, 2008; Vapnik, 2000, 116]. In short, scientific realism and minimizing predictive risk are two very different motivations. The question posed by the similarity between Popper's philosophy of science and statistical learning theory is whether there is some important connection between these two different aims. In this section, I consider how insights on this topic follow from a three way comparison between Popper, statistical learning theory, and a recent account of empirical simplicity from the literature on formal learning theory [2004; 2007a; 2007b].

Kevin Kelly proposes a general account of Ockham's razor — the claim simpler theories should be preferred to more complex ones — in terms of efficient convergence to the truth. Efficiency in Kelly's account is understood primarily in terms of minimizing the maximum number of retractions of conjecture, or "mind changes." The central idea is that a reliable learner who conjectures a more complex theory before a simpler one can be forced by the data to revert to the simpler theory and then to return again to the more complex one. In contrast, a learner who goes with the simpler theory first cannot be forced to loop around in this way, but can only be forced by the data to move directly from simpler to more com-

⁵Popper seems to have recognized that this problem [1959, 283-284]. For a suggestion about how to eliminate language variance from the concept of VC dimension, see [Steel, 2009].

plex. For instance, consider the example of drawing red or blue balls from an urn discussed above. According to Kelly's approach, Ockham's razor would require selecting the hypothesis that all the balls are red in this example. For suppose that we chose instead to conjecture the anti-inductivist alternative that the balls start out red and then permanently switch to blue at some future point. Then if learner is reliable, she will eventually have to revert to conjecturing that all the balls are red given a sufficiently large number of further observations of red and no blue balls. (If she did not, she could continue conjecturing the false hypothesis forever and hence not be reliable.) Yet once she has reverted to conjecturing that all the balls are red, subsequent observations of blue balls can force her to change her mind again. In contrast, if the learner begins with "all of the balls are red," then she cannot be forced by the data to loop around in this way.⁶

There are several conceptual links between Kelly's account of Ockham's razor and Popper's falsificationism. Like Popper, Kelly's approach is motivated by the aim of efficient convergence to truth. Moreover, in the ball drawing example, the Popper dimension of the hypothesis that all the balls are red is zero, while the Popper dimension of the hypothesis that they are all red until some future point and then permanently switch to blue is one. Thus, in this example, Kelly's account of Ockham's razor and Popper's claim that scientists should conjecture the most testable viable alternative yield the same recommendation. In addition, Popper and VC dimension coincide in the ball drawing example. An in fact, this example is just one instance of a more general class of cases in which Ockham's razor as construed by Kelly, Popper dimension, and VC dimension all lead to the same result [Steel, 2009].⁷

Given the connections between Kelly's account of Ockham's razor, Popper's concept of testability, and VC dimension described here, one might wonder whether objections to Popper's falsificationism would also be objections to either of these other two approaches. I will consider this question only in relation to the topic of simplicity. As is well known, Popper proposed that degrees of simplicity be equated with degrees of testability [1959, 126]. According to Popper, this approach had the advantage not only of providing a clearer definition of the concept of simplicity but also of leading to a principled explanation of why simplicity was important for science. One of Popper's arguments in favor of this proposal was that his concept of theory dimension corresponds to geometrical dimension [1959, 110-115, 129] and that geometrical dimension is often associated with simplicity. Geometrical dimension has to do with the number of free parameters needed to define a geometrical class (e.g. the straight lines or the parabolas). For instance, in the height-weight-gender example, the Popper dimension of the parabolic functions is three while the Popper dimension of the linear functions is two, and the equation

⁶This analysis of Goodman's new riddle of induction was first published in an essay by Schulte [1999].

⁷However, Kelly's analysis of Ockham's razor differs from both Popper and VC dimension on the issue of language variance. Kelly develops a language invariant definition of empirical complexity in terms of the number further mind changes that the data could force a reliable learner to make in an inductive problem [2007b].

used to define the set of parabolas (i.e. $y = a + bx + cx^2$) has three free parameters while the equation used to define the straight lines (i.e. $y = a + bx$) has two free parameters. Moreover, the linear functions are naturally thought of as simpler than the parabolic functions. However, subsequent authors have conclusively shown that Popper was mistaken in supposing that theory dimension corresponds to geometric dimension in general; indeed this correspondence does not even hold for some of the examples that Popper gave to illustrate it [Turner, 1991]. As a result of such criticisms, Popper's attempt to explain simplicity by way of testability is usually regarded as a failure.

Both statistical learning theory and Kelly's learning theoretic approach to Ockham's razor avoid the above difficulty, but in different ways. Statistical learning theory does not claim that lower VC dimension always corresponds to greater simplicity and hence does not commit itself to explaining simplicity in terms of VC dimension [Harman and Kulkarni, 2007, 69-73]. Thus, from the perspective of statistical learning theory, Popper was mistaken to think that testability is connected to simplicity in some systematic and general way. According to Kelly's approach, by contrast, Popper had the right intuition that efficient convergence to the truth is the central justification for a scientific preference for simpler theories. Popper's mistake, from Kelly's perspective, was to suppose that the concept of complexity relevant to Ockham's razor — which Kelly refers to as empirical complexity — corresponds to geometrical dimension.

As the survey given here shows, statistical learning theory offers many new insights regarding the concept of testability and such issues as predictive accuracy, efficient convergence to truth, and simplicity. It is, therefore, worthy of further sustained interest from philosophers of science.

ACKNOWLEDGMENTS

I would like to thank Malcolm Forster and an anonymous referee for helpful comments and suggestions on an earlier draft of this essay.

BIBLIOGRAPHY

- [Corfield *et al.*, 2005] D. Corfield, B. Schölkopf, and V. Vapnik. Popper, Falsification and the VC-Dimension. Max Plank Institute of Biological Cybernetics, Technical Report No. 145, 2005.
- [Forster and Sober, 1994] M. R. Forster and E. Sober. How to Tell when Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions. *British Journal for the Philosophy of Science* **45**: 1-35, 1994.
- [Goodman, 1946] N. Goodman. A Query on Confirmation. *Journal of Philosophy* **43**: 383-385, 1946.
- [Harman and Kulkarni, 2007] G. Harman and S. Kulkarni. *Reliable Reasoning: Induction and Statistical Learning Theory*, MIT Press, Cambridge, MA, 2007.
- [Kelly, 2004] K. Kelly. Justification as Truth-Finding Efficiency: How Ockham's Razor Works. *Minds and Machines* **14**: 485-505, 2004.
- [Kelly, 2007a] K. Kelly. A New Solution to the Puzzle of Simplicity. *Philosophy of Science* **74**: 561-573, 2007.

- [Kelly, 2007b] K. Kelly. Ockham's Razor, Empirical Complexity, and Truth-finding Efficiency. *Theoretical Computer Science* **383**: 270-289, 2007.
- [Kelly and Mayo-Wilson, 2008] K. Kelly and C. Mayo-Wilson. Reliable Reasoning: Induction and Statistical Learning Theory (Book Review). *Notre Dame Philosophical Reviews*, 2008. (<http://ndpr.nd.edu/reviews.cfm>).
- [Popper, 1959] K. Popper. *The Logic of Scientific Discovery*, Routledge, New York, 1959.
- [Popper, 1963] K. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, New York, 1963.
- [Schulte, 1999] O. Schulte. Means-Ends Epistemology. *British Journal for the Philosophy of Science* **50**: 1-31, 1999.
- [Steel, 2009] D. Steel. Testability and Ockham's Razor: How Formal and Statistical Learning Theory Converge in the New Riddle of Induction. *Journal of Philosophical Logic* **38**: 471-489, 2009.
- [Turney, 1991] P. Turney. A Note on Popper's Equation of Simplicity with Falsifiability. *The British Journal for the Philosophy of Science* **42**: 105-109, 1991.
- [Vapnik, 2000] V. Vapnik. *The Nature of Statistical Learning Theory*, Springer, New York, 2000.

This page intentionally left blank

Part XI

Different Approaches to
Simplicity Related to
Inference and Truth

This page intentionally left blank

LUCKINESS AND REGRET IN MINIMUM DESCRIPTION LENGTH INFERENCE

Steven de Rooij and Peter D. Grünwald

1 INTRODUCTION

Suppose we have been observing some phenomenon for a while, and we now consider a number of alternative hypotheses to explain how the data came about. We want to use the data somehow to evaluate these hypotheses and decide which is the “best” one. In case that one of the hypotheses exactly describes the true mechanism that underlies the phenomenon, then that is the one we hope to find. While this may already be a hard problem, available hypotheses are often merely approximations in practice. In that case the goal is to select a hypothesis that is useful, in the sense that it provides insight in previous observations, and matches new observations well. Of course, we can immediately reject hypotheses that are inconsistent with new experimental data, but hypotheses often allow for some margin of error; as such they are never truly inconsistent but they can vary in the degree of success with which they predict new observations. A quantitative criterion is required to decide between competing hypotheses. The Minimum Description Length (MDL) principle is such a criterion [Rissanen, 1978; Rissanen, 1989]. It is based on the intuition that, on the basis of a useful theory, it should be possible to *compress* the observations, i.e. to describe the data in full using fewer symbols than we would need using a literal description. According to the MDL principle, the more we can compress a given set of data, the more we have learned about it. The MDL approach to inference requires that all hypotheses are formally specified in the form of *codes*. A code is a function that maps possible outcomes to binary sequences; thus the length of the encoded representation of the data can be expressed in bits. We can encode the data by first specifying the hypothesis to be used, and then specifying the data with the help of that hypothesis. Suppose that $L(H)$ is a function that specifies how many bits we need to identify a hypothesis H in a countable set of considered hypotheses $\mathcal{H}$. Furthermore let $L_H(D)$ denote how many bits we need to specify the data D using the code associated with hypothesis H . The MDL principle now tells us to select that hypothesis H for which the total description length of the data, i.e. the length of the description of the hypothesis $L(H)$, plus the length of the description of data using that hypothesis $L_H(D)$, is shortest. The hypothesis yielding the minimum total description length as a function of the data is denoted H_{mdl} , and the resulting code length is L_{mdl} :

$$(1) \quad L_{\text{mdl}}(D) = L(H_{\text{mdl}}) + L_{H_{\text{mdl}}}(D) = \min_{H \in \mathcal{H}} (L(H) + L_H(D)).$$

Intuitively, the term $L(H)$ represents the complexity of the hypothesis while $L_H(D)$ represents how well the hypothesis is able to describe the data, often referred to as the goodness of fit. By minimising the sum of these two components, MDL implements a tradeoff between complexity and goodness of fit. Also note that the selected hypothesis only depends on the *lengths* of the used code words, and not on the binary sequences that make up the code words themselves. This will make things a lot easier later on.

By its preference for short descriptions, MDL implements a heuristic that is widely used in science as well as in learning in general. This is the *principle of parsimony*, also often referred to as Occam’s razor. A parsimonious learning strategy is sometimes adopted on the basis of a belief that the true state of nature is more likely to be simple than to be complex. We point out right at the outset that MDL is *not* based on such a belief. Instead, we try to avoid assumptions about the truth as much as possible, and focus on effective *strategies* for learning instead. As explained in Section 5.2, even if the truth is “complex”, it is often a very effective strategy to prefer “simple” hypotheses for small sample sizes. That said, a number of issues still need to be addressed if we want to arrive at a practical theory for learning:

1. What code should be used to identify the hypotheses?
2. What codes for the data are suitable as formal representations of the hypotheses?
3. After answering the previous two questions, to what extent can the MDL criterion actually be shown to identify a “useful” hypothesis, and what do we actually mean by “useful”?

These questions, which lie at the heart of MDL research, are illustrated in the following example.

EXAMPLE 1 Language Learning. Ray Solomonoff, one of the founding fathers of the concept of Kolmogorov complexity (to be discussed in Section 2.4), introduced the problem of language inference based on only positive examples as an example application of his new theory [Solomonoff, 1964]; it also serves as a good example for MDL inference.

We will model language using context-free grammars (introduced as “phrase structure grammars” by Chomsky in [Chomsky, 1956]). A grammar is a set of *production rules*, each of which maps a symbol from a set of *nonterminal symbols* N to a (possibly empty) *replacement sequence* consisting of both other nonterminals and *terminal symbols*, which are words from some dictionary Σ . A sentence is grammatical if it can be produced from the *starting symbol*, which is a particular nonterminal, by iteratively applying a production rule to one of the matching nonterminal symbols. The following grammar is an example. It uses the standard

abbreviation that multiple rules for the same nonterminal may be combined using a pipe symbol, i.e. two rules $n \rightarrow r_1$ and $n \rightarrow r_2$ are written $n \rightarrow r_1 \mid r_2$. The empty sequence is denoted ϵ .

Sentence	$\rightarrow$	Nounphrase Verbphrase
Nounphrase	$\rightarrow$	Determiner Adjectives Noun $\mid$ Adjectives Noun
Determiner	$\rightarrow$	the $\mid$ a
Adjectives	$\rightarrow$	Adjective Adjectives $\mid \epsilon$
Adjective	$\rightarrow$	big $\mid$ complex $\mid$ careful $\mid$ curvy
Noun	$\rightarrow$	men $\mid$ statistician $\mid$ model
Verbphrase	$\rightarrow$	Verb Nounphrase
(2) Verb	$\rightarrow$	prefers $\mid$ avoids

This grammar accepts such sentences as **The careful statistician avoids the complex model**, or **The curvy model prefers big men**. We proceed to infer such a grammar from a list of valid example sentences D . Assume that we already know the words of the language, i.e. Σ contains all words that occur in D and is background knowledge. The task is to use D to learn how the words may be ordered. For the set of our hypotheses $\mathcal{H}$ we take all context-free grammars that use only words from Σ . Note that $\mathcal{H}$ is a countably infinite set.

We need to define the relevant code length functions: $L(H)$ for the specification of the grammar $H \in \mathcal{H}$, and $L_H(D)$ for the specification of the example sentences with the help of that grammar. In this example we use very simple code length functions; later in this introduction, after describing in more detail what properties good codes should have, we return to language inference with a more in-depth discussion.

We use *uniform codes* in the definitions of $L(H)$ and $L_H(D)$. A uniform code on a finite set A assigns binary code words of equal length to all elements of the set. Since there are 2^l binary sequences of length l , a uniform code on A must have code words of length at least $\lceil \log |A| \rceil$. (Here and in the following, $\lceil \cdot \rceil$ denotes rounding up to the nearest integer, and $\log$ denotes binary logarithm.) To establish a baseline, we first use uniform codes to calculate how many bits we need to encode the data literally, without the help of any grammar. Namely, we can specify every word in D with a uniform code on $\Sigma \cup \diamond$, where $\diamond$ is a special symbol used to mark the end of a sentence. Two consecutive diamonds signal the end of the data. Let w denote the total number of words in D , and let s denote the number of sentences. With the uniform code we need $(w + s + 1) \lceil \log |\Sigma \cup \{\diamond\}| \rceil$ bits to encode the data. We are looking for grammars H which allow us to compress the data beyond this baseline value. We specify $L(H)$ as follows. For each production rule of H , we use $\lceil \log |N \cup \{\diamond\}| \rceil$ bits to uniformly encode the initial nonterminal symbol, and $\lceil \log |N \cup \Sigma \cup \{\diamond\}| \rceil$ bits for each symbol in the replacement sequence. The $\diamond$ symbol signals the end of each rule; an additional diamond marks the end of the entire grammar. If the grammar H has r rules, and the summed length of the replacement sequences is l , then we can calculate that the number of bits we need to encode the entire grammar is at most

$$(3) \quad (r+1)\lceil \log |N \cup \{\diamond\}| \rceil + (l+r)\lceil \log |N \cup \Sigma \cup \{\diamond\}| \rceil.$$

The first term is the length needed to describe the initial nonterminal symbols of each rule; the $+1$ is because, to signal the end of the grammar, we need to put an extra diamond. The second term is the length needed to describe the replacement sequences; the $+r$ is because, at the end of each rule, we put an extra diamond.

If a context-free grammar H is correct, i.e. all sentences in D are grammatical according to H , then its corresponding code L_H can help to compress the data, because it does not need to reserve code words for any sentences that are ungrammatical according to H . In this example we simply encode all words in D in sequence, each time using a uniform code on the set of words that *could* occur next according to the grammar. Again, the set of possible words is augmented with a $\diamond$ symbol to mark the end of the data. For example, based on the grammar (2) above, the 1-sentence data sequence $D = \text{the model avoids men}$ is encoded using $\lceil \log 10 \rceil + \lceil \log 7 \rceil + \lceil \log 2 \rceil + \lceil \log 9 \rceil + \lceil \log 10 \rceil$, because, by following the production rules we see that the first word can be a determiner, an adjective, or a noun; since we also allow for a diamond, this gives 10 choices in total. Given the first word **the**, the second word can be an adjective or a noun, so that there are 7 choices (now a diamond is not allowed); and so on. The final $\log 10$ is the number of bits needed to encode the final diamond.

Now we consider two very simple correct grammars, both of which only need one nonterminal symbol S (which is also the starting symbol). The “promiscuous” grammar (terminology due to Solomonoff [Solomonoff, 1964]) has rules $S \rightarrow S S$ and $S \rightarrow \sigma$ for each word $\sigma \in \Sigma$. This grammar generates *any* sequence of words as a valid sentence. It is very short: we have $r = 1 + |\Sigma|$ and $l = 2 + |\Sigma|$, so the number of bits $L(H_1)$ required to encode the grammar essentially depends only on the size of the dictionary and not on the amount of available data D . On the other hand, according to H_1 all words in the dictionary are allowed in all positions, so $L_{H_1}(D)$ requires as much as $\lceil \log(|\Sigma \cup \{\diamond\}|) \rceil$ bits for every word in D , which is equal to the baseline. Thus this grammar does not enable us to compress the data.

Second, we consider the “ad hoc” grammar H_2 . This grammar consists of a production rule $S \rightarrow d$ for each sentence $d \in D$. Thus according to H_2 , a sentence is only grammatical if it matches one of the examples in D exactly. Since this severely restricts the number of possible words that can appear at any given position in a sentence given the previous words, this grammar allows for very efficient representation of the data: $L_{H_2}(D)$ is small. However, in this case $L(H_2)$ is at least as large as the baseline, since in this case the data D appear literally in H_2 !

Both grammars are clearly useless: the first does not describe any structure in the data at all and is said to *underfit* the data. In the second grammar random features of the data (in this case, the selection of valid sentences that happen to be in D) are treated as structural information; this grammar is said to *overfit* the data. Consequently, both grammars $H \in \{H_1, H_2\}$ do not enable us to compress the data at all, since the *total* code length $L(H) + L_H(D)$ exceeds the baseline. In contrast, by selecting a grammar H_{mdl} that allows for the greatest total compression as

per (1), we avoid either extreme, thus implementing a natural tradeoff between underfitting and overfitting. Note that *finding* this grammar H_{mdl} may be a quite difficult search problem, but the computational aspects of finding the best hypothesis in large hypothesis spaces are not considered in this text. $\diamond$

1.1 Overview

The three fundamental questions of page 866 will be treated in more detail in the following text. First we discuss codes for the set of hypotheses in terms of the concepts of *luckiness* and *regret* in Section 2. Although most previous treatments put less emphasis on luckiness (or fail to mention it altogether), it is our position that both concepts are fundamental to the MDL philosophy. Then in Section 3 we discuss how the data should be encoded based on a hypothesis. This requires that we outline the close relationship between codes and probability distributions, linking information theory to statistics. We then focus on the special case that the hypothesis is formulated in terms of a *model*, which is a set of codes or probability distributions. We describe different methods of constructing *universal codes* to represent such models, which are designed to achieve low regret in the worst case. After describing the process of designing codes, in Section 4 we will try to convince you that it may be useful to try and compress your data, even if you are really interested in something different, such as truth finding or prediction of future events, which, arguably, are the primary goals of statistics. Finally in Section 5 we place Minimum Description Length in the context of other interpretations of learning.

2 ENCODING THE HYPOTHESIS: A MATTER OF LUCKINESS AND REGRET

The intuition behind the MDL principle is that “useful” hypotheses should help compress the data. For now, we will simply define “useful” to be the same as “useful to compress the data”. We postpone further discussion of what this means to Section 4. What remains is the task to find out which hypotheses are useful, by using them to compress the data.

Given many hypotheses, we could just test them one at a time on the available data until we find one that happens to allow for substantial compression. However, if we were to adopt such a methodology in practice, results would vary from reasonable to extremely bad. The reason is that among so many hypotheses, there could well be one that, by sheer force of luck, allows for significant compression of the data, even though it would probably do a very bad job of predicting future outcomes. This is the phenomenon of overfitting again, which we mentioned in Example 1. To avoid such pitfalls, we required that a *single* code L_{mdl} is proposed on the basis of all available hypotheses. The code has two parts: the first part, with length function $L(H)$, identifies a hypothesis to be used to encode the data,

while the second part, with length function $L_H(D)$, describes the data using the code associated with that hypothesis.

Intuitively, the smaller $L_H(D)$, the better H fits data D . For example, if $\mathcal{H}$ is a set of probability distributions, then the higher the likelihood that H achieves on D , the smaller L_H . Section 3 is concerned with the precise definition of $L_H(D)$. Until then, we will assume that we have already represented our hypotheses in the form of codes, and we will discuss some of the properties a good choice for $L(H)$ should have. Technically, throughout this text we use only length functions that correspond to prefix codes; we will explain what this means in Section 3.1.

2.1 Avoid Regret

Consider the case where the best candidate hypothesis, i.e. the hypothesis $\hat{H} \in \mathcal{H}$ that minimises $L_H(D)$, achieves substantial compression. It would be a pity if we did not discover the usefulness of $\hat{H}$ because we chose a code word with unnecessarily long length $L(\hat{H})$. The *regret* we incur on the data quantifies how bad this “detection overhead” can be, by comparing the total code length $L_{\text{mdl}}(D)$ to the code length achieved by the best hypothesis $L_{\hat{H}}(D)$. A general definition, which also applies if $\hat{H}$ is undefined, is the following: the *regret* of a code L on data D with respect to a set of alternative codes $\mathcal{M}$ is¹

$$(4) \quad \mathcal{R}(L, \mathcal{M}, D) := L(D) - \inf_{L' \in \mathcal{M}} L'(D).$$

The reasoning is now that, since we do not want to make a priori assumptions as to the process that generates the data, the code for the hypotheses $L(H)$ must be chosen such that the regret $\mathcal{R}(L_{\text{mdl}}, \{L_H \mid H \in \mathcal{H}\}, D)$ is small, *whatever data we observe*. This ensures that whenever $\mathcal{H}$ contains a useful hypothesis that allows for substantial compression of the data (more than this small regret), we detect this because L_{mdl} compresses the data as well.

EXAMPLE 2. Suppose that $\mathcal{H}$ is finite. Let L be a uniform code that maps every hypothesis to a binary sequence of length $l = \lceil \log_2 |\mathcal{H}| \rceil$. The regret incurred by this uniform code is always exactly l , whatever data we observe. This is the best possible worst-case guarantee: all other length functions L' on $\mathcal{H}$ incur a strictly larger regret for at least one possible outcome (unless $\mathcal{H}$ contains useless hypotheses H which have $L(H) + L_H(D) > L_{\text{mdl}}(D)$ for all possible D). In other words, the uniform code minimises the *worst-case regret* $\max_D \mathcal{R}(L, \mathcal{M}, D)$ among all code length functions L . (We discuss the exact conditions we impose on code length functions in Section 3.1.)

Thus, if we consider a finite number of hypotheses we can use MDL with a uniform code $L(H)$. Since in this case the $L(H)$ term is the same for all hypotheses, it cancels and we find $H_{\text{mdl}} = \hat{H}$ in this case. What we have gained from this possibly anticlimactic analysis is the following *sanity check*: since we equated

¹In this text $\inf$ and $\sup$ are used for infimum and supremum, generalisations of minimum and maximum respectively.

learning with compression, we should not trust $\hat{H}$, the hypothesis that best fits the data, to exhibit good performance on future data unless we were able to compress the data using H_{mdl} . $\diamond$

2.2 Try to Get Lucky

There are many codes L on $\mathcal{H}$ that guarantee small regret, so the next task is to decide which we should pick. As it turns out, given any particular code L , it is possible to select a special subset of hypotheses $\mathcal{H}' \subset \mathcal{H}$ and modify the code such that the code lengths for these special hypotheses are especially small, at the cost of increasing the code lengths for the other hypotheses only slightly. This can be desirable, because *if* a hypothesis in $\mathcal{H}'$ turns out to be useful, then we are lucky, and we can identify that hypothesis more readily to achieve superior compression. On the other hand, if all hypotheses in the special subset are poor, then we have not lost much. Thus, while a suitable code must always have small regret, there is quite a lot of freedom to favour such small subsets of special hypotheses.

Examples of this so-called *luckiness principle* are found throughout the MDL literature, although they usually remain implicit, possibly because it makes MDL inference appear subjective. Only recently has the luckiness principle been identified as an important part of code design. The concept of luckiness is introduced to the MDL literature in [Grünwald, 2007]; [Barron, 1998] already uses a similar concept but not under the same name. We take the stance that the luckiness principle introduces only a mild form of subjectivity, because it cannot substantially harm inference performance. Paradoxically, it can only really be harmful *not* to apply the luckiness principle, because that could cause us to miss out on some good opportunities for learning!

EXAMPLE 3. To illustrate the luckiness principle, we return to the grammar learning of Example 1. We will not modify the code for the data $L_H(D)$ defined there; we will only reconsider $L(H)$ that intuitively seemed a reasonable code for the specification of grammars. Note that $L(H)$ is in fact a luckiness code: it assigns significantly shorter code lengths to shorter grammars. What would happen if we tried to avoid this “subjective” property by optimising the worst-case regret without considering luckiness?

To keep things simple, we reduce the hypothesis space to finite size by considering only context-free grammars with at most $|N| = 20$ nonterminals, $|\Sigma| = 490$ terminals, $R = 100$ rules and replacement sequences summing to a total length of $L = 2000$. Using (3) we can calculate the luckiness code length for such a grammar as at most $101\lceil\log(21)\rceil + 2100\lceil\log(511)\rceil = 19405$ bits.

However, Example 2 shows that for finite $\mathcal{H}$, the worst-case regret is minimised for the uniform code. To calculate the uniform code length, we first count the number of possible grammars $H' \subset H$ within the given size constraints. One may

or may not want to verify that

$$|\mathcal{H}'| = \sum_{r=1}^R |N|^r \sum_{l=0}^L |N \cup \Sigma|^l \binom{l+r-1}{l}.$$

Calculation on the computer reveals that $\lceil \log |\mathcal{H}'| \rceil = 18992$ bits. Thus, we could compress the data 413 bits better than the code that we used before. This shows that there is room for improvement of the luckiness code, which may incur a larger regret than necessary. On the other hand, with the uniform code we *always* need 18992 bits to encode the grammar, even when the grammar is very short! Suppose that the best grammar uses only $r = 10$ and $s = 100$, then the luckiness code requires only $11 \lceil \log(21) \rceil + 110 \lceil \log(511) \rceil = 1045$ bits to describe that grammar and therefore out-compresses the uniform code by 17947 bits: this grammar is identified much more easily using the luckiness code.

A simple way to combine the advantages of both codes is to define a third code that uses one additional bit to specify whether the luckiness code or the uniform code is used. The regret of this third code is at most one bit more than minimal in the worst case, while many simple grammars can be encoded extra efficiently. This one bit is the “slight increase” that we referred to at the beginning of this section. $\diamond$

Note that we do not mean to imply that the hypotheses which get special luckiness treatment are *necessarily* more likely to be true than any other hypothesis. Rather, luckiness codes can be interpreted as saying that this or that special subset *might* be important, in which case we should like to know about it!

2.3 Infinitely Many Hypotheses

When $\mathcal{H}$ is countably infinite, there can be no upper bound on the lengths of the code words used to identify the hypotheses. Since any hypothesis *might* turn out to be the best one in the end, the worst-case regret is often infinite in this case. In order to retain MDL as a useful inference procedure, we are forced to embrace the luckiness principle. A good way to do this is to order the hypotheses such that H_1 is luckier than H_2 , H_2 is luckier than H_3 , and so on. We then need to consider only codes L for which the code lengths increase with the index of the hypothesis. This immediately gives a lower bound on the code lengths $L(H_n)$ for $n = 1, \dots$, because the nondecreasing code that has the shortest code word for hypothesis n is uniform on the set $\{H_1, \dots, H_n\}$ (and is unable to express hypotheses with higher indices). Thus $L(H_n) \geq \log |\{H_1, \dots, H_n\}| = \log n$. It is possible to define codes L with $L(H_n) = \log n + O(\log \log n)$, i.e. not much larger than this ideal. Rissanen describes one such code, called the “universal code for the integers”, in [Rissanen, 1983b] (where the restriction to monotonically increasing code word lengths is not interpreted as an application of the luckiness principle as we do here); in Example 5 we describe some codes for the natural numbers that are convenient in practical applications.

2.4 Ideal MDL

In MDL hypothesis selection as described above, it is perfectly well conceivable that the data generating process has a very simple structure, which nevertheless remains undetected because it is not represented by any of the considered hypotheses. For example, we may use hypothesis selection to determine the best Markov chain order for data which reads “110010010000111111...”, never suspecting that this is really just the beginning of the binary expansion of the number π . In “ideal MDL” such blind spots are avoided, by interpreting *any* code length function L_H that can be implemented in the form of a computer program as the formal representation of a hypothesis H . Fix a universal prefix Turing machine U . With some slight caveats, computer languages such as Java, C or LISP can be thought of as universal prefix mechanisms [Grünwald and Vitányi, 2008]. The result of running program T on U with input D is denoted $U(T, D)$. The *Kolmogorov complexity* of a hypothesis H , denoted $K(H)$, is the length of the shortest program T_H that implements L_H , i.e. $U(T_H, D) = L_H(D)$ for all binary sequences D . Now, the hypothesis H can be encoded by literally listing the program T_H , so that the code length of the hypotheses becomes the Kolmogorov complexity. For a thorough introduction to Kolmogorov complexity, see [M.Li and Vitányi, 2008].

In the literature the term “ideal MDL” is used for a number of approaches to model selection based on Kolmogorov complexity; for more information on the version described here, refer to [Barron and Cover, 1991; Grünwald and Vitányi, 2008]. The version of ideal MDL that we adopt here tells us to pick

$$(5) \quad \min_{H \in \mathcal{H}} K(H) + L_H(D),$$

which is (1), except that now $\mathcal{H}$ is the set of *all* hypotheses represented by computable length functions, and $L(H) = K(H)$.

In order for this code to be in agreement with MDL philosophy as described above, we have to check whether or not it has small regret. It is also natural to wonder whether or not it somehow applies the luckiness principle. The following property of Kolmogorov complexity is relevant for the answer to both questions. Let $\mathcal{H}$ be a countable set of hypotheses with computable corresponding length functions. Then for all computable length functions L on $\mathcal{H}$, we have

$$(6) \quad \exists c > 0 : \forall H \in \mathcal{H} : K(H) \leq L(H) + c.$$

Roughly speaking, this means that ideal MDL is ideal in two respects: first, the set of considered hypotheses is expanded to include all computable hypotheses, so that any computable concept is learned given enough data; since with any inference procedure that can be implemented as a computer algorithm — as they all can — we can only infer computable concepts, this makes ideal MDL “universal” in a very strong sense. Second, it matches all other length functions up to a constant, including all length functions with small regret as well as length functions with *any* clever application of the luckiness principle. Thus, we may think of the Kolmogorov

Complexity code length $K(H)$ as implementing a sort of “universal” luckiness principle.

On the other hand, performance guarantees such as (6) are not very specific, as the constant overhead may be so large that it completely dwarfs the length of the data. To avoid this, we would need to specify a particular universal Turing machine U , and give specific upper bounds on the values that c can take for important choices of $\mathcal{H}$ and L . While there is some work on such a concrete definition of Kolmogorov complexity for individual objects [Tromp, 2007], there are as yet no concrete performance guarantees for ideal MDL or other forms of algorithmic inference.

The more fundamental reason why ideal MDL is not practical, is that Kolmogorov complexity is uncomputable. Thus it should be understood as a theoretical ideal that can serve as an inspiration for the development of methods that can be applied in practice.

3 ENCODING THE DATA

On page 866 we asked how the codes that formally represent the hypotheses should be constructed. Often many different interpretations are possible and it is a matter of judgement how exactly a hypothesis should be made precise. There is one important special case however, where the hypothesis is formulated in the form of a set of probability distributions. Statisticians call such a set a *model*. Possible models include the set of all normal distributions with any mean and variance, or the set of all third order Markov chains, and so on. Model selection is a prime application of MDL, but before we can discuss how the codes to represent models are chosen, we have to discuss the close relationship between coding and probability theory.

3.1 Codes and Probability Distributions

We have introduced the MDL principle in terms of coding; here we will make precise what we actually mean by a code and what properties we require our codes to have. We also make the connection to statistics by describing the correspondence between code length functions and probability distributions.

A *code* $C : \mathcal{X} \rightarrow \{0,1\}^*$ is an injective mapping from a countable source alphabet $\mathcal{X}$ to finite binary sequences called *code words*. We consider only *prefix codes*, that is, codes with the property that no code word is the prefix of another code word. This restriction ensures that the code is *uniquely decodable*, i.e. any concatenation of code words can be decoded into only one concatenation of source symbols. Furthermore, a prefix code has the practical advantage that no look-ahead is required for decoding, that is, given any concatenation S of code words, the code word boundaries in any prefix of S are determined by that prefix and do not depend on the remainder of S . Prefix codes are as efficient as other uniquely decodable codes; that is, for any uniquely decodable code with length function

L_C there is a prefix code C' with $L_{C'}(x) \leq L_C(x)$ for all $x \in \mathcal{X}$, see [Cover and Thomas, 1991, Chapter 5]. Since we never consider non-prefix codes, from now on, whenever we say “code”, this should be taken to mean “prefix code”.

Associated with a code C is a *length function* $L : \mathcal{X} \rightarrow \mathbb{N}$, which maps each source symbol $x \in \mathcal{X}$ to the length of its code word $C(x)$.

Of course we want to use efficient codes, but there is a limit to how short code words can be made. For example, there is only one binary sequence of length zero, two binary sequences of length one, four of length three, and so on. The precise limit is expressed by the Kraft inequality:

LEMMA 4 Kraft inequality. *Let $\mathcal{X}$ be a countable source alphabet. A function $L : \mathcal{X} \rightarrow \mathbb{N}$ is the length function of a prefix code on $\mathcal{X}$ if and only if:*

$$\sum_{x \in \mathcal{X}} 2^{-L(x)} \leq 1.$$

Proof. See for instance [Cover and Thomas, 1991, page 82]. ■

If, for a given code C with length function L , the inequality holds strictly, then the code is called *defective*, otherwise it is called *complete*. (The term “defective” is usually reserved for probability distributions, but we apply it to code length functions as well.)

Let C be any prefix code on $\mathcal{X}$ with length function L , and define

$$(7) \quad \forall x \in \mathcal{X} : \quad P(x) := 2^{-L(x)}.$$

Since $P(x)$ is always positive and sums to at most one, it can be interpreted as a probability mass function that defines a distribution corresponding to C . This mass function and distribution are called *complete* or *defective* if and only if C is.

Vice versa, given a distribution P , according to the Kraft inequality there must be a prefix code L satisfying

$$(8) \quad \forall x \in \mathcal{X} : \quad L(x) := \lceil -\log P(x) \rceil.$$

To further clarify the relationship between P and its corresponding L , define the *entropy* of distribution P on a countable outcome space $\mathcal{X}$ by

$$(9) \quad H(P) := \sum_{x \in \mathcal{X}} -P(x) \log P(x).$$

(This H should not be confused with the H used for hypotheses.) According to Shannon’s noiseless coding theorem [Shannon, 1948], the mean number of bits used to encode outcomes from P using the most efficient code is at least equal to the entropy, i.e. for all length functions L' of prefix codes, we have $E_P[L'(X)] \geq H(P)$. The expected code length using the L from (8) stays within one bit of entropy. This one bit is a consequence of the requirement that code lengths have to be integers. Note that apart from rounding, (7) and (8) describe a one-to-one correspondence

between probability distributions and code length functions that are most efficient for those distributions.

Technically, probability distributions are usually more convenient to work with than code length functions, because their usage does not involve rounding. But conceptually, code length functions are often more intuitive objects than probability distributions. A practical reason for this is that the probability of the observations typically decreases exponentially as the number of observations increases, and such small numbers are hard to handle psychologically, or even to plot in a graph. Code lengths typically grow linearly with the number of observations, and have an analogy in the real world, namely the effort required to remember all the obtained information, or the money spent on storage equipment. A second disadvantage of probability theory is the philosophical controversy regarding the interpretation of probability, which is discussed in Section 5.3.

In order to get the best of both worlds: the technical elegance of probability theory combined with the conceptual clarity of coding theory, we generalise the concept of coding such that code lengths are no longer necessarily integers. While the length functions associated with such “ideal codes” are really just alternative representations of probability mass functions (or even densities), and probability theory is used under the hood, we will nonetheless call negative logarithms of probabilities “code lengths” to aid our intuition and avoid confusion. Since this difference between an ideal code length and the length using a real code is at most one bit, this generalisation should not require too large a stretch of the imagination. In applications where it is important to actually encode the data, rather than just compute its code length, there is a practical technique called arithmetic coding [Rissanen, 1976] which can usually be applied to achieve the ideal code length to within a few bits; for the purposes of MDL inference it is sufficient to compute code lengths, and we do not have to construct actual codes.

EXAMPLE 5. In Section 2.3 we remarked that a good code for the natural numbers always achieves code length close to $\log n$. Consider a probability distribution W defined as $W(n) = f(n) - f(n+1)$ for some function $f : \mathbb{N} \rightarrow \mathbb{R}$. If f is decreasing with $f(1) = 1$ and $f \rightarrow 0$, then W is an easy to use probability mass function that can be used as a basis for coding the natural numbers. To see this, note that $\sum_{i=1}^m W(i) = 1 - f(n+1) < 1$ and that therefore $\sum_{i=1}^{\infty} W(i) = 1$. For $f(n) = 1/n$ we get code lengths $-\log W(n) = \log(n(n+1)) \approx 2 \log n$, which, depending on the application, may be small enough. Even more efficient for high n are $f(n) = n^{-\alpha}$ for some $0 < \alpha < 1$ or $f(n) = 1/\log(n+1)$. $\diamond$

3.2 Universal Coding and Model Selection

We have discussed earlier what code L should be used to identify the hypotheses. Here we address the question what code L_H should represent a hypothesis $H \in \mathcal{H}$. So far we have treated the data as a monolithic block of information, but it is often convenient to think of the data as a sequence of observations $D = x_1, x_2, \dots, x_n$ instead. For instance, this makes it easier to discuss sequential prediction later on.

For simplicity we will assume each observation x_i is from some countable outcome space $\mathcal{X}$; we also abbreviate $x^n = x_1, \dots, x_n$.

In case that the hypothesis H is given in terms of a probability distribution P from the outset, $H = P$, then we identify it with the code satisfying (8) but without the integer requirement, so that $L_H(x^n) = -\log P(x^n)$. There are various reasons why this is the only reasonable mapping [Grünwald, 2007]. Now no more work needs to be done and we can proceed straight to doing MDL hypothesis selection as described above.

In the remainder of this section, we consider instead the case where the hypothesis is given in the form of a *model* $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$, which is a set of probability distributions parameterised by a vector θ from some set Θ of allowed parameter vectors, called the *parameter space*. For example, a model could be the set of all fourth order Markov chains, or the set of all normal distributions. In this case it is not immediately clear which single code should represent the hypothesis.

To motivate the code that represents such a model, we apply the same reasoning as we used in Section 2. Namely, the code should guarantee a small regret, but is allowed to favour some small subsets of the model on the basis of the luckiness principle. To make this idea more precise, we define a *maximum likelihood estimator* as a function $\hat{\theta} : \mathcal{X}^* \rightarrow \Theta$ satisfying

$$(10) \quad P_{\hat{\theta}(x^n)}(x^n) = \max\{P_\theta(x^n) \mid \theta \in \Theta\}.$$

Obviously, $\hat{\theta}(x^n)$ also minimises the code length. We abbreviate $\hat{\theta} = \hat{\theta}(x^n)$ if the sequence of outcomes x^n is clear from context. We start by associating with each individual $P_\theta \in \mathcal{M}$ the corresponding code L_θ with $L_\theta(x^n) = -\log P_\theta(x^n)$, and henceforth write and think of $\mathcal{M}$ as a set of *codes* $\{L_\theta \mid \theta \in \Theta\}$.

DEFINITION 6. Let $\mathcal{M} := \{L_\theta \mid \theta \in \Theta\}$ be a (countable or uncountably infinite) model with parameter space Θ . Let $f : \Theta \times \mathbb{N} \rightarrow [0, \infty)$ be some function. A code L is called *f-universal* for a model $\mathcal{M}$ if, for all $n \in \mathbb{N}$, all $x^n \in \mathcal{X}^n$, we have

$$\mathcal{R}(L, \mathcal{M}, x^n) \leq f(\hat{\theta}, n).$$

Our notion of universality is called *individual-sequence* universality in the information-theoretic literature [Grünwald, 2007]. When information theorists talk of universal codes, they usually refer to another definition in terms of expectation rather than individual sequences [Cover and Thomas, 1991]. Our formulation is very general, with a function f that needs further specification. Generality is needed because, as it turns out, different degrees of universality are possible in different circumstances. This definition allows us to express easily what we expect a code to live up to in all these cases.

For finite $\mathcal{M}$, a uniform code similar to the one described in Example 2 achieves *f-universality* for $f(\hat{\theta}, n) = \log |\mathcal{M}|$. Of course we incur a small overhead on top of this if we decide to use luckiness codes.

For countably infinite $\mathcal{M}$, the regret cannot be bounded by a single constant, but we can avoid dependence on the sample size. Namely, if we define a mass function

W on the parameter set, we can achieve f -universality for $f(\hat{\theta}, n) = -\log W(\hat{\theta})$ by using a Bayesian or two-part code (these are explained in subsequent sections).

Finally for uncountably infinite $\mathcal{M}$ it is often impossible to obtain a regret bound that does not depend on n . For parametric models however it is often possible to achieve f -universality for

$$(11) \quad f(\hat{\theta}, n) = \frac{k}{2} \log \frac{n}{2\pi} + g(\hat{\theta}),$$

where k is the number of parameters of the model and $g : \Theta \rightarrow [0, \infty)$ is some continuous function of the maximum likelihood parameter. Examples include the Poisson model, which can be parameterised by the mean of the distribution so $k = 1$, and the normal model, which can be parameterised by mean and variance so $k = 2$. Thus, for parametric uncountable models a logarithmic dependence on the sample size is the norm.

In MDL model selection, universal codes are used on two levels. On the inner level, each model H is represented by a universal code for that model. Then on the outer level, a code with length function $L(H) + L_H(D)$ is always used, see (1). This code is also universal, this time with respect to $\{L_H \mid H \in \mathcal{H}\}$, which is typically countable. This outer-level code is called a *two-part code*, explained in more detail below. Thus, MDL model selection proceeds by selecting the H (a model, i.e. a family of distributions) that minimises the two-part code $L(H) + L_H(D)$, where $L_H(D)$ itself is a universal code relative to model H .

We now describe the four most common ways to construct universal codes, and we illustrate each code by applying it to the model of Bernoulli distributions $\mathcal{M} = \{P_\theta \mid \theta \in [0, 1]\}$. This is the “biased coin” model, which contains all possible distributions on 0 and 1. The distributions are parameterised by the probability of observing one: $P_\theta(X = 1) = \theta$. Thus, $\theta = \frac{1}{2}$ represents the distribution of an unbiased coin. The distributions in the model are extended to n outcomes by taking the n -fold product distribution: $P_\theta(x^n) = P_\theta(x_1) \cdots P_\theta(x_n)$.

3.2.1 Two-Part Codes

In Section 2 we defined a code L_{mdl} that we now understand to be universal for the model $\{L_H \mid H \in \mathcal{H}\}$. This kind of universal code is called a two-part code, because it consists first of a specification of an element of the model, and second of a specification of the data using that element. Two-part codes may be defective: this occurs if multiple code words represent the same data sequence x^T . In that case one must ensure that the encoding function is well-defined by specifying exactly which representation is associated with each data sequence. Since we are only concerned with code *lengths* however, it suffices to adopt the convention that we always use one of the shortest representations.

We mentioned that, in model selection problems, universal coding is applied on two levels. At the outer level, two-part coding is always used because it allows us to associate a single hypothesis with the achieved code length $L_{\text{mdl}}(x^n)$. In model selection problems, two-part coding *can* also be used at the inner level. An

example of a two-part code that can be used as a universal code on the inner level follows below. We will see later that on this inner level, other universal codes are sometimes preferable because they are more efficient and do not require discretisation.

EXAMPLE 7. We define a two-part code for the Bernoulli model. In the first part of the code we specify a parameter value, which requires some discretisation since the parameter space is uncountable. However, as the maximum likelihood parameter for the Bernoulli model is just the observed frequency of heads, at a sample size of n we know that the ML parameter is one of $0/n, 1/n, \dots, n/n$. We discretise by restricting the parameter space to this set. A uniform code uses $L(\theta) = \log(n+1)$ bits to identify an element of this set. Therefore the resulting regret is always exactly $\log(n+1)$. By using a slightly more clever discretisation we can bring this regret down to about $\frac{1}{2} \log n + O(1)$, which we mentioned is usually achievable for uncountable single parameter models. $\diamond$

3.2.2 The Bayesian Universal Distribution

Let $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$ be a countable model; it is convenient to use mass functions rather than codes as elements of the model here. Now define a distribution with mass function W on the parameter space Θ . This distribution is called the *prior distribution* in the literature as it is often interpreted as a representation of a priori beliefs as to which of the hypotheses in $\mathcal{M}$ represents the “true state of the world”. More in line with the philosophy outlined above would be the interpretation that W is a *code* which should be chosen for practical reasons to optimise inference performance. At any rate, the next step is to define a *joint distribution* P on $\mathcal{X}^n \times \Theta$ by $P(x^n, \theta) = P_\theta(x^n)W(\theta)$. In this joint space, each outcome comprises a particular state of the world and an observed outcome.

In Bayesian statistics, inference is always based on this joint distribution. For our purpose, the one relevant quantity is the *marginal likelihood* of the data:

$$(12) \quad P(x^n) = \sum_{\theta \in \Theta} P_\theta(x^n)W(\theta).$$

Note that the marginal likelihood is a weighted average, satisfying

$$(13) \quad P_{\hat{\theta}}(x^n)W(\hat{\theta}) \leq P(x^n) \leq P_{\hat{\theta}}(x^n),$$

where the first inequality is obtained by underestimating (12) by a single term of the sum, and the second by overestimating by the largest item the weighted average is taken over. Together these bounds express that the Bayesian marginal likelihood is at most a factor $W(\hat{\theta})$ worse than the best element of the model, which is not very much considering that $P_{\hat{\theta}}(x^n)$ is often exponentially small in n .

Note that this Bayesian universal code yields a code length less than or equal to the code length we would have obtained with the two-part code with $L(\theta) = -\log W(\theta)$. Since we already found that two-part codes are universal, we can now conclude that Bayesian codes are at least as universal, with regret *at most*

$-\log W(\hat{\theta})$. On the flip side, the sum involved in calculating the Bayesian marginal distribution can be hard to evaluate in practice.

EXAMPLE 7 continued. Our definitions readily generalise to uncountable models with $\Theta \subseteq \mathbb{R}^k$, with the prior distribution given by a density w on Θ . Rather than giving explicit definitions we revisit our running example.

We construct a Bayesian universal code for the Bernoulli model. For simplicity we use a uniform prior density, $w(\theta) = 1$. Let n_0 and $n_1 = n - n_0$ denote the number of zeroes and ones in x^n , respectively. Now we can calculate the Bayes marginal likelihood of the data:

$$P(x^n) = \int_0^1 P_\theta(x^n) \cdot 1 \, d\theta = \int_0^1 \theta^{n_1} (1 - \theta)^{n_0} \, d\theta = \frac{n_0! n_1!}{(n+1)!}.$$

Using Stirling's approximation of the factorial, we find that the corresponding code length $-\log P(x^n)$ equals $-\log P_{\hat{\theta}}(x^n) + \frac{1}{2} \log n + O(1)$. Thus we find a regret similar to that achieved by a well-designed two-part code, but the constant will be slightly better. $\diamond$

3.2.3 Normalised Maximum Likelihood

The Normalised Maximum Likelihood universal code is preferred in the MDL literature, because it provably minimises the worst-case regret. It is not hard to show that the code minimising the worst-case regret must achieve equal regret for all possible outcomes x^n . In other words, the total code length must always be some constant longer than the code length achieved on the basis of the maximum likelihood estimator [Grünwald, 2007, Chapter 6]. The Normalised Maximum Likelihood (NML) distribution is defined to achieve exactly this:

$$(14) \quad P_{\text{nml}}(x^n) := \frac{P_{\hat{\theta}(x^n)}(x^n)}{\sum_{y^n \in \mathcal{X}^n} P_{\hat{\theta}(y^n)}(y^n)}.$$

For all outcomes x^n , the regret on the basis of this distribution is exactly equal to the logarithm of the denominator, called the *parametric complexity* of the model:

$$(15) \quad \inf_L \sup_{x^n} \mathcal{R}(L, \mathcal{M}, x^n) = \log \sum_{y^n \in \mathcal{X}^n} P_{\hat{\theta}(y^n)}(y^n).$$

Under some regularity conditions on the model it can be shown [Rissanen, 1996; Grünwald, 2007] that, if we restrict the parameters to an *ineccsi* set $\Theta_0 \subset \Theta$, i.e. a subset of Θ that is closed, bounded, has nonempty interior and excludes the boundary of Θ , then the parametric complexity is finite and given by

$$(16) \quad \frac{k}{2} \log \frac{n}{2\pi} + \log \int_{\Theta_0} \sqrt{\det I(\theta)} \, d\theta + o(1),$$

in accordance with (11) for constant function $g(\hat{\theta})$. Here $\det I(\theta)$ is the determinant of the Fisher information matrix, a fundamental quantity in statistics (see

[Grünwald, 2007, Chapter 6] for a precise statement). This is an important result: for example, it implies that we can never hope to construct universal codes with regret smaller than (16) on any substantial subset Θ_0 of Θ . Thus, if the parametric complexity is large, then with any coding method whatsoever we need a lot of bits to code the data, on top of the best-fitting element of the model. The model is thus “complex”, independently of the particular description method we use. We may also look at this in a different way: we see from the sum in (15) that a model is complex iff for many data y^n sequences, it contains a distribution that fits these sequences well in the sense that $P_{\hat{\theta}(y^n)}(y^n)$ is large. Thus, the parametric complexity of a model is related not so much to how many distributions it contains, but to how many patterns can be fit well by a distribution in it. It can be shown [Grünwald, 2007, Chapter 6] that this is related to the number of “essentially different” distributions the model contains.

We note that neither the parametric complexity (15) nor its approximation (16) depend on the particular chosen parameterisation, i.e. the function which maps θ to P_θ . While the parameterisation is arbitrary, the parametric complexity and its approximation are inherent properties of the model $\mathcal{M}$ under consideration.

EXAMPLE 7 continued. The parametric complexity (15) has exponentially many terms, but for the Bernoulli model the expression can be significantly simplified. Namely, we can group together all terms which have the same maximum likelihood estimator. Thus the minimal worst-case regret can be rewritten as follows:

$$(17) \quad \log \sum_{y^n \in \mathcal{X}^n} P_{\hat{\theta}(y^n)}(y^n) = \log \sum_{n_1=0}^n \binom{n}{n_1} \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n-n_1}{n}\right)^{n-n_1}.$$

This term has only linearly many terms and can usually be evaluated in practice. Approximation by Stirling’s formula confirms that the asymptotic regret is $\frac{1}{2} \log n + O(1)$, the same as for the other universal distributions. $\diamond$

The NML distribution has a number of significant practical problems. First, it is often undefined, because for many models the denominator in (14) is infinite, even for such simple models as the Poisson or geometric distributions. Second, $\mathcal{X}^n$ is exponentially large in n , so calculating the NML probability exactly is only possible in special cases such as the Bernoulli model above, where the number of terms is reduced using some trick. Something similar is possible for the more general multinomial model, see [Kontkanen and Myllymäki, 2007], but in most cases (15) has to be approximated, which introduces errors that are hard to quantify.

3.2.4 Prequential Universal Distributions and Predictive MDL

A distribution P (or its corresponding code), defined for a sequence of outcomes $x^n := x_1, x_2, \dots, x_n$ can be reinterpreted as a *sequential (probabilistic) prediction strategy* using the chain rule:

$$(18) \quad \begin{aligned} P(x^n) &= P(x^1) \cdot \frac{P(x^2)}{P(x^1)} \cdot \dots \cdot \frac{P(x^n)}{P(x^{n-1})} \\ &= P(x_1) \cdot P(x_2|x^1) \cdot \dots \cdot P(x_n|x^{n-1}). \end{aligned}$$

The individual factors are the actual probabilities $P(x_1), P(x_2 | x^1), \dots$ assigned to $x_1, x_2, \dots$ by the conditional “predictive” probability distributions $P(X_1 = \cdot | x_1), P(X_2 = \cdot | x^2), \dots$. Thus, to any distribution P for the data, there corresponds a sequential prediction strategy, where we predict x_1 by marginalising P before having seen any data, then we look at x_1 , condition P accordingly and marginalise again to obtain a prediction of x_2 , and so on. The chain rule shows that the probability of x^n must be equal to the product of the individual predictive probabilities. Vice versa, any prediction strategy that will issue a distribution on the next outcome x_{i+1} given any sequence of previous outcomes x^i , defines a distribution (or code) on the whole data via the chain rule. Together with the correspondence between code length functions and probability distributions described above, this links the domains of coding and sequential prediction.

An algorithm that, given a sequence of previous observations x^n , issues a probability distribution on the next outcome $P(X_{n+1} | x^n)$, may be viewed as a sequential (probabilistic) prediction strategy. In Dawid’s theory of prequential analysis such algorithms are called *prequential forecasting systems* [Dawid, 1984; Dawid, 1992b; Dawid, 1992a]. A prediction strategy defines a distribution on x^n by the chain rule (18). Vice versa, for any distribution P on $\mathcal{X}^n$, application of the chain rule yields a prediction strategy. We give two important prediction strategies here.

First, we may take a Bayesian approach and define a joint distribution on $\mathcal{X}^n \times \Theta$ based on some prior distribution W . As before, this induces a marginal distribution on $\mathcal{X}^n$ which in turn defines a prediction strategy. Thus, the Bayesian universal distribution can be reinterpreted as a prediction strategy: it is in fact an important special case of a prequential universal distribution.

Second, since the Bayesian predictive distribution can be hard to compute, it may be useful in practice to define prediction strategies that use simpler algorithms. Perhaps the simplest option is to predict an outcome X_{n+1} using a maximum likelihood estimator (10) evaluated on the previous outcomes. This “prequential maximum likelihood plug-in” approach was suggested independently by Dawid and by Rissanen who calls it *predictive MDL* [Rissanen, 1984; Rissanen, 1986a], and we will use it in our running example.

EXAMPLE 7 continued. The ML estimator for the Bernoulli model parameterised by $P_\theta(X = 1) = \theta$, equals the frequency of ones in the sample: $\hat{\theta}(x^n) = \frac{1}{n} \sum_{n=1}^n x_n$. We define a prediction strategy through $P(X_{n+1} = 1 | x^n) := P_{\hat{\theta}(x^n)}(X_{n+1} = 1) = \hat{\theta}(x^n)$.

This strategy is ill-defined for the first outcome. Another impractical feature is that it assigns probability 0 to the event that the second outcome is different from the first. To address these problems, we slightly tweak the estimator: rather than $\hat{\theta}$ we use $\tilde{\theta} = (1 + \sum_{n=1}^n x_n) / (n+2)$. Perhaps surprisingly, in this case the resulting prediction strategy is equivalent to the Bayesian universal distribution approach we defined in the previous section: $P_{\tilde{\theta}}$ turns out to be the Bayesian predictive distribution for the Bernoulli model if a uniform prior density $w(\theta) = 1$ is used. This is somewhat coincidental: for non-Bernoulli models, the distribution indexed

by such a “tweaked” ML estimator is usually quite different from the Bayesian predictive distribution [Grünwald, 2007, Chapter 9]. $\diamond$

The ML plug-in code is useful because it is often easier to implement than the other universal codes: it does not require discretisation, like the 2-part code, nor integration over the parameter space, like the Bayesian code, nor summation over all possible data sequences, like NML. Instead it only requires calculation of the ML estimator.

While the prequential ML code has been used successfully in practical inference problems, it is shown in [de Rooij and Grünwald, 2006] that (in expectation) it does not necessarily achieve the desired regret (11) of $(k/2) \log n + O(1)$ unless the data are actually sampled from a distribution in the model. Application of the ML plug-in code thus requires exactly the kind of assumption about the “truth” that we have been trying to avoid. Therefore the ML plug-in code should be used with care, especially in applications of model selection.

Another feature of prequential universal codes is that in general they are sensitive to the ordering of the data, which may be undesirable in case the data have no natural ordering. Interestingly, this does not apply to the Bayesian sequential prediction strategy, which is order-independent in i.i.d. settings. That is: if we change the order of the data x^n , then the individual predictions of x_i given x^{i-1} change, but the product (18) of these probabilistic predictions remains the same.

Unordered Data When prequential universal codes are defined for models that assume the data to be unordered (the most important case being i.i.d. data), the result often assigns different code lengths to different permutations of the data sequence, even if all codes that make up the models do not. Fortunately, the Bayesian prequential code does not suffer this effect. Neither does it occur for the 2-part and NML universal codes, which are not prequential. However the ML plug-in code length generally does depend on the order of the data. Also, several methods to turn NML codes into prediction strategies are studied in the literature [Rissanen and Roos, 2007]; the resulting prequential codes do typically depend on the order of the data. Similarly, recent attention to *model switching* (Section 5.5) has resulted in codes that introduce order dependence.

If the data are unordered, should we not encode the data as a *set* rather than as a *sequence*, in other words should we not try to avoid encoding the (arbitrary) order altogether? We do not favour such an approach, because once the order of the data is disregarded, we can no longer compare the code lengths we obtain to those we would get using other hypotheses that do assume order dependence! It would prevent us from testing our *assumption* that the data are unordered. A second, more practical argument against disregarding the order is that it is often easier to encode sequences than it is to encode sets. As long as all permutations of the sequence are assigned equal code lengths, no bias is introduced when comparing code lengths for sequences rather than sets.

However, as we mentioned many prequential codes do depend on the order. This means that we have to make sure that these codes still achieve low regret in

the worst case, i.e. that their performance does not deteriorate too much when the data appear in a particularly unlucky order. Unfortunately, for many prequential codes such worst-case results have not been proven, and we have to make do with performance guarantees either in expectation or with high probability. While this issue does not appear cause too much trouble in practice, it is slightly at odds with our interpretation of MDL philosophy.

4 THE PURPOSE OF MDL

While we have equated “learning” with “compressing the data” so far, in reality we often have a more specific goal in mind when we apply learning algorithms. This touches on the third question we asked on page 866: what do we mean by a “useful” hypothesis? In this section we will argue that MDL inference, while designed to achieve optimal compression of the data, is also suitable in some settings with a different objective, so that, at least to some extent, MDL provides a reliable one-size-fits-all solution.

EXAMPLE 1 continued. We return to grammar learning. The code we used for the data given the grammar is not very sophisticated. A better approach is the following: to encode each sentence, we start with the starting symbol, and then we repeatedly specify which production rule is to be applied to the first remaining nonterminal, until none are left. We still have to choose a code/distribution on the rules that match every nonterminal, to be able to specify which matching rule is applied. Of course, in practice, some production rules will be used much more often than others, and by specifying the “right” probabilities (approximately matching empirical frequencies) for each rule, we may achieve substantial compression of the data. Such a grammar, where each nonterminal has a distribution on the matching production rules is called a *stochastic context-free grammar*.

We generally do not know how we should set the probabilities to obtain optimal code lengths. However, we now have the tools to minimise the worst-case regret. Namely, each grammar defines a *model* on sentences, a set of distributions that is parameterised by the probabilities of applying production rules to matching nonterminals. For each model, we can subsequently define one of the universal codes of Section 3.2. This results in a code for each grammar that performs *almost as well* as the stochastic context-free grammar with optimally tuned probabilities of the production rules.

Each of the applications of MDL we describe below can be usefully applied to stochastic context free grammars. (1) We may use the two-part code we defined for grammar and data to *compress* the data (Section 4.1). (2) We may use the grammar to *predict* the continuation of the data (Section 4.3). Language models are used for prediction in applications such as speech recognition. (3) We may infer a grammar that matches the data well, without attempting to learn the model parameters as well. We then obtain a *model selection* problem (Section 4.2.2). Or alternatively (4) we may attempt to learn not only the grammar, but the

parameters as well. This is the *estimation problem* of Section 4.2.1. Note that in these last two cases, theoretical consistency results only carry over to real data in case those data are really sampled from a stochastic context free grammar, which may be unrealistic. $\diamond$

4.1 Data Compression

Until now we have considered MDL for compression, which is an important practical goal in itself. A substantial amount of progress in the data compression literature builds on MDL research. For example, the elegant Context Tree Weighting algorithm [Willems *et al.*, 1997] expands on earlier MDL-related work by Rissanen [Rissanen, 1983a]. The following subsections discuss in detail how compression is related to learning from data, first in the form of “truth finding”, which includes estimation and model selection, and then in the form of prediction. The MDL Principle is more general though: at least in principle, it can be applied to all problems involving inductive inference, not just compression, prediction and truth finding. For example, it is eminently suited for denoising and separating structure from noise, and it has also been used for similarity analysis, clustering, lossy data compression and DNA sequence alignment. We will not go into these applications here; for a non-exhaustive list with references, see [Grünwald, 2007, Chapter 1].

4.2 Truth Finding

Let $\mathcal{H}$ be a set of probability distributions, and let x^n be the observed data. The MDL principle provides a clear interpretation of learning without the assumption that the data are sampled from one of the distributions in $\mathcal{H}$, or indeed without any probabilistic assumptions at all. It can be applied in situations in which some $P \in \mathcal{H}$ is “adequate” (in the sense that it leads to valid insights) or “useful” (in the sense that it leads to reasonable predictions of future data; such P may be quite far off from the underlying “truth”). *Still*, as a sanity check, we want to make sure that MDL methods provably perform well in the special case in which some $P^* \in \mathcal{H}$ is true, in the sense that the data are sampled from P^* . Below we verify that MDL methods indeed behave reasonably in such ideal circumstances. First, in Section 4.2.1, we consider the case where the goal is simply to infer a good approximation of P^* . This includes the traditional statistical problems of parametric and nonparametric density estimation. In that situation, MDL performs well in the sense that, with high probability, the MDL methods output a good approximation to the true P^* even at modest sample sizes. In Section 4.2.2 we consider model selection, where $\mathcal{H}$ is carved up in finite or countably many parametric models $\mathcal{M}_1, \mathcal{M}_2, \dots$, and the goal is to identify the (smallest) $\mathcal{M}_{\delta^*}$ containing the true P^* . In that situation, MDL performs well in the sense that, with high probability, when input a large enough sample, the MDL methods outputs $\mathcal{M}_{\delta^*}$. Finally, in Section 4.3, we very briefly consider a version of truth finding more closely related to prediction.

4.2.1 Estimation

Here we assume that the true distribution P^* lies in the hypothesis space $\mathcal{H}$, and investigate whether we can approximate P^* using MDL. The hypothesis space $\mathcal{H}$ may be a parametric model (e.g., the set of Bernoulli distributions, or the set of second-order Markov chains); in that case our goal reduces to a most familiar topic of statistics: parameter estimation. Alternatively, $\mathcal{H}$ may be so large that it cannot be indexed by a finite set of real-valued parameters. An example is the set of all distributions for $0 < X < 1$ that have differentiable densities. In that case, our goal is what statisticians call “nonparametric density estimation”. Yet another possibility is that $\mathcal{H} = \bigcup_{k=1,2,\dots} \mathcal{M}_k$, where each $\mathcal{M}_k$ is a k -dimensional parametric model. In this last scenario we assume the truth to be “parametric”, but the dimension is unknown. The difference with model selection proper, which is discussed further below, is that in estimation we are interested in inferring a single element P^* residing in some of the $\mathcal{M}_k$, whereas in model selection we only try to find out which $\mathcal{M}_k$ contains P^* .

For all such estimation problems we use a two-part code; we first encode an element P of $\mathcal{H}$ using a code length function denoted L , and then we encode the data using the code that corresponds to P . In total we need $L(P) - \log P(x^n)$ bits to encode x^n via P . We then select the $\tilde{P} \in \mathcal{H}$ that minimises this total. This $\tilde{P}$ may be viewed as an estimator in the statistical sense: it maps data sequences of arbitrary length to distributions in $\mathcal{H}$. In general, the performance of statistical estimators is often expressed in terms of their *statistical risk*, the expected distance between $\tilde{P}$ and the true P^* . For technical reasons it is convenient to measure the risk in terms of the *squared Hellinger distance* D_{He} , a distance measure between probability distributions (see [Grünwald, 2007] for a definition). With respect to this distance, the risk of the 2-part MDL estimator relative to true distribution P^* at sample size n is defined as

$$(19) \quad R_n(P^*, \tilde{P}) = E_{X^n \sim P^*} [D_{\text{He}}(P^*, \tilde{P})].$$

Note that $\tilde{P}$ is really a function of the X^n over which we take expectation. As more and more data are gathered, the 2-part MDL estimator tends to get better and better in the sense that the risk tends to get smaller and smaller. The *rate of convergence* expresses how quickly the risk converges to 0. We have the following:

THEOREM 8. *Suppose that, for all $P^* \in \mathcal{H}$, data $X_1, X_2, \dots$ are i.i.d. under P^* . Then, under a mild regularity condition on the code length function $L(H)$, we have, for all $P^* \in \mathcal{H}$,*

$$(20) \quad R_n(P^*, \tilde{P}) \leq \frac{E_{X^n \sim P^*} [L(\tilde{P}) - \log \tilde{P}(X^n) - [-\log P^*(X^n)]]}{n}.$$

In words: *the statistical risk of two-part code MDL relative to P^* is bounded by the expected overhead of the two-part code relative to P^* , divided by n .* This provides

a deep link between data compression and truth finding: *if* our code happens to achieve a small code length using some approximation of the true distribution P^* (i.e. the right-hand side of (20) is small), then the expected performance of the two-part estimator is very good (i.e., the left-hand side of (20) is small). For example, if P^* has a finite code length $L(P^*)$ under the code L , then the right-hand side of (20) is bounded by

$$\begin{aligned} & n^{-1}[L(\ddot{P}) - \log \ddot{P}(X^n) - [-\log P^*(X^n)]] \\ \leq & n^{-1}[L(P^*) - \log P^*(X^n) + \log P^*(X^n)] = L(P^*)/n. \end{aligned}$$

Thus the risk converges to 0 at least at rate $1/n$, which, as compared to other statistical estimators, is quite fast. Note the role of luckiness: the smaller the code length we assigned to (good approximations of) P^* , the better the performance bound on two-part MDL. In many cases, the structure of $\mathcal{H}$ is such that we can design codes with good worst-case behaviour, even if $\mathcal{H}$ is uncountable. For example, if $\mathcal{H}$ is a parametric model and if a two-part code with a clever discretisation scheme is used, then it can be shown using (20) that $\max_{P^* \in \mathcal{H}} R_n(P^*, \ddot{P}) = O(1/n)$. This is the *minimax optimal rate*, i.e. in the worst case over all $P^* \in \mathcal{H}$, no statistical estimation method can have rate smaller than c/n for some $c > 0$. In the same way one can use (20) to show, for suitably chosen L , that two-part MDL achieves the minimax convergence rates for a range of nonparametric problems. Compare this to the classical maximum likelihood estimation method, which also achieves the minimax rate on parametric problems, but which may fail utterly on nonparametric problems.

Theorem 8 is due (in slightly different versions) to A. Barron, J. Li and T. Zhang, and appeared first in Li's Ph.D. thesis [Li, 1999]. Its precise statement and implications are discussed at length in Chapter 15 of [Grünwald, 2007]. It is of fundamental importance, since it directly links compression and learning: the better we can compress data from P^* , the faster we can identify a good approximation of P^* .

4.2.2 Model Selection

We have described models as formal representations of hypotheses; we will now try to determine which of the hypotheses is “true”, in the sense that the corresponding model contains the data generating distribution P^* . The goal is then to identify this true model on the basis of as little data as possible. In the MDL approach to this problem, we use a two-part code on the outer level (since, based on data x^n , we want to output a particular model $\mathcal{M}$, which therefore has to be encoded explicitly), and we can use any type of universal code at the inner level.

Since the universal code for the true model achieves a code length not much larger than the code length of the best code in the model (it was designed to achieve small regret), and the best code in the model achieves code length at most as large as the data generating distribution, it seems reasonable to assume that, as more data are being gathered, a true model will eventually be selected by MDL.

This intuition is confirmed in the form of the following consistency result, which is one of the pillars of MDL model selection. The result applies to Bayesian model selection as well.

We give a precise statement of the theorem, but also provide a more informal explanation for those unfamiliar with all the concepts involved.

THEOREM 9 Model Selection Consistency. *Let $\mathcal{H} = \{\mathcal{M}_1, \mathcal{M}_2, \dots\}$ be a countably infinite set of parametric models. For all $i \in \mathbb{Z}^+$, let $\mathcal{M}_i$ be 1-to-1 parameterised by $\Theta_i \subseteq \mathbb{R}^k$ for some $k \in \mathbb{N}$ and define a prior density w_i on Θ_i . Define a prior distribution W on the model indices. We require for all integers $j > i > 0$ that with w_j -probability 1, a distribution drawn from $\mathcal{M}_j$ is mutually singular with all distributions in $\mathcal{M}_i$. Define the MDL model selection criterion based on Bayesian universal codes to represent the models:*

$$\delta(x^n) = \arg \min_i \left(-\log W(i) - \log \int_{\theta \in \Theta_i} P_\theta(x^n) w_i(\theta) d\theta \right).$$

Then for all $\delta^* \in \mathbb{Z}^+$, for all $\theta^* \in \Theta_{\delta^*}$, except for a subset of Θ_{δ^*} of Lebesgue measure 0, it holds for $X_1, X_2, \dots \sim P_{\theta^*}$ that

$$P_{\theta^*}(\exists n_0 \forall n \geq n_0 : \delta(X^n) = \delta^*) = 1.$$

Proof. Proofs of various versions of this theorem can be found in [Barron, 1985; Dawid, 1992a; Barron *et al.*, 1998; Grünwald, 2007] and others. ■

Thus, the theorem can be applied to a sequence of models, as long as models do not contain a significant number of distributions that also occur in earlier models. The δ -function in the theorem indicates which model is selected by MDL given a specific sequence of observations. The theorem then states, roughly, that from some sample size n_0 onwards the true model will be always be selected by MDL.

The theorem uses a Bayesian code on the inner level (the distributions within each model). However as conjectured in [Grünwald, 2007], it can probably be extended to other universal codes on the inner level, such as NML (for some special cases, it was proven in [van Erven *et al.*, 2008] that this is indeed the case). Note that (as we keep emphasising) we do not make any assumptions about the true distribution P^* . However, if we are in the ideal situation where one of the models contains P^* , then the theorem guarantees that this model will be selected once sufficient data have been accumulated. This property of model selection criteria is called *consistency*. Unfortunately, the theorem says nothing about the more realistic scenario where the models are merely approximations of the true data generating process. If the true data generating process P^* is not an element of any of the models, we might still be interested in finding the best approximating model, for instance the model that contains the distribution $P_{\hat{\theta}}$ minimising the Kullback-Leibler divergence $D(P^* \| P_{\hat{\theta}})$. It is not known under what circumstances δ actually achieves this. On the other hand, some model

selection criteria that are often used in practice (such as maximum likelihood model selection, or AIC [Akaike, 1974]) are known *not* to be consistent, even in the restricted sense of Theorem 9. Therefore consistency of MDL and Bayesian model selection is reassuring.

At this point, we should add that MDL model selection, while being consistent, does not always achieve the minimax optimal convergence rate, in the following sense: suppose we first use MDL model selection to select a model $\mathcal{M}_k$, and then, within the chosen model, we use a second estimator (which may again be a two-part code MDL estimator), to select a distribution $\hat{P}_\theta \in \mathcal{M}_k$. Then, for some standard situations such as linear regression, using MDL model selection in the first stage, the Hellinger risk $R_n(P^*, \hat{P}_\theta)$ as defined in Section 4.2.1 will not converge to 0 at the fastest possible rate; in general, it is slower by a $\log n$ factor. The same holds for BIC and Bayes factor model selection: all these model selection methods are typically consistent, but do not achieve the optimal rate. In contrast, methods such as AIC and leave-one-out cross-validation (see Section 4.3 below) can be inconsistent, but do achieve the minimax rate. Quite recently, we have shown that with a different, more sophisticated code at the outer level of MDL selection — the so-called *switch code* — this problem can be overcome, and both consistency and optimal convergence rates are achieved; see Section 5.5.

4.3 Prediction

To use MDL for sequential prediction, we have to use codes that can be rendered as prequential forecasting systems. This includes any prequential universal code, in particular the Bayesian universal code and “predictive MDL” [Rissanen, 1984; Rissanen, 1986a]. See Section 3.2.4 for more information about prequential codes.

There are several ways to evaluate the performance of a probabilistic prediction strategy P . In the theoretical machine learning literature [Cesa-Bianchi and Lugosi, 2006], it is customary to look at the *accumulated loss* defined as

$$(21) \quad \sum_{i=1}^n \text{LOSS}(x_i, P(X_i = \cdot \mid x^{i-1})).$$

Here LOSS can be any function mapping pairs of outcomes $x \in \mathcal{X}$ and probability distributions on $\mathcal{X}$ to the real numbers, possibly including infinity. A loss function with particularly useful theoretical properties is the logarithmic loss, $\text{LOSS}(x, P) = -\log P(x)$. With this loss function, the accumulated loss reduces to

$$(22) \quad \sum_{i=1}^n -\log P(X_i = x_i \mid x^{i-1}) = -\log \prod_{i=1}^n P(x_i \mid x^{i-1}) = -\log P(x^n),$$

which is just the code length of the sample! This binds compression and predictive performance together directly: the more we compress, the better the predictive performance in terms of one step ahead prediction loss. In the context of MDL,

let P represent a universal code relative to some set of hypotheses $\mathcal{H}$. Then (22) implies that, if one of the distributions in $\mathcal{H}$, used as a prediction strategy, predicts x^n well, and P is guaranteed to have small regret relative to $\mathcal{H}$, then P is also guaranteed to predict x^n well. An important question is of course whether this result transfers from the logarithmic loss to the particular loss function that is of interest in the problem at hand. It turns out that, if, as in Section 4.2.1, we are prepared to assume that one of the distributions in $\mathcal{H}$ is true, then (with some caveats) this is indeed the case; a more precise statement can be found in [Grünwald, 2007, Chapter 17].

But even if we stick to logarithmic loss, the simple identity (22) has several important consequences. First of all, it is the basis of a fundamental theorem, due to A. Barron [Barron, 1998] and closely related to Theorem 8, that gives explicit guarantees about the convergence rate for prediction strategies based on MDL prequential universal codes. Analogously to Theorem 8, the better a code compresses data sampled from P^* , the faster it learns to predict outcomes from P^* well. This is discussed in Chapter 15 of [Grünwald, 2007]. The theorem also has consequences for estimation, because we may also think of the predictions $P(X_n = \cdot \mid X^{n-1})$ as an estimate of P^* rather than as a predictor of X_n . In this way, each prequential universal code in fact defines an estimator relative to $P^* \in \mathcal{H}$; this provides an alternative, “prequential” rather than “2-part”, MDL estimation method. Under this interpretation, Barron’s theorem now expresses something even more similar to Theorem 8: the better we can compress data from P^* using a prequential code, the faster the estimator based on this code converges to P^* .

Yet another consequence of (22) is that in some common situations, MDL model selection is a special case of Dawid’s theory of prequential model validation [Dawid, 1984; Dawid, 1992b; Dawid, 1992a]. His *weak prequential principle* states that we should evaluate a prequential forecasting system based solely on the predictions it made *for the data that were actually observed*, not the predictions that it might have made for other sequences of observations. This can be achieved by measuring the performance of a forecasting system in terms of its one step ahead prediction loss (21). Thus, suppose we want to choose between a finite number of models $\mathcal{M}_1, \dots, \mathcal{M}_k$. According to the prequential idea, we should associate each of these models with a prediction strategy, , and then pick the model whose associated prediction strategy achieves the smallest accumulated loss (21) on $x_1, \dots, x_n$. The prediction strategies could, for example, be maximum likelihood plug-in codes as in the example of Section 3.2.4. Such an approach is similar to leave-one-out cross validation in the sense that the models are evaluated using observed data only. The only difference is that in cross validation the prediction of each x_i is based on both the previous outcomes $x_1, \dots, x_{i-1}$ and the future outcomes $x_{i+1}, \dots, x_n$. In prequential validation, only the previous outcomes are used. As we can see from (22), if prequential validation is applied with the logarithmic loss function and prediction strategies $P_1, \dots, P_k$ associated with models $\mathcal{M}_1, \dots, \mathcal{M}_k$ respectively, then it is simply equivalent to MDL model selection between these models with a

uniform code (prior) on the outer level, and the codes based on $P_1, \dots, P_k$ on the inner level.

5 MDL IN PERSPECTIVE

In this section we link MDL to other approaches to learning, discussing both philosophical and technical differences and similarities. We also trace the development of the MDL principle back to its origin in algorithmic information theory and Kolmogorov complexity, and point out areas of recent development.

5.1 Basic Principles

The following advice summarises the basic principles of MDL as discussed above.

- Use codes to represent hypotheses about the data; measure the usefulness of the hypothesis by achieved compression.
- Use a code that achieves low regret whatever data are observed, and carefully consider its luckiness properties. This applies to both the inner and the outer level for model selection.
- Design codes with performance guarantees that depend only on the model, but *not on the data generating process given the model*.²

5.2 Occam's Razor

When two models fit the data equally well, MDL will choose the one that is the “simplest” in the sense that it allows for a shorter description of the data. As such, it implements a precise form of Occam's razor — even though as more and more data become available, the model selected by MDL may become more and more “complex”! Occam's razor is sometimes criticised for being either (1) arbitrary or (2) false [Webb, 1996; Domingos, 1999]. Do these criticisms apply to MDL as well?

1. “Occam's Razor (and MDL) Is Arbitrary” Because “description length” is a syntactic notion it may seem that MDL selects an arbitrary model: different codes would have led to different description lengths, and therefore, to different models. By changing the encoding method, we can make ‘complex’ things ‘simple’ and vice versa. This criticism overlooks the fact that we are not allowed to use just any code we like! MDL prescribes, for each model $\mathcal{M}_k$ under consideration, the use of a universal code L_k that achieves small worst-case regret. This

²This may seem to be at odds with our discussion of MDL consistency in Section 4.2: do we not assume that the true distribution P^* is in one of the models there? However, the two statements “We do not know anything about P^* ” and “If P^* is in one of the models, our method behaves especially nicely” are not contradictory.

prescription puts strong constraints on what models may be viewed as “simple” and what models may be viewed as “complex”. First, we note that, if the models $\mathcal{M}_k = \{P_\theta \mid \theta \in \Theta_k\}$ are parametric and the parameter sets are suitably bounded, then we can associate with each model an inherent complexity, namely its worst-case regret (15). We already explained why it makes sense to call this quantity “parametric complexity”. It does not depend on the chosen parameterisation, and is thus certainly not arbitrary. But in practice, we do not want to restrict the parameter set, and then we have to use codes L_k with different regret on different sequences. However, it can be shown that, whatever code we L_k we use, for all subsets Θ_0 of Θ , for “most” sequences x^n such that $\hat{\theta}(x^n) \in \Theta_0$, the regret of L_k is larger than (16), which we repeat here for convenience:

$$(23) \quad \frac{k}{2} \log \frac{n}{2\pi} + \log \int_{\Theta_0} \sqrt{\det I(\theta)} d\theta + o(1).$$

For a precise statement, see [Grünwald, 2007, Chapter 14]. Thus, the fraction of sequences for which L_k is “lucky” and achieves regret substantially smaller than (23) is vanishingly small. Therefore, it still makes sense to call the regret of L_k the “complexity of model $\mathcal{M}_k$ relative to sequence x^n ”. For most sequences, this complexity will be equal to $(k/2) \log n$ plus some small constant.

EXAMPLE 10. Let us consider a prototypical model selection example, regression with polynomial models. The data are given by $(x_1, y_1), \dots, (x_n, y_n)$ with $(x_i, y_i) \in \mathbb{R}^2$. The models are of form $\mathcal{M}_k = \{P_\theta \mid \theta \in \mathbb{R}^{k+1}\}$ where $P_\theta \in \mathcal{M}_k$, $\theta = (\theta_0, \dots, \theta_k)$, prescribes that $Y = f_\theta(X) + Z$, where $f_\theta(x) = \sum_{j=0}^k \theta_j x^j$ is a k -degree polynomial and Z is a normally distributed error (noise) term with mean 0. P_θ is extended to n outcomes by independence. For example, $\mathcal{M}_1$ represents the set of 1-dimensional polynomials, i.e. straight lines $Y = aX + b$, turned into probability distributions by adding a stochastic noise term. For now we restrict ourselves to model selection between $\mathcal{M}_1$ and $\mathcal{M}_2$. If we perform MDL model selection with these two models, we must associate each $\mathcal{M}_k$ with a universal code defined by its length function L_k , and we must also provide a universal code on the outer level with lengths $L(k)$ for $k \in \{1, 2\}$. We may interpret the regret achieved by L_k relative to $\mathcal{M}_k$ for a given sequence as the “complexity” of $\mathcal{M}_k$ relative to that sequence. Now the sometimes-heard argument “MDL is arbitrary since one could have used codes L_1 and L_2 such that $\mathcal{M}_2$ is always preferred over $\mathcal{M}_1$ ”, is simply wrong. Since the code on the outer level should be universal and achieve small worst-case regret, we prefer a uniform code here. The code L_2 we assign to $\mathcal{M}_2$ should be such that its regret is approximately $(3/2) \log n$ plus a small constant, since $\mathcal{M}_2$ has 3 free parameters. The regret of L_1 relative to $\mathcal{M}_1$ should be approximately $(2/2) \log n = \log n$ plus a small constant. As indicated by (23), codes such that for most sequences, the regret relative to $\mathcal{M}_2$ is substantially smaller than $(3/2) \log n$ do not exist. Codes such that for most sequences, the regret relative to $\mathcal{M}_1$ is substantially larger than $(2/2) \log n$ do exist, but they are highly wasteful, and the MDL principle does not allow us to use them. Hence, if proper MDL codes are used, then for most sequences, the “complexity” associated

with $\mathcal{M}_2$ relative to that sequence, which we identified with the regret of L_2 relative to that sequence, is substantially larger than the complexity associated with $\mathcal{M}_1$ relative to that sequence, the difference being about $(1/2)\log n$. It can be shown that this implies that, if proper MDL codes are used, then there *must* be sequences on which MDL chooses $\mathcal{M}_1$ rather than $\mathcal{M}_2$ and vice versa. There are no “MDL codes” for which $\mathcal{M}_1$ is always preferred over $\mathcal{M}_2$ or vice versa. $\diamond$

It is true that MDL leaves room for subjective input in the form of the luckiness principle. The statistician has the freedom to choose some special set A of hypotheses, or parameters within a model, and make them especially easy to learn. Of course, it may occur that these lucky few turn out *not* to improve compression performance, while giving another set B special treatment would have improved performance by a lot. This may seem arbitrary, but consider that the alternative is to make *everything* hard to learn. Also consider that subjective is not the same as arbitrary, there may be good reason to choose A rather than B as the lucky set, and if there is not, then we can even consider the union $A \cup B$ as a potential lucky set, at the modest price of increasing the codelength for sequences with maximum likelihood parameters in A slightly. It is completely impossible however to make *everything* lucky, so there are hard constraints on the freedom of the statistician. To illustrate, we turn back to the regression example for polynomials of degree one or two. Given enough data, we should be able to estimate the parameters of any second degree polynomial up to any desired precision, but this will occur a lot faster if the parameters happen to be in the special subset $\mathcal{M}_1$, which can thus be viewed as a lucky set. We could also have considered an alternative lucky set $\mathcal{M}'_1$, containing all polynomials with $Y_i = aX_i + b + 13X^2$. The choice between $\mathcal{M}_1$ and $\mathcal{M}'_1$ is subjective (not arbitrary), but whichever we pick, we can only make a code such that $\mathcal{M}_1$ is really lucky if $\mathcal{M}_1$ is a very small fraction of $\mathcal{M}_2$, for which the worst-case regret is set in stone. Here “really lucky” means that the regret of sequences with maximum likelihood distributions close to or in $\mathcal{M}_1$ is significantly smaller than the regret of the remaining sequences.

2. “Occam’s Razor Is False” It is often claimed that Occam’s razor is false — we often try to model real-world situations that are arbitrarily complex, so why should we favour simple models? In the words of Webb [Webb, 1996], “What good are simple models of a complex world?”³

The short answer is: even if the true data-generating machinery is very complex, it may be a good strategy to prefer simple models for small sample sizes. Thus, MDL (and the corresponding form of Occam’s razor) is a *strategy* for inferring models from data (“choose simple models at small sample sizes”), not a statement about how the world works (“simple models are more likely to be true”) — indeed, a strategy cannot be true or false; it is “clever” or “stupid.” And the strategy of preferring simpler models is clever even if the data-generating process is highly complex, as illustrated by the following example.

³Quoted with permission from *KDD Nuggets 96,28*, 1996.

EXAMPLE 11 Simple Models of Complex Sources. We return to the polynomial regression setting of the previous example, but now we suppose that the data are sampled from a very complex source, say, a polynomial of degree 1,000. Then, if we use two-part code MDL to learn a polynomial for data $((x_1, y_1), \dots, (x_n, y_n))$, the model $\mathcal{M}_k$ selected by MDL at sample size n will have $k < 1,000$ for small n , but k will typically increase with n , and almost surely it will settle on $\mathcal{M}_{1000}$ for all n larger than some n_0 (by consistency Theorem 9). Now, one could claim that MDL infers the “wrong” degree polynomial for $n < n_0$. Following this line of reasoning, suppose that we somehow knew the “right” degree, and redefined our method to report: the degree is 1,000. If we then try to fit this model for small n , say $n < 1000$, we would have far too few data to accurately estimate all 1001 parameters, which could well lead to disastrous predictions of future outcomes from the same source! On the other extreme, a low degree polynomial, say a parabola, may not be able to fit the data so well, but it has so few parameters that 1,000 observations suffice to estimate the optimal parameter values for this simple approximation quite accurately. Therefore, the mean squared error achieved on the available data should be a good indication of the squared error that should be expected on future data. So even though the model is “wrong” in the sense that it is much simpler than the true source, it yields much more reliable predictions. MDL implements a tradeoff between these two extremes: given enough data, the complex truth will be discovered, but if there is insufficient data to do so, MDL will report a simple, but useful, approximation.

The source we considered so far was complex, but at least it was in one of the considered models. It is far from uncommon however that the source is not in *any* of the models. The same reasoning applies in this case: a polynomial of sufficiently low degree will give a correct impression of how well it will predict future data, even if the data are not polynomial and none of the considered models is “true”! $\diamond$

The ideas in this example appear not only in the MDL literature, but is also the basis of Vapnik’s [Vapnik, 1998] structural risk minimisation approach and many standard statistical methods for nonparametric inference. In such approaches one acknowledges that the data-generating machinery can be very complex, and might not even be consistent with the considered models. Nevertheless, it is still a good strategy to approximate it by simple hypotheses (low-degree polynomials) as long as the sample size is small. Summarising:

The Inherent Difference between Under- and Overfitting (A) If we choose an overly simple (small) model for our data, then the best-fitting point hypothesis within the model is likely to be almost the best predictor, within the simple model, of future data coming from the same source. On the other hand, (B), if we overfit (choose a very complex, i.e., large, model) and there is noise in our data, then, *even if the complex model contains the “true” point hypothesis*, the best-fitting point hypothesis within the model is likely to lead to very bad predictions of future data coming from the same source. This statement is imprecise and is meant to convey the general idea, but it becomes provably true if we use MDL’s measure

of model complexity; we measure prediction quality by logarithmic loss; and we assume that one of the distributions in $\mathcal{H}$ actually generates the data. In fact, statement (A) is an implication of Barron's theorem about MDL prediction that we briefly referred to in Section 4.3.

5.3 MDL, Bayesian Inference and Frequentist Statistics

The MDL, Bayesian and Frequentist schools of thought differ in their interpretation of how the concept of probability relates to the real world.

The frequentist school of thought holds that probability can only express something about the real world in the context of a repeatable experiment. The frequency of a particular observation converges as more observations are gathered; this limiting value is then called the probability. This interpretation is too restrictive for many applications; for example the probability that a suspect is guilty (as often required in legal reasoning), or the probability that it will rain tomorrow, are not frequentist, because no repeatable experiments are involved.

According to the subjective Bayesian school on the other hand, a probability expresses a degree of belief in a certain proposition, which can make sense even outside the context of a repeatable experiment. However, this interpretation is also problematic, because in practice people of necessity often work with very crude models, which everybody agrees have no real truth to them. In such cases, probabilities are often used to represent “beliefs” which one knows a priori to be false! (In order to avoid offending people, we have to point out here that there are actually many different Bayesian schools of thought, and many practitioners would probably hesitate to subscribe completely to the subjectivist interpretation.) For further discussion, see e.g. [Savage, 1954; Berger, 1985].

In stark contrast, from an MDL point of view a probability distribution is really just another word for a code, which allows us to avoid this philosophical can of worms altogether. We do not think of codes in terms of “truth” or “belief”; instead, a code is judged by its *efficiency* when it is applied in practice: a good code achieves a short code length on the data that we observe, whether it is based on valid assumptions about the truth or not.

In spite of these quite profound differences in interpretation, *technically* MDL and Bayesian inference are very closely related. In fact, because Bayesian codes are usually mathematically elegant and, with a suitable choice of prior, can be made to achieve low regret in the worst case, they more often than not turn out to be the code of preference in MDL settings. (There is then a close correspondence between “luckiness functions” and prior distributions, but as explained in [Grünwald, 2007, Chapter 17], some differences remain.) However, MDL research sometimes employs non-Bayesian codes, while Bayesians sometime use priors which do not guarantee low regret. In fact, there are some known inconsistency problems with Bayesian inference in nonparametric settings [Diaconis and Freedman, 1986]. These problems are invariably due to the use of priors which achieve larger regret than necessary. By adopting nonparametric priors with small regret, as

prescribed by the MDL principle, the problems disappear. This is a direct consequence of results such as Theorem 8, which show that small coding regret implies fast learning.

5.4 A Brief History of MDL

The practical MDL principle that we discuss in this book has mainly been developed by Rissanen in a series of papers starting in 1978 with [Rissanen, 1978]. It has its roots in the theory of Kolmogorov complexity [M.Li and Vitányi, 2008], developed in the 1960s by Solomonoff [Solomonoff, 1964], Kolmogorov [Kolmogorov, 1965] and Chaitin [Chaitin, 1966; Chaitin, 1969]. Among these authors, Solomonoff (a former student of the famous philosopher of science, Rudolf Carnap) was explicitly interested in inductive inference. His 1964 paper contains explicit suggestions on how the underlying ideas could be made practical, thereby foreshadowing some of the later work on two-part MDL. While Rissanen was not aware of Solomonoff's work at the time, Kolmogorov's 1965 paper [Kolmogorov, 1965] did serve as an inspiration for Rissanen's [Rissanen, 1978] first development of MDL.

Another important inspiration for Rissanen was Akaike's (1973) AIC method for model selection (Section 4.2), essentially the first model selection method based on information-theoretic ideas [Akaike, 1973; Akaike, 1974]. Even though Rissanen was inspired by AIC, both the actual method and the underlying philosophy are substantially different from MDL.

Minimum Message Length MDL is much closer related to the *Minimum Message Length (MML) Principle* [Wallace, 2005], developed by Wallace and his coworkers in a series of papers starting in 1968 with the groundbreaking [Wallace and Boulton, 1968]; other milestones are [Wallace and Boulton, 1975] (1975) and [Wallace and Freeman, 1987] (1987). Remarkably, Wallace developed his ideas without being aware of the notion of Kolmogorov complexity. Although Rissanen became aware of Wallace's work before the publication of the first MDL paper [Rissanen, 1978], he developed his ideas mostly independently, being influenced more by Akaike and Kolmogorov. Indeed, despite the close resemblance of both methods in practice, the underlying philosophy is very different. A 10-page discussion of the precise relationship between MML and MDL, both technically and philosophically, can be found in Chapter 17 of [Grünwald, 2007].

The first publications on MDL only mention two-part codes. Important progress was made in 1984 when Rissanen used prequential codes for the first time [Rissanen, 1984], and again in 1987 [Rissanen, 1987], when Rissanen introduced Bayesian mixture codes in MDL. This led to the development of the notion of stochastic complexity as the shortest code length of the data given a model [Rissanen, 1986b; Rissanen, 1987]. However, the full development of the notion of "parametric complexity" (15) had to wait until 1996, when (again) Rissanen made the connection to Shtarkov's *normalised maximum likelihood code* [Rissanen, 1996]. In the mean time, A. Barron [Barron, 1985] showed in his impressive Ph.D. thesis how a specific

version of the two-part code criterion has excellent frequentist statistical consistency properties. This was extended in 1991 by Barron and Cover [Barron and Cover, 1991] who achieved a breakthrough for two-part codes: they gave clear prescriptions on how to design codes for hypotheses, relating codes with good minimax code length properties to rates of convergence in statistical consistency theorems, both for parametric and nonparametric problems. Some of the ideas of Rissanen’s 1987 and Barron and Cover’s 1991 paper were, as it were, unified when in 1996 Rissanen [Rissanen, 1996] introduced the normalised maximum likelihood code. The resulting theory was summarised for the first time by Barron, Rissanen and Yu [Barron *et al.*, 1998], and is the main subject of the first (2007) comprehensive overview of the field, [Grünwald, 2007]. In this book, the modern versions of MDL where code design is based on explicit rather than ad hoc goals, are called “refined MDL”.

5.5 Recent Developments

While this text is intended to be introductory rather than exhaustive, there have been a couple of recent developments in MDL research that shed new light on its interpretation, and which we consider relevant for anyone who currently uses or plans to use MDL or Bayesian inference in practice.

Luckiness Functions Although we emphasised the importance of luckiness in the design of the code for hypotheses in Section 2, it did not really play a part in the definition of the universal codes later in Section 3.2. This is for historical reasons: it was not known until recently (and nor was it considered important) how universal codes could be constructed with *both* low-worst case regret *and* a freely chosen luckiness structure. In [Grünwald, 2007] it is shown how this can be achieved using *luckiness functions*. A luckiness function $a : \Theta \rightarrow \mathbb{R}$ roughly expresses which parameter values in θ are especially easy to learn, should they provide good fit. While luckiness-related concepts were always implicit in MDL-like inference, the idea of using a luckiness function seems due to Barron [Barron, 1998], although he did not call it this way. The luckiness terminology was first adopted by Grünwald [Grünwald, 2007], who imported it from the computational learning theory literature [Shawe-Taylor *et al.*, 1998] where it plays a somewhat similar role.

Model Switching The outer-level codes that are used in model selection problems are designed to achieve low worst-case regret, where regret is normally measured against the “best” model, i.e. the one that allows for the most compression. However, in predictive settings, practitioners often find that *there is no single best model*. Instead, simpler models tend to predict well as long as the sample size is small, while more complex models predict better as more data are gathered. This so-called *catch-up phenomenon* leads to the idea of measuring regret not against the model with the best *overall* predictive performance, but against a sequence of

models, where each model in the sequence is used in turn to predict a number of outcomes. There are efficient and simple algorithms to achieve this, which often leads to significant improvements in model selection performance, both in theory and in practice. For example, in many situations, one can achieve both consistency and optimal convergence rates; see Section 4.2.2. For more details, please refer to [van Erven *et al.*, 2008].

BIBLIOGRAPHY

- [Akaike, 1973] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademiai Kiado.
- [Akaike, 1974] H. Akaike. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [Barron and Cover, 1991] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.
- [Barron *et al.*, 1998] A. Barron, J. Rissanen, and B. Yu. The Minimum Description Length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.
- [Barron, 1985] A.R. Barron. *Logically Smooth Density Estimation*. PhD thesis, Dept. of Electrical Engineering, Stanford University, Stanford, CA, 1985.
- [Barron, 1998] A.R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics 6*, pages 27–52. Oxford University Press, 1998.
- [Berger, 1985] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer-Verlag, New York, revised and expanded second edition, 1985.
- [Cesa-Bianchi and Lugosi, 2006] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, UK, 2006.
- [Chaitin, 1966] G.J. Chaitin. On the length of programs for computing finite binary sequences. *Journal of the ACM*, 13:547–569, 1966.
- [Chaitin, 1969] G.J. Chaitin. On the length of programs for computing finite binary sequences: statistical considerations. *Journal of the ACM*, 16:145–159, 1969.
- [Chomsky, 1956] N. Chomsky. Three models for the description of language. *IEEE Transactions on Information Theory*, 2(3), September 1956.
- [Cover and Thomas, 1991] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Series in telecommunications. John Wiley, 1991.
- [Dawid, 1984] A.P. Dawid. Present position and potential developments: Some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society, Series A*, 147(2):278–292, 1984.
- [Dawid, 1992a] A.P. Dawid. Prequential analysis, stochastic complexity and Bayesian inference. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 4*, pages 109–125. Oxford University Press, 1992.
- [Dawid, 1992b] A.P. Dawid. Prequential data analysis. In M. Ghosh and P.K. Pathak, editors, *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, Lecture Notes-Monograph Series, pages 113–126. Institute of Mathematical Statistics, 1992.
- [de Rooij and Grünwald, 2006] Steven de Rooij and Peter Grünwald. An empirical study of minimum description length model selection with infinite parametric complexity. *Journal of Mathematical Psychology*, 50:180–190, 2006.
- [Diaconis and Freedman, 1986] P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986.
- [Domingos, 1999] P. Domingos. The role of Occam’s razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425, 1999.
- [Grünwald and Vitényi, 2008] P.D. Grünwald and P.M.B. Vitényi. Algorithmic information theory. In P. Adriaans and J. van Benthem, editors, *Handbook of the Philosophy of Science*, volume 8: Philosophy of Information, pages 289–325. Elsevier Science, 2008.

- [Grünwald, 2007] P.D. Grünwald. *The Minimum Description Length Principle*. MIT Press, June 2007. Chapter 17, containing further discussion of philosophical and conceptual issues, can be freely downloaded from www.grunwald.nl.
- [Kolmogorov, 1965] A.N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.
- [Kontkanen and Myllymäki, 2007] P. Kontkanen and P. Myllymäki. A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, 103:227–233, September 2007.
- [Li, 1999] J.Q. Li. *Estimation of Mixture Models*. PhD thesis, Yale University, New Haven, CT, 1999.
- [M.Li and Vitányi, 2008] M.Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, New York, 3rd edition, 2008.
- [Rissanen and Roos, 2007] J. Rissanen and T. Roos. Conditional NML universal models. In *2007 Information Theory and Applications Workshop (ITA-07)*, pages 337–344, 2007.
- [Rissanen, 1976] J. Rissanen. Generalized Kraft inequality and arithmetic coding. *IBM Journal of Research and Development*, 20(3), 1976.
- [Rissanen, 1978] J. Rissanen. Modeling by the shortest data description. *Automatica*, 14:465–471, 1978.
- [Rissanen, 1983a] J. Rissanen. A universal data compression system. *IEEE Transactions on Information Theory*, IT-29(5):656–664, 1983.
- [Rissanen, 1983b] J. Rissanen. A universal prior for integers and estimation by Minimum Description Length. *Annals of Statistics*, 11:416–431, 1983.
- [Rissanen, 1984] J. Rissanen. Universal coding, information, prediction and estimation. *IEEE Transactions on Information Theory*, 30:629–636, 1984.
- [Rissanen, 1986a] J. Rissanen. A predictive least squares principle. *IMA Journal of Mathematical Control and Information*, 3:211–222, 1986.
- [Rissanen, 1986b] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.
- [Rissanen, 1987] J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society, Series B*, 49:223–239, 1987. Discussion: 252–265.
- [Rissanen, 1989] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15 of *Series in Computer Science*. World Scientific, 1989.
- [Rissanen, 1996] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.
- [Savage, 1954] L.J. Savage. *The Foundations of Statistics*. Dover Publications, 1954.
- [Shannon, 1948] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–656, 1948.
- [Shawe-Taylor et al., 1998] J. Shawe-Taylor, P. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimisation over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.
- [Solomonoff, 1964] R.J. Solomonoff. A formal theory of inductive inference, part 1 and part 2. *Information and Control*, 7:1–22, 224–254, 1964.
- [Tromp, 2007] J. Tromp. Binary lambda calculus and combinatory logic. Available at <http://www.cwi.nl/~tromp/cl/cl.html>, 2007.
- [van Erven et al., 2008] T. van Erven, P.D. Grünwald, and S. de Rooij. Catching up faster by switching sooner: a prequential solution to the AIC-BIC dilemma. Available at <http://arxiv.org/abs/0807.1005>, 2008. Submitted to Journal of the Royal Statistical Society, Series B. A much shorter version appears in the proceedings of NIPS 2008.
- [Vapnik, 1998] V.N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
- [Wallace and Boulton, 1968] C.S. Wallace and D.M. Boulton. An information measure for classification. *Computing Journal*, 11:185–195, 1968.
- [Wallace and Boulton, 1975] C.S. Wallace and D.M. Boulton. An invariant Bayes method for point estimation. *Classification Society Bulletin*, 3(3):11–34, 1975.
- [Wallace and Freeman, 1987] C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society B*, 49:240–251, 1987. Discussion: pages 252–265.
- [Wallace, 2005] C.S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer-Verlag, New York, 2005.

- [Webb, 1996] G.I. Webb. Further experimental evidence against the utility of Occam's razor. *Journal of Artificial Intelligence Research*, 4:397–417, 1996.
- [Willems *et al.*, 1997] F.M.J. Willems, Y.M. Shtarkov, and Tj.J. Tjalkens. Reflections on 'the Context-Tree Weighting method: Basic properties'. *Newsletter of the IEEE Information Theory Society*, 1997.

MML, HYBRID BAYESIAN NETWORK GRAPHICAL MODELS, STATISTICAL CONSISTENCY, INVARIANCE AND UNIQUENESS

David L. Dowe

1 INTRODUCTION

The problem of statistical — or inductive — inference pervades a large number of human activities and a large number of (human and non-human) actions requiring ‘intelligence’. Human and other ‘intelligent’ activity often entails making inductive inferences, remembering and recording observations from which one can make inductive inferences, learning (or being taught) the inductive inferences of others, and acting upon these inductive inferences.

The Minimum Message Length (MML) approach to machine learning (within artificial intelligence) and statistical (or inductive) inference gives us a trade-off between simplicity of hypothesis (H) and goodness of fit to the data (D) [Wallace and Boulton, 1968, p. 185, sec 2; Boulton and Wallace, 1969; 1970, p. 64, col 1; Boulton, 1970; Boulton and Wallace, 1973b, sec. 1, col. 1; 1973c; 1975, sec 1 col 1; Wallace and Boulton, 1975, sec. 3; Boulton, 1975; Wallace and Georgeff, 1983; Wallace and Freeman, 1987; Wallace and Dowe, 1999a; Wallace, 2005; Comley and Dowe, 2005, secs. 11.1 and 11.4.1; Dowe, 2008a, sec 0.2.4, p. 535, col. 1 and elsewhere]. There are several different and intuitively appealing ways of thinking of MML. One such way is to note that files with structure compress (if our file compression program is able to find said structure) and that files without structure don’t compress. The more structure (that the compression program can find), the more the file will compress.

Another, second, way to think of MML is in terms of Bayesian probability, where $Pr(H)$ is the prior probability of a hypothesis, $Pr(D|H)$ is the (statistical) likelihood of the data D given hypothesis H , $-\log Pr(D|H)$ is the (negative) log-likelihood, $Pr(H|D)$ is the posterior probability of H given D , and $Pr(D)$ is the marginal probability of D — i.e., the probability that D will be generated (regardless of whatever the hypothesis might have been). Applying Bayes’s theorem twice, with or without the help of a Venn diagram, we have $Pr(H|D) = Pr(H \& D) / Pr(D) = (1/Pr(D)) Pr(H)Pr(D|H)$.

Choosing the most probable hypothesis (a posteriori) is choosing H so as to maximise $Pr(H|D)$. Given that $Pr(D)$ and $1/Pr(D)$ are independent of the choice of

hypothesis H , this is equivalent to choosing H to maximise $Pr(H) \cdot Pr(D|H)$. By the monotonicity of the logarithm function, this is in turn equivalent to choosing H so as to minimise $-\log Pr(H) - \log Pr(D|H)$. From Shannon's information theory (see sec. 2.1), this is the amount of information required to encode H (in the first part of a two-part message) and then encode D given H (in the second part of the message). And this is, in turn, similar to our first way above of thinking about MML, where we seek H so as to give the optimal two-part file compression.

We have shown that, given data D , we can variously think of the MML hypothesis H in at least two different ways: (a) as the hypothesis of highest posterior probability and also (b) as the hypothesis giving the two-part message of minimum length for encoding H followed by D given H ; and hence the name Minimum Message Length (MML).

Historically, the seminal Wallace and Boulton paper [1968] came into being from Wallace's and Boulton's finding that the Bayesian position that Wallace advocated and the information-theoretic (conciseness) position that Boulton advocated turned out to be equivalent [Wallace, 2005, preface, p. v; Dowe, 2008a, sec. 0.3, p. 546 and footnote 213]. After several more MML writings [Boulton and Wallace, 1969; 1970, p. 64, col. 1; Boulton, 1970; Boulton and Wallace, 1973b, sec. 1, col. 1; 1973c; 1975, sec. 1, col. 1] (and an application paper [Pilowsky *et al.*, 1969], and at about the same time as David Boulton's PhD thesis [Boulton, 1975]), their paper [Wallace and Boulton, 1975, sec. 3] again emphasises the equivalence of the probabilistic and information-theoretic approaches. (And all of this work on Minimum Message Length (MML) occurred prior to the later Minimum Description Length (MDL) principle discussed in sec. 6.7 and first published in 1978 [Rissanen, 1978].)

A third way to think about MML is in terms of *algorithmic* information theory (or Kolmogorov complexity), the shortest input to a (Universal) Turing Machine [(U)TM] or computer program which will yield the original data string, D . This relationship between MML and Kolmogorov complexity is formally described — alongside the other two ways above of thinking of MML (probability on the one hand and information theory or concise representation on the other) — in [Wallace and Dowe, 1999a]. In short, the first part of the message encodes H and causes the TM or computer program to read (without yet writing) and prepare to output data, emulating as though it were generated from this hypothesis. The second part of the input then causes the (resultant emulation) program to write the data, D .

So, in sum, there are (at least) three equivalent ways of regarding the MML hypothesis. It variously gives us: (i) the best two-part compression (thus best capturing the structure), (ii) the most probable hypothesis (a posteriori, after we've seen the data), and (iii) an optimal trade-off between structural complexity and noise — with the first-part of the message capturing all of the structure (no more, no less) and the second part of the message then encoding the noise.

Theorems from [Barron and Cover, 1991] and arguments from [Wallace and Freeman, 1987, p241] and [Wallace, 2005, chap. 3.4.5, pp. 190-191] attest to the

general optimality of this two-part MML inference — converging to the correct answer as efficiently as possible. This result appears to generalise to the case of model misspecification, where the model generating the data (if there is one) is not in the family of models that we are considering [Grünwald and Langford, 2007, sec. 7.1.5; Dowe, 2008a, sec. 0.2.5]. In practice, we find that MML is quite conservative in variable selection, typically choosing less complex models than rival methods [Wallace, 1997; Fitzgibbon *et al.*, 2004; Dowe, 2008a, footnote 153, footnote 55 and near end of footnote 135] while also appearing to typically be better predictively.

Having introduced Minimum Message Length (MML), throughout the rest of this chapter, we proceed initially as follows. First, we introduce information theory, Turing machines and algorithmic information theory — and we relate all of those to MML. We then move on to Ockham’s razor and the distinction between inference (or induction, or explanation) and prediction. We then continue on to relate MML and its relevance to a myriad of other issues.

2 INFORMATION THEORY — AND VARIETIES THEREOF

2.1 Elementary information theory and Huffman codes

Tossing a fair unbiased coin n times has 2^n equiprobable outcomes of probability 2^{-n} each. So, intuitively, it requires n bits (or binary digits) of information to encode an event of probability 2^{-n} , so (letting $p = 2^{-n}$) an event of probability p contains $-\log_2 p$ bits of information. This results holds more generally for bases $k = 3, 4, \dots$ other than 2.

The Huffman code construction (for base k), described in (e.g.) [Wallace, 2005, chap. 2, especially sec. 2.1; Dowe, 2008b, p. 448] and below ensures that the code length l_i for an event e_i of probability p_i satisfies $-\log_k p_i \approx l_i < -\log_k p_i + 1$.

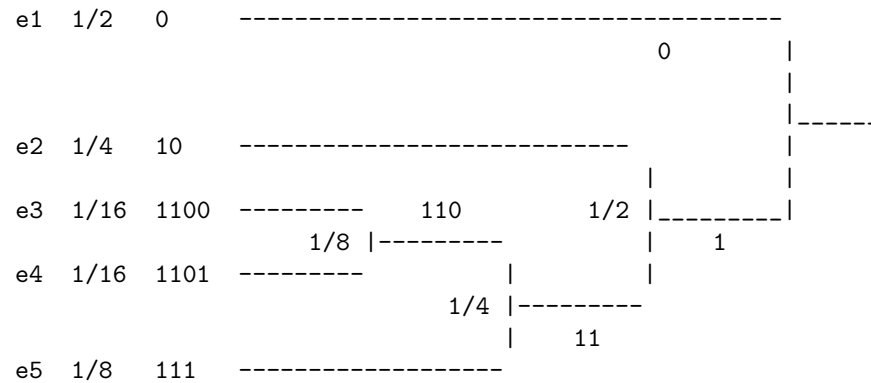
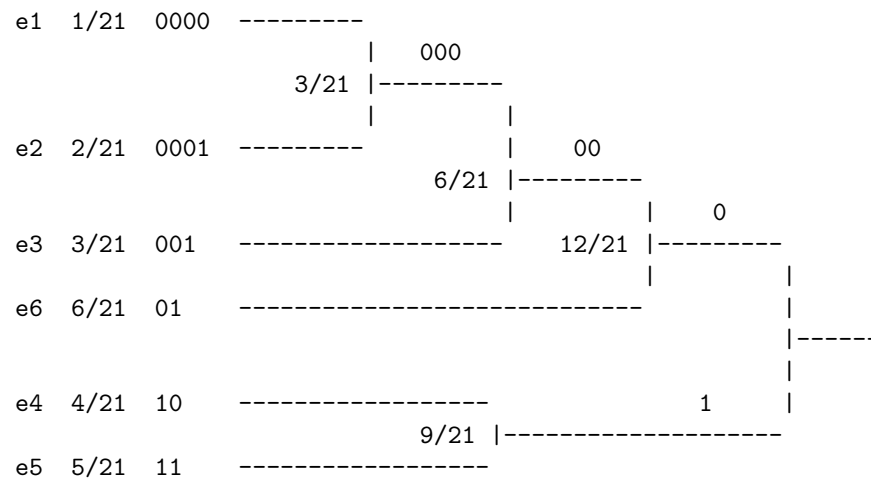
Huffman code construction proceeds by taking m events $e_1, \dots, e_m$ of probability $p_1, \dots, p_m$ respectively and building a code tree by successively (iteratively) joining together the events of least probability. So, with $k = 2$, the binary code construction proceeds by joining together the two events of least probability (say e_i and e_j) and making a new event $e_{i,j}$ of probability $p_{i,j} = p_i + p_j$. (For a k -ary code construction of arity k , we join the k least probable events together — see, e.g., fig. 3, with arity $k = 3$. We address this point a little more later.) Having joined two events into one event, there is now 1 less event left. This iterates one step at a time until the tree is reduced to its root.

An example with $k = 2$ from [Dowe, 2008b, p. 448, Fig. 1] is given in Figure 1.

Of course, we can not always expect all probabilities to be of the form k^{-n} , as they are in the friendly introductory example of fig. 1.

One example with $k = 2$ (binary) and where the probabilities are not all some k raised to the power of a negative (or zero) integer is $1/21, 2/21, 3/21, 4/21, 5/21, 6/21$, as per fig. 2, which we now examine.

Immediately below, we step through the stages of the binary Huffman code construction in fig. 2. The two events of smallest probability are e_1 and e_2 of

Figure 1. A simple (binary) Huffman code tree with $k = 2$ Figure 2. A not so simple (binary) Huffman code tree with $k = 2$

probabilities $1/21$ and $2/21$ respectively, so we join them together to form $e_{1,2}$ of probability $3/21$. The two remaining events of least probability are now $e_{1,2}$ and e_3 , so we join them together to give $e_{1,2,3}$ of probability $6/21$. The two remaining events of least probability are now e_4 and e_5 , so we join them together to give $e_{4,5}$ of probability $9/21$. Three events now remain: $e_{1,2,3}$, $e_{4,5}$ and e_6 . The two smallest probabilities are $p_{1,2,3} = 6/21$ and $p_6 = 6/21$, so they are joined to give $e_{1,2,3,6}$ with probability $p_{1,2,3,6} = 12/21$. For the final step, we then join $e_{4,5}$ and $e_{1,2,3,6}$. The code-words for the individual events are obtained by tracing a path from the root of the tree (at the right of the code-tree) left across to the relevant event at the leaf of the tree. For a binary tree ($k = 2$), every up branch is a 0 and every down branch is a 1. The final code-words are e_1 : 0000, e_2 : 0001, e_3 : 001, etc. (For the reader curious as to why we re-ordered the events e_i , putting e_6 in the middle and not at an end, if we had not done this then some of the arcs of the Huffman code tree would cross — probably resulting in a less elegant and less clear figure.)

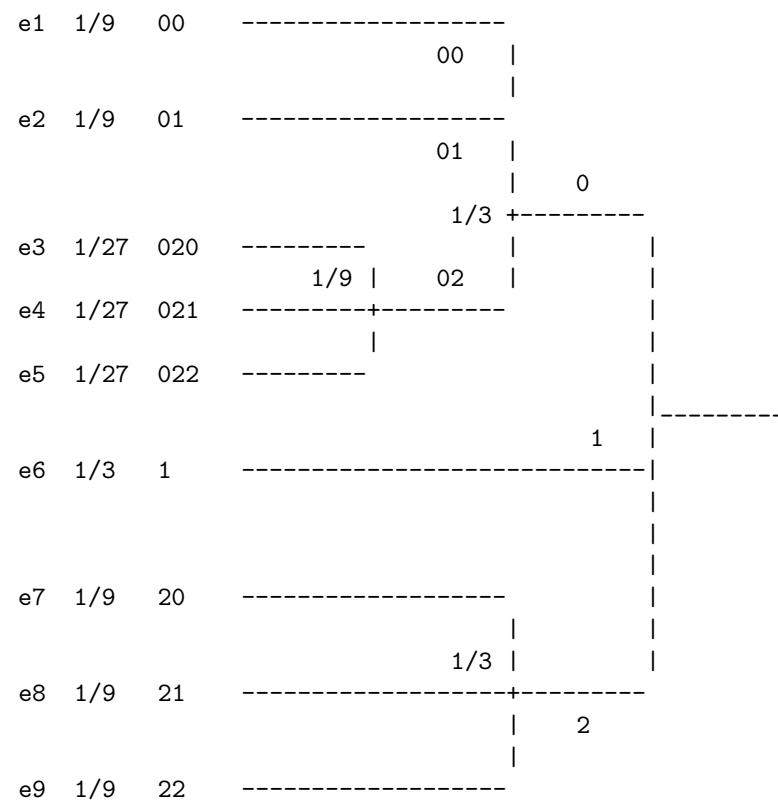
For another such example with $k = 2$ (binary) and where the probabilities are not all some k raised to the power of a negative (or zero) integer, see the example (with probabilities $1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 7/36, 8/36$) from [Wallace, 2005, sec. 2.1.4, Figs. 2.5–2.6].

An example with $k = 3$ is given in Fig. 3. The Huffman code construction for $k = 3$ is very similar to that for $k = 2$, but it also sometimes has something of a small pre-processing step. Each step of the k -ary Huffman construction involves joining k events together, thus reducing the number of events by $(k - 1)$, which is equal to 2 for $k = 3$. So, if the number of events is even, our initial pre-processing step is to join the two least probable events together. That done, we now have an odd number of events and our code tree construction is, at each step, to join the three least probable remaining events together. We continue this until just one event is left, when we are left with just the root node and we have finished. The assignment of code-words is similar to the binary case, although the ternary construction (with $k = 3$) has 3-way branches. The top branch is 0, the middle branch is 1 and the bottom branch is 2. The reader is invited to construct and verify this code construction example (in fig. 3) and the earlier examples referred to above.

For higher values of k , the code construction joins k events (or nodes) into 1 node each time, reducing the number of nodes by $(k - 1)$ at each step. If the number of nodes is $q(k - 1) + 1$ for some $q \geq 0$, then the Huffman code construction does not need a pre-processing step. Otherwise, if the number of nodes is $q(k - 1) + 1 + r$ for some $q \geq 0$ and some r such that $1 \leq r \leq k - 2$, then we require a pre-processing step of first joining together the $(r + 1)$ least probable events into one, reducing the number of nodes by r to $q(k - 1) + 1$.

The result mentioned earlier that

$$-\log_k p_i \approx l_i < -\log_k p_i + 1 \quad (1)$$

Figure 3. A simple (ternary) Huffman code tree with $k = 3$

follows from the Huffman construction. It is customary to make the approximation that $l_i = -\log_k p_i$.

Because of the relationship between different bases of logarithm a and b that $\forall x > 0 \log_b x = (\log_a x)/(\log_a b) = (\log_b a) \log_a x$, changing base of logarithms has the effect of scaling the logarithms by a multiplicative constant, $\log_b a$. As such, the choice of base of logarithm is somewhat arbitrary. The two most common bases are 2 and e . When the base is 2, the information content is said to be in bits. When the base is e , the information content is said to be in nits [Boulton and Wallace, 1970, p. 63; Wallace, 2005, sec. 2.1.8; Comley and Dowe, 2005, sec. 11.4.1; Dowe, 2008a, sec. 0.2.3, p. 531, col. 1], a term which I understand to have had its early origins in (thermal) physics. Alternative names for the nit include the *natural ban* (used by Alan Turing (1912-1954) [Hodges, 1983, pp. 196-197]) and (much much later) the *nat*.

2.2 Prefix codes and Kraft's inequality

Furthermore, defining a prefix code to be a set of (k -ary) strings (of arity k , i.e., where the available alphabet from which each symbol in the string can be selected is of size k) such that no string is the prefix of any other, then we note that the 2^n binary strings of length n form a prefix code. Elaborating, neither of the $2^1 = 2$ binary strings 0 and 1 is a prefix of the other, so the set of code words $\{0, 1\}$ forms a prefix code. Again, none of the $2^2 = 4$ binary strings 00, 01, 10 and 11 is a prefix of any of the others, so the set of code words $\{00, 01, 10, 11\}$ forms a prefix code. Similarly, for $k \geq 2$ and $n \geq 1$, the k^n k -ary strings of length n likewise form a prefix code. We also observe that the fact that the Huffman code construction leads to a (Huffman) *tree* means that the result of any Huffman code construction is always a prefix code. (Recall the examples from sec. 2.1.)

Prefix codes are also known as (or, perhaps more technically, are equivalent to) *instantaneous codes*. In a prefix code, as soon as we see a code-word, we instantaneously recognise it as an intended part of the message — because, by the nature of prefix codes, this code-word can not be the prefix of anything else. Non-prefix (and therefore non-instantaneous) codes do exist, such as (e.g.) $\{0, 01, 11\}$. For a string of the form 01^n , we need to wait to the end of the string to find out what n is and whether n is odd or even before we can decode this in terms of our non-prefix code. (E.g., 011 is 0 followed by 11, 0111 is 01 followed by 11, etc.) For the purposes of the remainder of our writings here, though, we can safely and do restrict ourselves to (instantaneous) prefix codes.

A result often attributed to Kraft [1949] but which is believed by many to have been known to at least several others before Kraft is Kraft's inequality — namely, that in a k -ary alphabet, a prefix code of code-lengths $l_1, \dots, l_m$ exists if and only if $\sum_{i=1}^m k^{-l_i} \leq 1$. The Huffman code construction algorithm, as carried out in our earlier examples (perhaps especially those of figs. 1 and 3), gives an informal intuitive argument as to why Kraft's inequality must be true.

2.3 Entropy

Let us re-visit our result from equation (1) and the standard accompanying approximation that $l_i = -\log p_i$.

Let us begin with the 2-state case. Suppose we have probabilities p_1 and $p_2 = 1 - p_1$ which we wish to encode with code-words of length $l_1 = -\log q_1$ and $l_2 = -\log q_2 = -\log(1 - q_1)$ respectively. As per the Huffman code construction (and Kraft's inequality), choosing such code lengths gives us a prefix code (when these code lengths are non-negative integers).

The negative of the expected code length would then be

$$p_1 \log q_1 + (1 - p_1) \log(1 - q_1),$$

and we wish to choose q_1 and $q_2 = 1 - q_1$ to make this code as short as possible on average — and so we differentiate the negative of the expected code length with respect to q_1 .

$$\begin{aligned} 0 &= \frac{d}{dq_1} (p_1 \log q_1 + (1 - p_1) \log(1 - q_1)) = (p_1/q_1) - ((1 - p_1)/(1 - q_1)) \\ &= (p_1(1 - q_1) - q_1(1 - p_1))/(q_1(1 - q_1)) = (p_1 - q_1)/(q_1(1 - q_1)) \end{aligned}$$

and so $(p_1 - q_1) = 0$, and so $q_1 = p_1$ and $q_2 = p_2$.

This result also holds for p_1, p_2, p_3, q_1, q_2 and q_3 in the 3-state case, as we now show. Let $P_2 = p_2/(p_2 + p_3)$, $P_3 = p_3/(p_2 + p_3) = 1 - P_2$, $Q_2 = q_2/(q_2 + q_3)$ and $Q_3 = q_3/(q_2 + q_3) = 1 - Q_2$, so $p_2 = (1 - p_1)P_2$, $p_3 = (1 - p_1)P_3 = (1 - p_1)(1 - P_2)$, $q_2 = (1 - q_1)Q_2$ and $q_3 = (1 - q_1)Q_3 = (1 - q_1)(1 - Q_2)$.

Encoding the events of probability p_1, p_2 and p_3 with code lengths $-\log q_1$, $-\log q_2$ and $-\log q_3$ respectively, the negative of the expected code length is then $p_1 \log q_1 + (1 - p_1)P_2 \log((1 - q_1)Q_2) + (1 - p_1)(1 - P_2) \log((1 - q_1)(1 - Q_2))$. To minimise, we differentiate with respect to both q_1 and Q_2 in turn, and set both of these to 0.

$$\begin{aligned} 0 &= \frac{\partial}{\partial q_1} (p_1 \log q_1 + (1 - p_1)P_2 \log((1 - q_1)Q_2) + \\ &\quad (1 - p_1)(1 - P_2) \log((1 - q_1)(1 - Q_2))) \\ &= (p_1/q_1) - ((1 - p_1)P_2)/(1 - q_1) - ((1 - p_1)(1 - P_2))/(1 - q_1) \\ &= (p_1/q_1) - (1 - p_1)/(1 - q_1) \\ &= (p_1 - q_1)/(q_1(1 - q_1)) \end{aligned}$$

exactly as in the 2-state case above, where again $q_1 = p_1$.

$$\begin{aligned} 0 &= \frac{\partial}{\partial Q_2} (p_1 \log q_1 + (1 - p_1)P_2 \log((1 - q_1)Q_2) + \\ &\quad (1 - p_1)(1 - P_2) \log((1 - q_1)(1 - Q_2))) \end{aligned}$$

$$\begin{aligned}
&= ((1-p_1)P_2)/Q_2 - ((1-p_1)(1-P_2)/(1-Q_2)) \\
&= (1-p_1) \times ((P_2/Q_2) - (1-P_2)/(1-Q_2)) \\
&= (1-p_1) \times (P_2 - Q_2)/(Q_2(1-Q_2))
\end{aligned}$$

In the event that $p_1 = 1$, the result is trivial. With $p_1 \neq 1$, we have, of very similar mathematical form to the two cases just examined, $0 = (P_2/Q_2) - (1-P_2)/(1-Q_2)$, and so $Q_2 = P_2$, in turn giving that $q_2 = p_2$ and $q_3 = p_3$.

One can proceed by the principle of mathematical induction to show that, for probabilities $(p_1, \dots, p_i, \dots, p_{m-1}, p_m = 1 - \sum_{i=1}^{m-1} p_i)$ and code-words of respective lengths $(-\log q_1, \dots, -\log q_i, \dots, -\log q_{m-1}, -\log q_m = -\log(1 - \sum_{i=1}^{m-1} q_i))$, the expected code length $-(p_1 \log q_1 + \dots + p_i \log q_i + \dots + p_{m-1} \log q_{m-1} + p_m \log q_m)$ is minimised when $\forall i \ q_i = p_i$.

This expected (or average) code length,

$$\sum_{i=1}^m p_i \times (-\log p_i) = -\sum_{i=1}^m p_i \log p_i \quad (2)$$

is called the *entropy* of the m -state probability distribution $(p_1, \dots, p_i, \dots, p_m)$.

Note that if we sample randomly from the distribution p with code-words of length $-\log p$, then the (expected) average long-term cost is the entropy.

Where the distribution is continuous rather than (as above) discrete, the sum is replaced by an integral and (letting x be a variable being integrated over) the entropy is then defined as

$$\begin{aligned}
\int f \times (-\log f) \, dx &= -\int f \log f \, dx = -\int f(x) \log f(x) \, dx \\
&= \int f(x) \times (-\log f(x)) \, dx \quad (3)
\end{aligned}$$

And, of course, entropy can be defined for hybrid structures of both discrete and continuous, such as Bayesian network graphical models (of sec. 7.6) — see sec. 3.6, where it is pointed out that for the hybrid continuous and discrete Bayesian net graphical models in [Comley and Dowe, 2003; 2005] (emanating from the current author's ideas in [Dowe and Wallace, 1998]), the log-loss scoring approximation to Kullback-Leibler distance has been used [Comley and Dowe, 2003, sec. 9].

The next section, sec. 2.4, introduces Turing machines as an abstract model of computation and then discusses the formal relationship between MML and minimising the length of some (constrained) input to a Turing machine. The section can be skipped on first reading.

2.4 Turing machines and algorithmic information theory

The area of algorithmic information theory was developed independently in the 1960s by Solomonoff [1960; 1964], Kolmogorov [1965] and Chaitin [1966], independently of and slightly before the seminal Wallace & Boulton paper on MML [1968].

Despite the near-simultaneous independent work of the young Chaitin [1966] and the independent earlier work of Solomonoff [1960; 1964] pre-dating Kolmogorov, the area of algorithmic information theory is often referred to by many as Kolmogorov complexity (e.g., [Wallace and Dowe, 1999a; Li and Vitányi, 1997]). Before introducing the notion of the algorithmic complexity (or Kolmogorov complexity) of a string s , we must first introduce the notion of a Turing machine [Turing, 1936; Wallace, 2005, sec. 2.2.1; Dowe 2008b, pp. 449-450]. Following [Dowe, 2008b, pp. 449-450], a *Turing machine* (TM) [Wallace, 2005, sec. 2.2.1; Dowe, 2008b, pp. 449-450] is an abstract mathematical model of a computer program. It can be written in a language from a certain alphabet of symbols (such as 1 and (blank) “ ”, also denoted by “ $\sqcup$ ”). We assume that Turing machines have a read/write head on an infinitely long tape, finitely bounded to the left and infinitely long to the right. Turing machines have a set of instructions — or an instruction set — as follows. A Turing machine in a given state (with the read/write head) reading a certain symbol either moves to the left (L) or to the right (R) or stays where it is and writes a specified symbol. The instruction set for a Turing machine can be written as: $f : States \times Symbols \rightarrow States \times (\{L, R\} \cup Symbols)$.

So, the definition that we are using is that a Turing Machine M is a set of quadruples $\{Q_n\} = \{\langle q_i, q_j, s_k, \{s_l, H\}\rangle\}$ where

- $q_i, q_j \in \{1, \dots, m\}$ (the machine states)
- $s_k, s_l \in \{s_0, \dots, s_r\}$ (the symbols)
- $H \in \{R, L\}$ (tape head direction)

(such that no two quadruples have the same first and third elements). The Turing machine in state q_i given input s_k goes into state q_j and either stays where it is and writes a symbol (s_l) or moves to the left or right (depending upon the value of H) without writing a symbol.

An alternative equivalent definition of a Turing Machine M which we could equally well use instead is a set of quintuples $\{Q_n\} = \{\langle q_i, q_j, s_k, s_l, H \rangle\}$ where

- (the machine states) $q_i, q_j \in \{1, \dots, m\}$
- (the symbols) $s_k, s_l \in \{s_0, \dots, s_r\}$
- (tape head direction) $H \in \{R, L\}$

and the Turing machine in state q_i given input s_k then goes into state q_j , writes symbol s_l and moves the head in direction H (and, again, we require that no two quintuples have the same first and third elements).

Note that the Turing Machine (TM) in the first definition *either* writes a (new) symbol *or* moves the head at each step whereas the TM in the second of these two equivalent definitions *both* writes a (new) symbol *and* moves the head.

Without loss of generality we can assume that the alphabet is the binary alphabet $\{0, 1\}$, whereupon the instruction set for a Turing machine can be written as: $f : States \times \{0, 1\} \rightarrow States \times (\{L, R\} \cup \{0, 1\})$.

Any known computer program can be represented by a Turing Machine. *Universal* Turing Machines (UTMs) are like (computer program) compilers and can be made to emulate *any* Turing Machine (TM).

An example of a Turing machine would be the program from fig. 4, which, given two inputs, x_0 and x_1 , adds them together, writing $x_0 + x_1$ and then stopping¹. This machine adds two unary numbers (both at least 1), terminated by blanks (and separated by a single blank). In unary, e.g., 4 is represented by “1111 $\sqcup$ ”. In general in unary, n is represented by n 1s followed by a blank.

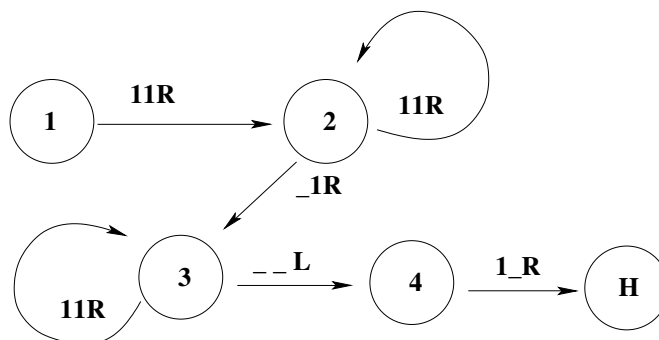


Figure 4. A Turing machine program for adding two numbers

Alternatively, recalling our notation of quintuples, $\langle q_i, q_j, s_k, s_l, H \rangle$, this Turing machine adding program from fig. 4 can be represented as:

$$\{\langle 1, 2, 1, 1, R \rangle, \langle 2, 2, 1, 1, R \rangle, \langle 2, 3, \sqcup, 1, R \rangle, \langle 3, 3, 1, 1, R \rangle, \langle 4, 5, 1, \sqcup, R \rangle\}$$

(where state 5 is the Halting — or stop — state, also referred to as H).

(This Turing machine program over-writes the blank ($\sqcup$) in the middle with a 1 and removes a 1 from the right end of the second number — and, in so doing, leaves behind the unary representation of the sum.)

Another example of a Turing machine would be a program which, for some a_0 and a_1 , when given any input x , calculates (or outputs) $a_0 + a_1x$. In this case, x would input in binary (base 2), and the output would be the binary representation of $a_0 + a_1x$.

A *Universal Turing machine* (UTM) [Wallace, 2005, sec. 2.2.5] is a Turing machine which can simulate any other Turing machine. So, if U is a UTM and M is

¹Wherever he might or might not have inherited it from, I acknowledge obtaining the figure in fig. 4 from Kevin B. Korb.

a TM, then there is some input c_M such that for any string s , $U(c_M s) = M(s)$ and the output from U when given the input $c_M s$ is identical to the output from M when given input s . In any other words, given any TM M , there is an emulation program [or translation program] (or code) c_M so that once U is input c_M it forever after behaves as though it were M .

The algorithmic complexity (or Kolmogorov complexity), $K_U(x)$, of a string x is the length of the shortest input (l_x) to a Universal Turing Machine such that, given input l_x , U outputs x and then stops. (This is the approach of Kolmogorov [1965] and Chaitin [1966], referred to as stream one in [Wallace and Dowe, 1999a].)

Algorithmic information theory can be used to give the algorithmic probability [Solomonoff, 1960; 1964; 1999; 2008] of a string (x) or alternatively also to insist upon the two-part MML form [Wallace and Dowe, 1999a; Wallace, 2005, secs. 2.2–2.3].

Let us elaborate, initially by recalling the notion of a prefix code (from sec. 2.2) and then by considering possible inputs to a UTM. Let us consider the two binary strings 0 and 1 of length 1, the four binary strings 00, 01, 10 and 11 of length 2, and (in general) the 2^n binary strings of length n . Clearly, if a Turing Machine stops on some particular input (of length n), then it will stop on that input with any suffix appended.

The (unnormalized) probability that a UTM, U , will generate x from random input is $P_U(x) = \sum_{s:U(s)=x} 2^{-\text{length}(s)}$, summing over the strings s such that U taking input s will output x and then stop. In Solomonoff's original predictive specification [Solomonoff, 1960; 1964] (stream two from [Wallace and Dowe, 1999a]), the (unnormalized) summation actually includes more strings (and leads to a greater sum), including [Wallace, 2005, sec. 10.1.3] those strings s such that U on input s produces x and possibly a suffix — i.e., outputs a string for which x is a prefix. For this sum to be finite, we must add the stipulation that the strings s (over which we sum) must form a prefix code. In choosing the strings s to form a prefix code, the sum is not affected by insisting that the strings s are all chosen so that (e.g.) for no prefix s' of s does $U(s') = x$ and then halt. And, for the sum $\sum_x P_U(x)$ to be useful, we must again make sure that the strings x are prefix-free — i.e., that the strings x all together form a prefix code - so as to avoid double-counting.

Clearly, $2^{-K_U(x)} < P_U(x)$, since $K_U(x)$ takes only the shortest (and biggest) [input] term outputting x , whereas $P_U(x)$ takes all the terms which output x (whether or not we wish to also include terms which append a suffix to x). The earlier mention above of “(unnormalized)” is because, for many inputs, the UTM will not halt [Turing, 1936; Chaitin, 2005; Dowe, 2008a, footnote 70]. For the purposes of prediction, these considerations just discussed are sufficient. But, for the purposes of *inference* (or, equivalently, *explanation* or *induction*), we need a *two-part* construction — as per theorems from [Barron and Cover, 1991] and arguments from [Wallace and Freeman, 1987, p. 241; Wallace, 2005, sec. 3.4.5, pp. 190–191] (and some examples of what can go wrong when we don't have a two-part construction [Wallace and Dowe, 1999a, sec. 6.2; 1999b, secs. 1.2, 2.3 and 3; 1999c,

sec. 2]). Our two-part input to the (Universal) Turing Machine will be such that [Wallace and Dowe, 1999a; Wallace, 2005, secs. 2.3.6–2.3.9] the first part results in no output being written but rather the Turing machine is programmed with the hypothesis, H . Now programmed with the hypothesis, H , the Turing machine now looks at the second part of the message (which is possibly the output of a Huffman code) and uses H to write out the data, D . The MML inference will be the hypothesis, H , which is represented by the first part of the shortest two-part input giving rise to the data, D .

The other thing to mention here is the Bayesianism inherent in all these approaches. The Bayesian and (two-part [file compression]) information-theoretic interpretations to MML are both clearly Bayesian. And, although some authors have been known to neglect this (by sweeping Order 1, $O(1)$, terms under the carpet or otherwise neglecting them), the choice of (Universal) Turing Machine in algorithmic information theory is (obviously?) also a Bayesian choice [Wallace and Dowe, 1999a, secs. 2.4 and 7; 1999c, secs. 1–2; Comley and Dowe, 2005, p. 269, sec. 11.3.2; Dowe, 2008a, footnotes 211, 225 and (start of) 133, and sec. 0.2.7, p. 546; 2008b, p. 450].

2.5 Digression on Wallace non-universality probability

This section is a digression and can be safely skipped without any loss of continuity or context, but it does follow on from sec. 2.4 — which is why it is placed here.

The Wallace non-universality probability [Dowe, 2008a, sec. 0.2.2, p. 530, col. 1 and footnote 70] of a UTM, U , is the probability that, given a particular infinitely long random bit string as input, U will become non-universal at some point. Quite clearly, the Wallace non-universality probability (WNUP) equals 1 for all non-universal TMs. Similarly, $\text{WNUP}(U)$ is greater than 0 for all TMs, U ; and WNUP equals 1 for some UTM if and only if it equals 1 for all UTMs. Wallace, others and I believed it to equal 1. In unpublished private communication, George Barmpalias argues that it isn't equal to 1, appealing to a result of Kucera. George is correct (and Chris and I mistaken) if and only if $\inf_{\{U:U \text{ a UTM}\}} \text{WNUP}(U) = 0$.

This section was a digression and could be safely skipped without any loss of continuity or context.

3 PROBABILISTIC INFERENCE, LOG-LOSS SCORING AND KULLBACK-LEIBLER DISTANCE — AND UNIQUENESS

There are many measures of predictive accuracy. The simplest of these, such as on a quiz show, is the number of correct answers (or “right”/“wrong” scoring). There are likewise many measures of how close some estimated function is to the true function from which the data is really coming.

We shall present the notions of probabilistic scoring and of measuring a distance between two functions.

From the notion of probabilistic scoring, we shall present our new apparent uniqueness property of log-loss scoring [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437–438]. From the notion of measuring a distance between two functions, we shall present a related result showing uniqueness (or two versions of uniqueness) of Kullback-Leibler distance [Dowe, 2008a, p. 438]. For those interested in causal decision theory and scoring rules and to those simply interested in scoring rules and scoring probabilities, I highly recommend log-loss scoring and Kullback-Leibler distance — partly for their invariance and partly for their apparent uniqueness in having this invariance.

3.1 “Right”/“wrong” scoring and re-framing

Imagine two different quizzes which are identical apart from their similar but not quite identical beginnings. Quiz 1 begins with a multiple-choice question with 4 possible answers: 0, 1, 2, 3 or (equivalently, in binary) 00, 01, 10, 11.

Quiz 2 begins with 2 multiple-choice questions:

- Q2.1: is the 2nd last bit a 0 or a 1?, and
- Q2.2: is the last bit a 0 or a 1?

Getting a score of 1 correct at the start of quiz 1 corresponds to getting a score of 2 at the start of quiz 2. Getting a score of 0 at the start of quiz 1 corresponds to a score of either 0 or 1 at the start of quiz 2. This seems unfair — so we might try to attribute the problem to the fact that quiz 1 began with 1 4-valued question where quiz 2 began with 2 2-valued questions and explore whether all is fair when (e.g.) all quizzes have 2 2-valued questions.

But consider now quiz 3 which, like quiz 2, begins with 2 2-valued questions, as follows:

- Q3.1: is the 2nd last bit a 0 or a 1?, and
- Q3.2: are the last two bits equal or not equal?

Getting Q3.2 correct means that on quiz 2 we either get 0 (if we get Q2.1 wrong) or 2 (if we get Q2.2 correct and therefore all questions correct).

We see that no matter how we re-frame the question — whether as one big question or as lots of little questions — we get all answers correct on one quiz if and only if we get all answers correct on all quizzes. But, however, as the following example (of Quiz 4 and Quiz 5) demonstrates, we also see that even when all questions are binary (yes/no), it is possible to have two different framings of n questions such that in one such quiz (here, Quiz 4) we have $(n - 1)$ questions answered correctly (and only 1 incorrectly) and in the re-framing to another quiz (here, Quiz 5) all n questions are answered incorrectly.

Quiz 4 (of n questions):

- Q4.1: What is the 1st of the n bits?
- Q4.i ($i = 2, \dots, n$): Is the 1st bit equal to the i^{th} bit?

Quiz 5 (of n questions):

- Q5.1: What is the 1st of the n bits?
- Q5.i ($i = 2, \dots, n$): What is the i^{th} bit?

If the correct bit string is $0^n = 0\dots 0$ and our guess is $1^n = 1\dots 1$, then on Quiz 4 we will get $(n - 1)$ correct (and 1 wrong) whereas on quiz 5 we will get 0 correct (and all n wrong).

This said by way of motivation, we now look at forms of prediction that remain invariant to re-framing — namely, probabilistic prediction with $\log(\text{arithm})$ -loss — and we present some recent uniqueness results [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437-438] here.

3.2 *Scoring predictions, probabilistic predictions and log-loss*

The most common form of prediction seems to be a prediction without a probability or anything else to quantify it. Nonetheless, in some forms of football, the television broadcaster sometimes gives an estimated probability of the kicker scoring a goal — based on factors such as distance, angle and past performance. And, of course, if it is possible to wager a bet on the outcome, then accurately estimating the probability (and comparing this with the potential pay-out if successful) will be of greater interest.

Sometimes we don't care overly about a probability estimate. On some days, we might merely wish to know whether or not it is more probable that it will rain or that it won't. On such occasions, whether it's 52% probable or 97% probable that it will rain, we don't particularly care beyond noting that both these numbers are greater than 50% and we'll take our umbrella with us in either case.

And sometimes we most certainly want a good and reliable probability estimate. For example, a patient reporting with chest pains doesn't want to be told that there's only a 40% chance that you're in serious danger (with a heart attack), so you can go now. And nor does an inhabitant of an area with the impending approach of a raging bush-fire want to be told that there's only a 45% chance of your dying or having serious debilitation if you stay during the fire, so you might as well stay. The notion of "reasonable doubt" in law is pertinent here — and, without wanting to seem frivolous, so, too, is the notion of when a cricket umpire should or shouldn't give the "benefit of the doubt" to the person batting (in l.b.w. or other contentious decisions).

Now, it is well-known that with logarithm-loss function ($\log p$) for scoring probabilistic predictions, the optimal strategy is to give the true probability, if known.

This property also holds true for quadratic loss $((1-p)^2)$ and has also been shown to be able to hold for certain other functions of probability [Deakin, 2001]. What we will show here is our new result that the logarithm-loss (or log-loss) function has an apparent uniqueness property on re-framing of questions [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437–438].

Let us now consider an example involving (correct) diagnosis of a patient. (With apologies to any and all medically-informed human earthlings of the approximate time of writing, the probabilities in the discussion(s) below might be from non-earthlings, non-humans and/or from a different time.) We'll give four possibilities:

1. no diabetes, no hypothyroidism
2. diabetes, but no hypothyroidism
3. no diabetes, but hypothyroidism
4. both diabetes and hypothyroidism.

Of course, rather than present this as one four-valued diagnosis question, we could have presented it in a variety of different ways.

As a second possibility, we could have also asked, e.g., the following two two-valued questions:

1. no diabetes
2. diabetes

and

1. hypothyroidism
2. no hypothyroidism.

As another (third) alternative line, we could have begun with

1. no condition present
2. at least one condition present,

and then finished if there we no condition present but, if there were at least one condition present, instead then continued with the following 3-valued question:

1. diabetes, but no hypothyroidism
2. no diabetes, but hypothyroidism
3. both diabetes and hypothyroidism.

To give a correct diagnosis, in the original setting, this requires answering one question correctly. In the second setting, it requires answering exactly two questions correctly — and in the third setting, it might require only answering one question correctly but it might require answering two questions correctly.

Clearly, then, the number of questions answered correctly is not invariant to the re-framing of the question. However, the sum of logarithms of probabilities *is* invariant, and — apart from (quite trivially, multiplying it by a constant, or) adding a constant multiple of the entropy of the prior distribution - would appear to be unique in having this property.

Let us give two examples of this. In the first example, our conditions will be independent of one another. In the second example, our conditions will be dependent upon one another.

So, in the first case, with the conditions independent of one another, suppose the four estimated probabilities are

1. no diabetes, no hypothyroidism; probability $1/12$
2. diabetes, but no hypothyroidism; probability $2/12 = 1/6$
3. no diabetes, but hypothyroidism; probability $3/12 = 1/4$
4. both diabetes and hypothyroidism; probability $6/12 = 1/2$.

Then, in the second possible way we had of looking at it (with the two given two-valued questions), in the first case we have

1. no diabetes; probability $1/3$
2. diabetes; probability $2/3$

and — because the diseases are supposedly independent of one another in the example — we have

1. no hypothyroidism; probability $1/4$
2. hypothyroidism; probability $3/4$.

The only possible way we can have an additive score for both the diabetes question and the hypothyroid question separately is to use some multiple of the logarithms. This is because the probabilities are multiplying together and we want some score that adds across questions, so it must be (a multiple of) the logarithm of the probabilities.

In the third alternative way that we had of looking at it, $\Pr(\text{no condition present}) = 1/12$. If there is no condition present, then we do need need to ask the remaining question. But, in the event (of probability $11/12$) that at least one condition is present, then we have

1. diabetes, but no hypothyroidism; probability $2/11$

2. no diabetes, but hypothyroidism; probability $3/11$
3. both diabetes and hypothyroidism; probability $6/11$.

And the logarithm of probability score again works.

So, our logarithm of probability score worked when the conditions were assumed to be independent of one another.

We now consider an example in which they are not independent of one another.

Suppose the four estimated probabilities are

1. no diabetes, no hypothyroidism; probability 0.1
2. diabetes, but no hypothyroidism; probability 0.2
3. no diabetes, but hypothyroidism; probability 0.3
4. both diabetes and hypothyroidism; probability 0.4.

Then, in the second possible way we had of looking at it (with the two given two-valued questions), for the first of the two two-valued questions we have

1. no diabetes; probability 0.4
2. diabetes; probability 0.6

and then for the second of which we have either

1. no hypothyroidism; $\text{prob}(\text{no hypothyroidism} \mid \text{no diabetes}) = 0.1/(0.1 + 0.3) = 0.1/0.4 = 0.25$
2. hypothyroidism; $\text{prob}(\text{hypothyroidism} \mid \text{no diabetes}) = 0.3/(0.1 + 0.3) = 0.3/0.4 = 0.75$

or

1. no hypothyroidism; $\text{prob}(\text{no hypothyroidism} \mid \text{diabetes}) = 0.2/(0.2 + 0.4) = 0.2/0.6 = 1/3$
2. hypothyroidism; $\text{prob}(\text{hypothyroidism} \mid \text{diabetes}) = 0.4/(0.2+0.4) = 0.4/0.6 = 2/3$.

And in the third alternative way of looking at this, $\text{prob}(\text{at least one condition present}) = 0.9$. If there is no condition present, then we do need need to ask the remaining question. But, in the event (of probability $9/10$) that at least one condition is present, then we have the following three-way question:

1. diabetes, but no hypothyroidism; probability $2/9$
2. no diabetes, but hypothyroidism; probability $3/9$
3. both diabetes and hypothyroidism; probability $4/9$.

We leave it to the reader to verify that the logarithm of probability score again works, again remaining invariant to the phrasing of the question. (Those who would rather see the above examples worked through with general probabilities rather than specific numbers are referred to a similar calculation in sec. 3.6.)

Having presented in sec. 3.1 the problems with “right”/“wrong” scoring and having elaborated on the uniqueness under re-framing of log(arithm)-loss scoring [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437–438] above, we next mention at least four other matters.

First, borrowing from the spirit of an example from [Wallace and Patrick, 1993], imagine we have a problem of inferring a binary (2-class) output and we have a binary choice (or a binary split in a decision tree) with the following output distributions. For the “no”/“left” branch we get 95 in class 1 and 5 in class 2 (i.e., 95:5), and for the “yes”/“right” branch we get 55 in class 1 and 45 in class 2 (i.e., 55:45).

Because both the “no”/“left” branch and the “yes”/“right” branch give a majority in class 1, someone only interested in “right”/“wrong” score would fail to pick up on the importance and significance of making this split, simply saying that one should always predict class 1. Whether class 1 pertains to heart attack, dying in a bush-fire or something far more innocuous (such as getting wet in light rain), by reporting the probabilities we don’t run the risk of giving the wrong weights to (so-called) type I and type II errors (also known as false positives and false negatives). (Digressing, readers who might incidentally be interested in applying Minimum Message Length (MML) to hypothesis testing are referred to [Dowe, 2008a, sec. 0.2.5, p. 539 and sec. 0.2.2, p. 528, col. 1; 2008b, p. 433 (Abstract), p. 435, p. 445 and pp. 455–456; Musgrave and Dowe, 2010].) If we report a probability estimate of (e.g.) 45%, 10%, 5%, 1% or 0.1%, we leave it to someone else to determine the appropriate level of risk associated with a false positive in diagnosing heart attack, severe bush-fire danger or getting caught in light drizzle rain.

And we now mention three further matters.

First, some other uniqueness results of log(arithm)-loss scoring are given in [Milne, 1996; Huber, 2008]. Second, in binary (yes/no) multiple-choice questions, it is possible (and not improbable for a small number of questions) to serendipitously fluke a good “right”/“wrong” score, even if the probabilities are near 50%-50%, and with little or no risk of downside. But with log(arithm)-loss scoring, if the probabilities are near 50%-50% (or come from random noise and are 50%-50%), then predictions with more extreme probabilities are fraught with risk. And, third, given all these claims about the uniqueness of log(arithm)-loss scoring in being invariant to re-framing (as above) [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437–438] and having other desirable properties [Milne, 1996; Huber, 2008], we discuss in sec. 3.3 how a quiz show contestant asked to name (e.g.) a city or a person and expecting to be scored on “right”/“wrong” could instead give a probability distribution over cities or people’s names and be scored by

$\log(\text{arithm})$ -loss.

3.3 Probabilistic scoring on quiz shows

Very roughly, one could have probabilistic scoring on quiz shows as follows. For a multiple-choice answer, things are fairly straightforward as above. But what if, e.g., the question asks for a name or a date (such as a year)?

One could give a probability distribution over the length of the name. For example, the probability that the name has length l might be $1/2^l = 2^{-l}$ for $l = 1, 2, \dots$. Then, for each of the l places, there could be 28 possibilities (the 26 possibilities a, ..., z, and the 2 possibilities space “ ” and hyphen “-”) of probability $1/28$ each. So, for example, as a default, “Wallace” would have a probability of $2^{-7} \times (1/28)^7$. Call this distribution Default. (Of course, we could refine Default by, e.g., noticing that the first character will neither be a space “ ” nor a hyphen “-” and/or by (also) noticing that both the space “ ” and the hyphen “-” are never followed by a space “ ” or a hyphen “-”.) For the user who has some idea rather than no proverbial idea about the answer, it is possible to construct hybrid distributions. So, if we wished to allocate probability $1/2$ to “Gauss” and probability $1/4$ to “Fisher” and otherwise we had no idea, then we could give a probability of $1/2$ to “Gauss”, $1/4$ to “Fisher” and for all other answers our probability we would roughly be $1/4 \times \text{Default}$. A similar comment applies to someone who thinks that (e.g.) there is a 0.5 probability that the letter “a” occurs at the start and a 0.8 probability that the letter “c” appears at least once or instead that (e.g.) a letter “q” not at the end can only be followed by an “a”, a “u”, a space “ ” or a hyphen “-”.

If a question were to be asked about the date or year of such and such an event, one could give a ([truncated] Gaussian) distribution on the year in the same way that entrants having been giving (Gaussian) distributions on the margin of Australian football games since early 1996 as per sec. 3.5.

It would be nice to see a (television) quiz (show) one day with this sort of scoring system. One could augment the above as follows. Before questions were asked to single contestants — and certainly before questions were thrown upon to multiple contestants and grabbed by the first one pressing the buzzer — it could be announced what sort of question it was (e.g., multiple-choice [with n options] or open-ended) and also what the bonus in bits was to be. The bonus in bits (as in the constant to be added to the logarithm of the contestant’s allocated probability to the correct answer) could relate to the ([deemed] prior) difficulty of the question, (perhaps) as per sec. 3.4 — where there is a discussion of adding a term corresponding to (a multiple of) the $\log(\text{arithm})$ of the entropy of some Bayesian prior over the distribution of possible answers. And where quiz shows have double points in the final round, that same tradition could be continued.

One comment perhaps worth adding here is that in typical quiz shows, as also in the probabilistic and Gaussian competitions running on the Australian Football League (AFL) football competitions from sec. 3.5, entrants are regularly updated

on the scores of all their opponents. In the AFL football competitions from sec. 3.5, this happens at the end of every round. In quiz shows, this often also happens quite regularly, sometimes after every question. The comment perhaps worth adding is that, toward the end of a probabilistic competition, contestants trying to win who are not winning and who also know that they are not winning might well nominate “aggressive” choices of probabilities that do not represent their true beliefs and which they probably would not have chosen if they had not been aware that they weren’t winning.

Of course, if (all) this seems improbable, then please note from the immediately following text and from sec. 3.5 that each year since the mid-1990s there have been up to hundreds of people (including up to dozens or more school students, some in primary school) giving not only weekly probabilistic predictions on results of football games but also weekly predictions of probability distributions on the margins of these games. And these have all been scored with $\log(\text{arithm})$ -loss.

3.4 Entropy of prior and other comments on log-loss scoring

Before finishing with some references to papers in which we have used $\log(\text{arithm})$ -loss probabilistic (“bit cost”) scoring and mentioning a log-loss probabilistic competition we have been running on the Australian Football League (AFL) since 1995 in sec. 3.5, two other comments are worth making in this section.

Our first comment is to return to the issue of quadratic loss $((1 - p)^2$, which is a favourite of many people) and also the loss functions suggested more recently by Deakin [2001]. While it certainly appears that only $\log(\text{arithm})$ -loss (and multiples thereof) retains invariance under re-framing, we note that

$$-\log(p_1^n p_2^n \dots p_m^n) = -n \sum_{i=1}^m \log p_i \quad (4)$$

So, although quadratic loss $((1 - p)^2)$ does not retain invariance when adding scores between questions, if we were for some reason to want to *multiply* scores between questions (rather than add, as per usual), then the above relationship in equation (4) between a power score (quadratic score with $n = 2$) and $\log(\text{arithmic})$ score — namely, that the log of a product is the sum of the logs — might possibly enable some power loss to be unique in being invariant under re-framing (upon *multiplication*).

The other comment to make about the uniqueness of $\log(\text{arithm})$ -loss (upon addition) under re-framing is that we can also add a term corresponding to (a multiple of) the entropy of the Bayesian prior [Tan and Dowe, 2006, sec. 4.2; Dowe, 2008a, footnote 176; 2008b, p. 438]. (As is hinted at in [Tan and Dowe, 2006, sec. 4.2, p. 600] and explained in [Dowe, 2008a, footnote 176], the idea for this arose in December 2002 from my correcting a serious mathematical flaw in [Hope and Korb, 2002]. Rather than more usefully use the fact that a logarithm of probability ratios is the difference of logarithms of probabilities, [Hope and Korb,

2002] instead suggests using a ratio of logarithms — and this results in a system where the optimal score will rarely be obtained by using the true probability.)

Some of many papers using log(arithm)-loss scoring include [Good, 1952] (where it was introduced for the binomial distribution) and [Dowe and Krusel, 1993, p. 4, Table 3; Dowe *et al.*, 1996d; Dowe *et al.*, 1996a; Dowe *et al.*, 1996b; Dowe *et al.*, 1996c; Dowe *et al.*, 1998, sec. 3; Needham and Dowe, 2001, Figs. 3–5; Tan and Dowe, 2002, sec. 4; Kornienko *et al.*, 2002, Table 2; Comley and Dowe, 2003, sec. 9; Tan and Dowe, 2003, sec. 5.1; Comley and Dowe, 2005, sec. 11.4.2; Tan and Dowe, 2004, sec. 3.1; Kornienko *et al.*, 2005a, Tables 2–3; Kornienko *et al.*, 2005b; Tan and Dowe, 2006, secs. 4.2–4.3] (and possibly also [Tan *et al.*, 2007, sec. 4.3]), [Dowe, 2007; 2008a, sec. 0.2.5, footnotes 170–176 and accompanying text; 2008b, pp. 437–438]. The 8-class multinomial distribution in [Dowe and Krusel, 1993, p. 4, Table 3] (from 1993) is the first case we are aware of in which log-loss scoring is used for a distribution which is not binomial, and [Dowe *et al.*, 1996d; 1996a; 1996b; 1996c] (from 1996) are the first cases we are aware of in which log-loss scoring was used for the Normal (or Gaussian) distribution.

3.5 Probabilistic prediction competition(s) on Australian football

A log(arithm)-loss probabilistic prediction competition was begun on the outcome of Australian Football League (AFL) matches in 1995 [Dowe and Lentin, 1995; Dowe, 2008b, p. 48], just before Round 3 of the AFL season. In 1996, this was extended by the author to a log(arithm)-loss Gaussian competition on the margin of the game, in which competition entrants enter a μ and a σ — in order to give a predictive distribution $N(\mu, \sigma)$ on the margin — for each game [Dowe *et al.*, 1996d; 1996a; 1996b; 1996c; 1998, sec. 3; Dowe, 2008a, sec. 0.2.5]. These log-loss compression-based competitions, with scores in bits (of information), have been running non-stop ever since their inceptions, having been put on the WWW in 1997 and at their current location of www.csse.monash.edu.au/~footy since 1998 [Dowe, 2008a, footnote 173]. (And thanks to many, especially Torsten Seemann for all the unsung behind the scenes support in keeping these competitions going [Dowe, 2008a, footnote 217].) The optimal long-term strategy in the log(arithm)-loss probabilistic AFL prediction competition would be to use the true probability if one knew it. Looking ahead to sec. 3.6, the optimal long-term strategy in the Gaussian competition would be to choose μ and σ so as to minimise the Kullback-Leibler distance from the (true) distribution on the margin to $N(\mu, \sigma^2)$. (In the log-loss probabilistic competition, the “true” probability is also the probability for which the Kullback-Leibler distance from the (true) distribution is minimised.) Competitions concerned with minimising sum of squared errors can still be regarded as compression competitions motivated by (expected) Kullback-Leibler distance minimisation, as they are equivalent to Gaussian competitions with σ fixed (where σ can be presumed known or unknown, as long as it’s fixed).

3.6 Kullback-Leibler “distance” and measuring “distances” between two functions

Recall from sec. 2.3 that the optimal way of encoding some probability distribution, f , is with code-words of length $-\log f$, and that the average (or expected) cost is $\sum_{i=1}^m f_i \times (-\log f_i)$, also known as the entropy. The log(arithm)-loss score is obtained by sampling from some real-world data. If we are sampling from some true (known) distribution, then the optimal long-run average that we can expect will be the entropy.

One way of thinking about the Kullback-Leibler divergence (or “distance”) between two distributions, f and g , is as the inefficiency (or sub-optimality) of encoding f using g rather than (the optimal) f . Equivalently, one can think about the Kullback-Leibler distance between two distributions, f and g , as the average (or expected) cost of sampling from distribution f and coding with the corresponding cost of $-\log g$ minus the entropy of f — and, of course, the entropy of f (namely, $-\sum f \log f$) is independent of g .

Recalling equations (2) and (3) for the entropy of discrete and continuous models respectively, the Kullback-Leibler distance from f to g is

$$\left(\sum_{i=1}^m f \times (-\log g_i)\right) - \left(\sum_{i=1}^m f \times (-\log f_i)\right) = \sum_{i=1}^m f \times \log(f_i/g_i) \quad (5)$$

for the discrete distribution. As one might expect, for a continuous distribution, the Kullback-Leibler distance from f to g is

$$\begin{aligned} & \left(\int f(x) \times (-\log g(x)) \, dx\right) - \left(\int f(x) \times (-\log f(x)) \, dx\right) \\ &= \int f \times \log(f(x)/g(x)) \, dx = \int dx \, f \times \log(f/g) \end{aligned} \quad (6)$$

We should mention here that many refer to the Kullback-Leibler “distance” as Kullback-Leibler *divergence* because it is not — in general — symmetrical. In other words, there are plenty of examples when $KL(f, g)$, or, equivalently, $\Delta(g||f)$, is not equal to $KL(g, f) = \Delta(f||g)$.

In sec. 3.2, we showed a new result about uniqueness of log(arithm)-loss scoring in terms of being invariant under re-framing of the problem [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437–438]. It turns out that there is a similar uniqueness about Kullback-Leibler distance in terms of being invariant to re-framing of the problem [Dowe, 2008b, p. 438]. Despite some mathematics to follow which some readers might find slightly challenging in places, this follows intuitively because (the entropy of the true distribution, f , is independent of g and) the $-\log g$ term in the Kullback-Leibler distance is essentially the same as the log(arithm)-loss term in sec. 3.2.

Before proceeding to this example, we first note that the Kullback-Leibler distance is quite clearly invariant under re-parameterisations such as (e.g.) transforming from polar co-ordinates (x, y) to Cartesian co-ordinates $(r = \text{sign}(x))$.

$\sqrt{x^2 + y^2}$, $\theta = \tan^{-1}(y/x)$) and back from Cartesian co-ordinates (r, θ) to polar co-ordinates $(x = r \cos \theta, y = r \sin \theta)$.

We give an example of our new result (about the uniqueness of Kullback-Leibler distance in terms of being invariant to re-framing) below, letting f and g both have 4 states, with probabilities $f_{1,1}, f_{1,2}, f_{2,1}, f_{2,2}, g_{1,1}, g_{1,2}, g_{2,1}$ and $g_{2,2}$ respectively. The reader is invited to compare the example below with those from sec. 3.2. The reader who would prefer to see specific probabilities rather than these general probabilities is suggested more strongly to compare with sec. 3.2.

$$\begin{aligned} KL(f, g) &= \Delta(g||f) = \sum_{i=1}^2 \sum_{j=1}^2 f_{i,j} (\log f_{i,j} - \log g_{i,j}) \\ &= \sum_{i=1}^2 \sum_{j=1}^2 f_{i,j} \log(f_{i,j}/g_{i,j}) \end{aligned} \quad (7)$$

Another way of looking at the Kullback-Leibler “distance” (or Kullback-Leibler divergence, or KL-distance) is to say that a proportion $f_{1,\cdot} = f_{1,1} + f_{1,2}$ of the time, we have the Kullback-Leibler distance to the corresponding cross-section $(g_{1,\cdot})$ of g , and then the remaining proportion $1 - f_{1,\cdot} = f_{2,\cdot} = f_{2,1} + f_{2,2}$ of the time, we have the Kullback-Leibler distance to the corresponding (other) cross-section $(g_{2,\cdot})$ of g .

To proceed down this path, we have to do the calculations at two levels. (Analogously with sec. 3.2, we could do the calculation in one step involving four terms or break it up into two levels of parts each involving two terms.) At the top level, we have to calculate the KL-divergence from the binomial distribution $(f_{1,\cdot}, f_{2,\cdot})$ to the binomial distribution $(g_{1,\cdot}, g_{2,\cdot})$. This top-level KL-divergence is

$$f_{1,\cdot} \log(f_{1,\cdot}/g_{1,\cdot}) + f_{2,\cdot} \log(f_{2,\cdot}/g_{2,\cdot}) \quad (8)$$

It then remains to go to the next level (or step) and first look at the binomial distribution $f_{1,1}/(f_{1,1} + f_{1,2})$ and $f_{1,2}/(f_{1,1} + f_{1,2})$ versus $g_{1,1}/(g_{1,1} + g_{1,2})$ and $g_{1,2}/(g_{1,1} + g_{1,2})$ (on the first or left branch), and then to look at the binomial distribution $f_{2,1}/(f_{2,1} + f_{2,2})$ and $f_{2,2}/(f_{2,1} + f_{2,2})$ versus $g_{2,1}/(g_{2,1} + g_{2,2})$ and $g_{2,2}/(g_{2,1} + g_{2,2})$ (on the second or right branch). Note that the first of these (KL-divergences or) coding inefficiencies will only occur a proportion $f_{1,1} + f_{1,2} = f_{1,\cdot}$ of the time, and the second of these (KL-divergences or) coding inefficiencies will only occur a proportion $f_{2,1} + f_{2,2} = f_{2,\cdot} = 1 - f_{1,\cdot}$ of the time.

The first of these KL-divergences is the (expected or) average coding inefficiency when encoding $(f_{1,1}/(f_{1,1} + f_{1,2}), f_{1,2}/(f_{1,1} + f_{1,2}))$ not using itself, but rather instead (sub-optimally) using $(g_{1,1}/(g_{1,1} + g_{1,2}), g_{1,2}/(g_{1,1} + g_{1,2}))$. This first KL-divergence is

$$\begin{aligned} &f_{1,1}/(f_{1,1} + f_{1,2}) \log((f_{1,1}/(f_{1,1} + f_{1,2}))/ (g_{1,1}/(g_{1,1} + g_{1,2}))) \\ &+ f_{1,2}/(f_{1,1} + f_{1,2}) \log((f_{1,2}/(f_{1,1} + f_{1,2}))/ (g_{1,2}/(g_{1,1} + g_{1,2}))) \end{aligned}$$

$$\begin{aligned}
&= (f_{1,1}/(f_{1,1} + f_{1,2})) \times ((\log f_{1,1}/g_{1,1}) - (\log((f_{1,1} + f_{1,2})/(g_{1,1} + g_{1,2})))) \\
&+ (f_{1,2}/(f_{1,1} + f_{1,2})) \times ((\log f_{1,2}/g_{1,2}) - (\log((f_{1,1} + f_{1,2})/(g_{1,1} + g_{1,2})))) \\
&= (f_{1,1}/f_{1,\cdot}) \times (\log(f_{1,1}/g_{1,1}) - \log(f_{1,\cdot}/g_{1,\cdot})) \\
&+ (f_{1,2}/f_{1,\cdot}) \times (\log(f_{1,2}/g_{1,2}) - \log(f_{1,\cdot}/g_{1,\cdot})) \tag{9}
\end{aligned}$$

Changing the very first subscript from a 1 to a 2, we then get that the second of these KL-divergences, namely the KL-divergence from the binomial distribution $(f_{2,1}/(f_{2,1} + f_{2,2}), f_{2,2}/(f_{2,1} + f_{2,2}))$ to the binomial distribution $(g_{2,1}/(g_{2,1} + g_{2,2}), g_{2,2}/(g_{2,1} + g_{2,2}))$, is

$$\begin{aligned}
&f_{2,1}/(f_{2,1} + f_{2,2}) \log((f_{2,1}/(f_{2,1} + f_{2,2}))/((g_{2,1}/(g_{2,1} + g_{2,2})))) \\
&+ f_{2,2}/(f_{2,1} + f_{2,2}) \log((f_{2,2}/(f_{2,1} + f_{2,2}))/((g_{2,2}/(g_{2,1} + g_{2,2})))) \\
&= (f_{2,1}/(f_{2,1} + f_{2,2})) \times ((\log f_{2,1}/g_{2,1}) - (\log((f_{2,1} + f_{2,2})/(g_{2,1} + g_{2,2})))) \\
&+ (f_{2,2}/(f_{2,1} + f_{2,2})) \times ((\log f_{2,2}/g_{2,2}) - (\log((f_{2,1} + f_{2,2})/(g_{2,1} + g_{2,2})))) \\
&= (f_{2,1}/f_{2,\cdot}) \times (\log(f_{2,1}/g_{2,1}) - \log(f_{2,\cdot}/g_{2,\cdot})) \\
&+ (f_{2,2}/f_{2,\cdot}) \times (\log(f_{2,2}/g_{2,2}) - \log(f_{2,\cdot}/g_{2,\cdot})) \tag{10}
\end{aligned}$$

The first coding inefficiency, or KL-divergence, given in equation (9), occurs a proportion $f_{1,\cdot} = (f_{1,1} + f_{1,2})$ of the time. The second coding inefficiency, or KL-divergence, given in equation (10), occurs a proportion $f_{2,\cdot} = (f_{2,1} + f_{2,2}) = (1 - (f_{1,1} + f_{1,2})) = 1 - f_{1,\cdot}$ of the time.

So, the total expected (or average) coding inefficiency of using g when we should be using f , or equivalently the KL-divergence from f to g , is the following sum: the inefficiency from equation (8) + $(f_{1,\cdot} \times (\text{the inefficiency from equation (9)}))$ + $(f_{2,\cdot} \times (\text{the inefficiency from equation (10)}))$.

Before writing out this sum, we note that

$$\begin{aligned}
&(f_{1,\cdot} \times (\text{the inefficiency from equation (9)})) \\
&= f_{1,1} \times (\log(f_{1,1}/g_{1,1}) - \log(f_{1,\cdot}/g_{1,\cdot})) \\
&+ f_{1,2} \times (\log(f_{1,2}/g_{1,2}) - \log(f_{1,\cdot}/g_{1,\cdot})) \tag{11}
\end{aligned}$$

and similarly that

$$\begin{aligned}
&(f_{2,\cdot} \times (\text{the inefficiency from equation (10)})) \\
&= f_{2,1} \times (\log(f_{2,1}/g_{2,1}) - \log(f_{2,\cdot}/g_{2,\cdot})) \\
&+ f_{2,2} \times (\log(f_{2,2}/g_{2,2}) - \log(f_{2,\cdot}/g_{2,\cdot})) \tag{12}
\end{aligned}$$

Now, writing out this sum, summing equations (8), (11) and (12), it is

$$\begin{aligned}
&f_{1,\cdot} \log(f_{1,\cdot}/g_{1,\cdot}) + f_{2,\cdot} \log(f_{2,\cdot}/g_{2,\cdot}) \\
&+ f_{1,1} \times (\log(f_{1,1}/g_{1,1}) - \log(f_{1,\cdot}/g_{1,\cdot})) \\
&+ f_{1,2} \times (\log(f_{1,2}/g_{1,2}) - \log(f_{1,\cdot}/g_{1,\cdot}))
\end{aligned}$$

$$\begin{aligned}
& + f_{2,1} \times (\log(f_{2,1}/g_{2,1}) - \log(f_{2,\cdot}/g_{2,\cdot})) \\
& + f_{2,2} \times (\log(f_{2,2}/g_{2,2}) - \log(f_{2,\cdot}/g_{2,\cdot})) \\
= & f_{1,1} \log(f_{1,1}/g_{1,1}) + f_{1,2} \log(f_{1,2}/g_{1,2}) \\
& + f_{2,1} \log(f_{2,1}/g_{2,1}) + f_{2,2} \log(f_{2,2}/g_{2,2}) \\
= & \sum_{i=1}^2 \sum_{j=1}^2 f_{i,j} \log(f_{i,j}/g_{i,j}) \\
= & \sum_{i=1}^2 \sum_{j=1}^2 f_{i,j} (\log f_{i,j} - \log g_{i,j}) = KL(f, g) = \Delta(g||f) \quad (13)
\end{aligned}$$

thus reducing to our very initial expression (7).

Two special cases are worth noting. The first special case is when events are independent — and so $f_{i,j} = f_{i,\cdot} \times \phi_j = f_i \times \phi_j$ and $g_{i,j} = g_{i,\cdot} \times \gamma_j = g_i \times \gamma_j$ for some ϕ_1, ϕ_2, γ_1 and γ_2 (and $f_1 + f_2 = 1, g_1 + g_2 = 1, \phi_1 + \phi_2 = 1$ and $\gamma_1 + \gamma_2 = 1$). In this case, following from equation (7), we get

$$\begin{aligned}
KL(f, g) &= \Delta(g||f) = \sum_{i=1}^2 \sum_{j=1}^2 f_{i,j} \log(f_{i,j}/g_{i,j}) = \sum_{i=1}^2 \sum_{j=1}^2 f_i \phi_j \log(f_i \phi_j / (g_i \gamma_j)) \\
&= \sum_{i=1}^2 \sum_{j=1}^2 (f_i \phi_j \log(f_i/g_i) + f_i \phi_j \log(\phi_j/\gamma_j)) \\
&= \left(\sum_{i=1}^2 f_i \log(f_i/g_i) \right) + \left(\sum_{j=1}^2 \phi_j \log(\phi_j/\gamma_j) \right) \quad (14)
\end{aligned}$$

$$\begin{aligned}
&= \left(\sum_{i=1}^2 f_i \log(1/g_i) \right) + \left(\sum_{j=1}^2 \phi_j \log(1/\gamma_j) \right) \\
&\quad - \left(\sum_{i=1}^2 \sum_{j=1}^2 f_i \phi_j \log(f_i \phi_j) \right) \quad (15)
\end{aligned}$$

$$= \left(\sum_{i=1}^2 f_i \log(1/g_i) \right) + \left(\sum_{j=1}^2 \phi_j \log(1/\gamma_j) \right) - (\text{Entropy of } f) \quad (16)$$

We observe in this case of the distributions being independent that the Kullback-Leibler scores de-couple in exactly the same uniquely invariant way as they do for the probabilistic predictions in sec. 3.2.

The second special case of particular note is when probabilities are correct in both branching paths — i.e., $f_{1,1}/f_{1,\cdot} = g_{1,1}/g_{1,\cdot}$, $f_{1,2}/f_{1,\cdot} = g_{1,2}/g_{1,\cdot}$, $f_{2,1}/f_{2,\cdot} = g_{2,1}/g_{2,\cdot}$ and $f_{2,2}/f_{2,\cdot} = g_{2,2}/g_{2,\cdot}$. In this case, starting from equation (7), we get

$$KL(f, g) = \Delta(g||f) = \sum_{i=1}^2 \sum_{j=1}^2 f_{i,j} \log(f_{i,j}/g_{i,j})$$

$$\begin{aligned}
&= \sum_{i=1}^2 \sum_{j=1}^2 f_{i,\cdot}(f_{i,j}/f_{i,\cdot}) \log((f_{i,\cdot}(f_{i,j}/f_{i,\cdot}))/g_{i,\cdot}(g_{i,j}/g_{i,\cdot})) \\
&= \sum_{i=1}^2 \sum_{j=1}^2 f_{i,\cdot}(f_{i,j}/f_{i,\cdot}) \log((f_{i,\cdot}/g_{i,\cdot}) \times [(f_{i,j}/f_{i,\cdot})/(g_{i,j}/g_{i,\cdot})]) \\
&= \sum_{i=1}^2 \sum_{j=1}^2 f_{i,\cdot}(f_{i,j}/f_{i,\cdot}) \log(f_{i,\cdot}/g_{i,\cdot}) \\
&= \sum_{i=1}^2 f_{i,\cdot} \log(f_{i,\cdot}/g_{i,\cdot}) = \sum_{i=1}^2 f_{i,\cdot} \log(1/g_{i,\cdot}) - (\text{Entropy of } f) \quad (17)
\end{aligned}$$

We observe in this second special case of probabilities being correct in both branching paths that the divergence between the two distributions is the same as that between the binomial distributions $(f_{1,\cdot}, f_{2,\cdot} = 1 - f_{1,\cdot})$ and $(g_{1,\cdot}, g_{2,\cdot} = 1 - g_{1,\cdot})$, exactly as it should and exactly as it would be for the uniquely invariant $\log(\text{arithm})$ -loss (“bit cost”) scoring of the probabilistic predictions in sec. 3.2.

Like the invariance of the $\log(\text{arithm})$ -loss scoring of probabilistic predictions under re-framing (whose uniqueness is introduced in [Dowe, 2008a, footnote 175 (and 176)] and discussed in [Dowe, 2008b, pp. 437–438] and sec. 3.2), this invariance of the Kullback-Leibler divergence to the re-framing of the problem is [Dowe, 2008b, p. 438] due to the fact(s) that (e.g.) $\log(f/g) = \log f - \log g$ and $-\log(f_{1,1}/(f_{1,1} + f_{1,2})) + \log f_{1,1} = \log(f_{1,1} + f_{1,2})$.

A few further comments are warranted by way of alternative measures of “distance” or divergence between probability distributions. First, where one can define the notion of the distance remaining invariant under re-framing for these, the Bhattacharyya distance, Hellinger distance and Mahalanobis distance are all not invariant under re-framing. Versions of distance or divergence based on the Rényi entropy give invariance in the trivial case that $\alpha = 0$ (where the distance will always be 0) and the case that $\alpha = 1$ (where we get the Shannon entropy and, in turn, the Kullback-Leibler distance that we are currently advocating).

A second further comment is that, just as in [Tan and Dowe, 2006, sec. 4.2], we can also add a term corresponding to (a multiple of) the entropy of the Bayesian prior. Just like the Kullback-Leibler divergence (and any multiple of it), this (and any multiple of it) will also remain invariant under re-parameterisation or other re-framing [Dowe, 2008a, footnote 176; 2008b, p. 438].

A third — and important — further comment is that it is not just the Kullback-Leibler distance from (say, the true distribution) f to (say, the inferred distribution) g , $KL(f, g) = \Delta(g||f)$, that is invariant and appears to be uniquely invariant under re-framing, but clearly also $KL(g, f) = \Delta(f||g)$ is invariant, as will also be a sum of any linear combination of these, such as (e.g.) $\alpha KL(f, g) + (1 - \alpha) KL(g, f)$ (with $0 \leq \alpha \leq 1$, although this restriction is not required for invariance) [Dowe, 2008b, p. 438]. The case of $\alpha = 1/2$ gives the symmetric Kullback-Leibler distance.

The notion of Kullback-Leibler distance can be extended quite trivially — as

above — to hybrid continuous and discrete Bayesian net graphical models [Tan and Dowe, 2006, sec. 4.2; Dowe 2008a, sec. 0.2.5; 2008b, p. 436] (also see sec. 7.6) or mixture models [Dowe, 2008b, p. 436], etc. For the hybrid continuous and discrete Bayesian net graphical models in [Comley and Dowe, 2003; 2005] (which resulted at least partly from theory advocated in [Dowe and Wallace, 1998]), the log-loss scoring approximation to Kullback-Leibler distance has been used [Comley and Dowe, 2003, sec. 9].

4 OCKHAM'S RAZOR (AND MISUNDERSTANDINGS) AND MML

Let us recall Minimum Message Length (MML) from secs. 1 and 2.4, largely so that we can now compare and contrast MML with Ockham's razor (also written as Occam's razor). Ockham's razor, as it is commonly interpreted, says that if two theories fit the data equally well then prefer the simplest (e.g., [Wallace, 1996b, sec. 3.2.2, p. 48, point b]). Re-phrasing this in statistical speak, if $Pr(D|H_1) = Pr(D|H_2)$ and $Pr(H_1) > Pr(H_2)$, then Ockham's razor advocates that we prefer H_1 over H_2 — as would also MML, since $Pr(H_1)Pr(D|H_1) > Pr(H_2)Pr(D|H_2)$. It is not clear what — if anything — Ockham's razor says in the case that $Pr(H_1) > Pr(H_2)$ but $Pr(D|H_1) < Pr(D|H_2)$, although MML remains applicable in this case by comparing $Pr(H_1) \times Pr(D|H_1)$ with $Pr(H_2) \times Pr(D|H_2)$. In this sense, I would at least contend that MML can be thought of as a generalisation of Ockham's razor — for MML tells us which inference to prefer regardless of the relationships of $Pr(H_1)$ with $Pr(H_2)$ and $Pr(D|H_1)$ with $Pr(D|H_2)$, but it is not completely clear what Ockham's razor *per se* advocates unless we have that $Pr(D|H_1) = Pr(D|H_2)$.

Our earlier arguments (e.g., from sec. 1) tell us why the MML theory (or hypothesis) can be thought of as the most probable hypothesis. Informal arguments of Chris Wallace's from [Dowe, 2008a, footnote 182] (in response to questions [Dowe and Hajek, 1997, sec. 5.1; 1998, sec. 5]) suggest that, if $Pr(D|H_1) = Pr(D|H_2)$ and $Pr(H_1) > Pr(H_2)$, then we expect H_1 to be a better predictor than H_2 . But, in addition, there is also an alternative, more general and somewhat informal argument for, in general, preferring the predictive power of one hypothesis over another if the former hypothesis leads to a shorter two-part message length. This argument is simply that the theory of shorter two-part message length contributes more greatly (i.e., has a greater Bayesian weighting) in the optimal Bayesian predictor. In particular, the MML model will essentially have the largest weight in the predictor. In those cases where the optimal Bayesian predictor is statistically consistent (i.e., converges to any underlying data when given sufficient data), the optimal Bayesian predictor and the MML hypothesis appear always to converge.

Several papers have been written with dubious claims about the supposed ineffectiveness of MML and/or of Ockham's razor. Papers using inefficient (Minimum Description Length [MDL] or MML) coding schemes lead quite understandably to sub-optimal results — but a crucial point about minimum *message length* is to make sure that one has a reliable message length (coding scheme) before one

sets about seeking the minimum of this “message length”. For corrections to such dubious claims in such papers by using better coding schemes to give better results (and sometimes vastly better coding schemes to get vastly better results), see (e.g.) examples [Wallace and Dowe, 1999a, secs. 5.1 and 7; 1999c, sec. 2; Wallace, 2005, sec. 7.3; Viswanathan *et al.*, 1999; Wallace and Patrick, 1993; Comley and Dowe, 2005, secs. 11.3 and 11.4.3; Needham and Dowe, 2001; Wallace, 2005, sec. 5.1.2; Grünwald, 2007, sec. 17.4, An Apologetic Remark; Dowe, 2008a, p. 536] such as those in [Dowe, 2008a, footnote 18]. Not unrelatedly, a variety of misconceptions have led a variety of authors to make ill-founded criticisms of Ockham’s razor. One (such) interpretation (I think I should say, *misinterpretation*) of Ockham’s razor seems to go along the lines that Ockham’s razor supposedly advocates the simplest hypothesis, regardless of any data — and so (e.g.) DNA should supposedly be shaped in a single-helix rather than a double-helix. [And it seems a pity to misinterpret Ockham’s razor so — especially in a biological framework — because interpreting Ockham’s razor more properly using MML enables us to make a strong case that proteins fold with the Helices (and Extendeds) forming first and then the “Other” turn classes forming subsequently to accommodate these structures [Edgoose *et al.*, 1998, sec. 6; Dowe *et al.*, 1996, sec. 5, p. 253] (see also [Dowe *et al.*, 1995]) [Wallace, 1998a, sec. 4.2; Dowe, 2008a, footnote 85; 2008b, p. 454].

What seems like a variation of this misconception is an argument in one paper that if we fit data from within some model family (such as fitting the data with a decision tree) and then subsequently find that a more complicated model predicts better, then this is somehow supposedly empirical evidence against Ockham’s razor. (See also a comment here from [Jorgensen and Gentleman, 1998, Some Criticisms].)

Using MML as our (more general) form of Ockham’s razor, these supposed criticisms based on using overly simple models and paying insufficient attention to the data seem somewhat silly. For a discussion of the adverse consequences of not giving equals weights to the lengths of the two parts of an MML message, see, e.g., [Dowe, 2008a, footnote 130].

For those who would like every function — both the simple functions and the more complicated functions — to have the same prior probability, not only does this seem counter-intuitive, but — furthermore — it is not possible when there are infinitely many theories. When there are infinitely many theories, it necessarily follows that, as we look at progressively more and more complicated theories, it must necessarily follow that the prior probability must tend asymptotically to zero so that the countable sum over the prior probabilities of the theories can equal 1 (or unity).

Another criticism of the Bayesian approach — and therefore of our Bayesian MML interpretation of Ockham’s razor — is that this approach can be undone by a pathological (sabotaging) form of prior. If we look at Bayes’s theorem (from sec. 1) and its consequences in terms of MML, we see what our intuition tells us — that we get a reasonable posterior distribution over the hypotheses if we start off with a

reasonable prior distribution. Our Bayesian priors (as we use them in problems of inference) should be somewhere between what we genuinely suspect *a priori* and (partly politically, so that we are less open to being accused of fudging our results, and perhaps partly to protect ourselves from ourselves) something innocuous (and seemingly “objective”).

For problems where the number of parameters is bounded above — or grows sufficiently slowly when the amount of data increases — Bayesian inference will converge to the underlying model given sufficient data. To criticise Bayesian MML and/or Ockham’s razor after being sabotaged by a counter-intuitive and misrepresentative pathological prior is somewhat akin to criticising any inference method when the bulk of the relevant explanatory variables are not made available (or at least not made available until after much data has been seen) but in their stead is a plethora of essentially irrelevant variables.

4.1 *Inference (or explanation) and prediction*

Inference — also variously known as *explanation* [Wallace, 2005, sec. 1.1, first sentence and sec. 1.5], *induction* and/or *inductive inference* — pertains to finding the single best explanation for a body of data. *Prediction* pertains to the activity of anticipating the future, whether this is done using a single inference or a combination of more than one inference. To give an example, someone doing inference would be interested in a model of stock market prices which gives a theory of how the stock market works. An investor would certainly find that useful, but an investor would perhaps be more interested in whether prices are expected to be going up or down (and a probability distribution over these events and the magnitude of movement). To give a second example [Dowe *et al.*, 2007, sec. 6.1.4], when two models of slightly different (Bayesian posterior) probabilities give substantially different answers, inference would advocate going with the more probable theory where prediction would advocate doing some sort of averaging of the theories.

In the classical (non-Bayesian) approach, inference and prediction are perhaps the same thing. Certainly, an inference can be used to predict — and, to the classically (non-Bayesian) minded, prediction seems to be done by applying the single best inference. But, to the Bayesian, the best predictor will often result from combining more than one theory [Wallace, 1996b, sec. 3.6, p. 55; Oliver and Hand, 1996; Wallace and Dowe, 1999a, sec. 8; Tan and Dowe, 2006].

Herein lies a difference between the predictive approach of Solomonoff and the MML inductive inference approach of Wallace from sec. 2.4. By taking the single best theory, MML is doing induction. Despite the potentially confusing use of the term “Solomonoff induction” by some others, Solomonoff (is not doing induction [and not really inductive inference *per se*, either] but rather) is doing prediction [Solomonoff, 1996; Wallace, 1996b, sec. 3.6, p. 55; 2005, sec 10.1].

On the relative merits of induction (or [inductive] inference) vs prediction, there can be no doubting that humans acknowledge and reward the intelligence behind

inductive inferences. When we ask for a list of great human intellects, whoever else is on the list, there will be people who have made prominent inductive inferences. Examples of such people and theories include Isaac Newton for the theory of gravity, Charles Darwin and Alfred Russel Wallace for the theory of natural selection and evolution, Albert Einstein for the theories of special and general relativity, Alfred Wegener for the theory of “continental drift”, and countless Nobel laureates and/or others in a variety of areas for their theories [Sanghi and Dowe, 2003, sec. 5.2]. (And when a human is paid the compliment of being called “perceptive”, my understanding of this term is that one thing that is being asserted is that this “perceptive” person is good at making inductive inferences about human behaviour.) Of course, those such theories as are accepted and whose developers are rewarded usually are not just good pieces of induction but typically also lead to good predictions. And whether or not predictions are done with the single best available theory or with a combination of theories, people are certainly interested in having good predictors.

In trying to re-construct or restore a damaged image, the argument in support of inference is that we clearly want the single best inference rather than a probability distribution over all possible re-constructions. On the other hand, if there are a few inferences almost as good as the best (MML) inference, we would also like to see these alternative models [Dowe, 2008a, sec. 0.3.1]. Let me now essentially repeat this example but modify its context. If you think that the sentence you are currently reading is written in grammatically correct unambiguous English (and that you correctly understand the author’s intended meaning), then you are using several little innocuous inferences — such as (e.g.) the author and you (the reader) have at least sufficiently similar notions of English-language word meaning, English-language grammar and English-language spelling. However, if the writing were smudged, the spelling was questionable and the grammar and punctuation were poor (several of which can happen with the abbreviated form of a telegram, some e-mails or a mobile text message), inference would advocate going with the single best interpretation. A related case in point is automatic (machine) translation. Whether for the smudged poorly written (ambiguous) sentence or for automatic translation, prediction (in its pure platonic form) would advocate having a probability distribution over all interpretations. In reality, if there is one outstandingly clear interpretation to the sentence, then someone doing prediction would most probably be satisfied with this interpretation, (as it were) “beyond a reasonable doubt”. But, as with the damaged image, if there are a few inferences almost as good as the best (MML) inference, we would again also like to see these alternative models.

The distinction between explanation (or inference, or inductive inference, or induction) and prediction is something which at least some other authors are aware of [Wallace and Dowe, 1999a, sec. 8; Wallace 2005, sec. 10.1.2; Shmueli and Koppius, 2007, Dowe *et al.*, 2007, secs. 6.1.4, 6.3 and 7.2; Dowe, 2008b, pp. 439–440], and we believe that both have their place [Dowe *et al.*, 2007, sec. 6.3]. Whether or not because of our newly discussed uniqueness in invariance properties of Kullback-

Leibler distance (from [Dowe, 2008a, p. 438] and sec. 3.6), some authors regard prediction as being about minimising the expected log-likelihood error — or equivalently, minimising the expected Kullback-Leibler distance between the true model (if there is one) and the inferred model. While the reasons (of many) for doing this might be (more) about minimising the expected log-likelihood error, the uniqueness in invariance properties of Kullback-Leibler distance suggest it is certainly a worthy interpretation of the term “*prediction*” and that doing prediction this way is worthy of further investigation.

Recalling the invariance of the Kullback-Leibler distance (from, e.g., sec. 3.6), taking the Bayesian approach to minimising the expected Kullback-Leibler distance will be invariant under re-parameterisation (e.g., from polar to Cartesian co-ordinates) [Dowe *et al.*, 1998; Wallace, 2005, secs. 4.7–4.9; Dowe *et al.*, 2007, secs. 4 and 6.1.4; Dowe, 2008a, sec. 0.2.2]. Recalling α at the very end of sec. 3.6 [from the expression $\alpha KL(f, g) + (1 - \alpha) KL(g, f) = \alpha \Delta(g||f) + (1 - \alpha) \Delta(f||g)$], the extreme of $\alpha = 1$ sees us choose a function (g) so that the *expected* coding inefficiency of using our function (g) rather than the (ideal) truth (true function, f) is minimised, weighting over our posterior distribution on f ; and the other extreme of $\alpha = 0$ sees us choose a function (g) so that (under the hypothetical assumption that the data were being sampled from distribution, g) the expected inefficiency of using a function (f) sampled from the (actual) Bayesian posterior rather than using our function (g) is minimised. Although both of these are statistically invariant, convention is that we are more interested in choosing a function of minimal expected coding inefficiency relative to the (ideal) truth (true function) — equivalently minimising the expected log-likelihood error (and hence choosing $\alpha = 1$).

As a general rule of thumb, the MML estimator lies between the Maximum Likelihood estimator (which is given to over-fitting) on the one hand and the Bayesian minEKL estimator (which is, curiously, given to under-fitting) on the other hand [Wallace, 2005, secs. 4.7–4.9]. (Wallace makes an excellent intuitive case for this in [Wallace, 2005, sec. 4.9].) Four examples of this are the multinomial distribution, the Neyman-Scott problem (see sec. 6.4) [Wallace, 2005, sec. 4.2–4.8], the “gap or no gap” (“gappy”) problem [Dowe *et al.*, 2007, sec. 6.2.4 and Appendix B] and the bus number problem [Dowe, 2008a, footnote 116; 2008b, p. 440]. We outline these below, and then mention at the end not yet totally explored possible fifth and sixth (which would probably begin from [Schmidt and Makalic, 2009b]) examples.

For the multinomial distribution, with counts $s_1, \dots, s_m, \dots, s_M$ in classes $1, \dots, m, \dots, M$ respectively and $S = s_1 + \dots + s_m + \dots + s_M$, Maximum Likelihood gives $\hat{p}_m = s_m/S$. With a uniform prior, the minEKL estimator (also known as the Laplace estimate or the posterior mean) is $(s_m + 1)/(S + M)$, whereas the Wallace-Freeman MML approximation [Wallace and Freeman, 1987; Wallace, 2005, sec. 5.4] with this same prior is $(\hat{p}_m)_{MML} = (s_m + 1/2)/(S + M/2)$.

For the particular case of the “gap or no gap” (“gappy”) problem [Dowe *et al.*, 2007, sec. 6.2.4 and Appendix B], data $\{x_i : 0 \leq x_i \leq 1, i = 1, \dots, N\}$ for

increasing N) are being generated uniformly either from the closed interval $[0, 1]$ or from a sub-region $[0, a] \cup [b, 1]$ for some a and b such that $a < b$. We see Maximum Likelihood and Akaike’s Information Criterion (AIC) over-fitting here, surmising a gap even whether there isn’t one. At the other extreme, we see the minEKL estimator *under-fitting*, stating no gap even in extreme cases such as (e.g.) $[0, 0.001] \cup [0.999, 1]$ with $a = 0.001$ and $b = 0.999$. The curious behaviour of minEKL is due to the fact that the posterior probability that the region is $[0, 1]$ will get arbitrarily small for large N but never down to 0, and there is an infinite penalty in Kullback-Leibler distance for ascribing a probability of 0 to something which can actually happen. Unlike the over-fitting Maximum Likelihood and AIC, and unlike the under-fitting minEKL, MML behaves fine in both cases [Dowe *et al.*, 2007].

For the Neyman-Scott problem (of sec. 6.4), see [Wallace, 2005, sec. 4.2–4.8]. For the bus number problem [Dowe, 2008a, footnote 116; 2008b, p. 440] (where we arrive in a new town with θ buses numbered consecutively from 1 to θ , and we see only one bus and observe its number, x_{obs} , and are then asked to estimate the number of buses in the town), the Maximum Likelihood estimate is the number of the observed bus, x_{obs} , which is an absolute lower bound and seems like a silly under-estimate. At the other extreme, minEKL will behave in similar manner to how it did with the abovementioned “gappy” problem. It will choose the largest positive integer (no less than x_{obs}) for which the prior (and, in turn, the posterior) is non-zero. In the event that the prior never goes to 0, it will return infinity. It seems fairly trivial that the MML estimate must fall between the Maximum Likelihood estimate (the lowest possible value) and the minEKL estimate (from a Bayesian perspective, the highest possible estimate). For further discussion of the behaviour of MML here, see [Dowe, 2008a, footnote 116]. In addition to the four examples we have just given of the MML estimate lying between (over-fitting) Maximum Likelihood and (under-fitting) minEKL, of possible interest along these lines as potential fifth and sixth examples worthy of further exploration, see sec. 6.5 on panel data (as a probable fifth example) and the treatment of MML shrinkage estimation in [Schmidt and Makalic, 2009b].

5 DESIDERATA: STATISTICAL INVARIANCE, STATISTICAL CONSISTENCY, EFFICIENCY, SMALL-SAMPLE PERFORMANCE, ETC.

In this section, we look at several *desiderata* — or properties that we might desire — from statistical estimators.

5.1 Statistical invariance

Statistical invariance [Wallace, 2005, sec. 5.2; Dowe *et al.*, 2007, sec. 5.3.2; Dowe, 2008b, p. 435] says, informally, that we get the same answer no matter how we phrase the problem.

So, if we know that the relationship between the area A of a circle and its radius r is given by $A = \pi r^2$ (and, equivalently, $r = \sqrt{A/\pi}$), then statistical invariance requires that our estimate of the area is π times our the square of our estimate of the radius. The estimator function is often denoted by a hat (or circumflex), $\hat{\cdot}$, above. So, for a circle, statistical invariance in the estimator would require that $\hat{A} = \pi \hat{r}^2$.

If we replace r by κ in the Cartesian and polar co-ordinates example from sec. 3.6, ($\kappa = \text{sign}(x) \cdot \sqrt{x^2 + y^2}$, $\theta = \tan^{-1}(y/x)$) and ($x = \kappa \cos \theta$, $y = \kappa \sin \theta$). If we are estimating the strength and direction (κ , θ) of a magnetic field or equivalently the x and y co-ordinates (x , y) [Wallace and Dowe, 1993], then statistical invariance requires that $\hat{x} = \hat{\kappa} \cos \theta$, $\hat{y} = \hat{\kappa} \sin \theta$.

Statistical invariance is surely an aesthetic property of an estimate. In many problems, we are not committed to only one parameterisation - and, in those cases, statistical invariance is more useful than a simple aesthetic nicety.

Maximum Likelihood, Akaike's Information Criterion, Strict Minimum Message Length (SMML) [Wallace and Boulton, 1975; Wallace, 2005, chap. 3] and many MML approximations [Wallace and Freeman, 1987; Wallace, 2005, chaps. 4–5; Dowe 2008a, sec. 0.2.2 and footnote 159; Schmidt, 2008; Dowe, 2008b, p. 438 and p. 451] are statistically invariant, but there do exist approaches — such as the Bayesian Maximum A Posteriori (MAP) approach [Oliver and Baxter, 1994; Dowe *et al.*, 1996e; Wallace and Dowe, 1999b, secs. 1.2–1.3; 1999c, sec. 2, col. 1; 2000, secs. 2 and 6.1; Comley and Dowe, 2005, sec. 11.3.1; Dowe *et al.*, 2007, sec. 5.1, coding prior; Dowe 2008a, footnote 158; 2008b, p. 443 and pp. 448–449] — which are not statistically invariant.

5.2 Statistical consistency

For those who like collecting larger and larger data sets in the hope and belief that this will bring us closer and closer to whatever model or process underlies the data, statistical consistency — the notion of getting arbitrary close to any true underlying model given sufficient data [Dowe *et al.*, 2007, secs. 5.3.4, 6.1.3 and later; Dowe 2008b, pp. 436–437] — is of paramount importance.

More formally, we might write it as [Dowe, 2008b, p. 436]

$$\forall \theta \quad \forall \epsilon > 0 \quad \exists N_0 \quad \forall N \geq N_0 \quad \Pr(|\theta - \hat{\theta}| < \epsilon) > 1 - \epsilon$$

and we could even venture to write it (in a parameterisation-invariant way) as (e.g.)

$$\forall \theta \quad \forall \epsilon > 0 \quad \exists N_0 \quad \forall N \geq N_0 \quad \Pr(\Delta(\hat{\theta}||\theta) = KL(\theta, \hat{\theta}) < \epsilon) > 1 - \epsilon.$$

Of course, as highlighted by Grünwald and Langford [2004; 2007], cases of model misspecification do occur. In other words, it might be that the true model θ (if there is one) is not contained in the family (or class) of models over which we conduct our search for $\hat{\theta}$. In such cases, we can modify (or generalise) the notion

of statistical consistency to be that (as implicitly described in [Dowe, 2008a, sec. 0.2.5, p. 540, col. 1]), as we get more and more data, we get the Kullback-Leibler distance arbitrarily close to that of the closest available member in our model space. Or, more formally,

$$\forall \theta \quad \forall \epsilon > 0 \quad \exists N_0 \quad \forall N \geq N_0 \quad \Pr(\Delta(\hat{\theta}||\theta) = KL(\theta, \hat{\theta}) < KL(\theta, \hat{\theta}_{\text{best}}) + \epsilon) > 1 - \epsilon,$$

where $\hat{\theta}_{\text{best}}$ is as close as one can get in Kullback-Leibler distance to θ from within the space of models being considered.

I should (and do now) qualify this slightly. It is possible that $\hat{\theta}_{\text{best}}$ might not exist in the same sense that there is no number in the list $1, 1/2, 1/3, 1/4, \dots$ which is the list's smallest element. One of a few ways of dealing with this is simply to replace the start of the above with “ $\forall \theta'$ in our model space” and to replace the finish of the above with “ $\Pr(KL(\theta, \hat{\theta}) < KL(\theta, \theta') + \epsilon) > 1 - \epsilon$ ”, thus now making it

$\forall \theta'$ in our model space

$$\forall \epsilon > 0 \quad \exists N_0 \quad \forall N \geq N_0 \quad \Pr(KL(\theta, \hat{\theta}) < KL(\theta, \theta') + \epsilon) > 1 - \epsilon.$$

(As a second point, replacing the first ϵ by $\epsilon/2$ does not change the semantics of the definition. Similarly, replacing one of the ϵ terms by a δ and adding a quantifier $\forall \delta > 0$ out front also does not change the semantics, as we can (e.g.) re-set $\epsilon' = \min\{\delta, \epsilon\}$.)

With an eye to secs. 6.4 and 6.5 and this issue of statistical consistency under misspecification, it is worth bearing in mind that — even though there is misspecification — the class (or family) of models over which we conduct our search might be dense in the space of possible models. In other words, if you have a non-negative valued function (or probability distribution) on the real line which integrates to 1 (and can't be written as a finite mixture model), it can still be possible to find a sequence of finite (Gaussian) mixture models which fit it arbitrarily closely. (For reading on mixture models, see, e.g., [Jorgensen and McLachlan, 2008].)

5.3 Efficiency, small-sample performance, other considerations, etc.

The notion of *efficiency* is perhaps ambiguous in that it has been used in the literature with at least two different meanings. On the one hand, *efficiency* has been taken to mean that the message length calculations and approximations are both optimal or near-optimal (with $l_i \approx -\log p_i$) [Wallace, 2005, sec. 5.2.4]. On the other hand, *efficiency* of an estimator has been taken to mean the speed with which that estimator converges to the true model generating the data as the amount of data increases [Dowe *et al.*, 2007, secs. 5.3.4 and 8].

While these two notions are different, it should be pointed out that, insofar as reliable MML coding schemes lead to good inferences and less reliable coding schemes lead to less reliable inferences [Quinlan and Rivest, 1989; Wallace and

Patrick, 1993, Kearns *et al.*, 1997; Viswanathan *et al.*, 1999; Wallace, 2005, sec. 7.3; Murphy and Pazzani, 1994; Needham and Dowe, 2001; Wallace and Dowe, 1999a, secs. 5.1 and 7; 1999c, sec. 2; Comley and Dowe, 2005, secs. 11.3 and 11.4.3; Wallace, 2005, sec. 5.1.2; Dowe, 2008a, footnote 18], the two notions are very related.

As well as the notions of statistical invariance (from sec. 5.1), statistical consistency (from sec. 5.2) and efficiency, there are also issues of performing well on small-sample sizes [Dowe, 2008b, p. 436 and p. 456]. The issue of which likelihood function(s), sample size(s), parameterisation(s), Bayesian prior(s) and protocol(s) (or which parts of LNPPP-space) are important when comparing the efficacy of two estimators is discussed in [Dowe, 2008a, sec. 0.2.7, pp. 543–544].

6 MINIMUM MESSAGE LENGTH (MML) AND STRICT MML

As in sec. 1, historically, the seminal Wallace and Boulton paper [1968] came into being from Wallace's and Boulton's finding that the Bayesian position that Wallace advocated and the information-theoretic (conciseness) position that Boulton advocated turned out to be equivalent [Wallace, 2005, preface, p. v; Dowe, 2008a, sec. 0.3, p. 546 and footnote 213]. After several more MML writings [Boulton and Wallace, 1969; 1970, p. 64, col. 1; Boulton, 1970; Boulton and Wallace, 1973b, sec. 1, col. 1; 1973c; 1975, sec. 1, col. 1] (and an application paper [Pilowsky *et al.*, 1969], and at about the same time as David Boulton's PhD thesis [Boulton, 1975]), their paper [Wallace and Boulton, 1975, sec. 3] again emphasises the equivalence of the probabilistic and information-theoretic approaches. (Different but not unrelated histories are given by Solomonoff [1997a] and a review of much later work by Kontoyiannis [2008]. For those interested in the formative thinking of Wallace (and Boulton) leading up to the seminal Wallace and Boulton MML paper [1968], see evidence of the young Bayesian (but pre-MML) Wallace in his mid-20s in the 1950s [Brennan, 2008, sec. 4; Brennan *et al.*, 1958, Appendix] and see Wallace's accounts of his early discussions with David Boulton [Wallace, 2005, preface, p. v; Dowe, 2008a, sec. 0.3, p. 546, col. 2 and footnote 213 (and sec. 1)] which resulted in [Wallace and Boulton, 1968]. If you can obtain it, then I also commend [Wallace, 1992] for background.)

As in sec. 1 and following the principles of information theory from sec. 2.1, given data D , we wish to choose a hypothesis H so as to minimise the length of a two-part message conveying H (in part 1) followed (in part 2) by D given H . The length of this message is

$$-\log Pr(H) - \log Pr(D|H).$$

A one-part form of the message was examined in [Boulton and Wallace, 1969], but various pieces of theory and practice (e.g., [Barron and Cover, 1991]) point to the merits of the two-part form of the message.

We now point to the Strict Minimum Message Length (SMML) formulation from Wallace and Boulton [1975] in sec. 6.1, and then go on to talk about some

“MML” approximations to SMML, some conjectures about the possible uniqueness of Strict MML in being both statistically invariant and statistically consistent for certain classes of problems, and some applications of MML to a variety of problems in inference and other areas in science and philosophy.

6.1 Strict MML (SMML)

The Strict Minimum Message Length (SMML) formulation from Wallace and Boulton [Wallace and Boulton, 1975; Wallace and Freeman, 1987; Wallace, 1996c; Dowe *et al.*, 1998; Wallace and Dowe, 1999a; 1999b; Farr and Wallace, 2002; Fitzgibbon *et al.*, 2002b, Fitzgibbon, 2004; Agusta, 2005; Wallace, 2005, chap. 3; Comley and Dowe, 2005, sec. 11.2; Dowe *et al.*, 2007; Dowe, 2008a, footnotes 12, 153, 158 and 196, and sec. 0.2.2] shows how to generate a code-book whose expected two-part message length is minimised, but this turns out to be computationally intractable except in the simplest of cases — such as the binomial distribution [Farr and Wallace, 2002; Wallace, 2005, chap. 3].

Of historical interest is the fact [Dowe, 2008a, sec. 0.1, p. 524, col. 1] that, even though MML had been in print many times over since 1968 [Wallace and Boulton, 1968, p. 185, sec. 2; Boulton and Wallace, 1969; 1970, p. 64, col. 1; Boulton, 1970; Boulton and Wallace, 1973b, sec. 1, col. 1; 1973c; 1975, sec. 1, col. 1; Boulton, 1975], referees delayed the publication of Strict MML until Wallace and Boulton [1975].

Strict MML (SMML) partitions in *data*-space and optimises a formula of the form

$$\left(-\sum_j (q_j \log q_j)\right) + \left(-\sum_j \sum_{i \in c_j} \left(q_j \cdot \frac{r(x_i)}{q_j} \cdot \log f(x_i|\theta_j)\right)\right) \quad (18)$$

Note here first that i indexes over the data. This set must be countable, as all recorded measurements are truncated and recorded to some finite accuracy. (See [Dowe, 2008a, footnote 63] for a discussion of consequences of attempting to side-step such an insistence.) This point established, we now assign data to groups indexed by j . The number of groups will certainly be countable (and to date I am not aware of any cases where there are infinitely many groups). Letting $h(\cdot)$ be the prior and $f(\cdot|\cdot)$ denote the statistical likelihood, $r(x_i) = \int h(\theta) f(x_i|\theta) d\theta$ is the marginal probability of datum x_i . Note that $r(x_i)$ is a probability and *not* a density, and also that $\sum_i r(x_i) = 1$. The term $q_j = \sum_{i \in c_j} r(x_i)$ is the amount of prior probability associated with the data group c_j .

The groups c_j form a partition of the data, with each datum being assigned to exactly one group — from which it follows that $\sum_j q_j = 1$.

For each data group c_j , we choose the estimate θ_j which maximises the weighted log-likelihood $\sum_{i \in c_j} r(x_i) \log f(x_i|\theta_j)$.

As we have written equation (18), the first term is the expected length of encoding the hypothesis (see, e.g., sec. 2.3) and the second term is the expected length

of encoding the data given this hypothesis — namely the hypothesis that datum x_i lies in group c_j with (prior) probability q_j and estimate θ_j .

The computational intractability of Strict MML (except in the simplest of cases — such as the binomial distribution [Farr and Wallace, 2002; Wallace, 2005, chap. 3]) is largely due to its discrete nature — or its being “gritty” (as Chris Wallace once put it) — requiring shuffling of data between data groups, then re-estimating the q_j and θ_j for each data group c_j , and then re-calculating the message length. The code-book with the shortest expected message length as per equation (18) is the SMML code-book, and the SMML estimator for each datum x_i is the θ_j corresponding to the group c_j to which x_i is assigned.

6.2 Strict Strict MML (SSMML)

Recall the notion of (algorithmic information theory or) Kolmogorov complexity from sec. 2.4 and sec. 1. It could be said that the relationship between Strict MML and Kolmogorov complexity [Wallace and Dowe, 1999a; Wallace, 2005, chaps. 2–3] might be slightly enhanced if we turn the negative logarithms of the probabilities from equation (18) into integer code lengths — such as would seem to be required for constructing a Huffman code from sec. 2.1 (or other fully-fledged kosher code). From [Wallace, 2005, sec. 3.4, p. 191] and earlier writings (e.g., [Wallace and Freeman, 1987]), it is clear that Wallace was aware of this issue but chose to neglect and not be distracted by it.

Although it is hard to imagine it having anything other than the most minor effect on results, we take the liberty here of introducing here what I shall call Strict Strict MML (SSMML), where the constituent parts of both the first part (currently, for each j , of length $-\log q_j$) and the second part (currently, for each j , for each i , of length $-\log f(x_i|\theta_j)$) of the message have non-negative integer lengths.

One reason for preferring Strict MML to Strict Strict MML is that, as can be seen from inspecting equation (18), the Strict MML data groups, estimates and code-book will all be independent of the base of logarithm — be it 2, 10, e or whatever — and the (expected) message length will transform in the obvious invariant way with a change of base of logarithms. However, Strict Strict MML will require an integer greater than or equal to 2 to be the base of logarithms, and will not be independent of this choice of base. The simplest response to this objection is to insist that the base of logarithms is always 2 for Strict Strict MML.

My guess is that Strict Strict MML (with base of logarithms set to 2, although any larger positive integer base should work both fine and similarly as the amount of data increases) will typically be very similar to Strict MML. By construction, Strict Strict MML (with fixed base of logarithms, such as 2) will necessarily be statistically invariant, and Strict Strict MML will presumably share statistical consistency and other desirable properties of Strict MML.

There is another issue which arises when relating Strict MML (from sec. 6.1) and Strict Strict MML to Kolmogorov complexity (or algorithmic information

theory). As per equation (18) and associated discussion(s), both Strict MML and Strict Strict MML require the calculation of the marginal probability, $r(x_i) = \int h(\theta)f(x_i|\theta) d\theta$, of each datum, x_i . As in sec. 6.1, these marginal probabilities are then used to calculate what we call the “coding prior” [Dowe *et al.*, 2007, sec. 5.1], namely the discrete set of possible estimates $\{\theta_j\}$ and their associated prior probabilities, $\{q_j\}$, with $\sum_j q_j = 1$. (Strict MML is then equivalent to using the coding prior in combination with the given statistical likelihood function and doing conventional Bayesian Maximum A Posteriori (MAP).) As per [Wallace and Dowe, 1999a] and [Wallace, 2005, chaps. 2–3], the input to the Turing machine will be of a two-part form such that the first part of this input message (which conveys the hypothesis, theory or model) programs the Turing machine (without any output being written). The second part is then input to the program resulting from the first part of the message, and this input causes the desired Data to be output. (In the example with noise in sec. 7.2, the first part of the message would encode the program together with an estimate of the noise, and the second part would encode the data with code lengths depending upon the probabilities as per sec. 2.1.)

The particular additional issue which arises when relating Strict MML and Strict Strict MML to Kolmogorov complexity (or algorithmic information theory) occurs when dealing with *universal* Turing machines (UTMs) and the Halting problem (Entscheidungsproblem) — namely, we can get lower bounds on the marginal probability ($r(x_i)$) of the various data (x_i) but, due to the Halting problem, typically for at least many values of x_i we will not be able to calculate $r(x_i)$ exactly but rather only give a lower bound. If the Turing machine (TM) representing our prior is not universal (e.g., if we restrict ourselves to the family of multivariate polynomials with one of the Bayesian priors typically used in such a case), then we can calculate $r(x_i)$ to arbitrary precision for each x_i . But if the TM representing our prior is a UTM, then we might have to live with only having ever-improving lower bounds on each of the $r(x_i)$. If we stop this process after some finite amount of time, then we should note that the coding prior corresponding to the grouping arising from Strict MML (and ditto from Strict Strict MML) would appear to have the potential to be different from the prior emanating from our original UTM. That said, if we don’t go to the trouble of summing different terms contributed from different programs in the calculation of $r(x_i)$ but rather simply take the largest available such term, then we quite possibly get something very similar or identical to our intuitive notion of a two-part Kolmogorov complexity.

Finally, it is worth changing tack slightly and adding here that Strict MML is a function of the sufficient statistics in the data [Wallace, 2005, sec. 3.2.6], as also should be Strict Strict MML. When some authors talk of the Kolmogorov sufficient statistics, it is as though they sometimes forget or are unaware that sometimes — such as for the Student t distribution or the restricted cut-point segmentation problem from [Fitzgibbon *et al.*, 2002b] — the minimal sufficient statistic can be the entire data set [Comley and Dowe, 2005, sec. 11.3.3, p. 270].

6.3 Some MML approximations and some properties

Given the typical computational intractability of Strict MML from sec. 6.1 (which would only be worse for Strict Strict MML from sec. 6.2), it is customary to use approximations.

Given data, D , the MMLD (or I_{1D}) approximation [Dowe, 2008a, sec. 0.2.2; Fitzgibbon *et al.*, 2002a; Wallace, 2005, secs. 4.10 and 4.12.2 and chap. 8, p. 360; Dowe, 2008b, p. 451, eqn (4)] seeks a region R which minimises

$$-\log\left(\int_R h(\vec{\theta}) d\theta\right) - \frac{\int_R h(\vec{\theta}) \cdot \log f(\vec{D}|\vec{\theta}) d\theta}{\int_R h(\vec{\theta}) d\theta} \quad (19)$$

The length of the first part is the negative log of the probability mass inside the region, R . The length of the second part is the (prior-weighted) average over the region R of the log-likelihood of the data, D .

An earlier approximation similar in motivation which actually inspired Dowe's MMLD approximation from eqn (19) above is the Wallace-Freeman approximation [Wallace and Dowe, 1999a, sec. 6.1.2; Wallace, 2005, chap. 5; Dowe, 2008b, p. 451, eqn (5)]

$$\begin{aligned} & -\log(h(\vec{\theta})) \cdot \frac{1}{\sqrt{\kappa_d^d \text{Fisher}(\vec{\theta})}} - \log f(\vec{x}|\vec{\theta}) + \frac{d}{2} \\ = & -\log(h(\vec{\theta})) + L + (1/2)\log(\text{Fisher}(\vec{\theta})) + (d/2)(1 + \log(\kappa_d)) \end{aligned} \quad (20)$$

which was first published in the statistics literature [Wallace and Freeman, 1987].

(Digressing, note that if one approximates $-\log(1/\text{Fisher}(\vec{\theta}))$ in equation (20) very crudely as $k \log N$, then equation (20) reduces to something essentially equivalent to Schwarz's Bayesian Information Criterion (BIC) [Schwarz, 1978] and Rissanen's original 1978 version of Minimum Description Length (MDL) [Rissanen, 1978], although it can be strongly argued [Wallace and Dowe, 1999a, sec. 7, p. 280, col. 2] that the $-\log(1/\text{Fisher}(\vec{\theta}))$ term from equation (20) is best not idly approximated away.)

A very recent approximation certainly showing promise is due to Schmidt [2008] and in [Dowe, 2008a, footnotes 64–65]. This MMLFS estimator [Schmidt, 2008], upon close examination, would appear to be based on an idea in [Fitzgibbon *et al.*, 2002a, sec. 2, especially equation (7)] (which in turn uses Wallace's FS-MML Boundary rule as from [Wallace, 1998e] and [Fitzgibbon *et al.*, 2002b, sec. 4] (again using Wallace's FSMML Boundary rule as from [Wallace, 1998e] and [Fitzgibbon *et al.*, 2002b, sec. 3.2], but see also [Wallace, 2005, sec. 4.11]).

My MMLD estimator from equation (19) gives the message length (MsgLen) for a region. The MMLFS estimator just mentioned gives MsgLen for a point (as does the Wallace-Freeman [1987] estimator from equation (20)). Both MMLD and MMLFS are calculated using Markov Chain Monte Carlo (MCMC) methods.

These approximations above, together with the Dowe-Wallace Ideal Group (or IG) estimator [Wallace, 2005, secs. 4.1, 4.3 and 4.9; Agusta, 2005, sec. 3.3.3,

pp. 60–62; Fitzgibbon, 2004, sec. 5.2, p. 70, footnote 1; Dowe, 2008a, p. 529, col. 1 and footnote 62] and other estimators (e.g., TAIG) discussed in [Dowe, 2008a, footnotes 62–65] are all statistically invariant. Recalling the notion of statistical consistency from sec. 5.2, we now show (in secs. 6.4 and 6.5) that MML is statistically consistent where a variety of other estimators fail — either under-estimating (which is typical of most alternatives to MML) or over-estimating (which is typical of the Bayesian minEKL estimator) the degree of noise.

6.4 *Neyman-Scott problem and statistical consistency*

In the Neyman-Scott problem [Neyman and Scott, 1948; Dowe, 2008b, p. 453], we measure N people’s heights J times each (say $J = 2$) and then infer

1. the heights $\mu_1, \dots, \mu_N$ of each of the N people,
2. the accuracy (σ) of the measuring instrument.

We have JN measurements from which we need to estimate $N + 1$ parameters. $JN/(N + 1) \leq J$, so the amount of data per parameter is bounded above (by J).

It turns out that $\hat{\sigma}_{\text{MaximumLikelihood}}^2 \rightarrow \frac{J-1}{J}\sigma^2$, and so for fixed J as $N \rightarrow \infty$ we have that Maximum Likelihood is statistically inconsistent — under-estimating σ [Neyman and Scott, 1948] and “finding” patterns that aren’t there. The Bayesian Maximum A Posteriori (MAP) approach (from sec. 5.1) is likewise not statistically consistent here [Dowe, 2008a, footnote 158].

Curiously, the Bayesian minimum expected Kullback-Leibler distance (minEKL) estimator [Dowe *et al.*, 1998; Wallace, 2005, secs. 4.7–4.9; Dowe *et al.*, 2007, secs. 4 and 6.1.4; Dowe, 2008a, sec. 0.2.2; 2008b, p. 444] from sec. 4.1 is also statistically inconsistent for the Neyman-Scott problem, conservatively over-estimating σ [Wallace, 2005, sec. 4.8]. Recall a discussion of this (the Neyman-Scott problem) and of the “gappy” (“gap or no gap”) problem in sec. 4.1.

However, the Wallace-Freeman MML estimator from equation (20) and the Dowe-Wallace Ideal Group (IG) estimator have both been shown to be statistically consistent for the Neyman-Scott problem [Dowe and Wallace, 1996; 1997a; 1997b; Wallace, 2005, secs. 4.2–4.5 and 4.8; Dowe *et al.*, 2007, secs. 6.1.3–6.1.4; Dowe, 2008a, secs. 0.2.3 and 0.2.5; 2008b, p. 453]. An interesting discussion of the intuition behind these results is given in [Wallace, 2005, sec. 4.9].

We now use MML to re-visit a Neyman-Scott(-like) panel data problem from [Lancaster, 2002], as hinted at in [Dowe, 2008a, sec. 0.2.3, footnote 88].

6.5 *Neyman-Scott panel data problem (from Lancaster)*

Following the concise discussion in [Dowe, 2008a, sec. 0.2.3, footnote 88], we use MML here to re-visit the panel data problem from [Lancaster, 2002, 2.2 Example 2, pp. 651–652].

$$y_{i,t} = f_i + x_{i,t}\beta + u_{i,t} \quad (i = 1, \dots, N; t = 1, \dots, T) \quad (21)$$

where the $u_{i,t}$ are independently Normal($0, \sigma^2$) conditional on the regressor sequence, f_i , and $\theta = (\beta, \sigma^2)$.

We can write the (negative) log-likelihood as

$$L = \frac{NT}{2} \log 2\pi + \frac{NT}{2} \log \sigma^2 + \frac{1}{\sigma^2} \sum_{i=1}^N \sum_{t=1}^T (y_{i,t} - f_i - x_{i,t}\beta)^2 \quad (22)$$

Using the Wallace-Freeman approximation [Wallace and Dowe, 1999a, sec. 6.1.2; Wallace, 2005, chap. 5; Dowe, 2008b, p. 451, eqn (5)] from equation (20), we require a Bayesian prior (which does not have a great effect but which does, among other things, keep the estimator statistically invariant) and the determinant of the expected Fisher information matrix of expected second-order partial derivatives (with respect to the f_i [$i = 1, \dots, N$], β and σ^2).

Before taking the expectations, let us first take the second-order partial derivatives — starting with the diagonal terms.

$$\frac{\partial L}{\partial f_i} = -\frac{1}{\sigma^2} \sum_{t=1}^T (y_{i,t} - f_i - x_{i,t}\beta), \quad \text{and} \quad \frac{\partial^2 L}{\partial f_i^2} = T/(\sigma^2) \quad (23)$$

Not dissimilarly,

$$\frac{\partial L}{\partial \beta} = -\frac{1}{\sigma^2} \sum_{t=1}^T x_{i,t}(y_{i,t} - f_i - x_{i,t}\beta), \quad \text{and} \quad \frac{\partial^2 L}{\partial \beta^2} = \frac{1}{\sigma^2} \sum_{i=1}^N \sum_{t=1}^T x_{i,t}^2 \quad (24)$$

$$\frac{\partial L}{\partial (\sigma^2)} = \frac{NT}{2\sigma^2} - \frac{1}{2(\sigma^2)^2} \sum_{i=1}^N \sum_{t=1}^T (y_{i,t} - f_i - x_{i,t}\beta)^2, \quad \text{and} \quad (25)$$

$$\frac{\partial^2 L}{\partial (\sigma^2)^2} = -\frac{NT}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} \sum_{i=1}^N \sum_{t=1}^T (y_{i,t} - f_i - x_{i,t}\beta)^2 \quad (26)$$

Still looking at the second derivatives, let us now look at the off-diagonal terms and then return to take expectations.

$$\frac{\partial^2 L}{\partial f_i \partial f_j} = \frac{\partial^2 L}{\partial f_j \partial f_i} = \frac{\partial}{\partial f_j} \left(-\frac{1}{\sigma^2} \sum_{t=1}^T (y_{i,t} - f_i - x_{i,t}\beta) \right) = 0 \quad (\text{for } i \neq j) \quad (27)$$

$$\frac{\partial^2 L}{\partial f_i \partial \beta} = \frac{\partial^2 L}{\partial \beta \partial f_i} = \frac{1}{\sigma^2} \sum_{t=1}^T x_{i,t} \quad (28)$$

$$\frac{\partial^2 L}{\partial f_i \partial (\sigma^2)} = \frac{\partial^2 L}{\partial (\sigma^2) \partial f_i} = \frac{1}{2(\sigma^2)^2} \sum_{t=1}^T (y_{i,t} - f_i - x_{i,t}\beta) \quad (29)$$

$$\frac{\partial^2 L}{\partial \beta \partial (\sigma^2)} = \frac{\partial^2 L}{\partial (\sigma^2) \partial \beta} = \frac{1}{(\sigma^2)^2} \sum_{i=1}^N \sum_{t=1}^T x_{i,t} (y_{i,t} - f_i - x_{i,t}\beta) \quad (30)$$

Now taking expectations to get the terms contributing to the determinant of the expected Fisher information matrix, namely the expected Fisher information, let us first use equation (27) (dealing with the off-diagonal cases $i \neq j$) and equation (23) (dealing with the diagonal cases $i = j$) to give

$$E\left(\frac{\partial^2 L}{\partial f^2}\right) = \prod_{i=1}^N E\left(\frac{\partial^2 L}{\partial f_i^2}\right) = \prod_{i=1}^N \frac{\partial^2 L}{\partial f_i^2} = \prod_{i=1}^N T/(\sigma^2) = T/((\sigma^2)^N) \quad (31)$$

Re-visiting equation (29), we have that

$$E\left(\frac{\partial^2 L}{\partial f_i \partial (\sigma^2)}\right) = 0 \quad (\text{for } i = 1, \dots, N) \quad (32)$$

Equation (28) gives us a term proportional to $1/(\sigma^2)$, namely

$$E\left(\frac{\partial^2 L}{\partial f_i \partial \beta}\right) = \frac{1}{\sigma^2} \sum_{t=1}^T E(x_{i,t}), \quad (33)$$

and equation (26) gives us

$$\begin{aligned} E\left(\frac{\partial^2 L}{\partial (\sigma^2)^2}\right) &= -\frac{NT}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} NT E((y_{i,t} - f_i - x_{i,t}\beta)^2) \\ &= -\frac{NT}{2(\sigma^2)^2} + \frac{NT\sigma^2}{(\sigma^2)^3} = \frac{NT}{2(\sigma^2)^2} \end{aligned} \quad (34)$$

From [Lancaster, 2002, p. 651, 2.2 Example 2, equation (2.8)] and our equation (30), we have

$$E\left(\frac{\partial^2 L}{\partial \beta \partial (\sigma^2)}\right) = 0 \quad (35)$$

Looking at the $(N+2) \times (N+2)$ expected Fisher information matrix, we first note that the only non-zero entry in the σ^2 column is also the only non-zero entry in σ^2 row, namely that from equation (34).

Looking at the rest of the matrix, namely the $(N+1) \times (N+1)$ sub-matrix in the top left, we see that the only non-zero off-diagonal terms are the $E(\partial L / (\partial f_i \partial \beta))$ terms from equation (33) in the row and column corresponding to β . Looking at equations (31) and (33), we see that these few off-diagonal terms from equation (33) and all the diagonal terms are of the form $\text{Const}_1 / (\sigma^2)^2$.

Combining this with equation (34), we see that the Fisher information is given by $\text{Const}_2 \times (1/(\sigma^2)^{N+1}) \times (NT)/(2(\sigma^2)^2) = \text{Const}/((\sigma^2)^{N+3})$.

Anticipating what we need for the Wallace-Freeman (1987) MML approximation in equation (20), this expression for $\text{Fisher}(\vec{\theta})$ and equation (22) then give that

$$\begin{aligned}
 & L + (\log \text{Fisher}(\vec{\theta}))/2 \\
 = & \quad \frac{NT}{2} \log 2\pi + \frac{NT}{2} \log \sigma^2 + \frac{1}{\sigma^2} \sum_{i=1}^N \sum_{t=1}^T (y_{i,t} - f_i - x_{i,t}\beta)^2 \\
 & + \left(\frac{1}{2} \log(\text{Const}) - \frac{N-3}{2} \log \sigma^2 \right) \\
 = & \quad \frac{1}{2} \log((2\pi)^{NT} \text{Const}) + \frac{(N-1)T-2}{2} \log(\sigma^2) \\
 & + \frac{1}{\sigma^2} \sum_{i=1}^N \sum_{t=1}^T (y_{i,t} - f_i - x_{i,t}\beta)^2 \tag{36}
 \end{aligned}$$

Re-capping, leaving aside the Bayesian priors (which will give statistical invariance to the MML estimator) and some constant terms, we see that the Wallace-Freeman MML approximation gives us what [Lancaster, 2002, p. 652] calls “the ‘correct’ degrees of freedom, $N(T-1)$, apart from a negligible term”.

As per [Dowe, 2008a, sec. 0.2.3, footnote 88], this can be extended to also deal with the subsequent panel data problem from [Lancaster, 2002, 2.3 Example 2, pp. 652-653].

Having made these points about statistical invariance and statistical consistency of MML, we perhaps digress slightly and note that Grünwald and Langford [2004; 2007] have shown statistical inconsistency under model misspecification for various Bayesian estimators and various forms of the Minimum Description Length (MDL) principle, but we are not aware of any current evidence for a statistical inconsistency in MML [Grünwald and Langford, 2007, sec. 7.1.5; Dowe, 2008a, sec. 0.2.5].

The above — and other — evidence and experience has led to the following conjectures. The first two conjectures deal with the case that there is a true model in the space of models being examined, and the subsequent conjectures deal with the case of model misspecification.

Conjecture 1 [Dowe *et al.*, 1998, p. 93; Edwards and Dowe, 1998, sec. 5.3; Wallace and Dowe, 1999a, p. 282; 2000, sec. 5; Comley and Dowe, 2005, sec. 11.3.1, p. 269; Dowe, 2008a, sec. 0.2.5, pp. 539–540; 2008b, p. 454]: Only MML and very closely-related Bayesian methods are in general both statistically consistent and invariant.

(This first conjecture was once the subject of a panel discussion at a statistics conference [Dowe *et al.*, 1998a].)

Conjecture 2 (*Back-up Conjecture*) [Dowe *et al.*, 2007, sec. 8; Dowe, 2008a, sec. 0.2.5; 2008b, p. 454]: If there are (hypothetically) any such non-Bayesian methods, they will be far less efficient than MML.

Re the issue of statistical consistency under model misspecification as per sec. 5.2, first, suppose that the space where we conduct our search is dense in the space from which the true model comes (e.g., suppose the true model space is that of friendly non-Gaussian t distributions and our search space is the space of finite mixtures of Gaussian distributions). Then, in this case, if our inference method is statistically consistent when the true model comes from the search space (i.e., in this example, if our inference method is statistically consistent within the space of finite Gaussian mixture models) then we would expect our inference method to still be statistically consistent for the misspecified true model from the larger class (i.e., in this example, we would expect our inference method to remain statistically consistent when the true model is a friendly non-Gaussian t distribution from [“just”] outside our search space). Paraphrasing, if it’s consistent in the search space, it can get arbitrarily close within the search space — and, if the search space is dense in the true space, then it would appear that we can get arbitrarily close to something arbitrarily close, seemingly implying statistical consistency.

Still on this issue of statistical consistency under model misspecification from sec. 5.2, we know that MML will be statistically invariant and we further conjecture [Dowe, 2008a, sec. 0.2.5, especially p. 540] that MML will still — in this more challenging setting — be statistically consistent. If there are (hypothetically) any non-Bayesian methods which are statistically consistent in this setting, then we further conjecture that they will be less efficient than MML.

Note throughout that the statistical consistency is coming from the information-theoretic properties of MML and the statistical invariance is coming from the Bayesian priors.

If there is any truth to these conjectures — and I am yet to see anything I could legitimately call a counter-example — then it would seem to suggest that inference done properly must inherently be Bayesian. I make this claim because

- recalling sec. 5.1, *statistical invariance* says that we get the same answer whether we model in (e.g.) polar or Cartesian co-ordinates, or in (e.g.) side length, face area or volume of a cube, etc.

- recalling sec. 5.2, *statistical consistency* — whether for a properly specified model or a misspecified model — merely says that collecting extra data (as people seem very inclined to do) is a worthwhile activity.

For other related arguments in support of Bayesianism and the Bayesian MML approach, recall Wallace’s notions of Bayesian bias [Wallace, 1996c, sec. 4.1] and false oracles [Wallace, 1996c, sec. 3; Dowe *et al.*, 2007], Wallace’s intuitive but nonetheless important proof that sampling from the Bayesian posterior is a false oracle [Wallace, 1996c, sec. 3.4] and Wallace’s arguments that the Strict MML estimator (which is deterministic) approximates a false oracle [Wallace, 1996c, secs. 5–7]. For different arguments in support of Bayesianism and the Bayesian MML approach for those who like the notions of Turing machines and (Kolmogorov complexity or) algorithmic information theory from sec. 2.4, recall that (as per the end of sec. 2.4) the choice of (Universal) Turing Machine in algorithmic information theory is (obviously?) also a Bayesian choice [Wallace and Dowe, 1999a, secs. 2.4 and 7; 1999c, secs. 1–2; Comley and Dowe, 2005, p. 269, sec. 11.3.2; Dowe, 2008a, footnotes 211, 225 and (start of) 133, and sec. 0.2.7, p. 546; 2008b, p. 450].

6.6 Further MML work, such as MML Support Vector Machines

The relationship between MML and Kolmogorov complexity (or algorithmic information theory) [Wallace and Dowe, 1999a; Wallace, 2005, secs. 2.2–2.3] from sec. 2.4 means that MML can be applied universally across all inference problems and even compare and contrast two models from very different families.

Let us say something about the statistical learning theory of Vapnik and Chervonenkis, VC dimension, Support Vector Machines (SVMs) and Structural Risk Minimisation (SRM) [Vapnik, 1995] before discussing how this might be put in an MML framework.

The statistical learning theory of Vapnik and Chervonenkis uses the notion of the Vapnik-Chervonenkis (VC) dimension of a set to give a classical non-Bayesian way of doing regression (typically using a technique called Structural Risk Minimisation [SRM]) and classification (typically using Support Vector Machines [SVMs]). Recalling the distinction between inference and prediction from sec. 4.1, both statistical learning theory and Akaike’s Information Criterion (AIC) seem to be motivated by prediction, whereas MML is motivated by inference — a point noted in [Wallace, 1997] (ref. [281] in [Dowe, 2008a]).

It is not clear to this writer what statistical learning theory advocates for (comparing) models from different families (e.g., polynomial vs exponential), or for decision trees (classification trees) (e.g., [Wallace and Patrick, 1993; Wallace, 2005, sec. 7.2]), where each split corresponds to a conjunctive “AND”, discretely partitioning the data. It is less clear what statistical learning theory will advocate for the generalisation of decision trees (classification trees) called decision graphs (classification graphs) [Oliver and Wallace, 1991; 1992; Oliver, 1993; Tan and Dowe, 2002; 2003], in which a disjunctive “OR” in the formula allows selected branches of the tree to join — making the model space more general, as now we have two

discrete operators (both split and join) in addition to various continuous-valued parameters (such as the multinomial class probabilities in the leaves).

Problems where this writer is either not sure what statistical learning theory will advocate and/or where I suspect that it might advocate (and have to bear the statistical inconsistency of) Maximum Likelihood include those mentioned above and also the Neyman-Scott and Neyman-Scott panel data problems from secs. 6.4 — 6.5 and the “gappy” problem and the bus number problem of sec. 4.1.

Whether or not my ignorance of how statistical learning theory will behave in certain situations is a shortcoming in statistical learning theory or in me is something for the reader to decide. Meanwhile, MML is universal — due to its relationship with Kolmogorov complexity (as per secs. 1 and 6.2) and likewise because it always has a message length as an objective function.

The original treatments (e.g., [Vapnik, 1995]) of the Vapnik-Chervonenkis (VC) notion of statistical learning theory, support vector machines (SVMs) and structural risk minimisation (SRM) are not Bayesian. Efforts have been made to put the notions of Vapnik-Chervonenkis statistical learning theory into a Bayesian MML (or similar) framework (starting with [Vapnik, 1995, sec. 4.6]). At least one motivation for doing this is to be able to apply statistical learning theory to problems where it might not have otherwise been possible to do so.

Fleshing out ideas hinted at in [Vapnik, 1995, sec. 4.6], MML has been applied to Support Vector Machines (SVMs) in [Tan and Dowe, 2004] (where we do not just have SVMs, but we also have decision trees — and, in fact, we have a hybrid model with SVMs in the leaves of decision trees), with discussions on alternative and refined coding schemes given in [Dowe, 2007; 2008a, footnote 53; 2008b, p. 444] including [Dowe, 2008a, footnote 53, fourth way, pp. 527–528; 2008b, p. 444] explicitly modelling the distribution of *all* the variables, including the input variables. It is MML’s abovementioned relationship with Kolmogorov complexity (or algorithmic information theory) that enables us to consider alternative coding schemes. Explicitly modelling the distribution of all the variables (including the input variables) would amount to making generalized hybrid Bayesian network graphical models (as per sec. 7.6), some of whose properties are discussed in secs. 2.3 and 3.6. (Perhaps digressing, as per [Dowe, 2008a, footnote 56], [Rubinstein *et al.*, 2007] might also be of some use here.)

Staying with the Vapnik-Chervonenkis VC dimension but moving from SVMs to SRM, MML was compared with SRM for univariate polynomial regression in [Wallace, 1997]. See the discussion in [Dowe, 2008a, sec. 0.2.2., p. 528, col. 1, including also footnotes 57 and 58].

MML has also been applied (e.g.) to hierarchical classification in [Boulton and Wallace, 1973b; Dowe, 2008a, sec. 0.2.3, p. 531, col. 1 and sec. 0.2.4, p. 537, col. 2] (and elsewhere), with an application of hierarchical MML mixture modelling in [Wallace and Dale, 2005], to image recognition in [Torsello and Dowe, 2008b; 2008a], with other work on MML mixture modelling in sec. 7.6, and MML applied to James-Stein estimation in [Schmidt and Makalic, 2009b]. For some of many more examples, see also (e.g.) sec. 7.6 (in particular) and (in general) all of sec. 7.

6.7 *A note on Minimum Description Length (MDL)*

I forget how many times and how regularly I have been asked to summarise and/or highlight the similarities and differences between MML and the much later Minimum Description Length (MDL) principle. Because of this and also because of a request from at least one and possibly all of my referees, I include the current section. For want of somewhere to put it, I have placed it here, but the reader can probably safely skip from sec. 6.6 to sec. 6.8 with probably greater continuity and perhaps no great loss.

Historically, the Minimum Description Length (MDL) principle [Rissanen, 1978] (following formative ideas in [Rissanen, 1976]) was first published 10 years, 6 journal papers [Wallace and Boulton, 1968, p. 185, sec. 2; Boulton and Wallace, 1969; 1970, p. 64, col. 1; 1973b, sec. 1, col. 1; 1975, sec. 1, col. 1; Wallace and Boulton, 1975, sec. 3] (it would be 7 journal papers if we were hypothetically to count [Boulton and Wallace, 1973a]), 1 Master's thesis [Boulton, 1970], at least one conference abstract [Boulton and Wallace, 1973c] and 1 PhD thesis [Boulton, 1975] after the seminal Wallace and Boulton MML paper [1968], including 3 years after the Wallace and Boulton [1975] paper introducing Strict MML (whose original publication was delayed as per sec. 6.1 and [Dowe, 2008a, sec. 0.1, p. 524, col. 1]).

The ideas in MML of being Bayesian and of having a two-part message have been unwaveringly constant throughout since the original 1968 inception [Wallace and Boulton, 1968]. A variety of theoretical justifications for Bayesianism are given in (e.g.) sec. 6.5 and in [Wallace, 1996c, secs. 3 (especially 3.4), 4.1 and 5–7; Wallace and Dowe, 1999a, secs. 2.4 and 7; 1999c, secs. 1–2; Comley and Dowe, 2005, p. 269, sec. 11.3.2; Dowe *et al.*, 2007; Dowe, 2008a, footnotes 211, 225 and (start of) 133, and sec. 0.2.7, p. 546; 2008b, p. 450]. A variety of theoretical justifications for the two-part form of the MML message are given in (e.g.) [Wallace and Freeman, 1987, p. 241; Barron and Cover, 1991; Wallace, 2005, sec. 3.4.5, p. 190, note use of “agrees”; Dowe *et al.*, 2007, sec. 5.3.4].

The objectives — or at the least the way(s) of attempting to achieve the objectives — of the Minimum Description Length (MDL) principle would appear to have changed over the years since the first MDL paper in 1978 [Rissanen, 1978], where part of the motivation appears [Rissanen, 1978, p. 465] to be (algorithmic information theory or) Kolmogorov complexity, a term repeated in [Rissanen, 1999a, sec. 2, p. 261]. It is the prerogative of any scientist or any researcher to change and/or refine their ideas, and I attempt to survey various developments and changes in the presentations I have seen of MDL.

Rissanen appears throughout his MDL works to want to avoid being Bayesian. This seems slightly curious to me for a few reasons. First, there are countably infinitely many Universal Turing Machines (UTMs) and, as per secs. 2.4 and 6.5, the choice of a UTM is a Bayesian choice. As such, in relating MDL to Kolmogorov complexity, it seems difficult not to relate MDL to Bayesianism. Second, although Rissanen does not seem to want to use a Bayesian prior, his Normalised Maximum Likelihood (NML) uses the Jeffreys “prior” [Rissanen, 1996; 1999a], an approach

one might facetiously call “Bayesian”. The Jeffreys “prior” is not without issue — it is based on the data (thus apparently weakening the claimed relationship with Kolmogorov complexity), it doesn’t always normalise [Wallace and Dowe, 1999b, sec. 2], and it will typically depend upon things which we would not expect to be overly relevant to our prior beliefs — namely, the strength and location of our measuring instruments [Dowe *et al.*, 1996e, p. 217; Wallace and Dowe, 1999a, sec. 2.3.1; Comley and Dowe, 2005, sec. 11.4.3, p. 273]. There are also other concerns [Wallace and Freeman, 1987, sec. 1, p. 241; Wallace and Dowe, 1999a, sec. 5, p. 277, col. 2]. Efforts to normalise the Jeffreys prior by restricting its domain are said by other authors to be “unsatisfying” [Dawid, 1999, sec. 5, p. 325, col. 2] and would certainly appear to be reverting back to Bayesianism (rather than “Bayesianism”). Another opinion about the Bayesianism or otherwise in MDL is “... we see that Rissanen’s approach is not incompatible with a Bayesian approach” [Clarke, 1999, sec. 2, p. 338, col. 2]. And while discussing the Jeffreys “prior” and Normalised Maximum Likelihood (NML) approach(es) [Rissanen, 1996; 1999a], it is worth inviting the reader to compare with the approximately contemporaneous PIC (Phillips Information Criterion, Posterior Information Criterion) [Phillips and Ploberger, 1996] and the much earlier and very similar Wallace-Freeman approximation [Wallace, 1984a; Wallace and Freeman, 1987] from equation (20) of no later than 1987 — see also the discussion in [Wallace, 2005, sec. 10.2.1].

The MDL notion of ‘completing the code’ (or complete coding) [Rissanen, 1996; Grünwald *et al.*, 1998, sec. 4] seems to break down for a variety of relatively simple cases [Wallace and Dowe, 1999b, secs. 1.2 and 2.3] and would appear to be in violation of contravening the convergence conditions of the two-part message form from which the results in [Barron and Cover, 1991] emanate, a variety of theoretical justifications for which are cited above. The latest versions of MDL seem to advocate using Normalised Maximum Likelihood (NML) to select the “model class” (see [Wallace and Dowe, 1999b, sec. 2.1] re issues of ambiguity here) and the order of the model but not to do the point estimation of the parameters. Given the issues with Maximum Likelihood of over-fitting and statistical inconsistency raised in secs. 4.1 and 5.2, we endorse the avoidance of Maximum Likelihood. But then Normalised Maximum Likelihood (NML) starts to look quite similar to the earlier Wallace-Freeman [1987] approximation for model order selection but without necessarily easily being able to advocate a point estimate.

And, as in sec. 6.5, Grünwald and Langford [2004; 2007] have shown statistical inconsistency for various Bayesian estimators and various forms of the Minimum Description Length (MDL) principle (under model misspecification), but none of us are aware of any current evidence for a statistical inconsistency in MML [Grünwald and Langford, 2007, sec. 7.1.5; Dowe, 2008a, sec. 0.2.5]. (It is probably worth mentioning here an attack on MML [Grünwald *et al.*, 1998] which was later retracted [Grünwald, 2007, sec. 17.4, An Apologetic Remark; Dowe, 2008a, sec. 0.2.4, p. 536].)

As we near the conclusion of this sub-section, it is worth pointing out that many authors use MDL as a generic term for any MDL-like or MML-like coding

scheme based on some similar method. People should take as much care as they are able to here — [Comley and Dowe, 2005, sec. 11.4.3] gives plenty of examples of poor MDL-like coding schemes whose performances vastly improved when they were re-visited using MML [Wallace and Patrick, 1991; 1993, Viswanathan *et al.*, 1999; Needham and Dowe, 2001]. (See also sec. 4.) One can make bad wine from good grapes, and a poor coding scheme will not do justice to the likes of MDL and MML.

Despite the above challenges to and/or criticisms of much MDL work to date as it compares with earlier MML work, considering the issues which Rissanen raises — as he attempts to maintain all the niceties (of MML) while also attempting to avoid being Bayesian — and contemplating responses can certainly yield insights at the very least, and of course possibly much more. One such purported insight is described in part of sec. 7.1, a section in which objective Bayesianism is discussed. (Also worth mentioning here in passing is a way in which MDL could be re-visited stating parameter estimates “using whatever ‘code’ or representation was used in the presentation of the raw data” [Wallace and Dowe, 1999b, sec. 3, p. 336, col. 2].) And, finally, re comparing MDL and MML, as per the final sentence of this section, I refer the reader to [Wallace and Dowe, 1999c, abstract].

Given that when we take logarithms base 2 ($\log_2$) we typically refer to the unit as a bit, for some historical context on what to call the units when we take natural logarithms (base e , $\log_e$), see (e.g.) [Hodges, 1983, pp. 196–197] for early names (e.g., ‘*ban*’), see (e.g.) [Boulton and Wallace, 1970, p. 63; Comley and Dowe, 2005, p. 271, sec. 11.4.1] re ‘*nit*’, and see (e.g.) much later MDL writings for the term ‘*nat*’.

Other treatments of this topic of contrasting MDL and MML are given in (e.g.) [Wallace, 1999; Wallace and Dowe, 1999b, sec. 3; Wallace, 2005, sec. 10.2; Comley and Dowe, 2005, sec. 11.4.3, pp. 272–273; Baxter and Oliver, 1995] and — perhaps most especially in summary — [Wallace and Dowe, 1999c, abstract].

6.8 Comparing “Right”/“Wrong” and Probabilistic scores

The original idea behind the notion of boosting was to more heavily (penalise or) weight incorrect answers in a decision tree (or classification tree) so as to grow the tree and ultimately have less errors — that is, right/wrong errors. Sec. 3.1 showed us that “right”/“wrong” scoring is not invariant to re-framing of questions, and sec. 3.2 re-iterated some recent results on the uniqueness of log(arithm)-loss scoring in being invariant to the re-framing of questions. This said, before we examine boosting more closely in sec. 6.9, we might ask what a good “right”/“wrong” score tells us about the log(arithm)-loss score and vice versa.

By rights, in a multiple-choice question of c choices, even if $c \gg 2$, giving the correct answer a probability of just less than 0.5 can still result in a higher probability of at least 0.5 being given to an incorrect answer and so a “right”/“wrong” score of “wrong” (or 0). So, regardless of the number of choices, c , an incorrect answer guarantees a score of at least 1 bit. Correspondingly, a score of less than 1 bit

guarantees a “right” answer. At the other extreme, it is possible that a probability of just over $1/c$ allocated to the correct answer will give us a “right”/“wrong” score of “right” (or 1). Correspondingly, a score of more than $\log(c)$ will surely give us a “right”/“wrong” score of “wrong” (or 0). So, if we have n multiple-choice questions of $c_1, \dots, c_i, \dots, c_n$ options each, then a “right”/“wrong” score of 0 corresponds to a $\log(\text{arithm})$ -loss cost of at least $\log(2^n) = n$ bits, and a “right”/“wrong” score of (all correct) n correct corresponds to a $\log(\text{arithm})$ -loss cost of at most $\sum_{i=1}^n \log(c_i) = \log(\prod_{i=1}^n c_i)$.

So, slightly paradoxically, on a quiz of 1 ternary (3-valued) question, someone (with probabilities $\{0.498, 0.501, 0.001\}$) might get a wrong answer for a minimum total of 0 “right” with a log-loss penalty score of just over $\log(2)$ whereas someone else (with probabilities $\{0.334, 0.333, 0.333\}$) might get a correct answer for a maximum total of 1 “right” but with a worse log-loss penalty score of just under $\log(3)$. I put it to the reader that the person with the better log-loss score actually has a better claim to having been correct on this question than the person given a score of 1 “right”.

And, of course, more emphatically, on a quiz of n questions, someone with a “right”/“wrong” score of 0 “right” might have a $\log(\text{arithm})$ -loss penalty score of little over $\log(2^n)$ whereas someone who got $(n - 1)$ out of n correct might have an arbitrarily large (or infinite) $\log(\text{arithm})$ -loss penalty score by assigning an arbitrarily small (or zero) probability to the correct answer in the one question that this person got “wrong”. (And, similarly, as at least implicitly pointed out in sec. 3.2, one can use boosting to make all sorts of guesses and predictions in data which is just random noise, and although much damage could be done to the $\log(\text{arithm})$ -loss penalty score, no damage will be done to the “right”/“wrong” score.) We address this issue of getting good “right”/“wrong” scores without unduly damaging the $\log(\text{arithm})$ -loss penalty score again in sec. 6.9.

6.9 Boosting

One suggestion of Chris Wallace’s (in private communication) was that the right/wrong predictive accuracy of MML decision trees could be improved by going to each leaf in turn and doing one additional split beyond the MML split. I understand that this was the motivation behind the subsequent [Oliver and Hand, 1994]. Of course, as per sec. 4.1, optimal prediction is given not just by using the MML tree, but by combining several trees — ideally as many as possible — together [Oliver and Hand, 1996; Tan and Dowe, 2006].

However, given our new apparent uniqueness results for probabilistic $\log(\text{arithm})$ -loss scoring from [Dowe, 2008a, footnote 175 (and 176); 2008b, pp. 437–438] and sec. 3.2, it perhaps makes more sense to carefully focus on improving the probabilistic $\log(\text{arithm})$ -loss score.

One option is to do MML inference but with the “boosting priors” from [Dowe, 2008a, sec. 0.2.6]. The idea behind these “boosting priors” is that, rather than fix our Beta/Dirichlet prior [Wallace, 2005, p. 47 and sec. 5.4] to have $\alpha = 1$, we

can try “boosting priors”, whose rough form [Tan and Dowe, 2006, sec. 3.4, p. 598; Dowe, 2008a, sec. 0.2.6] on α could be, e.g., $3/(2\sqrt{\alpha}(1 + \sqrt{\alpha})^4)$ or $(e^{-\alpha/\pi})/(\pi\sqrt{\alpha})$. The idea is simply to retain a mean of (approximately) 1 but to have a large spike near $\alpha = 0$, which in turn increases our propensity to have pure classes.

Another option is to re-visit boosting and to think of it not as in its original form of minimising the number of right/wrong errors but rather instead in the similar form of trying to optimise the expected predictive score. This modification to boosting is related to the original form of boosting in that each individual (right/wrong) mistake will typically correspond to a poor right/wrong (yes/no) score. The predictive score should not be done using the log-likelihood, but rather should be done using the minimum expected Kullback-Leibler (minEKL) probability estimate [Dowe *et al.*, 1998; Wallace, 2005, secs. 4.7–4.9; Dowe *et al.*, 2007, secs. 4 and 6.1.4; Dowe, 2008a, sec. 0.2.2, 2008b, p. 444] from sec. 4.1. In other words, if there are m classes and a given leaf has counts $c_1, \dots, c_i, \dots, c_m$ for the m classes and $C = \sum_{i=1}^m c_i$, Maximum Likelihood would advocate a probability estimate of $\hat{p}_i^{\text{MaxLhood}} = c_i/C$ for each class. However, if we think of α_i from a Dirichlet prior as denoting a “pre-count” in class i (before any data is actually counted), then the probability in each class can be regarded as $\hat{p}_i = (c_i + \alpha_i)/(C + (\sum_{j=1}^m \alpha_j))$. Of course, we can set $\alpha_i = \alpha$ for each class and then use the so-called “boosting priors” on α as per [Tan and Dowe, 2006, sec. 3.4, p. 598; Dowe, 2008a, sec. 0.2.6].

Let us finish with three further comments. First, by way of digression, an attempt to give “objective” ways of choosing α is given in sec. 7.1. Second, for those who wish to boost to improve the right/wrong score or simply wish to get a good to excellent right/wrong score, given the (apparent uniqueness of) invariance of the log-loss score, we make the simple recommendation that predictors that give good right/wrong scores be checked so that they also give a good log-loss score — this might involve moving extreme probabilities away from the extremities of 0 and 1 (such as can arise from using Maximum Likelihood). (Some possible estimators for doing this are given in, e.g., [Wallace, 2005, secs. 4.8 and 5.4]. A prediction method which is good enough to genuinely get a good “right”/“wrong” score can surely be gently modified or toned down to give a good log-loss score.) As a third comment, for those not wishing to sacrifice statistical consistency in their efforts to improve predictive accuracy, it might be worth considering comments [Dowe, 2008a, footnote 130] about potential dangers of placing too much weight on the likelihood of the data.

7 MML AND SOME APPLICATIONS IN PHILOSOPHY AND ELSEWHERE

MML has been applied to a variety of problems in philosophy — including (e.g.) the philosophy of science [Dowe and Oppy, 2001], the philosophy of statistics and inference [Dowe *et al.*, 2007], the philosophy of mind (see, e.g., sec. 7.3), the philosophy of language and the philosophy of religion. We mention these and some other — mainly philosophical — issues here.

7.1 Objective Bayesianism (and Bertrand's paradox) and some new invariant "objective priors"

Bertrand's paradox is essentially concerned with the issue that we can not choose a uniform Bayesian prior in all parameterisations. Certainly, many authors would like an objective form of Bayesianism — or, equivalently, a parameterisation in which our Bayesian prior can be uniform.

Recalling the notion of Universal Turing Machine (UTM) from sec. 2.4, one can claim that a simplest UTM is one with the smallest product of its number of states and its number of symbols (as this is the number of rows in the instruction table) [Dowe *et al.*, to appear (b)].

Simplest UTMs have been used in inference [Wallace, 2005, sec. 2.3.12; Gammernan and Vovk, 2007a; 2007b; Martin-Löf, 1966; Dowe, 2007], and they are one way of attempting to be objective (while still being Bayesian) and — as a consequence — side-stepping Bertrand's paradox. Of course, such objectivity — where possible — will potentially be useful in places such as legal battles [Dowe, 2008a, pp. 438–439].

Much good and interesting work has been done in the area of objective Bayesianism by (e.g.) J. M. Bernardo [Bernardo and Smith, 1994] and others. Above, we follow Wallace in offering simplest UTMs as objective priors. Below, as per the end of sec. 6.7, we now change tack and re-visit the Jeffreys "prior" (whose use, incidentally, was not advocated by Jeffreys [1946]) as an objective invariant prior.

Some things that we know to be invariant upon re-parameterisation include the likelihood function, Maximum Likelihood, the marginal probability $r(x_i) = \int h(\theta)f(x_i|\theta) d\theta$ of datum x_i (from sec. 6.1), the message length and many of its variants (from secs. 6 — 6.3), Minimum Message Length (when using an invariant form), the Fisher information and (recalling secs. 4.1 and 6.9) the minimum Expected Kullback-Leibler divergence (minEKL) estimator.

Given these invariant building blocks, we now take the Jeffreys "prior" (which we recall from sec. 6.7 and pointers therein does not always normalise), and construct a family of other invariant priors. To kick off with an example, let us take a (m -state, $(M - 1)$ -dimensional, $M \geq 2$) multinomial distribution — as per [Wallace, 2005, sec. 5.4] — with prior of the form $\text{Const } p_1^{\alpha_1} \dots p_M^{\alpha_M}$ (where we use α_i where Wallace [2005, sec. 5.4] writes $\alpha_i - 1$). We will have counts s_i ($i = 1, \dots, M$) and we let $N = \sum_{i=1}^M s_i$ and $A = \sum_{i=1}^M \alpha_i$. The Wallace-Freeman (1987) MML estimate from equation (20) is $(\hat{p}_i)_{MMLWF1987} = (s_i + \alpha_i + 1/2)/(N + A + M/2)$. And, recalling sec. 6.7, the minEKL (or Laplace) estimate (equivalently here, the posterior mean [Boulton and Wallace, 1969]) is $(\hat{p}_i)_{MinEKL} = (s_i + \alpha_i + 1)/(N + A + M)$. As such, we observe that the Wallace-Freeman [1987] MML estimate from equation (20) with the Jeffreys "prior" ($\alpha_i = -1/2$) gives the Maximum Likelihood estimate. We similarly observe that the Wallace-Freeman estimate with the uniform prior ($\alpha_i = 0$) is equivalent to getting the minEKL estimate with the Jeffreys "prior" ($\alpha_i = -1/2$).

In this particular case of the multinomial distribution, we note that we can trans-

form from the (invariant) Wallace-Freeman [1987] MML estimate to (invariant) minEKL by adding $1/2$ to the α_i . As such, if the Jeffreys prior $h_0 = h_{Jeffreys} = h_{FisherInfo}$ (with $\alpha_i = -1/2$) is to be called objective, then a case can be made that so, too, is the uniform prior h_1 (with $\alpha_i = 0$). We can iterate again to get further invariant priors: h_2 (with $\alpha_i = 1/2$), h_3 (with $\alpha_i = 1$), etc. One could also iterate in the opposite direction: h_{-1} (with $\alpha_i = -1$), h_{-2} (with $\alpha_i = -3/2$), etc. All such priors — or at least those which normalise — are invariant (by construction) and can be regarded in some sense as “objective”. One could then choose the prior h_j for the smallest value of j for which h_j normalises.

This method of using the Jeffreys “prior” to generate further invariant objective priors (via invariant transformations) and then taking the “first” to normalise certainly generalises — well beyond the above example of the multinomial distribution — to other distributions. In general, for some given distribution, start again with $h_0 = h_{Jeffreys} = h_{FisherInfo}$ and then, given prior h_i , let h_{i+1} be the prior such that (some invariant form of) the MML estimate with prior h_{i+1} is as close as possible in Kullback-Leibler distance (and ideally equal) to the minEKL estimate with prior h_i . With however much ease or difficulty, we can then generate this sequence of invariant priors $h_0 = h_{Jeffreys} = h_{FisherInfo}$, h_1 , h_2 , ... and perhaps also h_{-1} , h_{-2} , etc. (As a general rule, because of MML’s tendency to fit just right and minEKL’s tendency to under-fit as per sec. 4.1, we expect to see a corresponding progression in this sequence of priors — as is perhaps best seen from the above example with the multinomial distribution. In that case, h_i has $\alpha_i = (i - 1)/2$, meaning that, as i increases, it takes increasingly much data to move the estimate away from the centroid where for all i , $\hat{p}_i = 1/M$.) If there does exist some smallest j for which h_j normalises, a case could be made that this is an objective invariant prior which might be more suitable than the Jeffreys “prior”, h_0 .

Penultimately, cases could be made for investigating combining two such priors, as in considering (e.g.) $h_{hybrid} = \sqrt{h_{j_1} h_{j_2}}$. Cases could also be made for attempting to allow j not to be an integer but rather somehow to be fractional. We will not investigate these here.

And, finally, returning to issues from sec. 6.7, it would perhaps be nice if Normalised Maximum Likelihood (NML) could be re-visited but with use of one of these alternative invariant priors to the Jeffreys “prior”, h_0 . This should retain the statistical invariance but might reduce some of the vulnerabilities (such as over-fitting and statistical inconsistency, both mentioned earlier) associated with Maximum Likelihood.

7.2 Goodman’s “Grue” paradox (and choice of language)

Nelson Goodman’s “grue” paradox raises the issue of why notions like “green” and “blue” should be more natural than notions of “grue” (green before time t_0 [say year 3000] and blue thereafter) and “bleen” (blue before time t_0 [say year 3000] and green thereafter). This has been discussed in the Solomonoff predictive

and Wallace MML inductive frameworks, with relevant writings being [Solomonoff, 1996; 1997b, sec. 5; Comley and Dowe, 2005, sec. 11.4.4; Dowe, 2008a, footnotes 128, 184 and 227]. Among other things, an adequate solution of when to arrive at a notion like “grue” and when to arrive at a notion like “green” (which is, after all, grue before time t_0 and bleen thereafter) is presumably necessary when trying to evolve language (for those beings not yet with language) or when trying to communicate with non-human terrestrials or extra-terrestrials [Dowe, 2008a, footnote 184]. Wallace’s approach from [Dowe, 2008a, footnote 128], elaborating upon [Comley and Dowe, 2005, sec. 11.4.4], was summarised as follows: “Suppose someone is growing and harvesting crops, commencing (much) before t_0 and finishing (much) after t_0 . We expect the grass and certain moulds to be green, and we expect the sky and certain weeds to be blue. The notions of grue and bleen here offer at most little in return other than sometimes to require (time-based) qualification and to make the language sometimes unnecessarily cumbersome.” This said, there are times of event changes which can be of interest. If t_0 were the time of the next expected reversal of the earth’s magnetic field, then in talking on such a time-scale we have reason to disambiguate between magnetic north and geographic north in our language — as these notions are approximately equal before t_0 and approximately antipodal (for at least some time) after t_0 [Dowe, 2008a, footnote 128]. But the terms ‘grue’ and ‘bleen’ cost us but seem to gain us nothing. By and large, languages will develop, import, qualify and/or abbreviate terms when these terms warrant (sufficient) use.

And, while on that very issue of abbreviation, the reader will note at least one place in this article where we have written “Minimum Message Length (MML)”. This convention of putting an acronym or other abbreviation in brackets immediately after the term it abbreviates enables us to use the abbreviation (rather than the full term) elsewhere — thus enabling us to shorten the length of our message.

And, digressing, while on the earlier issue of languages, MML has been used to model evolution of languages [Ooi and Dowe, 2005; Dowe, 2008a, sec. 0.2.4; 2008b, p. 455] (not to mention finite state automata [Wallace and Georgeff, 1983] and DNA string alignment [Allison *et al.*, 1990a; 1990b; 1990; 1991; 1992a; 1992b; Allison and Wallace, 1993, 1994a, 1994b]).

An able philosopher colleague, Toby Handfield, has told me in private communication — while discussing Lewis [1976] and the “laws of nature” — that if MML were able to recognise a number constructed as (say) the sum without carries of e (the base of natural logarithms) expanded in hexadecimal (base 16) and π expanded in decimal (base 10), then this would go a long way towards convincing him that MML can solve Goodman’s grue paradox. Using the relationship between MML and (algorithmic information theory or) Kolmogorov complexity [Wallace and Dowe, 1999a; Wallace, 2005, chaps. 2–3] discussed in sec. 6, we outline the argument below. In short, MML will have no difficulty with doing this (in principle) — the caveat being that the search might take quite some time.

We can specify e as $\sum_{i=0}^{\infty} 1/i!$ in a (Turing machine) program of length P_e , and we can specify the h^{th} hex(adecimal) digit of e in a program of length $P_e + C_1 + l(h)$

for some constant C_1 , where $l(h)$ is the length of some prefix code (recall sec. 2.2) over the positive integers (e.g., the unary code from sec. 2.4). We could use a code of length 1 for $h = 1$ and of length $\leq 1 + \lceil 1 + \log_2(h) + 2 \log_2(\log_2(h)) \rceil < 2 + 1 + \log_2(h) + 2 \log_2(\log_2(h))$ for $h \geq 2$. Similarly, we can specify π as (e.g.) $\sum_{i=0}^{\infty} (4(-1)^i)/(2i+1)$ in a (Turing machine) program of length P_π , and we can specify the h^{th} hex(adecimal) digit of π in a program of length $P_\pi + C_2 + l(h)$ for some constant C_2 , where $l(h)$ is as above.

The program for addition without carry/ies simply entails addition without carry (or modulo addition) in each place, h , for $h = 1, 2, 3, 4, \dots$. So, for the h^{th} hex(adecimal) digit, we can say that the h^{th} hex digit, composite_h , of our composite number is given as follows:

```

if      (e_{h, 16} + pi_{h, 10} < 15)
then composite_h = e_{h, 16} + pi_{h, 10}
else  composite_h = e_{h, 16} + pi_{h, 10} - 16;

```

Given that this is how the composite number is being generated, given sufficiently many hex digits of this number, the Minimum Message Length (MML) inference will be the algorithm for generating this composite number. Again, the search might be slow, but this will be found.

We can actually take this further by randomly adding noise. Let us suppose that, with probability p , hex digit h comes from some probability distribution $(q_1, q_2, \dots, q_{14}, q_{15}, q_{16} = 1 - \sum_{i=1}^{15} q_i)$ and with probability $1 - p$ this h^{th} hex digit will be composite_h . So, $Pr(\text{CompositeWithNoise}_h = \text{composite}_h) = pq_{\text{composite}_h} + (1 - p)$. For each $i \neq \text{composite}_h$, $Pr(\text{CompositeWithNoise}_h = i) = pq_i$. In the case that $p = 0$, this reduces to the noiseless case. Here, the search will be even slower, but with sufficiently many digits and with sufficient search time, we will converge upon the noiseless program above generating the digits in addition to having an increasingly good quantification of the noise.

7.3 MML, inductive inference, explanation and intelligence

As intimated in sec. 1, MML gives us the inductive inference (or induction, or inference, or explanation) part of intelligence [Dowe and Hajek, 1997; 1998, especially sec. 2 (and its title) and sec. 4; Sanghi and Dowe, 2003, sec. 5.2]. And Ockham's razor tells us that we should expect to improve on Searle's "Chinese room" look-up table [Searle, 1980] by having a compressed representation — as per our commonsense intuition and arguments in [Dowe and Hajek, 1997, sec. 5.1 and elsewhere; 1997, p. 105, sec. 5 and elsewhere; Dowe, 2008a, footnote 182 and surrounding text] and sec. 4.

Let us consider an assertion by Hutter [Legg and Hutter, 2007] that compression is equivalent to (artificial) intelligence (although subsequent work by Hutter now seems instead to equate intelligence with a weighted sum of reward scores across different environments). This assertion is later than a similar idea of Hernández-Orallo [Hernández-Orallo and Minaya-Collado, 1998; Hernández-Orallo, 2000]. It

is also stronger than an earlier idea [Dowe and Hajek, 1997; 1998, especially sec. 2 (and its title) and sec. 4; Sanghi and Dowe, 2003, sec. 5.2 and elsewhere; Dowe, 2008a, sec. 0.2.5, p. 542, col. 2 and sec 0.2.7, p. 545, col. 1] that (the part of intelligence which is) inductive inference (or inductive inference) is equivalent to (two-part) compression. Let us look at the two issues separately of

- (i) first, whether all of (artificial) intelligence or perhaps just inductive inference is equivalent to (two-part) compression, and
- (ii) second, whether it is satisfactory simply to talk about (one-part) compression or whether we should insist upon two-part compression.

First, the components of intelligence would appear to include (at least) memory, deductive inference, inductive inference and ability to receive direct instruction. (By deductive inference, we mean and include mathematical calculations and logical reasoning, such as *modus ponens* — Socrates is a man, all men are mortal, therefore Socrates is mortal. To illustrate the distinction with an example, inductive inference is more along the lines of all men are mortal, Socrates is mortal, therefore we assert some probability that Socrates is a man.) We need memory to store observations for making inductive inferences, for remembering inductive inferences, for remembering our progress through mathematical calculations or other (logical) deductions and for remembering those direct instructions (perhaps the deductions or inferences of others) that we receive. For example, a good human player of a game where the search space is too vast to be exhaustively searched (like chess or Go) needs inductive inference and direct instruction to help with an evaluation function (such as, in chess, the advantages of passed pawns, the weaknesses of isolated and backward pawns, and the approximate equivalence between a queen and three minor pieces), memory to remember these, memory to remember the rules, and deduction (and memory again) to do the lookahead calculations in the search tree. It is clear that all these aspects of intelligence are useful to a human player of such a game. However, to the mathematician, the logician, and especially a mathematician or a logician checking the validity of a proof (or someone double-checking that a Sudoku solution is correct), the main forms of intelligence required would surely appear to be deduction and memory. It is fair to say that the harder aspects of inductive learning and (two-part) compression also require memory and deductive inference. And we have argued elsewhere that we are more likely to attribute intelligence to someone performing an act of great memory if they have done this using a compressed representation [Dowe and Hajek, 1997; 1998]. But we ask the reader whether we should not attribute intelligence to the chess player or the mathematician (or the person checking a Sudoku solution) when performing (difficult) activities involving at most little inductive inference.

Second, in many cases, doing straight (one-part) compression rather than two-part compression can lead to an incorrect model (as in the statistical inconsistency of the minEKL estimator from sec. 4.1 for the “gappy” problem mentioned in sec. 4.1 and for the Neyman-Scott problem in sec. 6.4) — and this remains true asymptotically regardless of how much data we have.

As per [Dowe, 2008a, sec. 0.2.7, p. 545, col. 1], I have discussed with J. Hernández Orallo the notion of quantifying the intelligence of a system of agents and endeavouring to quantify how much of this comes from the individual agents (in isolation) and how much comes from their communication. Let us try to take this further in a couple of different (related) ways. First, it would be good to (artificially) evolve such a communal intelligence, including (perhaps inevitably) evolving a language. (As a tiny step, one of my 4th year Honours project students in 2009, Jeffrey R. Parsons, has made slight progress in evolving Mealy and/or Moore machines with the message length as a guiding fitness function. I do not wish to overplay his current progress, but it is in a useful direction.) And, second, re the topics of swarm intelligence and ant colony optimisation, perhaps only a very small range of parameter values (where the parameters describe the individual agents and/or their communication) permit the different parts to interact as an “intelligent” whole. This raises a couple of further issues: the issue of using MML to analyse data (as per sec. 7.6 and [Dowe, 2008a, sec. 0.2.7, p. 545]) and infer the parameter values (or setting) giving the greatest communal intelligence, and the additional issue(s) of whether or not greater prior weight should be given to those systems giving the interesting outcome of intelligence, and (similarly) — in the *fine tuning* argument of sec. 7.7 and [Dowe *et al.*, to appear (b)] — whether greater prior probability should be given to parameter settings in which interesting universes (like our own) result.

Having mentioned here the issues of intelligence, non-human intelligence and communication, it is worth mentioning some of Chris Wallace’s comments about trying to communicate with an alien intelligence [Dowe, 2008a, sec. 0.2.5, p. 542, col. 2, and also footnote 184 and perhaps text around footnote 200] (and possibly also worth recalling Goodman’s notion of “grue” from sec. 7.2 and [Dowe, 2008a, footnote 128]).

We conclude here by saying that further discussion on some of the topics in this sub-section will appear in [Hernández-Orallo and Dowe, 2010].

7.4 (So-called) Causality

Chris Wallace did much work on “causal nets” using MML, including doing the (MML) mathematics and writing the software behind several papers on this topic [Wallace and Korb, 1994; Wallace, 1996b; Wallace *et al.*, 1996a; 1996b; Dai *et al.*, 1996a; 1996b; 1997a; 1997b; Wallace and Korb, 1997; 1999; Korb and Wallace, 1997; 1999] (with the possible exception of [Neil *et al.*, 1999a; 1999b]). I have no objection to the quality of Wallace’s MML statistical inference — from the available data — in this work. Indeed, I have (at most little or) nothing but the highest praise for it. However, there are at least two or so matters about which one should express caution when interpreting the results from such inference.

One issue is that getting the wrong statistical model (because we didn’t have enough data, we hadn’t searched thoroughly enough and/or our statistical inference method was sub-optimal) can lead to having arrows pointing the wrong way.

And even if we did have enough data, our statistical inference method was ideal and we searched thoroughly, it could still be the case that the true (underlying) model (from which the data has been generated) is outside the family of models that we are considering — e.g., our model family might be restricted to linear regressions (on the parent “explanatory” variables to the “target” child variable) with Gaussian noise while the real data-generating process might be more complicated. In such cases, slightly modifying the family of models being considered might change the directions of arrows in the inference, suggesting that the directions of these arrows should not all be regarded as directly “*causal*” [Dowe, 2008a, footnote 169].

As a related case in point, we might have data of (at least) two variables, including (i) height/co-ordinates of (weather) (monitoring) station and (ii) (air) pressure reading. Our best statistical model might have arrows from pressure reading to height of monitoring station, but we surely shouldn’t interpret this arrow as being in any way causal.

Of course, temporal knowledge (of the order in which things occur) is also important for attributing causality.

Whether or not this is well-known and well-documented (and please pardon my medical ignorance, as per sec. 3.2), there would appear to be a substantial overlap between cancer patients and stroke patients. Let’s suppose that in many cases the patient has a stroke and then cancer is detected some months later. It could appear that the stroke caused the cancer, but it is perhaps more probable that cancer-induced changes in the tissue and/or the bloodstream caused the stroke — even if the primary cancer was not in the brain and the metastatic cancer did not present in the brain until after the stroke. If this is all true, then it would suggest that the actual mechanism is that the Cancer is causing the Stroke — despite the possibility that an analysis of the data might easily lead one to conclude that the Stroke is causing the (Brain) Cancer.

We also have to be careful about issues such as (e.g.) hidden (or unknown) variables. As an example, a hidden latent variable might cause both A (which takes place slightly before B) and B . B might do the same exam paper (of mathematical calculations) as A but starting and finishing slightly later. Or perhaps B is a newspaper which goes to print after newspaper A goes to print but before newspaper A appears on the stands. We expect B ’s answers and stories to be very similar to A ’s, but this is because A and B have common (hidden) causes; and it seems loose to say that A causes B .

As another example, standing on a podium of a Grand Prix tends to greatly increase one’s chances of winning at a subsequent Grand Prix event. But this wouldn’t be true of one of the scaffolding constructors who tested the podium before the race ceremony, and nor would it be true of some overly exuberant spectator who managed to somehow get access to the podium. Rather, there is a (not very) hidden cause of ability causing someone to do well in two races, and the result of doing well in the first of these races caused that racer to stand on the podium at the end of the first of the two races.

Lecturers (at university, college, or wherever), tutors, teachers and instructors who are able to give lectures without notes (of whom Chris Wallace is but one notable example) often give excellent lectures. Let us assume that this is the norm for such people. The cause of the good lecturing is surely the excellent memory and strong command of the subject of the lecturer rather than any supposed benefit that the average person might supposedly gain by trying to lecture without notes.

As another example, if A causes B and B causes C and we only know about A and C but have not yet even conceived of B (and, of course, we might be open to the existence of B but simply don't know), then I think we can say A "causations" C but we have to be careful about saying that (e.g., supposedly) A *causes* C . A specific case might be the old example of A being living alone, C being having (few or) no rodents at home. B is the owning of a pet cat — the single person keeps the pet for company (and has no flat-mate or house-mate to complain), and the cat keeps the rodents at bay.

Possibly see also [Dowe, 2008a, sec. 0.2.7, pp. 543–544] re LNPPP and causality.

7.5 *Elusive model paradox (and encryption)*

Gödel's incompleteness theorem consists of constructing a mathematical statement which can be interpreted as saying that "This statement is not provable" [Gödel, 1931]. Clearly, this statement can't be false, or it would be provable and hence true, leading to a logical contradiction. Hence, the statement must be both true (of the natural numbers) and not provable.

The original version of the elusive model paradox gives us a sequence where the next number is one (or unity) more than what we would expect it to be [Dowe, 2008a, footnote 211]. The subsequent version of the paradox essentially takes modulo 2 (so that even numbers are transformed to 0 and odd numbers are transformed to 1) and then gives us a binary sequence (or bit string) (of 0s and 1s) in which we can (paradoxically) be sure that the next bit is not the bit that we expect (or would have predicted) based on what we have seen so far (before it). This leads to a contradiction from which the only escape would appear to be the undecidability of the Halting problem (or *Entscheidungsproblem*), the notion that there are many calculations which will never terminate but for which we can never know that they will not terminate [Turing, 1936].

Whether one takes the elusive model paradox as being over a sequence of (increasing) positive integers (as per the original version [Dowe, 2008a, footnote 211]) or over a binary bit string sequence of 0s and 1s (as per the later version [Dowe, 2008b, p. 455]), each of these versions in turn can be thought of in two (essentially) equivalent ways. One of these ways is to play this as a game, where we have one agent (which can be represented by a Turing machine) generating the sequence and a group of one or more agents (which can also be represented by a Turing machine) trying to guess the next bit — while the (Turing machine) agent generating the sequence is attempting to generate the opposite bit to what (the Turing machine representing) those guessing will guess. (It might possibly help to

think of the generating agent as a soccer player taking a penalty kick, trying to kick the ball where the goalie won't be — or as a tennis player trying to serve the ball to where the receiver won't be; and, in turn, the guessing agent as the goalie trying to guess the location of the kick or the tennis receiver trying to anticipate the serve.) To give both the generating Turing machine agent and the guessing Turing machine agent the best chances to do their respective jobs properly, we will assume that — recalling sec. 2.4 — these Turing machines are universal. As such, among other things, there will be finite emulation programs (or translation programs) causing one machine to emulate the other, and vice versa. As the generating program and the guessing program start out on small sequences being the early short initial segments of the generated bits and the guess(ed) bits, the programs will quite possibly have different models of the data. But, as the sequences get longer and longer, after they become at least kilobits, megabits, gigabits, terabits, etc. long and vastly longer than the abovementioned translation programs, the models that these two UTMs have of the available data will start to converge. The guessing UTM will have had a *very* good look at the generating UTM and — given that the generating UTM is a finite deterministic machine — the guessing UTM would appear to be able at some stage to lock in on the behaviour of the generating UTM, thereafter guessing all subsequent bits correctly. Similarly, at some stage, the generating UTM would appear to be able at some stage to lock in on the behaviour of the guessing UTM, thereafter anticipating all its guesses and then flipping the bit before generating it. After both these stages have occurred, we have the contradiction that the guessing UTM always guesses correctly and the generating UTM anticipates the guess, flips the bit that it knows will be guessed and ensures that all subsequent guesses are incorrect. The Halting problem gets us out of this paradox (and seems like the only way out), as both the generating UTM and the guessing UTM can and very often want more time before they are content that they have modelled the other correctly. The second (essentially) equivalent way of thinking of the elusive model paradox is simply that the generating UTM agent and the guessing UTM agent are the same — as at the end of the previous paragraph. After starting off the sequence, we guess which bit should most probably come next, and then generate the bit which is least probable to come next — and then continue this indefinitely. We get (essentially) the same paradox, and again the Halting problem seems like the only way out of the paradox.

The above all said by way of introduction, we now present some variations on the elusive model paradox [Dowe, 2008a, footnote 211; 2008b, p. 455], including — recalling sec. 4.1 — one using inference and one using prediction. (Recall that inference uses the single best model whereas prediction weights over all available models.) One variation is that we can restrict ourselves to multinomial Markov models where the n^{th} order Markov model has (a maximum of) 2^n binomial distributions.

Let $j = j_m = j_m(i) \leq i$ be some unbounded non-decreasing computable function of i . At step i , having bits $b_1, b_2, \dots, b_i$, we choose b_{i+1} as follows, from the following two similar but (slightly) different methods — noting that both these constructions

are computable.

Method 1 (inference — using restricted “memory”): We infer the best (MML) Markov model of order $\leq j_m$ based on $b_1, b_2, \dots, b_i$. We then use the predictive distribution from this MML inference to give a probability distribution for b_{i+1} . We then choose b_{i+1} to be the bit with the least predicted probability.

Method 2 (prediction — using restricted “memory”): We use Bayesian model averaging over all the Markov models of order $\leq i$ to get a predictive probability distribution over b_{i+1} . Again, we choose b_{i+1} to be the bit which has the lowest predicted probability.

With both of these methods — method 1 (inference) and method 2 (prediction) — the resultant sequence is “random” in the sense that no Markov model of finite order is going to be able to compress it. And this is so because the construction of the sequence is to destroy any such structure at the first viable opportunity upon its detection.

Recall that both these constructions immediately above based on restricting “memory” are computable. Two (or more) alternative computable constructions — based on restricting computation *time* rather than “memory” — are given below. Let $j = j_t = j_t(i) > i$ be some strictly increasing computable function of i .

Method 3 (inference — with restricted computation time): We infer the best (Minimum Message Length [MML]) inference from all computable functions (that we search over) within a search time of $\leq j_t$ based on $b_1, b_2, \dots, b_i$. As in method 1, we then use the predictive distribution from this MML inference to give a probability distribution for b_{i+1} . We then choose b_{i+1} to be the bit with the least predicted probability.

Method(s) 4 (prediction — with restricted computation time): We use Bayesian model averaging. There are two ways of proceeding further in restricted finite computation time — method 4(a) and (with tighter restriction) method 4(b).

Method 4(a): We use Bayesian model averaging over all the Markov models of order $\leq i$ to get a predictive probability distribution over b_{i+1} . Here, time restriction of $\leq j_t$ is applied to each of the individual Markov models in turn. They are then averaged as per Method 2. Again, we choose b_{i+1} to be the bit which has the lowest predicted probability.

Method 4(b): We use Bayesian model averaging over all the Markov models of order $\leq i$ to get a predictive probability distribution over b_{i+1} . But, here, the time restriction of $\leq j_t$ is tighter in that it is applied to the entire calculation, including the final Bayesian model averaging. And, yet again, we choose b_{i+1} to be the bit which has the lowest predicted probability.

We might refer to these various sequences emerging from variations of our elusive model paradox as “*red herring*” sequences. Among other things, these (red herring sequences) have the potential to be used in encryption. If various people or agents studying the sequence have varying computational resources (e.g., different lag lengths in the Markov models they can consider), a variant of the sequence can be constructed in such a way as to guide some sub-population (perhaps those from whom we wish to conceal some data) to believe in the presence or absence

of a particular pattern while guiding a different sub-population (perhaps those to whom we wish to divulge the data) to be aware of the presence (or absence) of some (particular) pattern.

Finally, let us return to the notes at the start of this subsection about how one (apparently) needs the Halting problem (or *Entscheidungsproblem*) [Turing, 1936] to resolve the elusive model paradox [Dowe, 2008a, footnote 211; 2008b, p. 455]. The Halting problem is something which people do not normally encounter before their undergraduate university years. I put it to the reader that the elusive model paradox is something from which we can deduce the halting problem yet which should be accessible to school students.

7.6 *Some of many other issues which MML can address*

- In numerical integration (or numerical quadrature), we see a variety of approaches such as the rectangular rule, the trapezoidal rule and Simpson's rule, etc. Of course, where the function we are trying to integrate is not generated from a polynomial and especially when it is generated from a noisy process, then it will typically be better to use MML or a related method to guide the fitting process rather than use arbitrarily complicated polynomials and suffer from the over-fitting problems that come with Maximum Likelihood and similar methods;
- generalized hybrid Bayesian network graphical models [Dowe and Wallace, 1998; Comley and Dowe, 2003; Tan and Dowe, 2004, sec. 5; Comley and Dowe, 2005] deal with the issue of “discriminative vs generative” studied by Jebara [2003] and others (e.g., [Long and Servedio, 2006]). Many authors have claimed that discriminative learning can often outperform generative learning. However, if one follows the ideas in [Dowe and Wallace, 1998; Comley and Dowe, 2003; Tan and Dowe, 2004, sec. 5; Comley and Dowe, 2005] and carefully uses MML — recalling the discussion of poor coding schemes in sec. 6.7, taking care with the coding scheme — to construct one's generalized hybrid Bayesian network graphical model (of which inference of a logic program via Inductive Logic Programming [ILP] [Dowe *et al.*, to appear (a)] is one possible outcome, as can be SVMs from sec. 6.6 or also, e.g., the sort of model from [Oliver and Dowe, 1995]), then the statistical consistency results of MML from [Dowe, 2008a, sec. 0.2.5] and discussed in sec. 6 should guarantee that “generative” learning (when done like this) works fine.

Some properties of these generalized hybrid Bayesian network graphical models (which can include both continuous and discrete variables [Comley and Dowe, 2003; 2005]) are discussed in secs. 2.3 (where it is mentioned in passing that entropy can be defined on such hybrid structures) and 3.6 (where we mention that there is no difficulty in defining Kullback-Leibler divergence over such structures);

- following this point, where an unnormalised database is sufficiently large,

then MML inference will lead to database normalisation [Dowe, 2008a, sec. 0.2.6, footnote 187; 2008b, pp. 454–455; Dowe and Zaidi, 2010], often resulting in several tables, conveying a generalised Bayesian net. We can adjust our priors to require this process to be noise-free;

- experimental design [Dowe, 2008a, sec. 0.2.7, p. 544; 2008b, pp. 445–446];
- MML can be applied to statistical hypothesis testing [Dowe, 2008a, sec. 0.2.5, p. 539 and sec. 0.2.2, p. 528, col. 1; 2008b, p. 433 (Abstract), p. 435, p. 445 and pp. 455–456; Musgrave and Dowe, 2010], as can also MDL [Rissanen, 1999a, sec. 3]. (Perhaps see also [Dowe, 2008a, sec. 1].) Daniel F. Schmidt and Enes Makalic have recently presented work in front of an audience including me showing their desire to take this further. As per sec. 7.1, I harbour some concerns about associating the Maximum Likelihood estimate with Normalised Maximum Likelihood;
- association rules (from “data mining” and machine learning) can be incorporated within a generalised Bayesian network structure;
- re-visiting A. Elo’s Elo system and M. Glickman’s Glicko system for chess ratings. Whether for chess players with the advantage of the white pieces and the first move or whether for sports teams with a home ground advantage, MML can both select the relevant model and do the parameter estimation. Of interest would be a Neyman-Scott-like situation in which many games are being played but there are relatively few games per player. If similar interest would be a situation with several groups of players where there are many games played within each of the groups but very games played between members of different groups;
- directional angular data, such as the von Mises circular distribution [Wallace and Dowe, 1993; 1994a; Dowe *et al.*, 1995a; 1995b] and the von Mises-Fisher spherical distribution [Dowe *et al.*, 1996e; 1996f];
- inference of megalithic stone circle (or non-circle) geometries [Patrick and Wallace, 1977; Patrick, 1978; Patrick and Wallace, 1982];
- polynomial regression [Wallace, 1997; Wallace, 1998c; Viswanathan and Wallace, 1999; Rumantir and Wallace, 2001; Fitzgibbon *et al.*, 2002a; Rumantir and Wallace, 2003] (and perhaps also [Schmidt and Makalic, 2009c]);
- inference of MML neural nets [Makalic *et al.*, 2003];
- inference of MML decision trees (or classification trees) and decision graphs (or classification graphs) [Oliver and Wallace, 1991; Oliver *et al.*, 1992; Oliver and Wallace, 1992; Oliver, 1993; Uther and Veloso, 2000; Tan and Dowe, 2002; Tan and Dowe, 2003; Tan and Dowe, 2004], including (as per sec. 6.6) decision trees with support vector machines (SVMs) in their leaves [Tan and

Dowe, 2004] — with applications of MML decision trees and graphs in a variety of areas including (e.g.) protein folding [Dowe *et al.*, 1992; 1992a; 1993] and medical diagnosis [McKenzie *et al.*, 1993];

- MML clustering, mixture modelling and intrinsic classification via the *Snob* program [Wallace and Boulton, 1968; Wallace, 1984b; 1986; 1990b; 1990c; Wallace and Dowe, 1994b; 1996; 1997a; 1997b; 2000] was originally for the multinomial and Gaussian distributions, but this was extended to also include the Poisson and von Mises circular distributions [Wallace and Dowe, 1994b; 1996; 1997a; 1997b; 2000] — with applications in a variety of areas including (e.g.) spectral modelling [Papp *et al.*, 1993], protein folding [Zakis *et al.*, 1994], psychology and psychiatry [Kissane *et al.*, 1994; 1996; 1996a; Prior *et al.*, 1998]. Also of interest is determining whether our data appears to contain one line segment or a mixture of more than one line segment [Georgeff and Wallace, 1984a; 1984b; 1985] (and much later work on engineering bridge deterioration using a mixture of a Poisson distribution and a uniform distribution with total assignment [Maheswaran *et al.*, 2006]). (After the MML mixture modelling of multinomial, Gaussian, Poisson and von Mises circular distributions from 1994 [Wallace and Dowe, 1994b; 1996] came a slightly different paper doing only MML Gaussian mixture modelling [Oliver *et al.*, 1996] but emphasising the success of MML in empirical comparisons.) The MML single linear factor analysis from [Wallace and Freeman, 1992] was incorporate into [Edwards and Dowe, 1998] — although [Edwards and Dowe, 1998] did total (rather than partial) assignment and only did single rather than multiple [Wallace, 1995a; 1998b] factor analysis. This has also been extended to a variety of forms of sequential clustering [Edgoose and Allison, 1999; Molloy *et al.*, 2006], with an extension of [Edgoose and Allison, 1999] being (as in the next item) to MML *spatial* clustering. See also [Boulton and Wallace, 1973b; Dowe, 2008a, sec. 0.2.3, p. 531, col. 1 and sec. 0.2.4, p. 537, col. 2] for a discussion of MML hierarchical clustering. As well as the abovementioned multinomial, Gaussian, Poisson and von Mises circular distributions [Wallace and Dowe, 1994b; 1996; 1997a; 1997b; 2000], this work — without sequential and spatial clustering (following in the next item) has been extended to other distributions [Agusta and Dowe, 2002a, 2002b; 2003a; 2003b; Bouguila and Ziou, 2007];
- extensions of MML spatial clustering [Wallace, 1998a; Visser and Dowe, 2007] to tomographic [Visser *et al.*, 2009a] and climate [Visser *et al.*, 2009b] models. Variations on this work (and possibly other MML image analysis work [Torsello and Dowe, 2008a; 2008b]) should lend themselves both to training a robot to learn a model for and then recognise a particular class of object (such as a coloured shape, like a particular type of fruit) for robotic hand-eye co-ordination and also to analysing data in constructing a bionic eye;
- inference of systems of one or more probabilistic/stochastic (partial or) or-

dinary (difference or) differential equations (plus at least one noise term) from (presumably noisy) data (as per wishes from [Dowe, 2008a, sec. 0.2.7, p. 545]). Uses for this should include the likes of (e.g.) inferring parameter settings for ant colonies and other swarms so as to model them and/or suggesting settings giving better “intelligence” (as per sec. 7.3) and medical applications such as (e.g.) cardiac modelling or modelling stem cells;

- MML, particle physics and the analysis of the data in the search for the Higgs boson [Dowe, 2008a, sec. 0.2.7, p. 544, col. 2];
- etc.

Also of interest here might be

- the relationship between MML and the likelihood principle of statistics [Wallace, 2005, sec 5.8; Wallace and Dowe, 1999b, sec. 2.3.5], for which MML’s violation is “innocent enough — a misdemeanour rather than a crime” [Wallace, 2005, sec. 5.8, p. 254; Dowe, 2008a, sec. 0.2.4, p. 535, col. 2];
- the relationship between MML and Ed Jaynes’s notion of maximum entropy (or MaxEnt) priors [Jaynes, 2003; Wallace, 2005, secs. 1.15.5 and 2.1.11; Dowe, 2008a, sec. 0.2.4, p. 535, col. 1]. While MML and MaxEnt are different, while still on the topic of entropy and MML, it turns out that — within the MML mixture modelling literature — the term used to shorten the message length when going from (the inefficient coding scheme of) total assignment to (the efficient coding scheme of) partial assignment equates to the entropy of the posterior probability distribution of the class assignment probabilities [Visser *et al.*, 2009b; Wallace, 1998a; Visser and Dowe, 2007];
- etc.

7.7 *MML and other philosophical issues*

When time next permits, here are some of the many other philosophical issues to which MML pertains

- entropy is *not* time’s arrow [Wallace, 2005, chap. 8 (and p. vii); Dowe, 2008a, sec. 0.2.5; 2008b, p. 455], and note the ability of MML to detect thermal fluctuations and not over-fit them where some other statistical methods might be tempted to regard the standard noisy fluctuations as being something more [Wallace, 2005, chap. 8]. (One wonders whether the formation of these ideas might be evident in [Wallace, 1973a].) Recalling sec. 4.1 on inference (or explanation) and prediction, one interesting thing here is Wallace’s take on why it is that we wish to predict the future but (only) infer (or explain) the past [Wallace, 2005, chap. 8];

- being able to accord something the title of a “*miracle*” [Wallace, 2005, sec. 1.2, p. 7; Dowe, 2008a, sec. 0.2.7, p. 545, col. 1; 2008b, p. 455], the *fine tuning* argument in *intelligent design* [Dowe *et al.*, to appear (b)] (possibly see also sec. 7.3) and evidence that there is an intelligent supervisor/shepherd listening to our prayers and overseeing — and sometimes intervening in — our lives. Just as we can use MML to decide whether or not to accord something the title of a miracle, so, too, we can set about being objective and using MML to quantify the probability of certain coincidences and whether or not there could be an intelligent supervisor/shepherd intervening in our lives. Of course, such a shepherd might be able to make discrete minor changes effectively impossible to notice in one place which lead to substantial changes in other places. (I am not offering my opinion one way or another here, but rather merely raising how this issue might be addressed in an MML framework);
- Efficient Markets [Dowe and Korb, 1996; Dowe, 2008a, sec. 0.2.5; 2008b, p. 455] — due to the Halting problem, MML and Kolmogorov complexity essentially say (in short) that financial markets are very unlikely to be efficient and next to impossible to be proved efficient. Attempts to make this point more accessible by showing the effects of having a variety of trading approaches equal in all ways but one where one trader is better in terms of (e.g.) inference method, speed or memory, are given in [Collie *et al.*, 2005; 2005a];
- redundant Turing Machines (unlike those of sec. 2.4), for which pre- and post-processing can be used to effectively emulate a (Universal) Turing Machine by non-conventional means [Dowe, 2008a, sec. 0.2.7, p. 544];
- undecidability in (optimal) engineering tuning and design (ultimately due to the Halting problem);
- probabilities of conditionals and conditional probabilities [Dowe, 2008a, sec. 0.2.7, p. 546];
- information and MML re originality of an idea, degree of creativity of an act or design — or humour [Dowe, 2008a, sec. 0.2.7, p. 545] (the reader is welcome to inspect not unrelated ideas in [Solomonoff, 1995; Schmidhuber, 2007] in order to determine the originality of this idea). Puns typically entail finding commonality between at least two different subject matters. The finding of such commonality is crucial to the creation of the pun, and the recognising of such commonality is crucial to the understanding of the pun. A similar comment applies to the creation and solving of (cryptic) clues from a (cryptic) crossword. This said, in my experience, creating puns seems to be more difficult than understanding them — whereas solving cryptic crossword clues seems to be more difficult than creating them;

- mnemonics (or memory aids), whether or not this should be regarded as a philosophical issue. Certain mnemonics are compressions or contractions from which we can re-construct that “data” that we ultimately wish to recall. However, there is a seemingly slightly curious phenomenon here. People might recall (e.g.) the periodic table of elements or (e.g.) the base 10 decimal expansion of π by recalling a mnemonic sequence of words which tells a story. In the case of the periodic table, these words (can) start with the first one or so letters of the chemical elements in sequence. In the case of π , these words in sequence (can) have lengths corresponding to the relevant digit: so, the length of the i^{th} word is the i^{th} digit of π — e.g., “How I want a drink, alcoholic of course, after the heavy chapters involving quantum mechanics” (for 3.14159265358979). These little stories are fairly easily remembered — one might say that they are compressible, so one can re-construct them fairly easily, from where one can go on to re-construct what one was really trying to recall. However, the slight curiosity is that for all the easy compressible niceties of the mnemonic sequence, it is actually longer than the original. Perhaps the resolution is that whatever slight redundancies there are in the mnemonics serve as error corrections. So, perhaps such mnemonics are quite compressible in their own right so that they can easily be re-constructed but have sufficiently much redundancy to reduce errors. I think there is room for further discussion on this topic;
- fictionalism is an area of philosophy which (according to my understanding of it) is concerned about the sense in which we can talk about some fictional character (e.g., Elizabeth Bennett from “*Pride and Prejudice*”) as though they were real — and then go on to discuss how said character(s) might behave in some scenario not presented in the story in which said character(s) appear(s). This seems very much to be an MML matter. We form a model of the character(s) based on what we know about the(se) character(s). We have a model of how various types of real-world character behave in certain scenarios, and we go from there. In similar vein, MML has much to say about the philosophical notion of *counterfactuals* and *possible worlds*, although here there is a minor issue of (recall sec. 4.1) whether we are interested in inference as to how things would most probably be in the nearest possible world or instead a weighted prediction of how things might be — obtained by doing a weighted combination of predictions over a variety of possible worlds;
- etc.

Having made the above list, I now mention some issues to which I would like to be able to apply MML.

- virtue — Confucius (the Chinese philosopher) [Confucius, 1938] and (about 500 years later) Jesus (from whom we have the Christian religion) are well-known for their comments on consideration, and Confucius further for his

other comments on virtue. In game theory (e.g., prisoners' dilemma), perhaps virtue is being/doing as all would need to do to give the best all-round solution (in similar vein to the merits of co-operation described in, e.g., [Wallace, 1998d]), perhaps it is doing the optimum by the others on the presumption that the others will all act out of self-interest. Perhaps MML can offer some insight here; and

- the Peter principle — I hear of and sometimes see far too many shocking appointments of the undeserving, of a candidate who “peaked in the interview”. Perhaps in terms of some sort of experimental design or more properly knowing how to analyse the available data on candidates, it would be nice to put MML to use here.

8 ACKNOWLEDGEMENTS

I first thank Prasanta Bandyopadhyay, Malcolm Forster and John Woods for permitting me to contribute this piece. I thank the anonymous referee — whose helpful feedback clearly indicated that said referee had read my submission closely. I especially thank Prasanta for his regular patient editorial monitoring of my progress (or lack thereof), at least comparable in magnitude to that of an additional referee.

I next thank my mother and my family [Comley and Dowe, 2005, p. 287, sec. 11.6]. I now quote Confucius [1938, Book I, saying 15]: “Tzu-kung said, ‘Poor without cadging, rich without swagger.’ What of that? The Master said, Not bad. But better still, ‘Poor, yet delighting in the Way; rich, yet a student of ritual.’ ”

I thank Chris Wallace [Gupta *et al.*, 2004; Dowe, 2008a, footnote 218] (for a variety of matters — such as, e.g., the letter referred to in [Dowe, 2008a, footnote 218] and all the training he gave me over the years), for whose Christopher Stewart WALLACE (1933-2004) memorial special issue of the *Computer Journal* [Dowe, 2008a; Brennan, 2008; Solomonoff, 2008; Jorgensen and McLachlan, 2008; Brent, 2008; Colon-Bonet and Winterrowd, 2008; Castro *et al.*, 2008] it was an honour to be guest editor. (For other mention and examples of the range and importance of his work, see, e.g., [Parry, 2005] and [Clark and Wallace, 1970].) I talk there of his deserving at least one Turing Award [Dowe, 2008a, sec. 0.2.2, p. 526 and sec. 0.2.4, p. 533, col. 2] and of how his work on entropy not being time's arrow might have stood him in contention for the Nobel prize in Physics if it could be experimentally tested [Dowe, 2008a, sec. 0.2.5, footnote 144]. (And, possibly, as per [Dowe, 2008a, sec. 0.2.7, p. 544, col. 2] and sec. 7.6, MML will have a role to play in the discovery of the Higgs boson.) I forgot to point out that, if he'd lived a few decades longer, the increasing inclusion of his work in econometrics (see, e.g., [Fitzgibbon *et al.*, 2004; Dowe, 2008a, sec. 0.2.3, footnote 88] and sec. 6.5 — and possibly also these papers on segmentation and cut-points [Oliver *et al.*, 1998; Viswanathan *et al.*, 1999; Fitzgibbon *et al.*, 2002b] — for teasers) might have one day earned him the Nobel prize in Economics. (Meanwhile, it would be good to re-visit both ARCH [Autoregressive Conditional Heteroskedasticity] and

GARCH [Generalised ARCH] using MML.) For a list of his publications (MML and other), see either the reference list to [Dowe, 2008a] (which also lists the theses he supervised) or www.csse.monash.edu.au/~dld/CSWallacePublications.

And, finally, I thank Fran Boyce [Dowe, 2008a, footnote 217; Obituaries, 2009], one of the nicest, gentlest, most thoughtful and most gracious people one could meet or ever even hope to meet.

BIBLIOGRAPHY

- [Obituaries, 2009] Obituaries: Drive to learn despite struggle with lupus. *The Herald Sun*, page 79, Thu. 6 Aug., 2009. Melbourne, Australia; Author: M. Sonogan.
- [Agusta, 2005] Y. Agusta. *Minimum Message Length Mixture Modelling for Uncorrelated and Correlated Continuous Data Applied to Mutual Funds Classification*. PhD thesis, School of Computer Science and Software Engineering, Clayton School of I.T., Monash University, Clayton, Australia, 2005.
- [Agusta and Dowe, 2002b] Y. Agusta and D. L. Dowe. Clustering of Gaussian and t distributions using minimum message length. In *Proc. International Conference on Knowledge Based Computer Systems (KBCS 2002)*, pages 289–299. Vikas Publishing House, 2002. <http://www.ncst.ernet.in/kbcs2002>.
- [Agusta and Dowe, 2002a] Y. Agusta and D. L. Dowe. MML clustering of continuous-valued data using Gaussian and t distributions. In B. McKay and J. Slaney, editors, *Lecture Notes in Artificial Intelligence (LNAI), Proc. Australian Joint Conference on Artificial Intelligence*, volume 2557, pages 143–154, Berlin, Germany, December 2002. Springer-Verlag.
- [Agusta and Dowe, 2003b] Y. Agusta and D. L. Dowe. Unsupervised learning of correlated multivariate Gaussian mixture models using MML. In *Lecture Notes in Artificial Intelligence (LNAI) 2903 (Springer), Proc. 16th Australian Joint Conf. on Artificial Intelligence*, pages 477–489, 2003.
- [Agusta and Dowe, 2003a] Y. Agusta and D. L. Dowe. Unsupervised learning of Gamma mixture models using minimum message length. In M. H. Hamza, editor, *Proceedings of the 3rd IASTED conference on Artificial Intelligence and Applications*, pages 457–462, Benalmadena, Spain, September 2003. ACTA Press.
- [Allison and Wallace, 1993] L. Allison and C. S. Wallace. The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimisation of multiple alignments. Technical report CS 93/188, Dept Computer Science, Monash University, Melbourne, Australia, 1993.
- [Allison and Wallace, 1994a] L. Allison and C. S. Wallace. An information measure for the string to string correction problem with applications. *17th Australian Comp. Sci. Conf.*, pages 659–668, January 1994. Australian Comp. Sci. Comm. Vol 16 No 1(C) 1994.
- [Allison and Wallace, 1994b] L. Allison and C. S. Wallace. The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments. *J. Mol. Evol.*, 39(4):418–430, October 1994. an early version: TR 93/188, Dept. Computer Science, Monash University, July 1993.
- [Allison et al., 1990] L. Allison, C. S. Wallace, and C. N. Yee. Induction inference over macromolecules. Technical Report 90/148, Monash University, Clayton, Victoria, Australia, 3168, 1990.
- [Allison et al., 1990b] L Allison, C S Wallace, and C N Yee. Inductive inference over macromolecules. In *Working Notes AAAI Spring Symposium Series*, pages 50–54. Stanford Uni., Calif., U.S.A., 1990.
- [Allison et al., 1990a] L. Allison, C. S. Wallace, and C. N. Yee. When is a string like a string? In *International Symposium on Artificial Intelligence and Mathematics*, January 1990.
- [Allison et al., 1991] L Allison, C S Wallace, and C N Yee. Minimum message length encoding, evolutionary trees and multiple-alignment. Technical report CS 91/155, Dept Computer Science, Monash University, Melbourne, Australia, 1991.

- [Allison *et al.*, 1992b] L. Allison, C. S. Wallace, and C. N. Yee. Finite-state models in the alignment of macro-molecules. *J. Mol. Evol.*, 35(1):77–89, July 1992. extended abstract titled: Inductive inference over macro-molecules in joint sessions at AAAI Symposium, Stanford, Mar 1990 on (i) Artificial Intelligence and Molecular Biology, pp5-9 & (ii) Theory and Application of Minimal-Length Encoding, pp50-54.
- [Allison *et al.*, 1992a] L. Allison, C. S. Wallace, and C. N. Yee. Minimum message length encoding, evolutionary trees and multiple alignment. *25th Hawaii Int. Conf. on Sys. Sci.*, 1:663–674, January 1992. Another version is given in TR 91/155, Dept. Computer Science, Monash University, Clayton, Vic, Australia, 1991.
- [Barron and Cover, 1991] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.
- [Baxter and Oliver, 1995] R. A. Baxter and J. J. Oliver. MDL and MML: Similarities and differences. Technical report TR 94/207, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1995.
- [Bernardo and Smith, 1994] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, New York, 1994.
- [Bouguila and Ziou, 2007] N. Bouguila and D. Ziou. High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1716–1731, October 2007.
- [Boulton, 1970] D. M. Boulton. Numerical classification based on an information measure. Master's thesis, M.Sc. thesis, Basser Computing Dept., University of Sydney, Sydney, Australia, 1970.
- [Boulton, 1975] D. M. Boulton. *The Information Measure Criterion for Intrinsic Classification*. PhD thesis, Dept. Computer Science, Monash University, Clayton, Australia, August 1975.
- [Boulton and Wallace, 1969] D. M. Boulton and C. S. Wallace. The information content of a multistate distribution. *J. Theor. Biol.*, 23:269–278, 1969.
- [Boulton and Wallace, 1970] D. M. Boulton and C. S. Wallace. A program for numerical classification. *Computer Journal*, 13(1):63–69, February 1970.
- [Boulton and Wallace, 1973c] D. M. Boulton and C. S. Wallace. A comparison between information measure classification. In *Proc. of the Australian & New Zealand Association for the Advancement of Science (ANZAAS) Congress*, August 1973. abstract.
- [Boulton and Wallace, 1973b] D. M. Boulton and C. S. Wallace. An information measure for hierarchic classification. *Computer Journal*, 16(3):254–261, 1973.
- [Boulton and Wallace, 1973a] D. M. Boulton and C. S. Wallace. Occupancy of a rectangular array. *Computer Journal*, 16(1):57–63, 1973.
- [Boulton and Wallace, 1975] D. M. Boulton and C. S. Wallace. An information measure for single link classification. *Computer Journal*, 18(3):236–238, 1975.
- [Brennan, 2008] M. H. Brennan. Data processing in the early cosmic ray experiments in Sydney. *Computer Journal*, 51(5):561–565, September 2008.
- [Brennan *et al.*, 1958] M. H. Brennan, D. D. Millar, and C. S. Wallace. Air showers of size greater than 10^5 particles - (1) core location and shower size determination. *Nature*, 182:905–911, Oct. 4 1958.
- [Brent, 2008] R. P. Brent. Some comments on C. S. Wallace's random number generators. *Computer Journal*, 51(5):579–584, September 2008.
- [Castro *et al.*, 2008] M. D. Castro, R. D. Pose, and C. Kopp. Password-capabilities and the Walnut kernel. *Computer Journal*, 51(5):595–607, September 2008.
- [Chaitin, 1966] G. J. Chaitin. On the length of programs for computing finite sequences. *Journal of the Association for Computing Machinery*, 13:547–569, 1966.
- [Chaitin, 2005] G. J. Chaitin. *Meta Math! The Quest for Omega*. Pantheon, 2005. ISBN 0-375-42313-3 (978-0-375-42313-0).
- [Clark and Wallace, 1970] G. M. Clark and C. S. Wallace. Analysis of nasal support. *Archives of Otolaryngology*, 92:118–129, August 1970.
- [Clarke, 1999] B. Clarke. Discussion of the papers by Rissanen, and by Wallace and Dowe. *Computer J.*, 42(4):338–339, 1999.
- [Collie *et al.*, 2005] M. J. Collie, D. L. Dowe, and L. J. Fitzgibbon. Stock market simulation and inference technique. In *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, Rio de Janeiro, Brazil, Nov 2005.

- [Collie *et al.*, 2005a] M. J. Collie, D. L. Dowe, and L. J. Fitzgibbon. Trading rule search with autoregressive inference agents. Technical report CS 2005/174, School of Computer Science and Software Engineering, Monash University, Melbourne, Australia, 2005.
- [Colon-Bonet and Winterrowd, 2008] G. Colon-Bonet and P. Winterrowd. Multiplier evolution — a family of multiplier VLSI implementations. *Computer Journal*, 51(5):585–594, September 2008.
- [Comley and Dowe, 2003] Joshua W. Comley and David L. Dowe. General Bayesian networks and asymmetric languages. In *Proc. Hawaii International Conference on Statistics and Related Fields*, 5–8 June 2003.
- [Comley and Dowe, 2005] Joshua W. Comley and David L. Dowe. Minimum message length and generalized Bayesian nets with asymmetric languages. In P. Grünwald, M. A. Pitt, and I. J. Myung, editors, *Advances in Minimum Description Length: Theory and Applications (MDL Handbook)*, pages 265–294. M.I.T. Press, April 2005. Chapter 11, ISBN 0-262-07262-9. Final camera-ready copy submitted in October 2003. [Originally submitted with title: “Minimum Message Length, MDL and Generalised Bayesian Networks with Asymmetric Languages”].
- [Confucius, 1938] Confucius. *The Analects of Confucius (Lun Yü)*. Vintage Books, 1989. (Published earlier with Macmillan in 1938.) Translated by Arthur Waley. Online at <http://myweb.cableone.net/subru/Confucianism.html>.
- [Dai *et al.*, 1996a] H Dai, K B Korb, and C S Wallace. The discovery of causal models with small samples. In *1996 Australian New Zealand Conference on Intelligent Information Systems Proceedings ANZIS96*, pages 27–30. IEEE, Piscataway, NJ, USA, 1996.
- [Dai *et al.*, 1996b] H Dai, K B Korb, and C. S. Wallace. A study of causal discovery with weak links and small samples. Technical report CS 96/298, Dept Computer Science, Monash University, Melbourne, Australia, 1996.
- [Dai *et al.*, 1997b] H Dai, K B Korb, C S Wallace, and X Wu. A study of causal discovery with weak links and small samples. Technical report SD TR97-5, Dept Computer Science, Monash University, Melbourne, Australia, 1997.
- [Dai *et al.*, 1997a] Honghua Dai, Kevin B. Korb, C. S. Wallace, and Xindong Wu. A study of causal discovery with weak links and small samples. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI 97*, pages 1304–1309, 1997.
- [Dawid, 1999] A. P. Dawid. Discussion of the papers by Rissanen and by Wallace and Dowe. *Computer J.*, 42(4):323–326, 1999.
- [Deakin, 2001] M. A. B. Deakin. The characterisation of scoring functions. *J. Australian Mathematical Society*, 71:135–147, 2001.
- [Dowe, 2007] D. L. Dowe. Discussion following “Hedging predictions in machine learning, A. Gammernan and V. Vovk”. *Computer Journal*, 2(50):167–168, 2007.
- [Dowe, 2008a] D. L. Dowe. Foreword re C. S. Wallace. *Computer Journal*, Christopher Stewart WALLACE (1933–2004) memorial special issue, 51(5):523–560, September 2008.
- [Dowe, 2008b] D. L. Dowe. Minimum Message Length and statistically consistent invariant (objective?) Bayesian probabilistic inference — from (medical) “evidence”. *Social Epistemology*, 22(4):433–460, October–December 2008.
- [Dowe *et al.*, 1995] D. L. Dowe, L. Allison, T. I. Dix, L. Hunter, C. S. Wallace, and T. Edgoose. Circular clustering by minimum message length of protein dihedral angles. Technical report CS 95/237, Dept Computer Science, Monash University, Melbourne, Australia, 1995.
- [Dowe *et al.*, 1996] D. L. Dowe, L. Allison, T. I. Dix, L. Hunter, C. S. Wallace, and T. Edgoose. Circular clustering of protein dihedral angles by minimum message length. In *Pacific Symposium on Biocomputing '96*, pages 242–255. World Scientific, January 1996.
- [Dowe *et al.*, 1998] D. L. Dowe, R. A. Baxter, J. J. Oliver, and C. S. Wallace. Point estimation using the Kullback-Leibler loss function and MML. In X. Wu, Ramamohanarao Kotagiri, and K. Korb, editors, *Proceedings of the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, volume 1394 of *LNAI*, pages 87–95, Berlin, April 15–17 1998. Springer.
- [Dowe *et al.*, 1996a] D. L. Dowe, G. E. Farr, A. J. Hurst, and K. L. Lentin. Information-theoretic football tipping. *3rd Conf. on Maths and Computers in Sport*, pages 233–241, 1996. See also Technical Report TR 96/297, Dept. Computer Science, Monash University, Australia 3168, Dec 1996.
- [Dowe *et al.*, 1996b] D. L. Dowe, G. E. Farr, A. J. Hurst, and K. L. Lentin. Information-theoretic football tipping. Technical report TR 96/297, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1996.

- [Dowe *et al.*, to appear (a)] D. L. Dowe, C. Ferri, J. Hernandez-Orallo, and M. Jose Ramirez-Quintana. An MML Scheme for ILP. Submitted.
- [Dowe *et al.*, 2007] D. L. Dowe, S. Gardner, and G. R. Oppy. Bayes not bust! Why simplicity is no problem for Bayesians. *British Journal for the Philosophy of Science*, 58(4):709–754, December 2007.
- [Dowe *et al.*, to appear (b)] D. L. Dowe, S. Gardner, and G. R. Oppy. MML and the fine tuning argument. To appear, 2011.
- [Dowe and Hajek, 1997] D. L. Dowe and A. R. Hajek. A computational extension to the Turing test. Technical Report 97/322, Dept. Computer Science, Monash University, Australia 3168, October 1997.
- [Dowe and Hajek, 1998] D. L. Dowe and A. R. Hajek. A non-behavioural, computational extension to the Turing test. In *Proceedings of the International Conference on Computational Intelligence & Multimedia Applications (ICCIMA '98)*, pages 101–106, Gippsland, Australia, February 1998.
- [Dowe *et al.*, 1996c] D. L. Dowe, A.J. Hurst, K.L. Lentin, G. Farr, and J.J. Oliver. Probabilistic and Gaussian football prediction competitions - Monash. *Artificial Intelligence in Australia Research Report*, June 1996.
- [Dowe *et al.*, 1998a] D. L. Dowe, M Jorgensen, G McLachlan, and C S Wallace. Information-theoretic estimation. In W Robb, editor, *Proceedings of the Fourteenth Biennial Australian Statistical Conference (ASC-14)*, page 125, Queensland, Australia, July 1998.
- [Dowe and Korb, 1996] D. L. Dowe and K. B. Korb. Conceptual difficulties with the efficient market hypothesis: Towards a naturalized economics. In *Proc. Information, Statistics and Induction in Science (ISIS)*, pages 212–223, 1996. See also Technical Report TR 94/215, Dept. Computer Science, Monash University, Australia 3168, 1994.
- [Dowe and Krusel, 1993] D. L. Dowe and N. Krusel. A decision tree model of bushfire activity. Technical report TR 93/190, Dept. of Computer Science, Monash University, Clayton, Vic. 3800, Australia, September 1993.
- [Dowe and Lentin, 1995] D. L. Dowe and K. L. Lentin. Information-theoretic footy-tipping competition — Monash. *Computer Science Association Newsletter (Australia)*, pages 55–57, December 1995.
- [Dowe *et al.*, 1996d] D. L. Dowe, K.L. Lentin, J.J. Oliver, and A.J. Hurst. An information-theoretic and a Gaussian footy-tipping competition. *FCIT Faculty Newsletter, Monash University, Australia*, pages 2–6, June 1996.
- [Dowe *et al.*, 1992] D. L. Dowe, J. J. Oliver, L. Allison, T. I. Dix, and C. S. Wallace. Learning rules for protein secondary structure prediction. In C. McDonald, J. Rohl, and R. Owens, editors, *Proc. 1992 Department Research Conference*. Dept. Computer Science, University of Western Australia, July 1992.
- [Dowe *et al.*, 1992a] D. L. Dowe, J. J. Oliver, L. Allison, C. S. Wallace, and T. I. Dix. A decision graph explanation of protein secondary structure prediction. Technical report CS 92/163, Dept Computer Science, Monash University, Melbourne, Australia, 1992.
- [Dowe *et al.*, 1995a] D. L. Dowe, J. J. Oliver, R. A. Baxter, and C. S. Wallace. Bayesian estimation of the von Mises concentration parameter. In *Proc. 15th Int. Workshop on Maximum Entropy and Bayesian Methods, Santa Fe*, July 1995.
- [Dowe *et al.*, 1995b] D. L. Dowe, J. J. Oliver, R. A. Baxter, and C. S. Wallace. Bayesian estimation of the von Mises concentration parameter. Technical report CS 95/236, Dept Computer Science, Monash University, Melbourne, Australia, 1995.
- [Dowe *et al.*, 1993] D. L. Dowe, J. J. Oliver, T. I. Dix, L. Allison, and C. S. Wallace. A decision graph explanation of protein secondary structure prediction. *26th Hawaii Int. Conf. Sys. Sci.*, 1:669–678, January 1993.
- [Dowe *et al.*, 1996e] D. L. Dowe, J. J. Oliver, and C. S. Wallace. MML estimation of the parameters of the spherical Fisher distribution. In *Algorithmic Learning Theory, 7th International Workshop, ALT '96, Sydney, Australia, October 1996, Proceedings*, volume 1160 of *Lecture Notes in Artificial Intelligence*, pages 213–227. Springer, October 1996.
- [Dowe *et al.*, 1996f] D. L. Dowe, J. J. Oliver, and C. S. Wallace. MML estimation of the parameters of the spherical Fisher distribution. Technical report CS 96/272, Dept Computer Science, Monash University, Melbourne, Australia, 1996.
- [Dowe and Oppy, 2001] D. L. Dowe and G. R. Oppy. Universal Bayesian inference? *Behavioral and Brain Sciences (BBS)*, 24(4):662–663, Aug 2001.

- [Dowe and Wallace, 1996] D. L. Dowe and C. S. Wallace. Resolving the Neyman-Scott problem by minimum message length (abstract). In *Proc. Sydney Int. Stat. Congress*, pages 197–198, 1996.
- [Dowe and Wallace, 1997a] D. L. Dowe and C. S. Wallace. Resolving the Neyman-Scott problem by Minimum Message Length. In *Proc. Computing Science and Statistics — 28th Symposium on the interface*, volume 28, pages 614–618, 1997.
- [Dowe and Wallace, 1997b] D. L. Dowe and C. S. Wallace. Resolving the Neyman-Scott problem by Minimum Message Length. Technical report TR no. 97/307, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, February 1997. Also in *Proc. Sydney Int. Stat. Congr. (SISC-96)*, Sydney, pages 197–198; and in *IMS Bulletin* (1996), 25 (4), pp410–411.
- [Dowe and Wallace, 1998] D. L. Dowe and C. S. Wallace. Kolmogorov complexity, minimum message length and inverse learning. In W. Robb, editor, *Proceedings of the Fourteenth Biennial Australian Statistical Conference (ASC-14)*, page 144, Queensland, Australia, July 1998.
- [Dowe and Zaidi, 2010] D. L. Dowe and N. A. Zaidi. Database normalization as a by-product of minimum message length inference. In *Proc. 23rd Australian Joint Conference on Artificial Intelligence (AI'2010)*, Adelaide, Australia, 7–10 December 2010, pp. 82–91. Springer Lecture Notes in Artificial Intelligence (LNAI), vol. 6464, Springer, 2010.
- [Edgoose and Allison, 1999] T. Edgoose and L. Allison. MML Markov classification of sequential data. *Stats. and Comp.*, 9(4):269–278, September 1999.
- [Edgoose et al., 1998] T. Edgoose, L. Allison, and D. L. Dowe. An MML classification of protein structure that knows about angles and sequence. In *Pacific Symposium on Biocomputing '98*, pages 585–596. World Scientific, January 1998.
- [Edwards and Dowe, 1998] R. T. Edwards and D. L. Dowe. Single factor analysis in MML mixture modelling. In Xindong Wu, Ramamohanarao Kotagiri, and Kevin B. Korb, editors, *Proceedings of the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, volume 1394 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 96–109, Berlin, April 15–17 1998. Springer.
- [Farr and Wallace, 2002] G. E. Farr and C. S. Wallace. The complexity of strict minimum message length inference. *Computer Journal*, 45(3):285–292, 2002.
- [Fitzgibbon, 2004] L. J. Fitzgibbon. *Message from Monte Carlo: A Framework for Minimum Message Length Inference using Markov Chain Monte Carlo Methods*. PhD thesis, School of Computer Science and Software Engineering, Clayton School of I.T., Monash University, Clayton, Australia, 2004.
- [Fitzgibbon et al., 2002b] L. J. Fitzgibbon, D. L. Dowe, and Lloyd Allison. Change-point estimation using new minimum message length approximations. In *Lecture Notes in Artificial Intelligence (LNAI) 2417, 7th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pages 244–254. Springer, 2002.
- [Fitzgibbon et al., 2002a] L. J. Fitzgibbon, D. L. Dowe, and Lloyd Allison. Univariate polynomial inference by Monte Carlo message length approximation. In *19th International Conference on Machine Learning (ICML)*, pages 147–154, 2002.
- [Fitzgibbon et al., 2004] L. J. Fitzgibbon, D. L. Dowe, and F. Vahid. Minimum message length autoregressive model order selection. In *Proc. Int. Conf. on Intelligent Sensors and Information Processing*, pages 439–444, Chennai, India, January 2004.
- [Gammerman and Vovk, 2007a] Alex Gammerman and Vladimir Vovk. Hedging predictions in machine learning. *Computer Journal*, 2(50):151–163, 2007.
- [Gammerman and Vovk, 2007b] Alex Gammerman and Vladimir Vovk. Rejoinder: Hedging predictions in machine learning. *Computer Journal*, 2(50):173–177, 2007.
- [Georgeff and Wallace, 1984a] M. P. Georgeff and C. S. Wallace. A general selection criterion for inductive inference. In T. O'Shea, editor, *Advances in Artificial Intelligence: Proc. Sixth European Conference on Artificial Intelligence (ECAI-84)*, pages 473–482, Amsterdam, September 1984. Elsevier Science Publishers B.V. (North Holland).
- [Georgeff and Wallace, 1984b] M. P. Georgeff and C. S. Wallace. A general selection criterion for inductive inference. Technical report TR 44, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, June 1984.
- [Georgeff and Wallace, 1985] M. P. Georgeff and C. S. Wallace. Minimum information estimation of structure. In T. O'Shea, editor, *Advances in Artificial Intelligence*, pages 219–228. Elsevier, 1985.

- [Gödel, 1931] K. Gödel. On formally undecidable propositions of *Principia mathematica* and related systems I. *Monatshefte für Mathematik und Physik*, 38:173–198, 1931.
- [Good, 1952] I. J. Good. Rational decisions. *J. Roy. Statist. Soc. B*, B 14:107–114, 1952.
- [Grünwald, 2007] P. D. Grünwald. *The Minimum Description Length principle (Adaptive Computation and Machine Learning)*. M.I.T. Press, 2007.
- [Grünwald et al., 1998] P. D. Grünwald, P. Kontkanen, P. Myllymaki, T. Silander, and H. Tirri. Minimum encoding approaches for predictive modeling. In *Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence (UAI'98)*, pages 183–192, 1998.
- [Grünwald and Langford, 2004] Peter D. Grünwald and John Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. In *COLT*, pages 331–347, 2004.
- [Grünwald and Langford, 2007] Peter D. Grünwald and John Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66:119–149(31), March 2007.
- [Gupta et al., 2004] G. K. Gupta, I. Zukerman, and D. W. Albrecht. Obituaries: Australia's inspiring leader in the computing revolution — Christopher Wallace computer scientist. *The Age*, page 9, 2004. Melbourne, Australia; near and probably on Fri. 1 Oct. 2004.
- [Hernández-Orallo, 2000] José Hernández-Orallo. Beyond the Turing test. *Journal of Logic, Language and Information*, 9(4):447–466, 2000.
- [Hernández-Orallo and Dowe, 2010] José Hernández-Orallo and D. L. Dowe. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence Journal*, 174(18):1508–1539, 2010.
- [Hernández-Orallo and Minaya-Collado, 1998] José Hernández-Orallo and N. Minaya-Collado. A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In *Proceedings of the International Symposium of Engineering of Intelligent Systems, ICSC Press*, pages 146–163, 1998.
- [Hodges, 1983] Andrew Hodges. *Alan Turing: The Enigma*. Simon and Schuster, 1983.
- [Hope and Korb, 2002] L. R. Hope and K. Korb. Bayesian information reward. In R. McKay and J. Slaney, editors, *Proc. 15th Australian Joint Conference on Artificial Intelligence — Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, Germany, ISSN: 0302-9743, Vol. 2557*, number 2557 in Lecture Notes in Artificial Intelligence (LNAI), pages 272–283. Springer Verlag, 2002.
- [Huber, 2008] F. Huber. Milne's argument for the log-ratio measure. *Philosophy of Science*, pages 413–420, 2008.
- [Jaynes, 2003] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.
- [Jebara, 2003] Tony Jebara. *Machine Learning: Discriminative and Generative*. Kluwer Academic Publishers, Norwell, MA, U.S.A., 2003.
- [Jeffreys, 1946] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. of the Royal Soc. of London A*, 186:453–454, 1946.
- [Jorgensen and Gentleman, 1998] M. A. Jorgensen and R. Gentleman. Data mining. *Chance*, 11:34–39, 42, 1998.
- [Jorgensen and McLachlan, 2008] M. A. Jorgensen and G. J. McLachlan. Wallace's approach to unsupervised learning: the Snob program. *Computer Journal*, 51(5):571–578, September 2008.
- [Kearns et al., 1997] M. Kearns, Y. Mansour, A. Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27:7–50, 1997.
- [Kissane et al., 1994] D. W. Kissane, S. Bloch, W. I. Burns, J. D. Patrick, C. S. Wallace, and D. P. McKenzie. Perceptions of family functioning and cancer. *Psycho-oncology*, 3:259–269, 1994.
- [Kissane et al., 1996] D. W. Kissane, S. Bloch, D. L. Dowe, R. D. Snyder, P. Onghena, D. P. McKenzie, and C. S. Wallace. The Melbourne family grief study, I: Perceptions of family functioning in bereavement. *American Journal of Psychiatry*, 153:650–658, May 1996.
- [Kissane et al., 1996a] D. W. Kissane, S. Bloch, P. Onghena, D. P. McKenzie, R. D. Snyder, and D. L. Dowe. The Melbourne family grief study, II: Psychosocial morbidity and grief in bereaved families. *American Journal of Psychiatry*, 153:659–666, May 1996.
- [Kolmogorov, 1965] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:4–7, 1965.

- [Kontoyiannis, 2008] I. Kontoyiannis. Review of “Information and Complexity in Statistical Modeling by Jorma Rissanen”. *American Mathematical Monthly — Reviews*, 115:956–960, 2008.
- [Korb and Wallace, 1997] K B Korb and C. S. Wallace. In search of the philosopher’s stone: Remarks on Humphreys and Freedman’s critique of causal discovery. Technical report CS 97/315, Dept Computer Science, Monash University, Melbourne, Australia, 1997.
- [Korb and Wallace, 1999] K. B. Korb and C. S. Wallace. In search of the philosopher’s stone: Remarks on Humphreys and Freedman’s critique of causal discovery. *British Jnl. for the Philosophy of Science*, pages 543–553, 1999. TR 97/315, Mar 1997, Dept. Computer Science, Monash University, Australia 3168.
- [Kornienko *et al.*, 2005a] L. Kornienko, D. W. Albrecht, and D. L. Dowe. A preliminary MML linear classifier using principal components for multiple classes. In *Proc. 18th Australian Joint Conference on Artificial Intelligence (AI’2005)*, volume 3809 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 922–926, Sydney, Australia, Dec 2005. Springer.
- [Kornienko *et al.*, 2005b] L. Kornienko, D. W. Albrecht, and D. L. Dowe. A preliminary MML linear classifier using principal components for multiple classes. Technical report CS 2005/179, School of Computer Sci. & Softw. Eng., Monash Univ., Melb., Australia, 2005.
- [Kornienko *et al.*, 2002] Lara Kornienko, David L. Dowe, and David W. Albrecht. Message length formulation of support vector machines for binary classification — A preliminary scheme. In *Lecture Notes in Artificial Intelligence (LNAI), Proc. 15th Australian Joint Conf. on Artificial Intelligence*, volume 2557, pages 119–130. Springer-Verlag, 2002.
- [Kraft, 1949] L. G. Kraft, 1949. Master’s thesis, Dept. of Elec. Eng., M.I.T., U.S.A.
- [Lancaster, 2002] A. Lancaster. Orthogonal parameters and panel data. *Review of Economic Studies*, 69:647–666, 2002.
- [Legg and Hutter, 2007] S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, November 2007.
- [Lewis, 1976] David K. Lewis. Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, 85(3):297–315, July 1976.
- [Li and Vitányi, 1997] Ming Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its applications*. Springer Verlag, New York, 1997.
- [Long and Servedio, 2006] Phil Long and Rocco Servedio. Discriminative learning can succeed where generative learning fails. In *The 19th Annual Conference on Learning Theory*, Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A., 2006.
- [Maheswaran *et al.*, 2006] T. Maheswaran, J. G. Sanjayan, D. L. Dowe, and P. J. Tan. MML mixture models of heterogeneous Poisson processes with uniform outliers for bridge deterioration. In *Lecture Notes in Artificial Intelligence (LNAI) (Springer), Proc. 19th Australian Joint Conf. on Artificial Intelligence*, pages 322 – 331, Hobart, Australia, Dec. 2006.
- [Makalic *et al.*, 2003] E. Makalic, L. Allison, and D. L. Dowe. MML inference of single-layer neural networks. In *Proc. of the 3rd IASTED Int. Conf. on Artificial Intelligence and Applications*, pages 636–642, September 2003. See also Technical Report TR 2003/142, CSSE, Monash University, Australia Oct. 2003.
- [Martin-Löf, 1966] P. Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.
- [McKenzie *et al.*, 1993] D. P. McKenzie, P. D. McGorry, C. S. Wallace, L. H. Low, D. L. Copolov, and B. S. Singh. Constructing a minimal diagnostic decision tree. *Methods in Information in Medicine*, 32:161–166, 1993.
- [Milne, 1996] P. Milne. $\log[Pr(H|E \cap B)/Pr(H|B)]$ is the one true measure of confirmation. *Philosophy of Science*, 63:21–26, 1996.
- [Molloy *et al.*, 2006] S. B. Molloy, D. W. Albrecht, D. L. Dowe, and K. M. Ting. Model-Based Clustering of Sequential Data. In *Proceedings of the 5th Annual Hawaii International Conference on Statistics, Mathematics and Related Fields*, January 2006.
- [Murphy and Pazzani, 1994] P. Murphy and M. Pazzani. Exploring the decision forest: An empirical investigation of Occam’s razor in decision tree induction. *Journal of Artificial Intelligence*, 1:257–275, 1994.
- [Musgrave and Dowe, 2010] S. Musgrave and D. L. Dowe. Kinship, optimality and typology, *Behavioral and Brain Sciences (BBS)*, 33(5), 2010.
- [Needham and Dowe, 2001] S. L. Needham and D. L. Dowe. Message length as an effective Ockham’s razor in decision tree induction. In *Proc. 8th Int. Workshop on Artif. Intelligence and Statistics (AI+STATS 2001)*, pages 253–260, Jan. 2001.

- [Neil *et al.*, 1999a] J. R. Neil, C. S. Wallace, and K. B. Korb. Learning Bayesian networks with restricted causal interactions. In Kathryn B. Laskey and Henri Prade, editors, *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 486–493, S.F., Cal., July 30–August 1 1999. Morgan Kaufmann Publishers.
- [Neil *et al.*, 1999b] Julian R. Neil, C. S. Wallace, and K. B. Korb. Bayesian networks with non-interacting causes. Technical Report 1999/28, School of Computer Science & Software Engineering, Monash University, Australia 3168, September 1999.
- [Neyman and Scott, 1948] J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrika*, 16:1–32, 1948.
- [Oliver, 1993] J. J. Oliver. Decision graphs — an extension of decision trees. In *Proceedings of the Fourth International Workshop on Artificial Intelligence and Statistics*, pages 343–350, 1993. Extended version available as TR 173, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia.
- [Oliver and Baxter, 1994] J. J. Oliver and R. A. Baxter. MML and Bayesianism: similarities and differences. Technical report TR 94/206, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1994.
- [Oliver *et al.*, 1998] J. J. Oliver, R. A. Baxter, and C. S. Wallace. Minimum message length segmentation. In Xindong Wu, Ramamohanarao Kotagiri, and Kevin B. Korb, editors, *Proceedings of the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, volume 1394 of *LNAI*, pages 222–233, Berlin, April 15–17 1998. Springer.
- [Oliver *et al.*, 1996] J. J. Oliver, Rohan A. Baxter, and Chris S. Wallace. Unsupervised learning using MML. In *Proc. 13th International Conference on Machine Learning*, pages 364–372. Morgan Kaufmann, 1996.
- [Oliver and Dowe, 1995] J. J. Oliver and D. L. Dowe. Using unsupervised learning to assist supervised learning. In *Proc. 8th Australian Joint Conf. on Artificial Intelligence*, pages 275–282, November 1995. See also TR 95/235, Dept. Comp. Sci., Monash University, Australia 3168, Sep 1995.
- [Oliver *et al.*, 1992] J. J. Oliver, D. L. Dowe, and C. S. Wallace. Inferring decision graphs using the minimum message length principle. In *Proc. of the 1992 Aust. Joint Conf. on Artificial Intelligence*, pages 361–367, September 1992.
- [Oliver and Hand, 1996] J. J. Oliver and D. J. Hand. Averaging on decision trees. *Journal of Classification*, 1996. An extended version is available as Technical Report TR 5-94, Dept. of Statistics, Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.
- [Oliver and Hand, 1994] J. J. Oliver and D. J. Hand. Fanned decision trees. Technical report TR 5-94, Dept. of Statistics, Open University, Walton Hall, Milton Keynes, MK7 6AA, UK, 1994.
- [Oliver and Wallace, 1991] J. J. Oliver and C. S. Wallace. Inferring decision graphs. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91), workshop 8*, January 1991.
- [Oliver and Wallace, 1992] J. J. Oliver and C. S. Wallace. Inferring decision graphs. Technical report CS 92/170, Dept Computer Science, Monash University, Melbourne, Australia, 1992.
- [Ooi and Dowe, 2005] J. N. Ooi and D. L. Dowe. Inferring phylogenetic graphs of natural languages using minimum message length. In *CAEPIA 2005 (11th Conference of the Spanish Association for Artificial Intelligence)*, volume 1, pages 143–152, Nov. 2005.
- [Papp *et al.*, 1993] E. Papp, D. L. Dowe, and S. J. D. Cox. Spectral classification of radiometric data using an information theory approach. In *Proc. Advanced Remote Sensing Conf.*, pages 223–232, UNSW, Sydney, Australia, July 1993.
- [Parry, 2005] Leigh Parry. Midas touch. *The Age newspaper, Melbourne, Australia (Education section)*, page 6 (in Education section), Mon. 20 June 2005. www.TheAge.com.au, www.monash.edu.au/policy/midas.htm.
- [Patrick, 1978] J. D. Patrick. *An Information Measure Comparative Analysis of Megalithic Geometries*. PhD thesis, Department of Computer Science, Monash University, Australia, 1978.
- [Patrick and Wallace, 1977] J. D. Patrick and C. S. Wallace. Stone circles: A comparative analysis of megalithic geometry. In *Proc. 48th Australian & New Zealand Association for the Advancement of Science (ANZAAS) Conference*. 1977. abstract.

- [Patrick and Wallace, 1982] J. D. Patrick and C. S. Wallace. Stone circle geometries: an information theory approach. In D. Heggie, editor, *Archaeoastronomy in the New World*, pages 231–264. Cambridge University Press, 1982.
- [Phillips and Ploberger, 1996] P. C. B. Phillips and W. Ploberger. An asymptotic theory of Bayesian inference for time series. *Econometrica*, 64(2):240–252, 1996.
- [Pilowsky *et al.*, 1969] I. Pilowsky, S. Levine, and D.M. Boulton. The classification of depression by numerical taxonomy. *British Journal of Psychiatry*, 115:937–945, 1969.
- [Prior *et al.*, 1998] M. Prior, R. Eisenmajer, S. Leekam, L. Wing, J. Gould, B. Ong, and D. L. Dowe. Are there subgroups within the autistic spectrum? A cluster analysis of a group of children with autistic spectrum disorders. *J. Child Psychol. Psychiat.*, 39(6):893–902, 1998.
- [Quinlan and Rivest, 1989] J.R. Quinlan and R.L. Rivest. Inferring decision trees using the minimum description length principle. *Information and Computation*, 80:227–248, 1989.
- [Rissanen, 1976] J. J. Rissanen. Generalized Kraft inequality and arithmetic coding. *IBM J. Res. Develop.*, 20(3):198–203, May 1976.
- [Rissanen, 1978] J. J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.
- [Rissanen, 1996] J. J. Rissanen. Fisher Information and Stochastic Complexity. *IEEE Trans. on Information Theory*, 42(1):40–47, January 1996.
- [Rissanen, 1999a] J. J. Rissanen. Hypothesis selection and testing by the MDL principle. *Computer Journal*, 42(4):260–269, 1999.
- [Rubinstein *et al.*, 2007] B. Rubinstein, P. Bartlett, and J. H. Rubinstein. Shifting, one-inclusion mistake bounds and tight multiclass expected risk bounds. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19 (NIPS 2006)*. MIT Press, Cambridge, MA, U.S.A., 2007.
- [Rumantir and Wallace, 2001] G. W. Rumantir and C. S. Wallace. Sampling of highly correlated data for polynomial regression and model discovery. In *The 4th International Symposium on Intelligent Data Analysis (IDA)*, pages 370–377, 2001.
- [Rumantir and Wallace, 2003] G. W. Rumantir and C. S. Wallace. Minimum message length criterion for second-order polynomial model selection applied to tropical cyclone intensity forecasting. In *The 5th International Symposium on Intelligent Data Analysis (IDA)*, pages 486–496, 2003.
- [Sanghi and Dowe, 2003] P. Sanghi and D. L. Dowe. A computer program capable of passing I.Q. tests. In *4th International Conference on Cognitive Science (and 7th Australasian Society for Cognitive Science Conference)*, volume 2, pages 570–575, Univ. of NSW, Sydney, Australia, Jul 2003.
- [Schmidhuber, 2007] J. Schmidhuber. Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In *Lecture Notes in Computer Science (LNCS) 4755*, pages 26–38. Springer, 2007.
- [Schmidt, 2008] D. F. Schmidt. *Minimum Message Length Inference of Autoregressive Moving Average Models*. PhD thesis, Faculty of Information Technology, Monash University, 2008.
- [Schmidt and Makalic, 2009b] D. F. Schmidt and E. Makalic. Minimum message length shrinkage estimation. *Statistics & Probability Letters*, 79(9):1155–1161, 2009.
- [Schmidt and Makalic, 2009c] D. F. Schmidt and E. Makalic. MML invariant linear regression. In *Lecture Notes in Artificial Intelligence (Proc. 22nd Australian Joint Conf. on Artificial Intelligence [AI’09])*. Springer, December 2009.
- [Schwarz, 1978] G. Schwarz. Estimating dimension of a model. *Ann. Stat.*, 6:461–464, 1978.
- [Searle, 1980] J. R. Searle. Minds, brains and programs. *Behavioural and Brain Sciences*, 3:417–457, 1980.
- [Shmueli and Koppius, 2007] G. Shmueli and O. Koppius. Predictive vs. Explanatory Modeling in IS Research. In *Proc. Conference on Information Systems & Technology*, 2007. Seattle, Wa, U.S.A. (URL www.citi.uconn.edu/cist07/5c.pdf).
- [Solomonoff, 1960] R. J. Solomonoff. A preliminary report on a general theory of inductive inference. Report V-131, Zator Co., Cambridge, Mass., U.S.A., 4 Feb. 1960.
- [Solomonoff, 1964] R. J. Solomonoff. A formal theory of inductive inference. *Information and Control*, 7:1–22, 224–254, 1964.
- [Solomonoff, 1995] R. J. Solomonoff. The discovery of algorithmic probability: A guide for the programming of true creativity. In P. Vitányi, editor, *Computational Learning Theory: EuroCOLT’95*, pages 1–22. Springer-Verlag, 1995.

- [Solomonoff, 1996] R. J. Solomonoff. Does algorithmic probability solve the problem of induction? In D. L. Dowe, K. B. Korb, and J. J. Oliver, editors, *Proceedings of the Information, Statistics and Induction in Science (ISIS) Conference*, pages 7–8, Melbourne, Australia, August 1996. World Scientific. ISBN 981-02-2824-4.
- [Solomonoff, 1997a] R. J. Solomonoff. The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55(1):73–88, 1997.
- [Solomonoff, 1997b] R. J. Solomonoff. Does algorithmic probability solve the problem of induction? Report, Oxbridge Research, P.O.B. 400404, Cambridge, Mass. 02140, U.S.A., 1997. See <http://world.std.com/~rjs/isis96.pdf>.
- [Solomonoff, 1999] R. J. Solomonoff. Two kinds of probabilistic induction. *Computer Journal*, 42(4):256–259, 1999. Special issue on Kolmogorov Complexity.
- [Solomonoff, 2008] R. J. Solomonoff. Three kinds of probabilistic induction: Universal and convergence theorems. *Computer Journal*, 51(5):566–570, September 2008.
- [Tan and Dowe, 2002] P. J. Tan and D. L. Dowe. MML inference of decision graphs with multi-way joins. In R. McKay and J. Slaney, editors, *Proc. 15th Australian Joint Conference on Artificial Intelligence — Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, Germany, ISSN: 0302-9743, Vol. 2557*, number 2557 in Lecture Notes in Artificial Intelligence (LNAI), pages 131–142. Springer Verlag, 2002.
- [Tan and Dowe, 2003] P. J. Tan and D. L. Dowe. MML inference of decision graphs with multi-way joins and dynamic attributes. In *Lecture Notes in Artificial Intelligence (LNAI) 2903 (Springer), Proc. 16th Australian Joint Conf. on Artificial Intelligence*, pages 269–281, Perth, Australia, Dec. 2003.
- [Tan and Dowe, 2004] P. J. Tan and D. L. Dowe. MML inference of oblique decision trees. In *Lecture Notes in Artificial Intelligence (LNAI) 3339 (Springer), Proc. 17th Australian Joint Conf. on Artificial Intelligence*, volume 3339, pages 1082–1088, Cairns, Australia, Dec. 2004.
- [Tan and Dowe, 2006] P. J. Tan and D. L. Dowe. Decision forests with oblique decision trees. In *Lecture Notes in Artificial Intelligence (LNAI) 4293 (Springer), Proc. 5th Mexican International Conf. Artificial Intelligence*, pages 593–603, Apizaco, Mexico, Nov. 2006.
- [Tan et al., 2007] P. J. Tan, D. L. Dowe, and T. I. Dix. Building classification models from microarray data with tree-based classification algorithms. In *Lecture Notes in Artificial Intelligence (LNAI) 4293 (Springer), Proc. 20th Australian Joint Conf. on Artificial Intelligence*, Dec. 2007.
- [Torsello and Dowe, 2008a] A. Torsello and D. L. Dowe. Learning a generative model for structural representations. In *Lecture Notes in Artificial Intelligence (LNAI)*, volume 5360, pages 573–583, 2008.
- [Torsello and Dowe, 2008b] A. Torsello and D. L. Dowe. Supervised learning of a generative model for edge-weighted graphs. In *Proc. 19th International Conference on Pattern Recognition (ICPR2008)*. IEEE, 2008. 4pp. IEEE Catalog Number: CFP08182, ISBN: 978-1-4244-2175-6, ISSN: 1051-4651.
- [Turing, 1936] A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc.* 2, 42:230–265, 1936.
- [Uther and Veloso, 2000] W. T. B. Uther and M. M. Veloso. The Lumberjack Algorithm for Learning Linked Decision Forests. In *Proc. 6th Pacific Rim International Conf. on Artificial Intelligence (PRICAI'2000), Lecture Notes in Artificial Intelligence (LNAI) 1886 (Springer)*, pages 156–166, 2000.
- [Vapnik, 1995] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
- [Visser and Dowe, 2007] Gerhard Visser and D. L. Dowe. Minimum message length clustering of spatially-correlated data with varying inter-class penalties. In *Proc. 6th IEEE International Conf. on Computer and Information Science (ICIS) 2007*, pages 17–22, July 2007.
- [Visser et al., 2009a] Gerhard Visser, D. L. Dowe, and I. D. Svalbe. Information-theoretic image reconstruction and segmentation from noisy projections. In *Lecture Notes in Artificial Intelligence (Proc. 22nd Australian Joint Conf. on Artificial Intelligence [AI'09])*, pp. 170–179. Springer, December 2009.
- [Visser et al., 2009b] Gerhard Visser, D. L. Dowe, and J. Petteri Uotila. Enhancing MML clustering using context data with climate applications. In *Lecture Notes in Artificial Intelligence (Proc. 22nd Australian Joint Conf. on Artificial Intelligence [AI'09])*, pp. 350–359. Springer, December 2009.

- [Viswanathan and Wallace, 1999] M. Viswanathan and C. S. Wallace. A note on the comparison of polynomial selection methods. In D. Heckerman and J. Whittaker, editors, *Proceedings of Uncertainty 99: The Seventh International Workshop on Artificial Intelligence and Statistics*, pages 169–177, Fort Lauderdale, Florida, USA, January 1999. Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA.
- [Viswanathan *et al.*, 1999] Murlikrishna Viswanathan, C. S. Wallace, David L. Dowe, and Kevin B. Korb. Finding cutpoints in noisy binary sequences — a revised empirical evaluation. In *Proc. 12th Australian Joint Conference on Artificial Intelligence*, volume 1747 of *Lecture Notes in Artificial Intelligence*, pages 405–416. Springer Verlag, 1999.
- [Wallace, 1973a] C. S. Wallace. Simulation of a two-dimensional gas. In *Proc. of the Australian & New Zealand Association for the Advancement of Science (ANZAAS) Conf.*, page 19, August 1973. abstract.
- [Wallace, 1984b] C. S. Wallace. An improved program for classification. Technical Report 47, Department of Computer Science, Monash University, Clayton, Victoria 3168, Australia, Melbourne, 1984.
- [Wallace, 1984a] C. S. Wallace. Inference and estimation by compact coding. Technical Report 84/46, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, August 1984.
- [Wallace, 1986] C. S. Wallace. An improved program for classification. In *Proc. of the 9th Australian Computer Science Conference (ACSC-9)*, pages 357–366, February 1986. Published as Proc. of ACSC-9, volume 8, number 1.
- [Wallace, 1990c] C. S. Wallace. Classification by minimum-message-length encoding. In S. G. Akl *et al.*, editor, *Advances in Computing and Information — ICCI '90*, volume 468 of *Lecture Notes in Computer Science (LNCS)*, pages 72–81. Springer-Verlag, May 1990.
- [Wallace, 1990b] C. S. Wallace. Classification by minimum-message-length inference. In *Working Notes AAAI Spring Symposium Series*, pages 65–69. Stanford Uni., Calif., U.S.A., 1990.
- [Wallace, 1992] C. S. Wallace. A Model of Inductive Inference. Seminar, November 1992. Also on video, Dept. of Computer Science, Monash University, Clayton 3168, Australia, Wed. 25 Nov. 1992.
- [Wallace, 1995a] C. S. Wallace. Multiple Factor Analysis by MML Estimation. Technical report CS TR 95/218, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, Clayton, Melbourne, Australia, 1995.
- [Wallace, 1996c] C. S. Wallace. False oracles and SMML estimators. In D. L. Dowe, K. B. Korb, and J. J. Oliver, editors, *Proceedings of the Information, Statistics and Induction in Science (ISIS) Conference*, pages 304–316, Melbourne, Australia, August 1996. World Scientific. ISBN 981-02-2824-4. Was previously Tech Rept 89/128, Dept. Comp. Sci., Monash Univ., Australia, June 1989.
- [Wallace, 1996b] C. S. Wallace. MML inference of predictive trees, graphs and nets. In A. Gammerman, editor, *Computational Learning and Probabilistic Reasoning*, chapter 3, pages 43–66. Wiley, 1996.
- [Wallace, 1997] C. S. Wallace. On the selection of the order of a polynomial model. Technical report, Royal Holloway College, England, U.K., 1997. Chris released this in 1997 (from Royal Holloway) in the belief that it would become a Royal Holloway Tech Rept dated 1997, but it is not clear that it was ever released there. Soft copy certainly does exist, though. Perhaps see www.csse.monash.edu.au/~dld/CSWallacePublications.
- [Wallace, 1998d] C. S. Wallace. Competition isn't the only way to go, a Monash FIT graduation address, April 1998. (Perhaps see www.csse.monash.edu.au/~dld/CSWallacePublications).
- [Wallace, 1998a] C. S. Wallace. Intrinsic classification of spatially correlated data. *Computer Journal*, 41(8):602–611, 1998.
- [Wallace, 1998b] C. S. Wallace. Multiple factor analysis by MML estimation. In *Proceedings of the Fourteenth Biennial Australian Statistical Conference (ASC-14)*, page 144, Queensland, Australia, July 1998.
- [Wallace, 1998c] C. S. Wallace. On the selection of the order of a polynomial model. In W. Robb, editor, *Proc. of the 14th Biennial Australian Statistical Conf.*, page 145, Queensland, Australia, July 1998.
- [Wallace, 1998e] C. S. Wallace. PAKDD-98 Tutorial: Data Mining, 15-17 April 1998. Tutorial entitled “Data Mining” at the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD-98), held in Melbourne, Australia. This

- partly constituted an early draft of Chris Wallace's 2005 book "*Statistical and Inductive Inference by Minimum Message Length*".
- [Wallace, 1999] C. S. Wallace. *The MIT Encyclopedia of the Cognitive Sciences (MITECS)*, chapter Minimum description length, (major review), pages 550–551. The MIT Press, London, England, ISBN: 0-262-73124-X, 1999.
 - [Wallace, 2005] C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Information Science and Statistics. Springer Verlag, May 2005. ISBN 0-387-23795X.
 - [Wallace and Boulton, 1968] C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.
 - [Wallace and Boulton, 1975] C. S. Wallace and D. M. Boulton. An invariant Bayes method for point estimation. *Classification Society Bulletin*, 3(3):11–34, 1975.
 - [Wallace and Dale, 2005] C. S. Wallace and M. B. Dale. Hierarchical clusters of vegetation types. *Community Ecology*, 6(1):57–74, 2005. ISSN: 1585-8553.
 - [Wallace and Dowe, 1993] C. S. Wallace and D. L. Dowe. MML estimation of the von Mises concentration parameter. Technical Report 93/193, Dept. of Computer Science, Monash University, Clayton 3168, Australia, December 1993.
 - [Wallace and Dowe, 1994a] C. S. Wallace and D. L. Dowe. Estimation of the von Mises concentration parameter using minimum message length. In *Proc. 12th Aust. Stat. Soc. Conf.*, 1994. 1 page abstract.
 - [Wallace and Dowe, 1994b] C. S. Wallace and D. L. Dowe. Intrinsic classification by MML — the Snob program. In *Proc. 7th Australian Joint Conf. on Artificial Intelligence*, pages 37–44. World Scientific, November 1994.
 - [Wallace and Dowe, 1996] C. S. Wallace and D. L. Dowe. MML mixture modelling of Multi-state, Poisson, von Mises circular and Gaussian distributions. In *Proc. Sydney International Statistical Congress (SISC-96)*, page 197, Sydney, Australia, 1996.
 - [Wallace and Dowe, 1997a] C. S. Wallace and D. L. Dowe. MML mixture modelling of multi-state, Poisson, von Mises circular and Gaussian distributions. *Proc 28th Symp. on the Interface*, pages 608–613, 1997.
 - [Wallace and Dowe, 1997b] C. S. Wallace and D. L. Dowe. MML mixture modelling of multi-state, Poisson, von Mises circular and Gaussian distributions. In *Sixth International Workshop on Artificial Intelligence and Statistics, Society for AI and Statistics*, pages 529–536, San Francisco, USA, 1997.
 - [Wallace and Dowe, 1999a] C. S. Wallace and D. L. Dowe. Minimum message length and Kolmogorov complexity. *Computer Journal*, 42(4):270–283, 1999.
 - [Wallace and Dowe, 1999b] C. S. Wallace and D. L. Dowe. Refinements of MDL and MML coding. *Computer Journal*, 42(4):330–337, 1999.
 - [Wallace and Dowe, 1999c] C. S. Wallace and D. L. Dowe. Rejoinder. *Computer Journal*, 42(4):345–347, 1999.
 - [Wallace and Dowe, 2000] C. S. Wallace and D. L. Dowe. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, 10:73–83, January 2000.
 - [Wallace and Freeman, 1987] C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society series B*, 49(3):240–252, 1987. See also Discussion on pp252–265.
 - [Wallace and Freeman, 1992] C. S. Wallace and P. R. Freeman. Single-factor analysis by minimum message length estimation. *J. Royal Stat. Soc. B*, 54(1):195–209, 1992.
 - [Wallace and Georgeff, 1983] C. S. Wallace and M. P. Georgeff. A general objective for inductive inference. Technical Report #83/32, Department of Computer Science, Monash University, Clayton, Australia, March 1983. Reissued in June 1984 as TR No. 44.
 - [Wallace and Korb, 1994] C. S. Wallace and K. B. Korb. A Bayesian learning agent. In C. S. Wallace, editor, *Research conference: Faculty of Computing and Information Technology*, page 19. Monash University Melbourne, 1994.
 - [Wallace and Korb, 1997] C. S. Wallace and K B Korb. Learning linear causal models by MML sampling. Technical report CS 97/310, Dept Computer Science, Monash University, Melbourne, Australia, 1997.
 - [Wallace and Korb, 1999] C. S. Wallace and K. B. Korb. Learning linear causal models by MML sampling. In A. Gammerman, editor, *Causal Models and Intelligent Data Management*, pages 89–111. Springer Verlag, 1999. see TR 97/310, Dept. Comp. Sci., Monash Univ., Australia, June 1997.

- [Wallace *et al.*, 1996b] C. S. Wallace, K B Korb, and H Dai. Causal discovery via MML. Technical report CS 96/254, Dept Computer Science, Monash University, Melbourne, Australia, 1996.
- [Wallace *et al.*, 1996a] C. S. Wallace, Kevin B. Korb, and Honghua Dai. Causal discovery via MML. In *13th International Conf. on Machine Learning (ICML-96)*, pages 516–524, 1996.
- [Wallace and Patrick, 1991] C. S. Wallace and J D Patrick. Coding decision trees. Technical report CS 91/153, Dept Computer Science, Monash University, Melbourne, Australia, 1991.
- [Wallace and Patrick, 1993] C. S. Wallace and J. D. Patrick. Coding decision trees. *Machine Learning*, 11:7–22, 1993.
- [Zakis *et al.*, 1994] J. D. Zakis, I. Cosic, and D. L. Dowe. Classification of protein spectra derived for the resonant recognition model using the minimum message length principle. *Australian Comp. Sci. Conf. (ACSC-17)*, pages 209–216, January 1994.

SIMPLICITY, TRUTH, AND PROBABILITY

Kevin T. Kelly

1 INTRODUCTION

Scientific theories command belief or, at least, confidence in their ability to predict what will happen in remote or novel circumstances. The justification of that trust must derive, somehow, from scientific method. And it is clear, both from the history of science and from the increasing codification and automation of the scientific method both in statistics and in machine learning, that a major component of that method is *Ockham's razor*, a systematic bias toward *simple* theories, where “simplicity” has something to do with minimizing free parameters, gratuitous entities and causes, independent principles and ad hoc explanations and with maximizing unity, testability, and explanatory power.

Ockham's razor is not a bloodless, formal rule that must be learned — it has a native, visceral grip on our credence. For a celebrated example, Copernicus was driven to move the earth to eliminate five epicycles from medieval astronomy [Kuhn, 1957]. The principal problem of positional planetary astronomy was to account for the apparently irregular, retrograde or backward motion of the planets against the fixed stars. According to the standard, Ptolemaic theory of the time, retrograde motion results from the planet revolving around an *epicycle* or circle whose center revolves, in turn, on another circle called the *deferent*, centered on the earth. Making the epicycle revolve in the same sense as the deferent implies that the planet should be closest or brightest at the midpoint of its retrograde motion, which agrees with observations. Copernicus explained retrograde motion in terms of the moving earth being lapped or lapping the other planets on a cosmic racetrack centered on the sun, which eliminates one epicycle per planet (figure 1). Copernicus still required many superimposed circles to approximate elliptical orbits, so the mere elimination of five such circles may not seem very impressive. But there is more to the story than just counting circles. It happens that the retrograde motions of Mars, Jupiter, and Saturn occur precisely when the respective planet is in solar opposition (i.e., is observed 180° from the sun) and that the retrograde motions of Mercury and Venus occur at solar conjunction (i.e., when the respective planet is 0° from the sun). Ptolemy's epicycles can be adjusted to recover the same effect, but only in a rather bizarre manner. Think of the line from the earth to the sun as the hand of a clock and think of the line from the center of Saturn's epicycle to Saturn as the hand of another clock. Then retrograde motion happens exactly at solar opposition if and only if Saturn's epicycle clock is

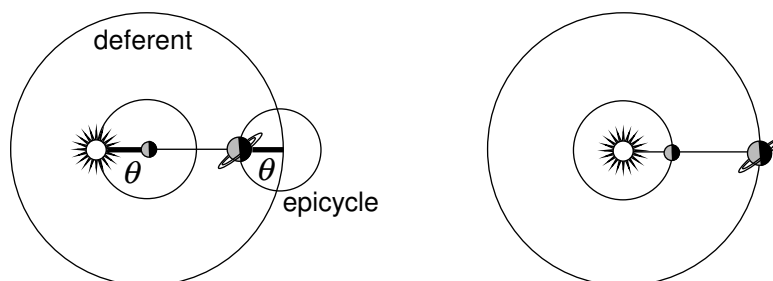


Figure 1. Ptolemy vs. Copernicus

perfectly synchronized with the sun's deferent clock. The same is true of Mars and Jupiter. Furthermore, Mercury and Venus undergo retrograde motion exactly at solar opposition just in case their deferent clocks are perfectly synchronized with the sun's deferent clock. In Ptolemy's theory, these perfect synchronies across vast distances in the solar system appear bizarre and miraculous. On Copernicus' theory, however, they are ineluctable, geometrical banalities: the earth passes an outer planet exactly when the earth crosses the line from the central sun to the planet passed. So Copernicus' theory crisply *explains* the striking synchronies. Copernicus' theory is also *severely tested* by the synchronies, since it would be refuted by any perceived deviation from exact synchrony, however slight. Ptolemy's theory, on the other hand, merely *accommodates* the data in an *ad hoc* manner by means of its adjustable parameters. It seems that Copernicus' theory should get some sort of reward for surviving a test shirked by its competitor. One could add clockwork gears to Ptolemy's theory to explain the synchronies, but that would be an *extra principle* receiving no *independent confirmation* from other evidence. Copernicus' explanation, on the other hand, recovers both retrograde motion and its correlation with solar position from the geometry of a circular racetrack, so it provides a *unified explanation* of the two phenomena. Empirical simplicity is more than mere notational brevity — it implies such red-blooded considerations as explanatory power [Harman, 1965], unity [Kitcher, 1982], independently confirmable principles [Friedman, 1983; Glymour, 1980] and severe testability [Popper, 1968; Mayo, 1996].

Another standard example of Ockham's razor in action concerns the search for empirical laws (figure 2). Any finite number of observations can be connected with a polynomial curve that passes through each, but we may still prefer a straight line that comes close to each point. It is, perhaps, more tempting in this case to identify simplicity with syntactic length or complexity of the law, since $\alpha_0 x^0 + \alpha_1 x^1 + \dots + \alpha_n x^n$ is obviously more verbose than $\alpha_0 x^0 + \alpha_1 x^1$. But one can also say that the complex law merely accommodates the data by having an independent, adjustable parameter for each data point, whereas when the two parameters of the simple law can be estimated with a few data, providing an explanation of the remaining data points. The complex law is also less unified than the simple law

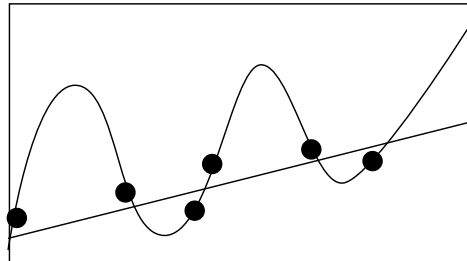


Figure 2. Inferring polynomial degree

(the several coefficients receive isolated support from the data points they are set to account for) and is less visually “uniform” than the simple law.

Ockham’s razor does the heavy lifting in scientific theory choice, for no other principle suffices to winnow the infinite range of possible explanations of the available data down to a unique one. And whereas simplicity was once the theorist’s personal prerogative, it is now a mathematically explicit and essential component of contemporary statistical and computational techniques for drawing conclusions from empirical data (cf. [Mitchell, 1977, Duda *et al.*, 2001]). The explicitness and indispensability of Ockham’s razor in scientific theory selection raises a natural question about its justification. Epistemic justification is not just a word or a psychological urge or a socially sanctioned, exculpatory ritual or procedure. It should imply some sort of *truth-conduciveness* of the underlying process by which one’s trust is produced. An attractively ambitious concept of truth-conduciveness is *reliable indication* of the truth, which means that the process has a high chance of producing the true theory, whatever the truth happens to be, the way a properly functioning thermometer indicates temperature. But Ockham’s razor is more like a trick thermometer whose reading never changes. Such a thermometer cannot be said to indicate the temperature even if its fixed reading happens to be true. Neither can a fixed bias toward simplicity immediately indicate the truth about nature — unless the truth is already known to be simple, in which case there would be no need to invoke Ockham’s razor by way of justification.¹

Ockham’s razor has a good excuse for failing to reliably indicate true theories, since theory choice requires inductive inference and no inductive inference method can be a truth-indicator: each finite set of data points drawn with bounded precision from a linear law is also compatible with a sufficiently flat parabola, so no possible data-driven process could reliably indicate, in the short run, whether the truth is linear or quadratic. A more feasible concept of truth-conduciveness for inductive inference is *convergence in the limit*, which means that the chance that the method produces the true theory converges to one, no matter what the true

¹This updated version of Plato’s Meno paradox is underscored in machine learning by the “no free lunch theorems” [Wolpert, 1996].

theory might be.² Convergence to the truth in the limit is far weaker than short-run truth-indication, since it is compatible with the choice of any finite number of false theories with arbitrarily high chance before settling on the correct one. Each time a new theory T_{n+1} is produced with high chance, the chance of producing the previous candidate T_n must drop precipitously and one may say that the output is *retracted*. So convergence in the limit differs from reliable indication by allowing any finite number of arbitrarily precipitous retractions prior to “locking on” to the right answer. Assuming that the true theory is polynomial, Ockham’s razor does converge in the limit to the true polynomial degree of $f(x)$ — each polynomial degree lower than the true degree is ruled out, eventually, by the data (e.g., when new bumps in the true law become noticable), after which the true theory is the simplest theory compatible with experience. Think of successively more complex theories as tin cans lined up on a fence, one of which (the true one) is nailed to the fence. Then, if one shoots the cans from left to right, eventually the nailed can becomes and remains the first can in line that has not yet been shot down. The familiar trouble with this explanation of Ockham’s razor is that convergence in the long run is compatible reliance on any alternative bias for any finite duration [Salmon, 1967]. For example, guess an equation of degree 10 with the hope that the coefficient is so large that the thousand bumps will be noticed early — say in a sample of size 1000. If they aren’t seen by then, revert back to Ockham’s razor, which succeed in the limit. Hence, convergence in the limit is feasible in theoretical inference, but it does not single out simple theories as the right theories to produce in the short run.

To summarize, the justification of Ockham’s razor poses a puzzle. Ockham’s razor can’t reliably indicate the true theory in the short run, due to the problem of induction. And although Ockham’s razor does converge to the truth in the ideal limit of inquiry, alternative methods producing very complex theories are also truth-conducive in that very weak sense as well [Salmon, 1967]. So short-run indication is too strong to be feasible and long-run convergence is too weak to single out Ockham’s razor. It remains, therefore, to define a sense of truth-conduciveness according to which it can be argued, without circularity, that Ockham’s razor helps one find the truth better than alternative methods that would produce arbitrarily complex theories *now*. Absent such a story, Ockham’s razor starts to look like an exercise in wishful thinking — the epistemic sin of inferring that reality is simple because the true theory of a simple world would have pragmatic virtues (e.g., explanatory power) that one would like to have. Such doubts motivate a skeptical or anti-realist attitude toward scientific theories in general [van Fraassen, 1981].

This paper reviews the standard explanations of Ockham’s razor, which fall into two main groups. The first group invokes a tacit, prior bias toward simplicity, which begs the question in favor of Ockham’s razor. The second group substitutes a particular notion of predictive accuracy for truth, based on the surprising fact that a false theory may make more accurate predictions than the true one when

²This concept is called *convergence in probability* in probability theory and *consistency* in statistics.

the truth is complex. That evidently fails to explain how Ockham's razor finds true theories (as opposed to useful models). Furthermore, when predictions concern the outcomes of interventions on the world, even the argument for predictive accuracy fails.³ Since neither approach really explains how Ockham's razor leads to true theories or even to accurate policy predictions, the second part of the paper develops an entirely new explanation: Ockham's razor does not point at the truth, even with high probability, but it does help one arrive at the truth with uniquely optimal *efficiency*, where efficiency is measured in terms of such epistemically pertinent considerations as the total number of errors and retractions of prior opinions incurred before converging to the truth and the elapsed times by which the retractions occur. Thus, in a definite sense, Ockham's razor is demonstrably the uniquely most truth-conducive method for inferring general theories from particular facts — even though no possible method can be guaranteed to point toward the truth with high probability in the short run.

2 THE ARGUMENT FROM BAYES FACTORS

Bayesian statisticians assign probability-valued degrees of belief to all the propositions in some language and then “rationally” update those degrees of belief by a universal rule called *conditionalization*.⁴ If $p_t(T)$ is your prior degree of belief that T at stage t and if E is new evidence received at stage $t+1$, then conditionalization says that your updated degree of belief that T at $t+1$ should be:

$$p_{t+1}(T) = p_t(T \mid E).$$

It follows from the conditionalization rule that:

$$p_{t+1}(T) = (p_t(T) \cdot p_t(E \mid T)) / p_t(E).$$

An important feature of the rule is that one's updated degree of belief p_{t+1} depends on one's prior degree of belief $p_t(T)$, which might have been strongly biased for or against T prior to collecting any evidence about T whatever. That feature suggests an easy “justification” of Ockham's razor — just start out with prior probabilities biased toward simple theories. Then, if simple theories explain the data about as well as complex ones, the prior bias toward the simple theory passes through the updating procedure. But to invoke a prior bias toward simplicity to explain a prior bias toward simplicity evidently begs the main question at hand.

A more promising Bayesian argument for Ockham's razor centers not on the prior probability $p_t(T)$, but on the term $p_t(E \mid T)$, which corresponds to the rational credence conferred on E by theory T . (cf. [Jeffreys, 1961; Rosenkrantz,

³For candid discussions of the shortcomings of the usual explanations of Ockham's razor as it is used in machine learning, cf., for example, [Domingos, 1999] and [Mitchell, 1997].

⁴Not all Bayesians accept updating by conditionalization. Some Bayesians recommend accepting hypotheses altogether, in which case the degree of belief goes to one. Others recommend updating on partially believed evidence. Others recommend updating interval-valued degrees of belief, etc. Others reject its coherentist justification in terms of diachronic Dutch books.

1983; Myrvold, 2003]). According to this explanation, Ockham's razor does not demand that the simpler theory T_1 start out ahead of its complex competitor T_2 ; it suffices that T_1 pull ahead of T_2 when evidence E compatible with T_1 is received. That sounds impressive, for the conditional probability $p_t(E | T)$ is often thought to be more objective than the prior probability $p_t(T)$, because $p_t(E | T)$ reflects the degree to which T "explains" E . But that crucially overstates the case when T has free parameters to adjust, as when Ockham's razor is at issue. Thoroughly subjective Bayesians interpret "objective" probabilities as nothing more than relatively inter-subjective degrees of belief, but a more popular, alternative view ties objectivity to *chances*. Chances are supposed to be natural, objective probabilities that apply to possible outcomes of random experiments. Chance will be denoted by a capital P , in contrast with the lower-case p denoting degrees of belief. Bayesian statisticians link chances to evidence and to action by means of the *direct inference principle* [Kyburg, 1977; Levi, 1977; Lewis, 1987], which states that degrees of belief should accord with known chances, given only *admissible*⁵ information E' :

$$p_t(E | P(E) = r \wedge E') = r.$$

If theory T says exactly that the true chance distribution of X is P , then by the direct inference principle:

$$p_t(E | T) = P(E),$$

which is, indeed, objective. But if T is complex, then T has adjustable parameters and, hence, implies only that the true chance distribution lies in some set, say: $\{P_1, \dots, P_n\}$. Then the principle of direct inference yields the weighted average:

$$p_t(E | T) = \sum_{i=1}^n P_i(E) \cdot p_t(P_i | T),$$

in which the weights $p_t(P_n | T)$ are prior degrees of belief, not chances. So the objective-looking quantity $p_t(E | T)$ is *loaded* with prior opinion when T is complex and that potentially hidden fact is crucial to the Bayesian explanation of Ockham's razor.

A standard technique for comparing the posterior probabilities of theories is to look at the *posterior ratio*:

$$\frac{p_t(T_1 | E)}{p_t(T_2 | E)} = \frac{p_t(T_1)}{p_t(T_2)} \cdot \frac{p_t(E | T_1)}{p_t(E | T_2)}.$$

The first quotient on the right-hand-side is the *prior ratio*, which remains constant as new evidence E is received. The second quotient is the *Bayes factor*, which accounts for the entire impact of E on the relative credence of the two theories [Kass and Raftery, 1995].

To guarantee that $p(T_1 | E) > p(T_2 | E)$, one must impose some further, material restrictions on coherent degrees of belief, but it can be argued that the constraints

⁵Defining admissibility is a vexed question that will be ignored here.

are presupposed by the very question whether Ockham's razor should be used when choosing between a simple and a complex theory. That places the Bayesian explanation of Ockham's razor in the same class of a priori metaphysical arguments that includes Descartes' *cogito*, according to which the thesis "I exist" is evidently true each time one questions it. First of all, a Bayesian wouldn't think of herself as *choosing* between T_1 and T_2 if she started with a strong bias toward one theory or the other, so let $p_t(T_1) \approx p_t(T_2)$. Second, she wouldn't be choosing between two theories *compatible* with E unless simple theory T_1 explains E , so that $P(E) \approx 1$. Third, she wouldn't say that T_2 is *complex* unless T_2 has a free parameter i to adjust to save the data. She would not say that the parameter of T_2 is *free* unless she were fairly uncertain about which chance distribution P_i would obtain if T_2 were true: e.g. $p_t(P_i | T_2) \approx 1/n$. Furthermore, she would not say that the parameter must be *adjusted* to save E unless the chance of E is high only over a narrow range of possible chance distributions compatible with T_2 : e.g., $P_0(E) \approx 1$ and for each alternative i such that $0 < i \leq n$, $p_i(E) \approx 0$. It follows from the above assumptions that the prior probability ratio is approximately 1 and the Bayes' factor is approximately n , so:

$$\frac{p_t(T_1 | E)}{p_t(T_2 | E)} \approx n.$$

Thus, the simple theory T_1 ends up far more probable than the complex theory T_2 in light of evidence E , as the complex theory T_2 becomes more "adjustable", which is the argument's intended conclusion. When the set of possible chance distributions $\{P_\theta : \theta \in \mathbb{R}\}$ is continuously parameterized, the argument is similar, except that the (discrete) weighted sum expressing $p_t(E | T_2)$ becomes a (continuous) integral:

$$p_t(E | T_2) = \int P_\theta(E) \cdot p_t(P_\theta | T_2) d\theta,$$

which, again, is weighted by the subjective degrees of belief $p_t(P_\theta | T_2)$.

Each of the above assumptions can be weakened. It suffices that the prior ratio not favor T_2 too much, that the explanation of E by T_1 not be too vague, that the explanation of E by T_2 not be too robust across parameter values and that the distribution of degrees of belief over free parameters of T_2 not be focused too heavily on the parameter values that more closely mimic the predictions of T_1 .

The Bayes factor argument for Ockham's razor is closely related to standard paradoxes of indifference. Suppose that someone is entirely ignorant about the color of a marble in a box. Indifference over the various colors implies a strong bias against blue in the partition blue vs. non-blue, whereas indifference over blue vs. non-blue implies a strong bias against yellow. The Bayes factor argument amounts to plumping for the former bias. Think of the simple theory T_0 as "blue" and of the complex theory T_2 as "non-blue" with a "free parameter" ranging over red, green, yellow, etc. and assume, for example, that the evidence E is "either blue or red". Then, by the above calculation, the posterior ratio of "blue" over "non-blue" is the number n of distinguished non-blue colors. Now consider the

underlying prior probability over the refined partition blue, red, green, yellow, etc. It is apparent that “blue” is assigned prior probability $1/2$, whereas each alternative color is assigned $1/2n$, where $n > 1$. Hence, the complex *theory* starts out even with the simple theory, but each complex *possibility* starts out with a large disadvantage. Thus, although “red” objectively “explains” E just as well as “blue” does, the prior bias for “blue” over “red” gets passed through the Bayesian updating formula and begs the question in favor of “blue”. One could just as well choose to be “ignorant” over blue, red, green, yellow, etc., in which case “blue” and “red” end up locked in a tie after E is observed and “non-blue” remains more probable than “blue”. So the Bayes factor argument again comes down to a question-begging prior bias in favor of simple possibilities.

One can attempt to single out the simplicity bias by expanding the Bayesian notion of rationality to include “objective” constraints on prior probability: e.g., by basing them on the length of Turing machine programs that would produce the data or type out the hypothesis [Jeffreys, 1961; Rissanen, 2007; Li and Vitanyi, 1993]. But that strategy is an epistemological red herring. Even if “rationality” is augmented to include an intuitively appealing, formal rule for picking out some prior biases over others, the real question regarding Ockham’s razor is whether such a bias helps one find the truth better than alternative biases (cf. [Mitchell, 1997]). To answer that question relevantly, one must explain, without circular appeal to the very bias in question, whether and in what sense Bayesians who start with a prior bias toward simplicity find the truth better than Bayesians starting with alternative biases would. There are two standard strategies for justifying Bayesian updating. Dutch Book arguments show that violating the Bayesian updating rule would result in preference for combinations of diachronic bets that result in a sure loss over time [Teller, 1976]. But such arguments do not begin to establish that Bayesian updating leads to higher degrees of belief in true theories in the short run. In fact, Bayesian updating can result in a huge short-run boost of credence in a false theory: e.g., when the parameters of the true, complex theory are set very close to values that mimic observations fitting a simple alternative. Perhaps the nearest that Bayesians come to taking theoretical truth-conduciveness seriously is to argue that iterated Bayesian updating *converges* to the true theory in the limit, in the sense that $p(T \mid E_n)$ converges to the truth value of T as n increases.⁶ But the main shortcoming with that approach has already been discussed: both Ockham and non-Ockham initial biases are compatible with convergent success in the long run. In sum, Bayesians either beg the question in favor of simplicity by assigning higher prior probability to simpler possibilities, or they ignore truth-conduciveness altogether in favor of arguments for coherence, or they fall back upon the insufficient strategy of appealing to long-run convergence.

⁶Even then, convergence is guaranteed only with unit probability in the agent’s prior probability. The non-trivial consequences of that slip are reviewed in [Kelly, 1996].

3 THE ARGUMENT FROM OVER-FITTING

Classical statisticians seek to justify scientific method entirely in terms of objective chances, so the Bayesian explanation of Ockham's razor in terms of Bayes factors and prior probabilities is not available to them. Instead, they maintain a firm focus on truth-conduciveness but lower their sights from choosing the true theory to choosing the theory that yields the most accurate predictions. If theory T is deterministic and observation is perfectly reliable and T has no free parameters, then prediction involves deducing what will happen from T . If T has a free parameter θ , then one must use some empirical data to fix the true value of θ , after which one deduces what will happen from T (e.g., two observed points determine the slope and intercept of a linear law). More generally, fixing the parameter values of T results only in a chance distribution P_θ over possible experimental outcomes. In that case, it is natural to use past experimental data E' to arrive at an empirical estimate $\hat{\theta}(T, E')$ of parameter θ . A standard estimation technique is to define $\hat{\theta}(T, E')$ to be the value of θ that maximizes $P_\theta(E')$. Then $\hat{\theta}(T, E')$ is called the *maximum likelihood estimate* or MLE of T (given outcome E') and the chance distribution $P_{\hat{\theta}(T, E')}$ is a guess at the probability of future experimental outcomes E . The important point is that theory T is not necessarily *inferred* or *believed* in this procedure; The aim in choosing T is not to choose the true T but, rather, the T that maximizes the accuracy of the estimate $P_{\hat{\theta}(T, E')}$ of P^* . Classical statisticians underscore their non-inferential, instrumentalistic attitude toward statistical theories by calling them *models*.

It may seem obvious that no theory predicts better than the true theory, in which case it would remain mysterious why a fixed bias toward simplicity yields more accurate predictions. However, if the data are random, the true theory is complex, the sample is small, and the above recipe for using a theory for predictive purposes is followed, then *a false, overly simplified theory can predict more accurately than the true theory* — e.g., even if God were to inform one that the true law is a degree 10 polynomial, one might prefer, on grounds of predictive accuracy, to derive predictions from a linear law. That surprising fact opens the door to an alternative, non-circular explanation of Ockham's razor in terms of predictive accuracy. The basic idea applies to accuracy in general, not just to accurate prediction. Consider, for example, a marksman shooting at a target. To keep our diagrams as elementary as possible, assume that the marksman is a Flatlander who exists entirely in a two-dimensional plane, so that the target is one-dimensional. There is a wall (line) in front of the marksman and the bull's eye is a distinguished point θ^* on that line. Each shot produced by the marksman hits the wall at some point $\hat{\theta}$, so it is natural to define the *squared error* of shot $\hat{\theta}$ as $(\hat{\theta} - \theta^*)^2$ (figure 3.a). Then for n shots, the average of the squared errors of the n points is a reflection of the marksman's accuracy, because the square function keeps all the errors positive, so none of them cancel.⁷ If one thinks of

⁷One could also sum the absolute values of the errors, but the square function is far more commonly used in statistics.

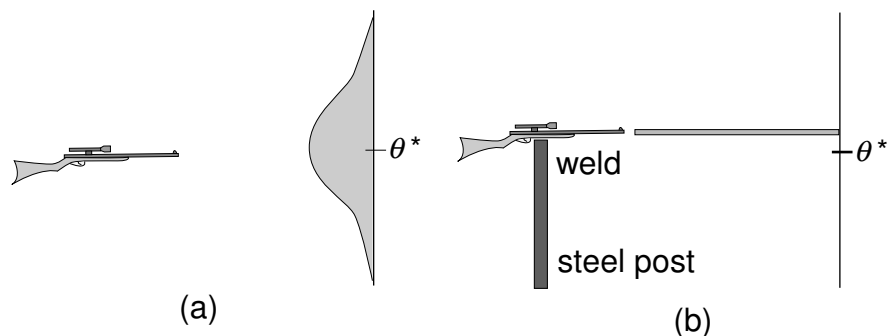


Figure 3. Firing range

the marksman's shots as being governed by a probability distribution reflecting all the stray causes that affect the marksman on a given shot, then one can explicate the marksman's *accuracy* as the expected or *mean* squared error (MSE) of a single shot with respect to distribution P :

$$MSE_P(\hat{\theta}, \theta^*) = \text{Exp}_P(\hat{\theta} - \theta^*)^2.$$

The MSE is standardly factored in a revealing way into a formula known as the *bias-variance trade-off* [Wasserman, 2004]:

$$MSE_P(\hat{\theta}, \theta^*) = \text{Bias}_P(\hat{\theta}, \theta^*)^2 + \text{Var}_P(\hat{\theta}),$$

where $\text{Bias}_P(\hat{\theta}, \theta^*)$ is defined as the deviation of the marksman's average or expected shot from the bull's eye θ^* :

$$\text{Bias}_P(\hat{\theta}, \theta^*) = \text{Exp}_P(\hat{\theta}) - \theta^*;$$

and the *variance* $\text{Var}_P(\hat{\theta})$ is defined as the expected distance of a shot from the average shot:

$$\text{Var}_P(\theta) = \text{Exp}_P((\hat{\theta} - \text{Exp}_P(\hat{\theta}))^2).$$

Bias is a systematic tendency to hit to a given side of the bull's eye, whereas variance reflects spread around the marksman's expected or average shot. Even the best marksman is subject to some variance due to pulse, random gusts of wind, etc., and the variance is amplified systematically as distance from the target increases. In contrast, diligent aim, proper correction of vision, etc. can virtually eliminate bias, so it seems that a marksman worthy of the name should do everything possible to eliminate bias. But that argument is fallacious. Consider the extreme strategy of welding the rifle to a steel post to eliminate variance altogether (figure 3.b). In light of the bias-variance trade-off, the welded rifle is more accurate than honest aiming as long as the squared bias of the welded rifle is less than the variance of the marksman's unconstrained aim. If variance is sufficiently high (due to distance

from the target, for example), the welded rifle can be more accurate, in the MSE sense, than skillful, unrestricted aim even if the weld *guarantees a miss*. That is the key insight behind the over-fitting argument.

Welding the rifle to a post is draconian. One can imagine a range of options, from the welded rifle, through various, successively less constraining clamps, to unconstrained aim. For a fixed position θ^* of the bull's eye, squared bias goes down and variance goes up as aiming becomes less constrained. The minimum MSE (among options available) occurs at a "sweet spot" where the sum of the curves achieves a minimum. Aiming options that are sub-optimal due to high bias are said to *under-aim* (the rifle's aim is too constrained) and aiming options that are sub-optimal due to high variance are said to *over-aim* (the rifle's aim is not constrained enough).

So far, the welded rifle strategy looks like a slam-dunk winner over all competing strategies — just hire an accurate welder to obtain a perfect score! But to keep the contest sporting, the target can be concealed behind a curtain until all the welders complete their work. Now the welded rifles still achieve zero variance or spread, but since bias depends on the bull's eye position θ^* , which might be anywhere, the welding strategy cannot guarantee any bound whatever on bias. The point generalizes to other, less draconian constraints on aim — prior to seeing the target there is no guarantee how much extra bias such constraints would contribute to the shot. One could lay down a prior probability reflecting about where the organizers might have positioned the target, but classical statisticians refuse to consider them unless they are grounded in knowledge of objective chance.

Empirical prediction of random quantities is closely analogous to a shooting contest whose target is hidden in advance. The maximum likelihood estimate $\hat{\theta}(T, E')$ is a function of random sample E' and, hence, has a probability distribution P^* that is uniquely determined by the true, sampling distribution P_{θ^*} . Thus, $\hat{\theta}(T, E')$ is like a stochastic shot $\hat{\theta}$ at bull's eye θ^* . When the MLE is taken with respect to the completely unconstrained theory $T_1 = \{P_\theta : \theta \in \Theta\}$, it is known in many standard cases that the MLE is unbiased: i.e., $\text{Bias}_{P^*}(\hat{\theta}(T_1, E'), \theta^*) = 0$. Thus, the MLE based on the complex, unconstrained theory is like the marksman's free aim at the bull's eye. How can that be, when the scientist can't see the bull's eye θ^* she is aiming at? The answer is that *nature* aims the rifle straight at θ^* ; the scientist merely chooses whether the rifle will be welded or not and then records the result of the shot. Similarly, the MLE with respect to constrained theory $T_0 = \{P_{\theta_0}\}$ is like shooting with the welded rifle — it has zero variance but no guarantee whatever regarding bias. For a fixed parameter value θ^* and for theories ordered by increasing complexity, there is a "sweet spot" theory T that maximizes accuracy by optimally trading bias for variance. Using a theory simpler than T reduces accuracy by adding extra bias and is called *under-fitting* whereas using a theory more complex or unconstrained than T reduces accuracy by adding variance and is called *over-fitting*. Note that over-fitting is defined in terms of the bias-variance trade-off, which is relative to sample size, and definitely *not* in terms of distinguishing genuine trends from mere noise, as some motivational discussions

seem to suggest (e.g., [Forster and Sober, 1994]).

To assume a priori that θ_0 is sufficiently close to θ^* for the MLE based on T_0 to be more accurate than the MLE based on T_1 is just another way to beg the question in Ockham's favor. But the choice between basing one's MLE on T_0 or on T_1 is a false dilemma — Ockham's razor says to presume no more complexity than necessary, rather than to presume no complexity at all, so it is up to Ockham's razor to say how much complexity *is* necessary to accommodate sample E' . To put the same point another way, Ockham's razor is not well-defined in statistical contexts until one specifies a formula that *scores* theories in a manner that rewards fit but taxes complexity. One such formula is the Akaike [1973] information criterion (AIC), which ranks theories (lower is better) relative to a given sample E' in terms of the remarkably tidy and suggestive formula:

$$\text{AIC}(T, E) = \text{badness of fit of } T \text{ to } E + \text{complexity of } T,$$

where theoretical complexity is the number of free parameters in T and badness of fit is measured by: $-\ln(P_{\hat{\theta}(T, E')}(E'))$.⁸

Choosing T so as to minimize the AIC score computed from sample E' is definitely one way to strike a precise balance between simplicity and fit. The official theory behind AIC is that the AIC score is an unbiased estimate of a quantity whose minimization would minimize MSE [Wasserman, 2004]. That sounds remotely comforting, but it doesn't cut to the chase. Ultimately, what matters is the MLE of the whole strategy of *using* AIC to choose a model and then computing the MLE of the model so chosen. To get some feel for the MLE of the AIC strategy, itself, it is instructive to return to the firing line. Recall that the MLE based on T_0 is like a shot from the welded rifle that always hits point θ_0 and the MLE based on T_1 is like honest, unbiased aiming at the bull's eye after the curtain rises. Using AIC to decide which strategy to employ has the effect of *funneling* shots that fall within a fixed distance r from θ_0 *exactly* to θ_0 — call r the *funnel radius*. So on the firing range, AIC could be implemented by making a sturdy funnel of radius r out of battleship plate and mounting it on a firm post in the field so that its spout lines up with the point θ_0 (figure 4). The funnel is a welcome sight when the curtain over the target rises and θ_0 is seen to line up with the bull's eye θ^* , because all shots caught by the funnel are deflected to more accurate positions. In that case, one would like the funnel to have an infinite radius so as to redirect every shot to the bull's eye (which is decision-theoretically identical to welding the rifle to hit point θ_0). The funnel is far less welcome, however, if the intended target is barely obscured by the edge of the funnel, for then then accurate shots get deflected or biased away from the bull's eye, with possibly dire results if the target happens to be hostile (fig. 5). In that case, one would prefer the funnel to have radius 0 (i.e., to get rid of it altogether).

⁸Recall that the MLE $\hat{\theta}(T, E')$ is the value of free parameter θ in theory T that maximizes $P_{\theta}(E')$, so $P_{\hat{\theta}(T, E')}(E')$ is the best likelihood that can be obtained from T for sample E' . Now recall that $-\ln$ drops monotonically from ∞ to 0 over the half-open interval $(0, 1]$.

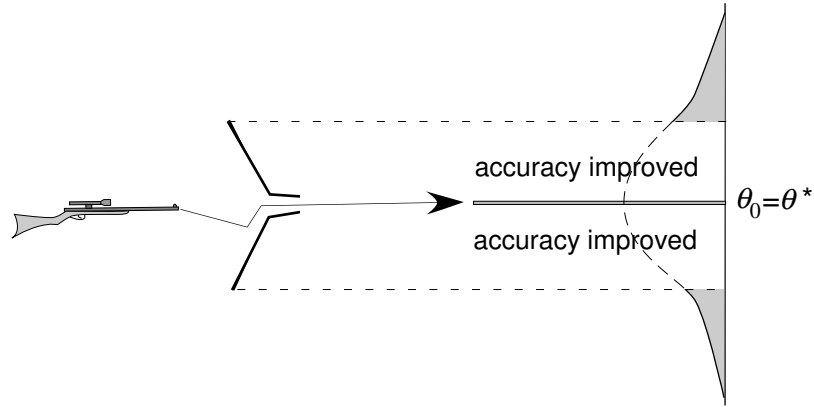


Figure 4. Ockham funnel, best case

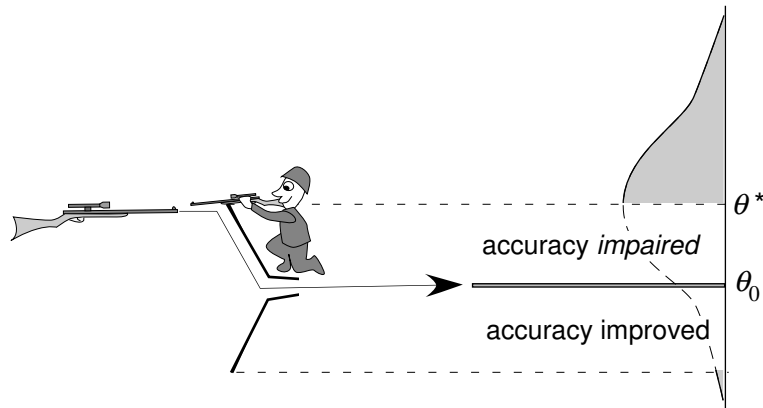


Figure 5. Ockham funnel, worst case

More generally, for each funnel radius r from 0 to infinity, one can plot the funnel's MSE over possible bull's eye positions θ^* in order to portray the methods as decision-theoretic acts with MSE as the loss and θ as the state of the world (fig. 6).⁹ How, then, does one choose a funnel radius r ? Proponents of AIC sometimes speak of typical or anomalous performance, but that amounts to a tacit appeal to prior probabilities over parameter values, which is out of bounds for classical statisticians when nothing is known a priori about the prior location of the bull's eye. One prior-free decision rule is to eliminate *dominated* alternatives, but none of the options in figure 6 is dominated — larger funnels do better as θ^* approaches θ_0 and smaller ones do better as θ^* diverges from θ_0 . Another prior-free decision rule is to choose a *minimax* strategy, i.e., a strategy whose maximum MSE, over all

⁹For computer plots of such curves, cf. [Forster, 2001].

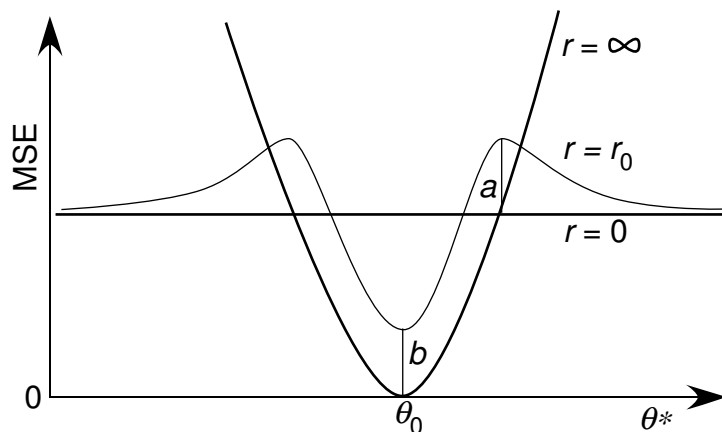


Figure 6. Ockham funnel decision problem

possible values of θ^* is minimal, over all alternative strategies under consideration. Alas, from figure 6, it is clear that the unique minimax solution among the available options is $r = 0$, which corresponds to estimation using the most *complex* theory — hardly a ringing endorsement for Ockham’s razor. There is, however, at least one prior-free decision rule that favors a non-extremal funnel diameter $0 < r < \infty$. The *regret* of an option at θ is the difference between the MSE of the option at θ and the minimum MSE over all alternative options available at θ . The *minimax regret* option minimizes worst-case regret. As r goes to infinity, regret a goes up against $r = 0$ and as r goes to 0 the regret b goes up against $r = \infty$. So there must be a “sweet” value r^* of r that minimizes a, b jointly and that yields a minimax regret solution. Then r^* can be viewed as the right balance between simplicity and fit, so far as minimax regret with respect to predictive inaccuracy is concerned. In some applications, it can be shown that AIC is approximately the same as the minimax regret solution when the difference in model complexity is large [Goldenschluger and Greenshtein, 2000]. AIC is just one representative of a broad range of funnel-like techniques motivated by the over-fitting argument, including cross-validation [Hjorth, 1994], Mallows’ [1973] statistic, minimum description length [Grünwald, 2007], minimum message length, and structural risk minimization [Vapnik, 1995].

There are, of course, some objections to the over-fitting argument. (1) The argument irrevocably ties Ockham’s razor to randomness. Intuitively, however, Ockham’s razor has to do with uniformity of nature, conservation laws, symmetry, sequential patterns, and other features of the universe that may be entirely deterministic and discretely observable without serious concerns about measurement error.

(2) Over-fitting arguments are sometimes presented vaguely in terms of “minimizing” MSE, without much attention to the awkward decision depicted in figure 6 and the consequent need to invoke either prior probabilities or minimax regret as

a decision rule.¹⁰ In particular, figure 6 should make it clear that computer simulations of Ockham strategies at “typical” parameter values should not be taken seriously by classical statisticians, who reject prior probabilistic representations of ignorance.

(3) MSE can be challenged as a correct explication of accuracy in some applications. For an extreme example, suppose that an enemy soldier is aiming directly at you. There happens to be a rifle welded to a lamp post that would barely miss your opponent and another, perfectly good rifle is lying free on the ground. If you value your life, you will pick up the rifle on the ground and aim it earnestly at your opponent even if you know that the welded rifle has lower MSE with respect to the intended target. For that reason, perhaps, military marksmanship is scored in terms of hits vs. misses on a human silhouette [U.S. Army, 2003] rather than in terms of MSE from a geometrical bull’s eye.

(4) Finally, the underlying sense of accurate prediction does not extend to predicting the results of novel policies that alter the underlying sampling distribution and, therefore, is too narrow to satisfy even the most pragmatic instrumentalist. That important point is developed in detail in the following section on causal discovery and prediction.

4 OCKHAM’S CAUSAL RAZOR

Suppose that one employs a model selection technique justified by the over-fitting argument to accurately estimate the incidence of lung cancer from the concentration of nicotine on teeth and suppose that a strong statistical “link” is found and reported breathlessly in the evening news. Nothing in the logic of over-fitting entails that the estimated correlation would accurately predict the cancer-reducing efficacy of a public tooth-brushing subsidy, for enactment of the policy would *change* the underlying sampling distribution so as to sever the “link”. Getting the underlying causal theory wrong can make even the most accurate predictions about the actual population useless for predicting the counterfactual results of enacting new policies that alter the population.

¹⁰Readers familiar with structural risk minimization (SRM) may suspect otherwise, because SRM theory is based on a function $b(\alpha, n, c)$ such that with worst-case chance $1 - \alpha$, the true MSE of using model T of complexity c for predictive purposes is less than $b(\alpha, n, c)$ [Vapnik, 1995]. The SRM rule starts with a fixed value $\alpha > 0$ and sample size n and a fixed sequence of models of increasing complexity and then chooses for predictive purposes (at sample size n) the model whose worst-case MSE bound $b(\alpha, n, c)$ is least. Note, however, that the bound is valid only when the model in question is selected and used for predictive purposes *a priori*. Since b can be expressed as a sum of a measure of badness of fit and a term taxing complexity, SRM is just another version of an Ockham funnel (albeit with a diameter larger than that of AIC). Therefore, the MSE of SRM will be higher than that of the theory SRM selects at the “bumps” in MSE depicted in figure 6. So the (short-run) decision theory for SRM ultimately poses the same problems as the decision theory for AIC. In the long run, SRM converges to the true model and AIC does not but, as has already been explained, long-run convergence does not explain Ockham’s razor.

A possible response is that causal conclusions require controlled, randomized trials, in which case the sample is already taken from the modified distribution and the logic of over-fitting once again applies. But controlled experiments are frequently too expensive or too immoral to perform. Happily, there is an alternative to the traditional dilemma between infeasible experiments and causal skepticism: recent work on causal discovery [Spirtes *et al.*, 2000; Verma and Pearl, 1991] has demonstrated that there *is*, after all, a sense in which *patterns* of correlations among several (at least three) variables can yield conclusions about causal orientation. The essential idea is readily grasped. Let $X \rightarrow Y$ abbreviate the claim that X is a direct cause of Y . Consider the causal situations depicted in figure 7. It is helpful to think of variables as measurements of flows in pipes and of causal

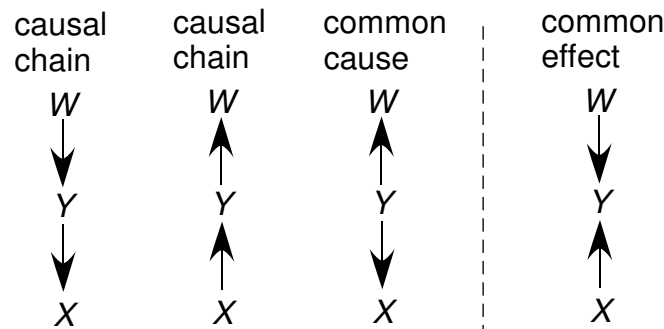


Figure 7. Causal situations

relation $X \rightarrow Y$ as a pipe with water flowing from flow meter X to flow meter Y [Heise, 1973]. In the causal chain $W \rightarrow Y \rightarrow X$, we have three meters connected by a straight run of pipe, so it is clear that information about one meter's reading would provide some information about the other meter readings. But since W informs about X only in virtue of providing information about Y , knowledge of X provides no *further* information about W than Y does — in jargon, X is independent of W *conditional on* Y . By symmetrical logic, the same holds for the inverted chain $X \rightarrow Y \rightarrow W$. The common cause situation $W \leftarrow Y \rightarrow X$ is the same: W provides information about X only in virtue of providing information about the common cause Y so, conditional on Y , W is independent of X . So far, the situation is pretty grim — all three situations imply the same conditional dependence relations. But now consider the common effect $W \rightarrow Y \leftarrow X$. In that case, W provides no information about X , since the two variables are causally independent and could be set in any combination. But *conditional on* Y , the variable X does provide some information about W because both W and X must collaborate in a specific manner to produce the observed value of Y . Thus, the common effect implies a pattern of dependence and conditional dependence distinct from the pattern shared by the remaining three alternatives. Therefore, common effects and their consequences can be determined from observable conditional dependencies

holding in the data.

There is more. A standard skeptical concern is the possibility that apparent causal relation $W \rightarrow X$ is actually produced by a latent or unobserved common cause $W \leftarrow C \rightarrow X$ (just as a puppeteer can make one puppet appear to speak to another). Suppose, for example, that Z is a direct effect of common effect Y . Consider the skeptical alternative in which $Y \rightarrow Z$ is actually produced by a hidden common cause C of Y and Z (fig. 8). But the skeptical alternative leaves

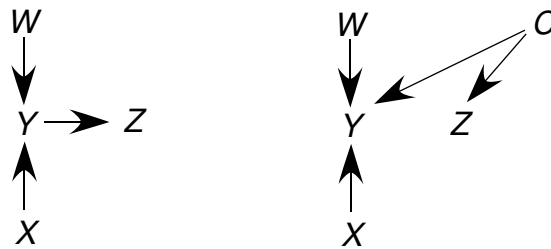


Figure 8. Confounding hidden cause

a footprint in the data, since in the confounded situation Z and W are dependent given Y (since W provides some information about C given Y and C , as a direct cause of Y , provides some information about Y). In the non-confounded situation, the reverse pattern of dependence obtains: W is independent of Z given Y because Z yields information about W only in virtue of the information Z yields about Y . So it is possible, after all, to obtain non-confoundable causal conclusions from non-experimental data.

Given the *true causal theory* relating some variables of interest and given an accurate estimate of the free parameters of the theory, one can obtain accurate *counterfactual* predictions according to a natural rule: to predict the result of *intervening* on variable X to force it to assume value x , first *erase* all causal arrows into X , holding other theory parameters fixed at their prior values and now use the modified theory to predict the value of the variable of interest, say Y . Thus, for example, if X is itself an effect, forcing X to assume a value will break all connections between X and other variables, so the values of other variables will be predicted not to change, whereas if X is a cause, forcing X to assume a value will alter the values of the effects of X . The moral is that accurate counterfactual predictions depend on inferring the causal model corresponding to the true causal relations among the variables of interest — causal hypotheses are not merely a way to constrain noise in actual empirical estimates.

Causal discovery from non-experimental data depends crucially on Ockham's razor in the sense that causal structure is read off of patterns of conditional correlations and there is a bias toward assuming that a conditional correlation is zero. That is a version of Ockham's razor, because non-zero conditional correlations are free parameters that must be estimated in order to arrive at predictions. Absent

any bias toward causal theories with fewer free parameters, one would obtain no non-trivial causal conclusions, since the most complex theory entails a causal connection between each pair of variables and all such causal networks imply exactly the same patterns of conditional statistical dependence. But since the over-fitting argument does not explain how such a bias conduces to the identification of true causal structure, it fails to justify Ockham's razor in causal discovery from non-experimental data. The following, novel, alternative explanation does.

5 EFFICIENT PURSUIT OF THE TRUTH

To summarize the preceding discussion, the puzzle posed by Ockham's razor is to explain how a fixed bias toward simplicity is conducive to finding true theories. The crux of the puzzle is to specify a concept of truth-conduciveness according to which Ockham's razor is more truth-conducive than competing strategies. The trouble with the standard explanations is that the concepts of truth-conduciveness they presuppose are respectively either too weak or too strong to single out Ockham's razor as the most truth-conducive inferential strategy. Mere convergence to the truth is too weak, since alternative strategies would also converge to the truth. Reliable indication or tracking of the truth in the short run, on the other hand, is so strict that Ockham's razor can be shown to achieve it only by circular arguments (the Bayes factor argument) or by substituting accurate, non-counterfactual predictions for theoretical truth (over-fitting argument).

There is, however, a third option. A natural conception of truth-conduciveness lying between reliable indication of the truth and mere convergence to the truth is *effective pursuit* of the truth. Effective pursuit is not necessarily direct or even

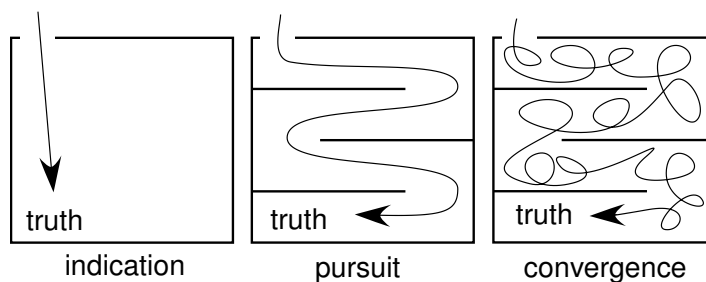


Figure 9. Three concepts of truth-conduciveness

bounded in time or complexity (e.g., pursuit through a labyrinth of unknown extent). But neither is effective pursuit entirely arbitrary — gratuitous course reversals and cycles should evidently be avoided. Perhaps, then, Ockham's razor is the best possible way to pursue theoretical truth, even though simplicity cannot point at or indicate the true theory in the short run and even though alternative methods would have converged to the truth eventually.

In the pursuit of truth, a course reversal occurs when one *retracts* or takes back an earlier belief, as when Ptolemaic theory was replaced by Copernican theory. It caused a sensation when Thomas Kuhn [1962] argued that scientific change essentially involves losses or retractions of content and invoked the tremendous cognitive cost of retooling entailed by such changes to explain retention of one's theoretical ideas in the face of anomalies. Emphasis on cognitive retooling may suggest that retractions are a merely "pragmatic" cost, but deeper considerations point to their epistemic relevance. (1) Potential retractions have been invoked in philosophical analyses of the concept of knowledge since ancient times. Plato traced the essential difference between knowledge and true belief to the *stability* of knowledge in his dialogue *Meno* and subsequent authors have expanded upon that theme in attempts to provide indefeasibility accounts of knowledge. For example, suppose that one has good but inconclusive evidence E that Jones owns a Ford when, in fact, only Smith has one and that one believes, on the basis of E that either Smith or Jones owns a Ford [Gettier, 1963]. It seems that the inferred belief is not known. Indefeasibility analyses of knowledge (e.g., [Lehrer, 1990]) attempt to explain that judgment in terms of the the potential for retracting the disjunctive belief when the grounds for the false belief are retracted. (2) Deductive logic is *monotonic*, in the sense that additional premises never yield fewer conclusions. Inductive logic is non-monotonic, in the sense that additional premises (new empirical evidence) can undermine conclusions based on earlier evidence. Non-monotonicities are retractions of earlier conclusions, so to minimize retractions as far as finding the truth allows is to approximate deduction as closely as finding the truth allows. (3) In mathematical logic, a formal proof system provides a computable, positive test for theorem-hood — i.e., a Turing machine that *halts* with "yes" if and only if the given statement is a theorem. The halting condition essentially bounds the power of sound proof systems. But nothing other than convention requires a Turing machine to halt when it produces an answer — like human scientists and mathematicians, a Turing machine can be allowed to output a sequence of revised answers upon receipt of further inputs, in an unending loop. Hilary Putnam [1965] showed that Turing machines that are allowed to retract prior answers at most $n + 1$ times prior to convergence to the truth can do more than Turing machines that are allowed to retract at most n times. Furthermore, formal verifiability (halting with "yes" if and only if ϕ is a theorem) is computationally equivalent to finding the right answer with one retraction starting with "no" (say "no" until the verifier halts with "yes" and then retract to "yes"), refutation is computationally equivalent to finding the right answer with one retraction starting with "yes" and formal decidability is computationally equivalent with finding the right answer with no retractions. So retraction bounds are a natural and fundamental *generalization* of the usual computational concepts of verifiability, refutability, and decidability [Kelly, 2004]. The idea is so natural from a computational viewpoint that theoretical computer scientists interested in inductive inference have developed an elaborate theory of inductive retraction complexity [Case and Smith, 1983; Freivalds and Smith, 1993]. (4) Fi-

nally, and most importantly, the usual reason for distinguishing epistemic from merely pragmatic considerations is that the former are truth-conducive and the latter conduce to some other aim (e.g., wishful thinking is happiness-conducive but not truth-conducive). Retraction-minimization (i.e., optimally direct *pursuit* of the truth) is part of what it *means* for an inductive inference procedure to be truth-conducive, so retractions are a properly epistemic consideration.

Additional costs of inquiry may be considered in addition to retractions: e.g., the number and severity of erroneous conclusions are a natural epistemic cost, and the times elapsed until errors and/or retractions are finally avoided. But retractions are crucial for elucidating the elusive, truth-finding advantages of Ockham's razor, for reasons that will become apparent below.

6 EMPIRICAL SIMPLICITY DEFINED

In order to prove anything about Ockham's razor, a precise definition of empirical simplicity is required. The basic approach adopted here is that empirical complexity is a reflection of *empirical effects* relevant to the theoretical inference problem addressed. Thus, empirical complexity is not a mere matter of notation, but it is relative to the kind of truth one is trying to discover. An *empirical effect* is just a verifiable proposition — a proposition that might never be known to be false, but that comes to be known, eventually, if it is true. For example, Newton [1726] tested the identity of gravitational and inertial mass by swinging large pendula filled with identical weights of different kinds of matter and then watching to see if they ever went noticeably out of phase. If they were not identical in phase, the accumulating phase difference would have been noticeable eventually. Particle reactions are another example of empirical effects that may be very difficult to produce but that, once observed, are known to occur. Again, two open intervals through which no constant curve passes constitute a *first-order effect*, three open intervals through which no line passes constitute a *second-order effect*, and so forth (fig. 10.a-c).¹¹ Effects can be arbitrarily small or arbitrarily arcane, so they can take arbitrarily long to notice.

Let E be a countable set of possible effects.¹² Let the *empirical presupposition* K be a collection of finite subsets of E . It is assumed that each element of K is a possible candidate for the set of all effects that will ever be observed. The *theoretical question* Q is a partition of K into sets of finite effect sets. Each partition cell in Q corresponds to an *empirical theory* that might be true. Let T_S denote the (unique) theory in Q that corresponds to finite effect set S in K . For example, the hypotheses of interest to Newton can be identified, respectively, with the absence

¹¹That is very close to Karl Popper's [1968] discussion of degrees of falsifiability, except that his approach assumed exact data rather than intervals. The difference is crucial to the following argument.

¹²Effects are here assumed to be primitive. A more ambitious and explanatory approach, in which the problem (K, Q) is presented in terms of mere observations and effects are *constructed* from the topological structure of (K, Q) is developed in [Kelly, 2007; 2008].

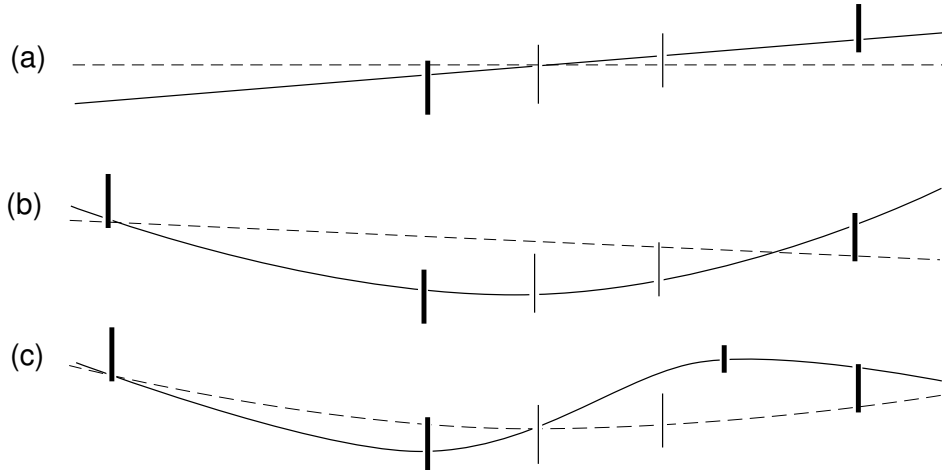


Figure 10. First, second, and third order effects

of an out-of-phase effect or the eventual appearance of an out-of-phase effect. The hypothesis that the degree of an unknown polynomial law is n can similarly be identified with an effect — refutation of all polynomial degrees $< n$. In light of the above discussion of causal inference, each linear causal network corresponds to a pattern of partial correlation effects (note that conditional dependence is noticeable, whereas independence implies only absence of verification of dependence). Each conservation theory of particle interactions can be identified with a finite set of effects corresponding to the discovery of reactions that are not linearly dependent on known reactions [Schulte, 2000; Luo and Schulte, 2006].¹³ The pair (K, Q) then represents the scientist's *theoretical inference problem*. The scientist's aim is to infer the true answer to Q from observed effects, assuming that the true effect set is in K .

Now empirical simplicity will be defined with respect to inference problem (K, Q) . Effect set S *conflicts* with S' in Q if and only if T_S is distinct from $T_{S'}$. Let π be a finite sequence of sets in K . Say that π is a *skeptical path* in (K, Q) if and only if for each pair S, S' of successive effect sets along π , effect set S is a subset of S' and S conflicts with S' in Q . Define the *empirical complexity* $c(S)$ of effect set S relative to (K, Q) to be $a - 1$, where a denotes the length of the

¹³Ptolemy's theory can be tuned to duplicate Copernican observations for eternity, so the two theories share an effect set. The proposed framework does not apply to that case unless it is assumed that a Ptolemaic universe would not duplicate Copernican appearances for eternity. One reason for ruling out the possibility of an eternally perfect illusion is that no possible method could converge to the truth in such an empirical world, so even *optimally* truth-conducive methods fail in such worlds. The proposed account focuses, therefore, on *empirical adequacy* (i.e., consistency with all possible experience), rather than on inaccessible truths transcending all possible experience.

longest skeptical path through (K, Q) that terminates in S .¹⁴ Let the empirical complexity $c(T)$ of theory T denote the empirical complexity of the least complex effect set in T .

A skeptical path through (K, Q) poses an *iterated* problem of induction to a would-be solver of problem (K, Q) , since every finite sequence of data received from a given state on such a path might have been produced by a state for which some alternative answer to Q is true. That explains why empirical complexity ought to be relevant to the problem of finding the true theory. Problem-solving effectiveness always depends on the intrinsic difficulty of the problem one is trying to solve and the depth of embedding of the problem of induction determines how hard it is to find the truth by inductive means. Since syntactically defined simplicity (e.g., [Li and Vitanyi, 1993]) can, but need not, latch onto skeptical paths in (K, Q) , it does not provide such an explanation.

Let e be some input information. Let S_e denote the set of all effects verified by e . Define the *conditional* empirical complexities $c(S | e)$, $c(T | e)$ in (K, Q) just as before, but with respect to the *restricted problem* (K_e, Q) , where K_e denotes the set of all effect sets S in K such that S_e is a subset of S .

7 INQUIRY AND OCKHAM'S RAZOR

The next step is to provide a precise model of inquiry concerning the problem (K, Q) . A *stream of experience* is an input sequence that presents some finite set (possibly empty) of empirical effects at each stage. Let S_w denote the effect set whose effects are exactly those presented by stream of experience w . An *empirical world* for K is an infinite stream of experience such that S_w is an element of K . Let T_w denote T_{S_w} . An *empirical strategy* or *method* for the scientist is a mapping M from finite streams of experience to theories in Q or to '?', which corresponds to a skeptical refusal to choose any theory at the moment. Let $w|i$ be the initial segment of length i of empirical world w . Method M *converges to the truth* in problem (K, Q) if and only if for each empirical world w for K :

$$\lim_i M(w|i) = T_w.$$

Methodological principles can be viewed as restrictions on possible scientific strategies. For example, strategy M is *logically consistent* if and only if S_e is a subset of $M(e)$, for each finite input sequence e . Strategy M satisfies *Ockham's razor* if and only if M chooses no theory unless it is the uniquely simplest theory compatible with experience, where simplicity is relative to (K, Q) , as described above. As stated, Ockham's razor allows for any number of counter-intuitive vacillations between some theory T and '?'. A natural, companion principle requires that one hang onto one's current theory choice T as long as T remains uniquely

¹⁴The reason for subtracting 1 is to assign complexity 0 to the simplest states, since each such state S is reached by a path (S) of length 1. There is a maximum precedence path to S because of the assumption that S is finite.

simplest among the theories compatible with experience.¹⁵ Call that principle *stalwartness*. A third principle is *eventual informativeness*, which says that the method cannot stall with ‘?’ for eternity. A *normal* Ockham method is a method that satisfies Ockham’s razor, stalwartness, and eventual informativeness. The first bit of good news is:

PROPOSITION 1. *Normal Ockham methods are logically consistent and converge to the truth.*

Proof. Let M be a method that is normally Ockham for (K, Q) . Logical consistency follows immediately from Ockham’s razor. For convergence, let w be an empirical world for K . Since the effect set S_w presented by w is finite, it follows that only finitely many effect sets in K are simpler than S_w . After some finite stage of inquiry, the finitely many effects in S_w are presented by w and from that point onward, S_w is the uniquely simplest state compatible with experience. At some later stage along w , method M produces some answer to Q , by eventual informativeness. Ockham’s razor implies that the answer produced is T_S . Stalwartness guarantees that T_S is never again dropped along w . $\dashv$

8 A BASIC OCKHAM EFFICIENCY THEOREM

The trouble with proposition 1 is that Ockham’s razor is not *necessary* for mere convergence to the truth: e.g., start out guessing theory T_{1000} of complexity 1000 without even looking at the data for 1000 stages of inquiry and then switch to a normal Ockham strategy. *Efficient* convergence rules out all such alternative strategies.

Let $r(M, w)$ denote the total number of times along w that M produces an output that does not entail the output produced at the immediately preceding stage (assume that ‘?’ is entailed by every output). If e is a finite stream of experience, define $C_j(e)$ to be the set of all worlds w for K that extend e and that satisfy $c(S_w \mid e) = j$. Call $C_j(e)$ the j th *empirical complexity set* for (K, Q) given e . Define $r_j(M \mid e)$ to be the least upper bound of $r(M, w)$ with respect to all worlds w in complexity set $C_j(e)$ (the least upper bound is ∞ if no finite upper bound exists). Thus, $r(M \mid e)$ is the *worst-case* retraction cost of M given e and given that the actual empirical complexity of the world is exactly j .

Next, compare alternative, convergent, logically consistent strategies in terms of worst-case retractions, over all possible world complexities. Let e_- denote the result of deleting the last entry in e (if e is the empty sequence, then $e_- = e$). Let M, M' be two strategies. Say that M is as *efficient* as M' given e if and only if $r_j(M \mid e) \leq r_j(M' \mid e)$, for each complexity set $C_j(e)$. Say that convergent, logically consistent strategy M is *efficient* given e if and only if M is as efficient as an arbitrary convergent, logically consistent strategy M' that agrees with M

¹⁵Since theories are linearly ordered by empirical complexity in this introductory sketch, uniqueness is trivial, but the argument can be extended to the non-unique case, with interesting consequences discussed below.

along e_- . Inefficiency is a weak property — it entails only that M does worse than some convergent, logically consistent competitor over some complexity set $C_j(e)$. A much more objectionable situation obtains when $r_j(M' | e) > r_j(M | e)$, for each non-empty $C_j(e)$. In that case, say that M *strongly beats* M' given e . Strategy M' is *weakly beaten* by M when M does as well as M' over each non-empty complexity set and better in at least one. Then M' is strongly (weakly) beaten given e if and only if M' is strongly (weakly) beaten by some convergent, logically consistent competitor. A strong beating given e implies a weak beating which, in turn, implies inefficiency. Each of those properties is relative to available information e . Say that such a property holds *always* just in case it holds for each e compatible with K . It is now possible to state the most basic Ockham efficiency theorem:

THEOREM 2. *Assume that (i) K is totally ordered by empirical precedence and (ii) each theory is satisfied by a unique effect state. Define efficiency and beating with respect to all convergent, logically consistent methods. Then the following are equivalent:*

1. M is always normally Ockham;
2. M is always efficient in terms of retractions;
3. M is never strongly beaten in terms of retractions.

The proof has three, straightforward steps.

Step I. Let O be a normal Ockham strategy. Suppose that the scientist always employs some fixed normal Ockham strategy O . Let e of length i be the finite sequence of input data received so far. Let $r \leq i$ be the number of retractions performed by O along e_- . Let w be an empirical world in $C_j(e)$. By stalwartness, O retracts at most j times after stage i along w . Thus, $r_j(O | e) \leq r + j$ if O does not retract at i and $r_j(O | e) \leq r + j + 1$ otherwise (figure 11).

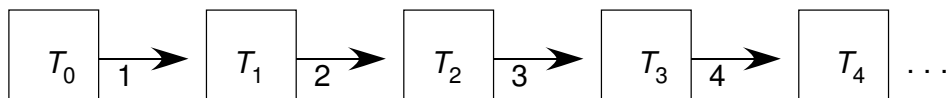


Figure 11. Sequential retractions of normal Ockham methods

Step II. Suppose that the scientist switches at stage i from normal Ockham strategy O to some arbitrary, convergent, logically consistent method M that agrees with O along e_- . Suppose that $C_j(e)$ is non-empty, so there exists skeptical path $(S_0, \dots, S_j)$ through (K, Q) . Nature can present M with an endless stream of data extending e that presents only effects true in S_0 until, on pain of failing to converge to the truth, M converges to T_{S_0} . Thus, if O happens to retract at stage i , then M retracts to T_{S_0} no sooner than i , since $M(e_-) = O(e_-)$. Thereafter,

nature can present just the effects true in S_1 followed by no more effects until, on pain of failing to converge to the truth, M switches to T_{S_1} . Iterate that argument until M produces T_{S_j} . Since the path is skeptical, it follows that M retracts at least j times after (possibly) retracting to T_{S_0} , so:

$$\begin{aligned} r_j(M \mid e) &\geq r + j + 1 \geq r_j(O \mid e) && \text{if } O \text{ retracts at } i; \\ r_j(M \mid e) &\geq r + j + 0 \geq r_j(O \mid e) && \text{otherwise.} \end{aligned}$$

So for each convergent, logically consistent M agreeing with O along e_- and for each j such that $C_j(e)$ is non-empty, we have that $r_j(O \mid e) \leq r_j(M \mid e)$. So O is *retraction efficient* given e . Since e is arbitrary in the preceding argument, O is *always* retraction efficient.

Step III. Finally, suppose that M violates Ockham's razor at the last entry of input sequence e compatible with K . Since M is logically consistent and effect sets are totally ordered, it follows that M produces a theory T more complex than the simplest theory T_{S_0} compatible with e . Since that is the first Ockham violation by M , we know that M did not also produce T_j at stage $i - 1$. Therefore, M retracts at i if O does. Suppose that $C_j(e)$ is non-empty. Let skeptical path $(S_0, \dots, S_j)$ witness that fact. Thereafter, as in the preceding paragraph, nature can force M to retract T back to T_{S_0} and can then force another j retractions. Note that O does not perform the (needless) retraction from T back to T_{S_0} (e.g., the retraction from T_4 to T_2 in figure 12), so:

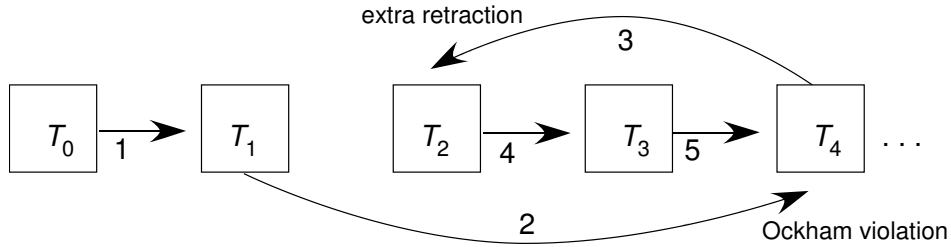


Figure 12. Ockham violator's extra retraction

$$\begin{aligned} r_j(M \mid e) &\geq r + j + 2 > r + j + 1 \geq r_j(O \mid e) && \text{if } O \text{ retracts at } i; \\ r_j(M \mid e) &\geq r + j + 1 > r + j + 0 \geq r_j(O \mid e) && \text{otherwise.} \end{aligned}$$

Thus, O *strongly beats* M at e in terms of retractions. Suppose, next, that M violates stalwartness given e . Then it is immediate that M retracts one extra time in each T_{S_i} compatible with e in comparison with O . Method M cannot violate eventual informativeness, since that would imply failure to converge to the truth. $\dashv$

Unlike over-fitting explanations, the Ockham efficiency theorem applies to deterministic questions. Unlike the Bayes factor explanation, the Ockham efficiency

theorem does not presuppose a question-begging prior bias in credence toward simple worlds — every world is as important as every other. The crux of any non-circular epistemic argument for Ockham’s razor is to explain why leaping to a needlessly complex theory makes one a bad truth-seeker *even if that theory happens to be true*. To see how the hard case is handled in the Ockham efficiency theorem, note that even if T_4 is true in figure 12, leaping straight to T_4 when experience refutes T_1 provides nature with a strategy to force one through the sequence of theories T_4, T_2, T_3, T_4 , which not only adds an extra retraction to the optimal sequence T_2, T_3, T_4 but also involves an embarrassing cycle away from T_4 and back to T_4 . In terms of the metaphor of pursuit, it is as if a heat-seeking missile *passed* its target and had to make a hairpin turn back to it — a performance likely to motivate some re-engineering.

Normal Ockham strategies do not *dominate* alternative strategies in the sense of having a better outcome in *every* possibility, since an Ockham violator can be saved from the embarrassment of an extra retraction (and look like a genius) if nature is kind enough to provide the anticipated empirical effects before she loses confidence in her complex theory. Nor are Ockham strategies *admissible*, in the sense of not being weakly dominated by an alternative method — indeed, *every* normal Ockham strategy is dominated in error times and retraction times by a strategy that stalls with ‘?’ for a longer time prior to producing an answer. That reflects the special structure of the problem of inductive inquiry — waiting longer to produce an informative answer avoids more possibilities for setbacks, but waiting forever precludes finding the truth at all. Nor are Ockham strategies *minimax* solutions, in the sense that they minimize worst-case overall cost, since the overall worst-case bound on each of the costs under consideration is infinity for arbitrary, convergent methods. The Ockham efficiency property is essentially a hybrid of admissibility and minimax reasoning. First, one partitions all problem instances according to empirical complexity and then one compares corresponding worst-case bounds over these complexity classes. The idea is borrowed from the standard practice for judging algorithmic efficiency [Gary and Johnson, 1979]. No interesting algorithm can find the answer for an arbitrarily large input under a finite resource bound, so inputs are routinely sorted by length and worst-case bounds over each size are compared. In the case of empirical inquiry, the inputs (worlds of experience) are all infinite, so length is replaced with empirical complexity.

9 STABILITY, ERRORS AND RETRACTION TIMES

Theorem 2 establishes that, in a specific sense, the normal Ockham path is the straightest path to the truth. But the straightest path also a narrow path that one might veer from inadvertently. Complex theories have been proposed because no simpler theory had yet been conceived of or because the advantages of a simpler theory were not yet recognized as such (e.g., Newton dismissed the wave theory of light, which was simpler than his particle theory, because he mistakenly thought it could not explain shadows). Theorem 2 does not entail that one should return to

the straightest path, having once departed from it. For example, suppose that at stage $i - 1$, method M violates Ockham's razor by producing needlessly complex theory T when T_{S_0} is the simplest theory compatible with experience. Let O be just like M prior to i and switch to a normal Ockham method thereafter. Then at stage i , method M *saves* a retraction compared to O by retaining T_k — nature can force a retraction back to T_m — but that is the same retraction O performs at i anyway. So the justification of normal Ockham strategies is *unstable* in the sense that retraction efficiency does not push an Ockham violator back onto the Ockham path after a violation has already occurred.

The persistent Ockham violator M does incur other costs. For example, M produces more false answers than O from stage i onward over complexity set $C_0(e)$, since O produces no false outputs after e_- along an arbitrary world in $C_0(e)$. Furthermore, both M and O commit unbounded errors, in the worst case over $C_j(e)$, if $C_j(e)$ is non-empty and $j > 0$. So returning to the normal Ockham fold weakly beats persistent violation, in terms of retractions and errors, at *each* violation.

It would be better to show that Ockham violators are strongly beaten at each violation. Such an argument can be given in terms of retraction *times*. The motivation is, again, both pragmatic and epistemic. Pragmatically, it is better to minimize the accumulation of applications of a theory prior to its retraction, even if that theory is true, since retraction occasions a reexamination of all such applications. Epistemically, belief that is retracted in the future does not count as knowledge even if it is true [Gettier, 1963]. It would seem, therefore, that more retractions in the future imply greater distance from knowledge than do fewer such retractions. Hence, in the sole interest of minimizing one's distance from the state of knowledge, one ought to get one's retractions over with as soon as possible.

Considerations of timing occasion the hard question whether a few very late retractions are worse than many early ones. Focus entirely on the easy (Pareto) comparisons in which total cost and lateness both agree. Let $(j_0, j_1, \dots, j_r)$ denote the sequence of times at which M retracts prior to stage i , noting that r also records the total number of retractions. Let σ, τ be such sequences. Say that σ is *as bad as* τ just in case there is a sub-sequence σ' of σ whose length is identical to the length of τ and whose successive entries are all at least as great as the corresponding entries in τ . Furthermore, σ is *worse than* τ if and only if σ is as bad as τ but τ is not as bad as σ . For example, $(2, 4, 8)$ is worse than $(3, 7)$, in light of the sub-sequence $(4, 8)$. The efficiency argument for O goes pretty much as before. Suppose that M violates Ockham's razor at e of length i . Let O be just like M along e_- and switch to a normal Ockham strategy from stage i onward. Let $(k_1, \dots, k_r)$ be the retraction times of both M and O along e_- . Suppose that $C_j(e)$ is non-empty, so there exists a skeptical path $(S_0, \dots, S_j)$ through (K_e, Q_e) . The hard case is the one in which O retracts at i and M does not. Since O is stalwart at i , it follows that $T = M(e_-) \neq T_{S_0}$. Nature can refuse to present new effects until M retracts T in favor of T_{S_0} , and can then force an arbitrarily late retraction for each step along the path $(S_0, \dots, S_j)$. Method O retracts at most j

times over $C_j(e)$ and retracts once at i in $C_0(e)$. Thus:

$$\begin{aligned} r_j(O \mid e) &\leq (k_0, k_1, \dots, k_r, i, \underbrace{\infty, \dots, \infty}_j) \\ &< (k_0, k_1, \dots, k_r, i+1, \underbrace{\infty, \dots, \infty}_j) \leq r_j(M \mid e). \end{aligned}$$

Note that the preceding argument never appeals to logical consistency, which may be dropped. The beating argument for stalwartness violators is easier, since one never saves a retraction by violating stalwartness. Again, violations of eventual informativeness are impossible for convergent methods, so now we have (cf. [Kelly, 2004]):

THEOREM 3. *Assume conditions (i) and (ii) of theorem 2. Define efficiency and beating with respect to the set of all convergent methods. Then the following are equivalent:*

1. *M is normally Ockham from e onward;*
2. *M is efficient in terms of retraction times and errors from e onward;*
3. *M is never weakly beaten in terms of retractions and errors from e onward;*
4. *M is never strongly beaten in terms of retraction times from e onward.*

A stronger version of Ockham's razor follows if one charges for expansions of belief or for elapsed time to choosing the true theory, for in that case one should avoid agnosticism and select the simplest theory at the very outset to achieve zero loss in the simplest theory compatible with experience. That conclusion seems too strong, however, confirming the intuition that when belief changes, the epistemically costly part is retracting the old belief rather than adopting the new one. This asymmetry between avoiding retractions as soon as possible and finding truth as soon as possible arises again, in a subtle way, when the Ockham Efficiency theorem is extended from theory choice to Bayesian degrees of belief.

10 EXTENSION TO BRANCHING SIMPLICITY

Sometimes, the theories of interest are not ordered sequentially by simplicity, in which case there may be more than one simplest theory compatible with experience. For example, suppose that the question is to find the true form of a polynomial law. For another example, let T_S be the theory that the true causal structure is compatible with exactly the partial statistical dependencies in set S . In the inference of linear causal structures with Gaussian error, the branching simplicity structure over models with three variables is exactly the lattice depicted in figure 13 (cf. [Chickering, 2003; Meek, 1995]).

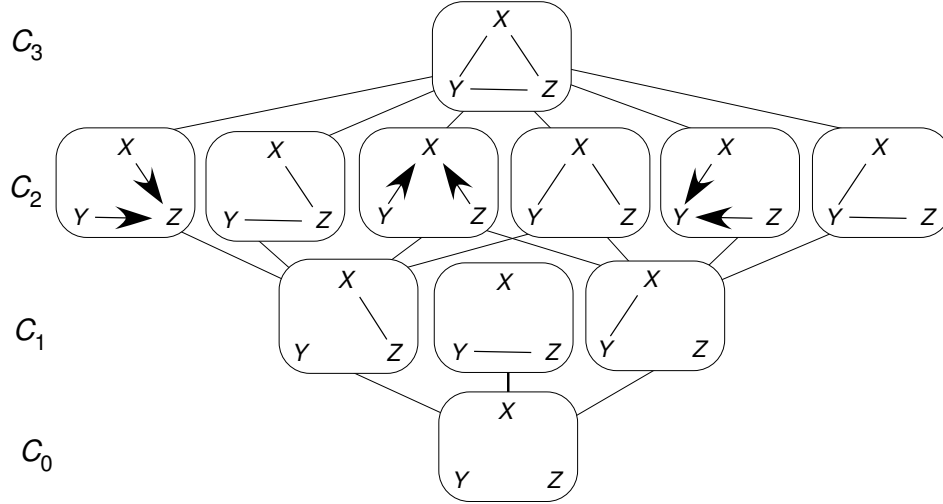


Figure 13. Simplicity for acyclic linear causal models

When there is more than one simplest theory compatible with experience, Ockham's razor seems to demand that one suspend judgment with '?' until nature winnows the field down to a unique theory. That judgment is enforced by efficiency considerations. Suppose that, as in the causal case (figure 13), no maximal, skeptical path is longer than another.¹⁶ Call that the *no short path* assumption. Then violating Ockham's razor by choosing one simplest theory over another incurs an extra retraction in every non-empty complexity set, since nature is free to make the *other* simplest theory appear true, forcing the scientist into an extra retraction. Thereafter, nature can force the usual retractions along a path that visits each non-empty complexity set $C_j(e)$, by the assumption that no path is short.

THEOREM 4 Kelly, 2007. *Theorem 3 continues to hold if (i) is replaced by the no short path assumption.*

Without the no short path assumption, methods that return to the Ockham path are no longer efficient, even in terms of retraction times. Suppose that T_0 and T_1 are equally simple, and that T_2 is more complex than T_1 but not more complex than T_0 . Then T_0 and T_1 both receive empirical complexity degree 0 and T_2 is assigned complexity degree 1. Suppose that method M has already violated Ockham's razor by choosing T_1 when T_0 is still compatible with experience. Alas,

¹⁶In the case of acyclic linear causal models with independently distributed Gaussian noise, it is a consequence of [Chickering, 2003] that the only way to add a new implied conditional dependence relationship is to add a new causal connection. Hence, each causal network with n causal connections can be extended by adding successive edges, so there are no short paths in that application and the strong argument for Ockham's razor holds.

sticking with the Ockham violation *beats* Ockham's retreating strategy in terms of retractions. For Ockham's retreat counts as a retraction in $C_0(e)$. Nature can still lure Ockham to choose T_0 and can force a further retraction to T_1 for a total of 2 retractions in $C_1(e)$. But strategy M retracts just once in $C_0(e)$ and once in $C_1(e)$. In terms of retraction times, there is a hard choice — the born-again Ockham strategy retracts early in $C_0(e)$ and retracts more times in $C_1(e)$.

One response to the short path problem is to question whether the short path really couldn't be extended — if all paths are infinite, there are no short paths. Polynomial theories can always be given another term. In causal networks, one can always study another variable that might have a weak connection with variables already studied. A second response is that the simplicity degrees assigned to theories along a short path are arbitrary as long as they preserve order along the path. The proposed definition of simplicity degrees ranks theories along a short complexity path as low as possible, but one might have ranked them as high as possible (e.g., putting T_0 in $C_1(e)$ rather than in $C_0(e)$), in which case the preceding counterexample no longer holds.¹⁷ That option is no longer available, however, if some path is infinite in length and another path is finite in length. The third and, perhaps, best response is to weaken Ockham's razor to allow for the selection of the theory at the root of the longer path. Violating that version of Ockham's razor still results in a strong beating in terms of retraction times and methods that satisfy it along with stalwartness at every stage are never strongly beaten. The third option becomes all the more compelling below, when it is entertained that some retractions count more than others due to the amount of *content* retracted.

11 WHEN DEFEAT DOES NOT IMPLY REFUTATION

The preceding efficiency theorems all assume that each theory is true of just one effect state. It follows that whenever an Ockham conclusion is *defeated* by new data, it is also *refuted* by that data. That is not necessarily the case, as when the question concerns whether polynomial degree is even or odd. A more important example concerns the status of a single causal relation $X \rightarrow Y$. Figure 14 presents a sequence of causal theories that nature can force every convergent method to produce. Focus on the causal relation between X and Y . Note that the orientation of the edge flips when the inferred common effect at Y is canceled through discovery of new causal connection $V - Z$ and is flipped in the opposite direction by the inference of a common effect at X . The process can be iterated by canceling the new common effect and re-introducing one at Y , etc. So, assuming an unlimited supply of potentially relevant variables, nature can force an arbitrary, convergent method to cycle any number of times between the opposite causal conclusions $X \rightarrow Y$ and $Y \rightarrow X$.¹⁸ The causal flips depicted in figure 14 have been elicited (in probability) from the PC causal discovery algorithm [Spirtes *et al.*, 2000] using

¹⁷That approach is adopted, for example, in earlier work by [Freivalds and Smith, 1993].

¹⁸In fact, it can be demonstrated that arbitrarily long causal chains can be flipped in this way.

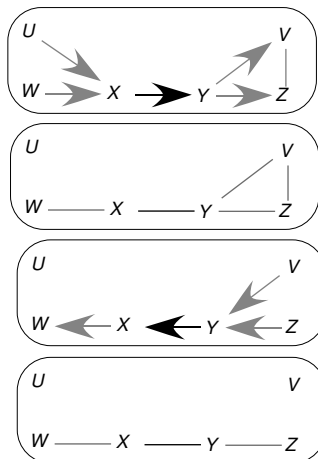


Figure 14. Causal flipping

computer simulated random samples of increasing size from a fixed causal model.

Note that one can no longer rely on logical consistency to force retractions of defeated theories, so the beating argument provided for theorem 2 fails when assumption (ii) is dropped. Happily, the beating arguments based on retraction times still work, which is yet another motive for considering retraction times in addition to total retractions.

THEOREM 5 Kelly, 2006. *Theorems 3 and 4 continue to hold without assumption (ii).*

12 EXTENSION TO RANDOMIZED SCIENTIFIC STRATEGIES

The preceding theorems assume that the scientist's method is a deterministic function of the input data. It is frequently the case, however, that randomized or "mixed" strategies achieve lower worst-case losses than deterministic strategies. For example, if the problem is to guess which way a coin lands inside of a black box and the loss is 0 or 1 depending on whether one is right or wrong, guessing randomly achieves a worst-case expected loss bound of 1/2, whereas the lowest worst-case loss bound achieved by either pure (deterministic) strategy is 1. Nonetheless, the Ockham efficiency argument can be extended to show that deterministically stalwart, Ockham strategies are efficient with respect to all convergent mixed scientific strategies, where convergence efficiency is defined in terms of *expected* retractions and convergence *in probability*, meaning that the objective *chance* (grounded in the method's internal coin-flipper) that the method produces the true theory goes to one as experience increases [Kelly and Mayo-Wilson, 2010].

THEOREM 6 Kelly and Mayo-Wilson, 2010. *All of the preceding theorems extend to random empirical methods when retractions are replaced with expected retractions and retraction times are replaced with expected retraction times.*

Here is how it works. A method is said to retract T *in chance* to degree r at stage $k + 1$ if the chance that T produces T goes down by r from k to $k + 1$. Total retractions in chance are summed over theories and stages of inquiry, so as the chance of producing one theory goes up, the chance of producing the remaining theories goes down. Therefore, nature is in a position to force a convergent method to produce total retractions arbitrarily close to i by presenting an infinite stream of experience w making T true. It is readily shown that the total retractions in chance along w are a lower bound on expected total retractions along w . It is also evident that for deterministic strategies, the total expected retractions are just the total deterministic retractions. So, since deterministically Ockham strategies retract at most i times given that T is true, they are efficient over all mixed strategies as well, and violating either property results in inefficiency.

The extension of the Ockham efficiency theorem to random methods and expected retraction times suggests a further extension to probabilistic theories and evidence (i.e., statistical theoretical inference). It remains an open question to obtain a result exactly analogous to theorem 6 in the case of statistical theory choice. It is no problem to obtain lower bounds on expected retractions and retraction times that agree with those in the proof of theorem 6. The difficulties are on the positive side — to define appropriate analogues of $C_e(n)$, Ockham's razor, and stalwartness that allow for the fact that no statistical hypothesis is ever strictly incompatible with the data.

13 DISJUNCTIVE BELIEFS, RETRACTION DEGREES, AND A GETTIER EXAMPLE

Using ‘?’ to indicate refusal to choose a particular theory is admittedly crude. When there are two simplest theories T_1, T_2 compatible with the data, it is more realistic to allow retreat to the disjunction $T_1 \vee T_2$ than to a generic refusal to say anything at all — e.g., uncertainty between two equally simple orientations of a single causal arrow does not necessarily require (or even justify) retraction of all the other causal conclusions settled up to that time. Accordingly, method M will now be allowed to produce finite *disjunctions* of theories in Q . Suppose that there are mutually exclusive and exhaustive theories $\{T_i : i \leq n\}$ and let $\mathbf{x}$ be a Boolean n -vector. Viewing $\mathbf{x}$ as the indicator function of finite set $S_{\mathbf{x}} = \{i \leq n : x_i = 1\}$, one can associate with $\mathbf{x}$ the disjunction:

$$T_{\mathbf{x}} = \bigvee_{i \in S_{\mathbf{x}}} T_i.$$

A retraction now occurs whenever some disjunct is added to one's previous conclusion, regardless how many disjuncts are also removed. Charging one unit per

retraction, regardless of the total content retracted, amounts to the following rule:

$$\rho_{\text{ret}}(T_{\mathbf{x}}, T_{\mathbf{y}}) = \max_i y_i - x_i.$$

One could also charge one unit for each disjunct added to one's prior output, regardless how many disjuncts are removed, which corresponds to the slightly modified rule:

$$\rho_{\text{dis}}(T_{\mathbf{x}}, T_{\mathbf{y}}) = \sum_i y_i \dot{-} x_i,$$

where the cutoff subtraction $y \dot{-} x$ assumes value 0 when $x \geq y$.¹⁹ Assuming no short simplicity paths, charging jointly for the total number of disjuncts added and the times at which the disjuncts are added allows one to derive stronger versions of Ockham's razor and stalwartness from retraction efficiency. The strengthened version of Ockham's razor is that one should never produce a disjunction stronger than the disjunction of all currently simplest theories (disjunctions take the place of '?') and the strengthened version of stalwartness is that one should never disjoin a theory T to one's prior conclusion unless T is among the currently simplest theories.²⁰

When there are short simplicity paths, the Ockham efficiency argument can fail for both of the proposed retraction measures. The counterexample is reminiscent of Gettier's [1963] counterexample to the justified true belief analysis of knowledge (fig. 15). Suppose that T_0 is simpler than T_1 and T_2 and that T_2 is simpler than T_3 . Suppose that experience e is compatible with T_0 and that M produces the disjunction of $(T_0 \vee T_2)$ in response to e "because" M believes T_0 on the basis of Ockham's razor and the disjunction follows from T_0 . If T_1 true, then M has true belief $(T_0 \vee T_2)$ "for the wrong reason" — a Gettier case. Suppose that T_0 is refuted. An Ockham method should now retract to $(T_1 \vee T_2)$, but M *expands* to T_2 "because" M believed $(T_0 \vee T_2)$ and learned that $\neg T_0$. If the truth is T_1 , then both methods have 1 retraction on either retraction measure and Ockham incurs the retraction earlier, so Ockham (barely) wins in $C_1(e)$ after T_0 is refuted. But M wins by retracting only once in $C_2(e)$, when T_3 is true.²¹ Possible responses to the issue of short simplicity paths include those discussed above in section 10.

¹⁹The same formula takes a finite value for a countable infinity of dimensions as long as each disjunction has at most finitely many disjuncts.

²⁰It is still the case that nature can force at least n retractions in complexity set C_n and stalwart, Ockham methods retract no more than that. If M violates the strengthened version of Ockham's razor, M produces a disjunction missing some simplest theory T . Nature is now free to force M down a path of increasingly complex theories that begins with T . By the no short paths assumption, this path passes through each complexity set, so M incurs at least one extra retraction in each complexity set. If M violates the strengthened version of stalwartness, then M retracts by adding a complex disjunct T . Nature is free to present a world of experience for a simplest world, forcing M to retract disjunct T .

²¹To see why short paths are essential to the example, suppose that there were a theory T_4 more complex than T_1 . Then M would also retract twice in C_2 and Ockham would complete the retraction in C_1 earlier.

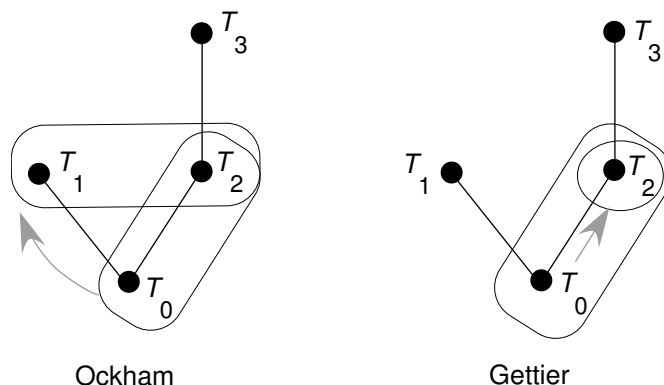


Figure 15. Gettier counterexample to Ockham efficiency

14 EXTENSION TO DEGREES OF BELIEF

Bayesian agents may use their degrees of belief to choose among potential theories [Levi, 1983], but they may also regard updated degrees of belief as the ultimate product of scientific inquiry. It is, therefore, of considerable interest to extend the logic of the Ockham efficiency theorems from problems of theory choice to problems of degree of belief assignment. Here are some recent ideas in that direction.

Suppose that the theories under consideration are just T_1, T_2, T_3 , in order of increasing complexity. Then each prior probability distribution p over these three theories can be represented uniquely as the ordered triple $\mathbf{p} = (p(T_1), p(T_2), p(T_3))$. The extremal distributions are the basis vectors $\mathbf{i}_1 = (1, 0, 0)$, $\mathbf{i}_2 = (0, 1, 0)$, and $\mathbf{i}_3 = (0, 0, 1)$ and all other coherent distributions lie on the *simplex* or triangle connecting these points in three-dimensional Euclidean space. A standard argument for distributing degrees of belief as probabilities [de Finetti, 1975; Rosenkrantz, 1983; Joyce, 1998] is that each point $\mathbf{x}$ off of the simplex is farther from the true corner of the simplex (whichever it might be) than the point $\mathbf{p}$ on the simplex directly below $\mathbf{x}$, so agents who seek immediate proximity to the truth should stay on the surface of the simplex — i.e., be coherent (fig. 16 (a)).

It is natural to extend that static argument to the active *pursuit* of truth in terms of total Euclidean distance traversed on the surface of the simplex prior to convergence to the truth (fig. 16 (b)). As in section 8, nature has a strategy to force each convergent Bayesian arbitrarily close to $\mathbf{i}_1$, then arbitrarily close to $\mathbf{i}_2$ and then all the way to $\mathbf{i}_3$. Each side of the triangular simplex has length $\sqrt{2}$, so if one adopts $\sqrt{2}$ as the unit of loss, then nature can force retraction bound k in complexity set $C_k(e)$, just as in the discussion of theory choice. Therefore, the path $(\mathbf{p}, \mathbf{i}_2, \mathbf{i}_3)$ is efficient, since it achieves that bound. Furthermore, suppose that method M favors complex theory T_2 over simpler theory T_1 by moving from $\mathbf{p}$ to $\mathbf{q}$ instead of to $\mathbf{i}_2$. Then nature can force M back to $\mathbf{i}_2$ by presenting simple data.

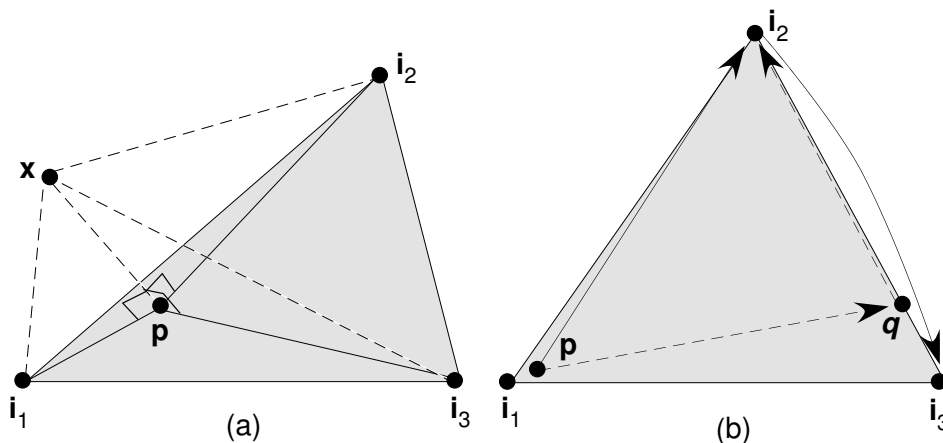


Figure 16. Distance from the truth vs. efficient pursuit of the truth

So the detour through $\mathbf{q}$ results, in the worst case, in the longer path $(\mathbf{p}, \mathbf{q}, \mathbf{i}_2, \mathbf{i}_3)$ that hardly counts as an efficient pursuit curve ($\mathbf{q}$ is passed *twice*, which amounts to a needless cycle).

An ironic objection to the preceding argument is that the conclusion seems too *strong* — efficiency measured by total distance traveled demands that one start out with full credence in the simplest theory and that one leap immediately and fully to the newly simplest theory when the previous simplest theory is refuted. Avoidance of that strong conclusion was one of the motives for focusing on retractions as opposed to expansions of belief in problems of theory choice, since movement from a state of suspension to a state of belief is not counted as a retraction. Euclidean distance charges equally for expansions and retractions of Bayesian credence, so it is of interest to see whether weaker results can be obtained by charging only for Bayesian retractions.

One approach is to define Bayesian retractions as increases in *entropy*, defined as:

$$M(q) = - \sum_i q_i \log_2 q_i.$$

That is wrong, however, since the circuit path $(\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_1)$ seems to incur two large retractions, but entropy remains constantly 0. A more sophisticated idea is to tally the cumulative increases in entropy along the entire path from p to q , rather than just at the endpoints. But that proposal still allows for “retraction-free” circuits around the entropy peak at the midpoint $(1/3, 1/3, 1/3)$ along a path of constant entropy. The same objection obtains if entropy is replaced with any alternative scalar field that plausibly represents informativeness.

Another idea is to measure the retractions from p to q in terms of a popular measure of separation for probability distributions called the *Kullback Leibler (KL)*

divergence from p to q :

$$KL(q|p) = \sum_i q_i \log_2 \frac{q_i}{p_i}.$$

KL divergence is commonly applied to measure motions on the simplex in Bayesian experimental design, where the idea is to design the experiment that maximizes the KL divergence from the prior distribution p to the posterior distribution q [Chaloner and Verdinelli, 1995]. It is well known that KL divergence is not a true distance measure or *metric* because it is asymmetrical and fails to satisfy the triangle inequality. It is interesting but less familiar that the asymmetry amounts to a bias against retractions: e.g., if $\mathbf{p} = (1/3, 1/3, 1/3)$ and $\mathbf{q} = (.999, .0005, .0005)$ then $KL(p|q) \approx 5.7$ and $KL(q|p) \approx 1.6$. Unfortunately, KL divergence cannot be used to measure retractions after a theory is refuted because it is undefined (due to taking $\log(0)$) for any motion terminating at the perimeter of the simplex. But even if one approximates such a motion by barely avoiding the perimeter, KL divergence still charges significantly more for hedging one's bets than for leaping directly to the current simplest theory. For example, if $\mathbf{p} = (.999, .0005, .0005)$, $\mathbf{q} = (.0001, .5, .4999)$, $\mathbf{r} = (.0005, .9995, .0005)$, then the KL divergence along path $(\mathbf{p}, \mathbf{r})$ is nearly 10.9, whereas the total KL divergence along path $(\mathbf{p}, \mathbf{q}, \mathbf{r})$ is around 17.7.

Here is a different approach, motivated by a fusion of logic and geometry, that yields Ockham efficiency theorems closely analogous to those in the disjunctive theory choice paradigm.²² The simplex of coherent probability distributions over T_0, T_1, T_2 is just the intersection of the unit cube with a plane through each of the unit vectors (fig. 17). The Boolean vectors labeling vertices of the unit cube are the labels of the possible disjunctions of theories (the origin $\mathbf{0} = (0,0,0)$ corresponds to the empty disjunction or contradiction). To extend that picture to the entire unit cube, think of $T_{\mathbf{x}}$ as a *fuzzy* disjunction in which theory T_i occurs to degree x_i . Say that $T_{\mathbf{x}}$ is *sharp* when $\mathbf{x}$ is Boolean and say that $\mathbf{y}$ is *sharp* when $\mathbf{y}$ is a unit vector. Each vector $\mathbf{y}$ in the unit cube can also be viewed as a fuzzy assignment of semantic values to the possible theories. Define the *valuation* of $T_{\mathbf{x}}$ in $\mathbf{y}$ to be the inner product: $\tau_{\mathbf{y}}(T_{\mathbf{x}}) = \mathbf{y} \cdot \mathbf{x} = \sum_i y_i \cdot x_i$. If $\mathbf{y}$ and $T_{\mathbf{x}}$ are both sharp, then $\tau_{\mathbf{y}}(T_{\mathbf{x}})$ is the classical truth value of $T_{\mathbf{x}}$ in $\mathbf{y}$ and if p is a probability and $T_{\mathbf{x}}$ is sharp, then $\tau_{\mathbf{p}}(T_{\mathbf{x}}) = p(T_{\mathbf{x}})$.²³ Entailment is defined by: $T_{\mathbf{x}} \models T_{\mathbf{y}}$ if and only if $\tau_{\mathbf{z}}(T_{\mathbf{x}}) \leq \tau_{\mathbf{z}}(T_{\mathbf{y}})$, for each vector $\mathbf{z}$ in the unit cube. Thus, $T_{\mathbf{x}} \models T_{\mathbf{y}}$ holds if and only if $x_i \leq y_i$, for each i . The resulting entailment relations are isomorphic to the subset relation over the fuzzy subsets of a 3-element set [Zadeh, 1965]. The *fully consistent* disjunctions are the fuzzy disjunctions that evaluate to 1 in some sharp assignment. They comprise exactly the upper three faces of the unit cube. The vertices of those faces are the consistent, sharp disjunctions of classical logic.

The formulas for retraction measures ρ_{ret} and ρ_{dis} are already defined over the entire unit cube and, hence, may be applied directly to probability assign-

²²The following definitions and results were developed in collaboration with Hanti Lin.

²³It is tempting, but not necessary for our purposes, to define $p(T_{\mathbf{x}}) = p \cdot \mathbf{x}$ for non-sharp $T_{\mathbf{x}}$ as well.

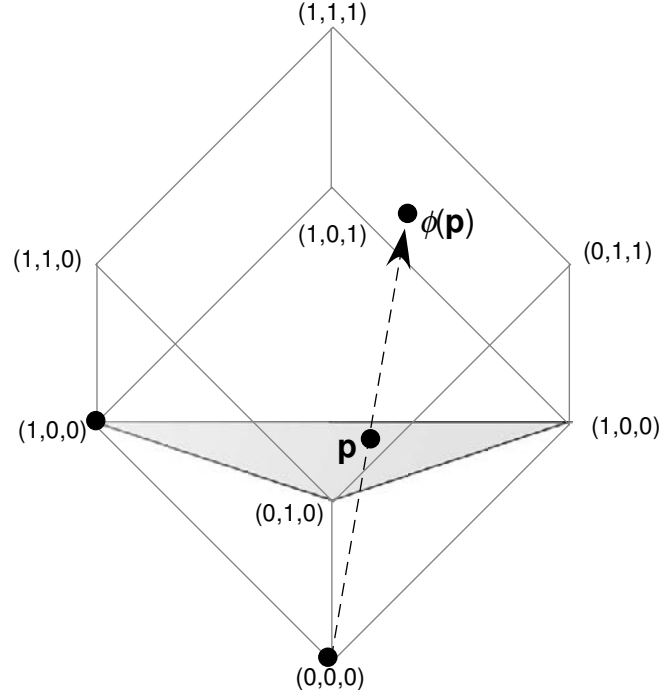


Figure 17. Simplex and unit cube

ments. That is not the right idea, however, for it is natural to view the move from $(0, 1/2, 1/2)$ to $(0, 1, 0)$ as a pure expansion of credence, but both retraction measures assign retraction $1/2$ in this case. As a result, efficiency once again demands that one move immediately to full credence in T_1 when T_0 is refuted.

Here is a closely related idea that works. The grain of truth behind probabilistic indifferentism is that the sharp disjunction $T_{(1,1,0)} = T_1 \vee T_2$ more faithfully summarizes or expresses the uniform distribution $(1/2, 1/2, 0)$ than the biased distribution $(1/3, 2/3, 0)$; a view that can be conceded without insisting, further, that uniform degrees of belief should be adopted. One explanation of the indifferentist intuition is geometrical — the components of $\mathbf{p} = (1/2, 1/2, 0)$ are proportional to the components of $\mathbf{x} = (1, 1, 0)$ in the sense that there exists constant c such that $\mathbf{x} = c\mathbf{p}$. To be assertible, a proposition should be fully consistent. T_p satisfies the proportionality condition for p but is not fully consistent. Accordingly, say that $T_{\mathbf{x}}$ expresses p just in case $T_{\mathbf{x}}$ is fully consistent and $\mathbf{x}$ is proportional to $\mathbf{p}$. Sharp propositions cannot express non-uniform distributions, but fuzzy propositions can: e.g., $T_{(1/2,1,0)}$ expresses $(1/3, 2/3, 0)$ in much the same, natural way that $T_{(1,1,0)}$ expresses $(1/2, 1/2, 0)$.²⁴ Each fully consistent disjunction has a unit

²⁴A disanalogy: $\tau_{(1/2,1/2,0)}(T_{(1,1,0)}) = 1$, but $\tau_{(1/3,2/3,0)}(T_{(1/2,1,0)}) = 5/6$, so the expression

component, which fixes the constant of proportionality at $1/\max_i p_i$. Thus, the unique, propositional expression of p is $T_{\phi(\mathbf{p})}$, where:

$$\phi(\mathbf{p})_i = \mathbf{p}_i / \max_i \mathbf{p}_i.$$

Geometrically, $\phi(\mathbf{p})$ can be found simply by drawing a ray from $\mathbf{0}$ through $\mathbf{p}$ to the upper surface of the unit cube (fig. 17).

One can now define probabilistic retractions as the logical retractions of the corresponding, propositional expressions:

$$\begin{aligned} \rho_{\text{ret}}(\mathbf{p}, \mathbf{q}) &= \rho_{\text{ret}}(T_{\phi(\mathbf{p})}, T_{\phi(\mathbf{q})}); \\ \rho_{\text{dis}}(\mathbf{p}, \mathbf{q}) &= \rho_{\text{dis}}(T_{\phi(\mathbf{p})}, T_{\phi(\mathbf{q})}). \end{aligned}$$

In passing, one can also define Bayesian *expansions* of belief by permuting $\mathbf{p}$ and $\mathbf{q}$ on the right-hand-sides of the above formulas. Revisions are then the sum of the expansions and retractions. Thus, one can extend the concepts of belief revision theory [Gärdenfors, 1988] to Bayesian degrees of belief — an idea that may have useful applications elsewhere, such as in Bayesian experimental design.

Both retraction measures have the natural property that if T_i is the most probable theory under p , then for each alternative theory T_j , the move from p to the conditional distribution $p(\cdot | \neg T_j)$ incurs no retractions [Lin, 2009]. Moreover, for purely retractive paths (paths that incur 0 expansions), the disjunctive measure is attractively *path-independent*:

$$\rho_{\text{dis}}(p, r) = \rho_{\text{dis}}(p, q) + \rho_{\text{dis}}(q, r).$$

Most importantly, both measures entail simplicity biases that fall short of the implausible demand that one must leap to the currently simplest theory immediately (fig. 18). For ρ_{ret} , the zone of efficient moves from $\mathbf{p}$ to the next simplest vertex $\mathbf{j}$ when nearby vertex $\mathbf{i}$ is refuted is constructed as follows. Let $\mathbf{c}$ be the center of the simplex, let $\mathbf{i}$ be the vertex nearest to $\mathbf{p}$, let $\mathbf{m}$ be the mid-point of the side nearest $\mathbf{p}$ and let $\mathbf{m}'$ be the midpoint of the side farthest from $\mathbf{p}$ (ties don't matter). Let $\mathbf{v}$ be the intersection of line $\overline{\mathbf{p}\mathbf{m}'}$ with line $\overline{\mathbf{c}\mathbf{m}}$. Let $\mathbf{o}$ be the intersection of line $\overline{\mathbf{i}\mathbf{v}}$ with the side of the simplex farthest from $\mathbf{p}$. Then assuming that credence in the refuted theory drops to 0 immediately, retraction-efficiency countenances moving anywhere on the line segment connecting $\mathbf{j}$ and $\mathbf{o}$. For retraction measure ρ_{dis} , the construction is the same, except that $\mathbf{v}$ is the intersection of $\overline{\mathbf{c}\mathbf{m}}$ with $\overline{\mathbf{p}\mathbf{j}}$. Note that when $\mathbf{p} \approx \mathbf{i}$, the Ockham zone for ρ_{ret} is nearly the entire half-side $\overline{\mathbf{j}\mathbf{m}'}$, whereas measure ρ_{dis} allows only for movement directly to the corner $\mathbf{j}$, as is already required in the disjunctive theory choice setting described in section 13. Thus, the extreme version of Ockham's razor is tied to the plausible aim of preserving as much content as possible. In practice, however, an open-minded Bayesian never puts *full* credence in the currently simplest theory and in that case

of a uniform distribution is also the support of the distribution, but that fails in the non-uniform case.

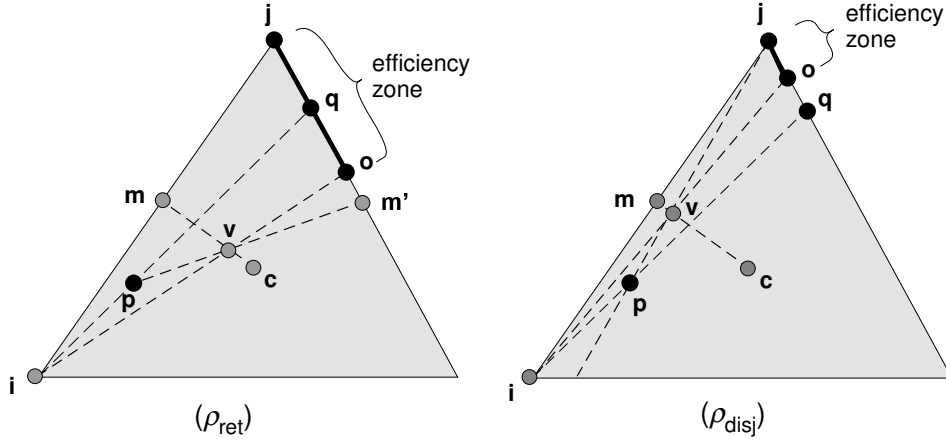


Figure 18. Two versions of Ockham's Bayesian razor

the Ockham zone for ρ_{ret} allows some leeway but is still not liberal enough for Bayesian updating to count as efficient. In both figures, the result $\mathbf{q}$ of updating $\mathbf{p}$ with the information that T_i is false can be found by drawing a ray from vertex $\mathbf{i}$ to the opposite side of the triangle. Note that $\mathbf{q}$ falls within the zone of efficiency for retraction measure ρ_{ret} but not for measure ρ_{disj} .

The Gettier-like counterexample presented in section 13 can also arise in 4 dimensions or more for Bayesian agents when the no short path assumption fails (just embed the example into the upper faces of the 4-dimensional unit cube and project it down onto the 3-dimensional simplex contained in that cube). The potential responses reviewed in section 13 apply here as well.

15 CONCLUSION

This study reviewed the major justifications of Ockham's razor in philosophy, statistics, and machine learning, and found that they fail to explain, in a non-circular manner, how Ockham's razor is more conducive to finding true theories than alternative methods would be. The failure of standard approaches to connect simplicity with methodical truth was traced to the concepts of truth-conduciveness underlying the respective arguments. Reliable indication of the truth is too strong to establish without (a) trading empirical truth for accurate prediction or (b) begging the question by means of a prior bias against complex possibilities. Convergence in the limit is too weak to single out simplicity as the right bias to have in the short run. An intermediate concept of truth-conduciveness is effective pursuit of the truth, where effectiveness is measured in terms of such costs as total retractions and errors prior to convergence. Then one can prove, without circularity or substituting predictive accuracy for theoretical truth, that Ockham's razor is the

best possible strategy for finding true theories. That result, called the Ockham efficiency theorem, can be extended to problems with branching paths of simplicity, to problems in which defeated theories are not refuted, to random strategies and, except in some interesting, Gettier-like cases, to Bayesian degrees of belief and to strategies that produce disjunctions of theories. The ultimate goal, which has not yet been reached, is to extend the Ockham efficiency argument to statistical inference.

ACKNOWLEDGEMENTS

The results on random strategies were obtained in close collaboration with Conor Mayo-Wilson, who also provided detailed comments on the draft. The definitions concerning Bayesian retractions were arrived at in close collaboration with Hanti Lin, who also formulated and proved many of the theorems. Unusually detailed comments were provided by Prasanta Bandyopadhyay and by the anonymous reviewer. They were greatly appreciated.

BIBLIOGRAPHY

- [Akaike, 1973] H. Akaike. A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19: 716-723, 1973.
- [Carnap, 1950] R. Carnap. *Logical Foundations of Probability*, Chicago: University of Chicago Press, 1950.
- [Case and Smith, 1983] J. Case and C. Smith. Comparison of identification criteria for machine inductive inference, *Theoretical Computer Science* 25:, 193-220, 1983.
- [Chickering, 2003] D. Chickering. Optimal Structure Identification with Greedy Search, *JMLR*, 3: 507-554, 2003.
- [Domingos, 1999] P. Domingos. The Role of Occam's Razor in Knowledge Discovery, *Data Mining and Knowledge Discovery*, vol. 3: 409-425.
- [Duda et al., 2001] R. Duda, P. Hart, and D. Stork. *Pattern Classification*, New York: Wiley.
- [Freivalds and Smith, 1993] R. Freivalds and C. Smith. On the Role of Procrastination in Machine Learning, *Information and Computation* 107: 237-271, 1993.
- [Forster, 2001] M. Forster. The New Science of Simplicity, in *Simplicity, Inference, and Modeling*, A. Zellner, H. Keuzenkamp, and M. McAleer, eds., Cambridge: Cambridge University Press, 2001.
- [Forster and Sober, 1994] M. Forster and E. Sober. How to Tell When Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions, *The British Journal for the Philosophy of Science* 45: 1-35, 1994.
- [Friedman, 1983] M. Friedman. *Foundations of Space-time Theories*, Princeton: Princeton University Press, 1983.
- [Gärdenfors, 1988] P. Gärdenfors. *Knowledge in Flux*, Cambridge: MIT, 1988.
- [Garey and Johnson, 1979] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*, New York: Wiley, 1979.
- [Gettier, 1963] E. Gettier. Is Justified True Belief Knowledge? *Analysis* 23: 121-123, 1963.
- [Glymour, 1980] C. Glymour. *Theory and Evidence*, Princeton: Princeton University Press, 1980.
- [Glymour, 2001] C. Glymour. Instrumental Probability, *Monist* 84: 284-300, 2001.
- [Goldenshluger and Greenshtein, 2001] A. Goldenshluger and E. Greenshtein. Asymptotically minimax regret procedures in regression model selection and the magnitude of the dimension penalty, *Annals of Statistics*, 28: 1620-1637, 2001.

- [Grünewald, 2007] P. Grünewald. *The Minimum Description Length Principle*, Cambridge, MIT Press, 2007.
- [Harman, 1965] G. Harman. The Inference to the Best Explanation, *Phil Review* 74: 88-95, 1965.
- [Heise, 1975] D. Heise. *Causal Analysis*, New York: John Wiley and Sons, 1975.
- [Hjorth, 1994] J. Hjorth. *Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap*, London: Chapman and Hall, 1994.
- [Jeffreys, 1961] H. Jeffreys. *Theory of Probability*, 3rd ed., London: Oxford University Press, 1961.
- [Joyce, 1998] J. Joyce. A Nonpragmatic Vindication of Probabilism, *Philosophy of Science* 65: 73-81, 1998.
- [Kass and Raftery, 1995] R. Kass and A. Raftery. Bayes Factors, *Journal of the American Statistical Association* 90: 773-795, 1995.
- [Kelly, 1996] K. Kelly. *The Logic of Reliable Inquiry*, New York: Oxford, 1996.
- [Kelly, 2007] K. Kelly. How Simplicity Helps You Find the Truth Without Pointing at it, V. Harazinov, M. Friend, and N. Goethe, eds. *Philosophy of Mathematics and Induction*, Dordrecht: Springer, pp. 321-360, 2007.
- [Kelly, 2008] K. Kelly. Ockham's Razor, Truth, and Information. In *Philosophy of Information*, Van Benthem, J. Adriaans, P. eds. Dordrecht: Elsevier, pp. 321-360, 2008.
- [Kelly and Mayo-Wilson, 2009] K. Kelly and C. Mayo-Wilson. Ockham Efficiency Theorem for Random Empirical Methods, Formal Epistemology Workshop, 2009, http://fitelson.org/few/kelly_mayo-wilson.pdf.
- [Kelly and Schulte, 1995] K. Kelly and O. Schulte. The Computable Testability of Theories with Uncomputable Predictions, *Erkenntnis* 43: 29-66, 1995.
- [Kitcher, 1982] P. Kitcher. Explanatory Unification, *Philosophy of Science* 48:507-531 1982.
- [Kyburg, 1977] H. Kyburg. Randomness and the Right Reference Class, *The Journal of Philosophy*, 74: 501-521, 1977.
- [Kuhn, 1957] T. Kuhn. *The Copernican Revolution*, Cambridge: Harvard University Press, 1957.
- [Kuhn, 1962] T. Kuhn. *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press, 1962.
- [Lehrer, 1990] K. Lehrer. *Theory of Knowledge*, Boulder: Westview Press, 1990.
- [Levi, 1974] I. Levi. On Indeterminate Probabilities, *Journal of Philosophy* 71: 397-418, 1974.
- [Levi, 1983] I. Levi. *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*, Cambridge: MIT Press, 1983.
- [Lewis, 1987] D. Lewis. A Subjectivist's Guide to Objective Chance, in *Philosophical Papers* Volume II, Oxford: Oxford University Press, pp. 83-133, 1987.
- [Li and Vitanyi, 1993] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and its Applications*, New York: Springer, 1993.
- [Luo and Schulte, 2006] W. Luo and O. Schulte. Mind change efficient learning, *Information and Computation*, 204:989-1011, 2006.
- [Mallows, 1973] C. Mallows. Some comments on Cp, *Technometrics* 15: 661-675, 1973.
- [Mayo, 1996] D. G. Mayo. *Error and the Growth of Experimental Knowledge*, Chicago: The University of Chicago Press, 1996.
- [Meek, 1995] C. Meek. Strong completeness and faithfulness in Bayesian networks, *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal*, pp. 411-418, 1995.
- [Mitchell, 1997] T. Mitchell. *Machine Learning*, New York: McGraw Hill, 1997.
- [Myrvold, 2003] W. Myrvold. A Bayesian Account of the Virtue of Unification, *Philosophy of Science*:399-423, 2003.
- [Newton, 1726] I. Newton. *Philosophiae Naturalis Principia Mathematica*, London, 1726.
- [Popper, 1968] K. Popper. *The Logic of Scientific Discovery*, New York: Harper, 1968.
- [Putnam, 1965] H. Putnam. Trial and Error Predicates and a Solution to a Problem of Mostowski, *Journal of Symbolic Logic* 30: 49-57, 1965.
- [Rissanen, 2007] J. Rissanen. *Information and Complexity in Statistical Modeling*, New York: Springer-Verlag, 2007.
- [Rosenkrantz, 1983] R. Rosenkrantz. Why Glymour is a Bayesian, in *Testing Scientific Theories*, Minneapolis: University of Minnesota Press, 1983.
- [Salmon, 1967] W. Salmon. *The Logic of Scientific Inference*, Pittsburgh: University of Pittsburgh Press, 1967.

- [Schulte, 1999] O. Schulte. Means-Ends Epistemology, *The British Journal for the Philosophy of Science*, 50: 1-31, 1999.
- [Schulte, 2000] O. Schulte. Inferring Conservation Principles in Particle Physics: A Case Study in the Problem of Induction, *The British Journal for the Philosophy of Science*, 51: 771-806, 2000.
- [Spirtes *et al.*, 2000] P. Spirtes, C. Glymour and R. Scheines. *Causation, Prediction, and Search*, second edition, Cambridge: MIT Press, 2000.
- [Teller, 1976] P. Teller. Conditionalization, Observation, and Change of Preference, in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, W. Harper and C. Hooker, eds., Dordrecht: D. Reidel, 1976.
- [US Army, 2003] U.S. Army. *Rifle Marksmanship M16A1, M16A2/3, M16A4, and M4 Carbine*, FM 3-22.9, Headquarters, Dept. of the Army, 2003.
- [Vapnik, 1995] V. Vapnik. *The Nature of Statistical Learning Theory*, Berlin: Springer, 1995.
- [Verma and Pearl, 1991] T. Verma and J. Pearl. Equivalence and Synthesis of Causal Models, *Uncertainty in Artificial Intelligence* 6:220-227, 1991.
- [Wasserman, 2004] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer, 2001.
- [Wolpert, 1996] D. Wolpert. The lack of a prior distinction between learning algorithms, *Neural Computation*, 8: pp. 1341-1390, 1996.
- [Zadeh, 1965] L. Zadeh. Fuzzy sets, *Information and Control* 8: 338-353, 1965.

Part XII

Special Problems in Statistics/Computer Science

This page intentionally left blank

NORMAL APPROXIMATIONS

Robert J. Boik

1 INTRODUCTION

It often occurs that an investigator is interested in computing probabilities for a random variable, Y , whose distribution is known only partially. If the mean and variance of the random variable are known, then the investigator could use a normal distribution with the same mean and variance as that of Y to approximate the probabilities of interest. Unfortunately, the magnitude of the approximation error generally is unknown and it can be unacceptably large. If Y is the sum or mean of a sequence of component variables and some regularity conditions are satisfied, however, then the magnitude of the approximation error is bounded and the normal approximation can be quite accurate. The conditions under which the distribution of a sum or a mean of a sequence of component variables approaches a normal distribution are described in so called “central limit” theorems.

This chapter describes the use of the normal distribution to approximate the probabilities of interest. Section 2 describes several versions of the central limit theorem (CLT). This theorem ensures that, under fairly mild conditions, the distribution of a suitably standardized sum of random variables approaches a normal distribution as the number of random variables that are summed increases. The scope of the CLT is broadened substantially by two additional theoretical results, namely the delta method and Slutsky’s theorem. These two results are described in §3. Sections 4–6 describe various applications of the CLT. In practice, the primary use of the CLT is to approximate the sampling distribution of a test statistic or the sampling distribution of an asymptotically pivotal quantity. It is frequentists rather than Bayesians who use sampling distributions to make inferences about parameter values. Accordingly, the applications in sections 4–6 are frequentist in nature. For example, a frequentist who is interested in the value of a population mean, μ , might obtain a random sample, $Y_1, Y_2, \dots, Y_n$, from the population and use the CLT to approximate the sampling distribution of the sample mean $\bar{Y}$. The approximate sampling distribution of $\bar{Y}$ then could be used to construct a confidence interval for μ . A Bayesian who is interested in a population mean, μ would adopt a different approach. The Bayesian would treat μ as a random variable, rather than as a fixed quantity. Inference about μ would be based on the distribution of μ conditional on the data, $Y_1, Y_2, \dots, Y_n$. This conditional distribution is called the posterior distribution. Conditional distributions are defined in Appendix A.1. Section 7 describes a Bayesian CLT for posterior distributions.

The last section describes higher-order expansions that improve on the accuracy of frequentist and Bayesian CLTs.

2 THE CENTRAL LIMIT THEOREM

The normal distribution with mean μ and variance σ^2 is denoted by $N(\mu, \sigma^2)$. Its probability density function (pdf) is

$$(1) \quad \varphi(x; \mu, \sigma^2) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} \text{ for } x \in (-\infty, \infty).$$

A random variable Z with distribution $N(0, 1)$ is said to have the standard normal distribution. The pdf of the standard normal distribution is $\varphi(z, 0, 1)$ and, for convenience, this pdf is denoted simply as $\varphi(z)$.

The normal distribution also is called the “Gaussian” distribution in honor of K.F. Gauss, even though Gauss was not the first to use this distribution. Nonetheless, it was Gauss [1809] who discovered the intimate connections between the normal distribution and the method of least squares. The recognition that many naturally occurring variables are approximately normally distributed is the reason that Peirce [1873], Galton [1877; 1895], and Lexis [1877] called the distribution “normal” in the first place (see [David, 1995; Kruskal and Stigler, 1997]).

The classical CLT usually is attributed to Laplace [1810, although substantial related work had been done before 1810. A description of this earlier work can be found in Adams [1974] and Le Cam [1986]. The CLT was named the central limit theorem by Polya [1920] because of its central role in probability and statistics. The basic form of the theorem is given as Theorem 1. A proof is outlined in Appendix B.1. A square is used to mark the end of a theorem.

THEOREM 1 Laplace Central Limit Theorem. *Suppose that $Y_1, Y_2, \dots, Y_\infty$ is a sequence of independent and identically distributed (iid) random variables each with mean μ and variance $\sigma^2 < \infty$. Define $\bar{Y}_n$, Z_n , and $F_{Z_n}(z)$ as*

$$\bar{Y}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n Y_i, \quad Z_n \stackrel{\text{def}}{=} \frac{\sqrt{n}(\bar{Y}_n - \mu)}{\sigma}, \text{ and } F_{Z_n}(z) \stackrel{\text{def}}{=} P(Z_n \leq z),$$

where z is any finite constant. Then,

$$\lim_{n \rightarrow \infty} F_{Z_n}(z) = \Phi(z), \text{ where } \Phi(z) = \int_{-\infty}^z \varphi(u) du,$$

$\varphi(u) = \varphi(u; 0, 1)$ is defined in (1), and $\Phi(z)$ is the cumulative distribution function (cdf) of the $N(0, 1)$ distribution. ■

The conclusion of the theorem also can be written as

$$\frac{\sqrt{n}(\bar{Y}_n - \mu)}{\sigma} \xrightarrow{\text{dist}} N(0, 1) \text{ or as } \frac{T_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\text{dist}} N(0, 1), \text{ as } n \rightarrow \infty,$$

where $T_n = \sum_{i=1}^n Y_i$ and the relationship operator $\xrightarrow{\text{dist}}$ denotes convergence in distribution or law. Convergence in distribution is defined in Appendix A.2. See [Ferguson, 1996, part 1; Sen and Singer, 1993, ch. 2] for additional details about modes of convergence. In practice, Theorem 1 is applied by using $N(\mu, \sigma^2/n)$ as the approximating distribution for $\bar{Y}_n$ or using $N(n\mu, n\sigma^2)$ as the approximating distribution for T_n .

One point of confusion for beginning statistics students is the distinction between the law of large numbers and the central limit theorem. The weak law of large numbers states that under fairly general conditions $\bar{Y} \xrightarrow{\text{prob}} \mu$ as $n \rightarrow \infty$, where the relationship operator $\xrightarrow{\text{prob}}$ denotes convergence in probability. Informally, $\bar{Y} \xrightarrow{\text{prob}} \mu$ means that the distribution of $\bar{Y}$ becomes more and more concentrated near μ as $n \rightarrow \infty$. In the limit, the entire distribution is concentrated at one point, namely μ . Convergence in probability is formally defined in Appendix A.2. Students also are taught that the limiting distribution of $\bar{Y}$, as $n \rightarrow \infty$, is a normal distribution. Some students (usually the better ones) become confused because these two claims are contradictory. The first says that $\bar{Y}$ approaches a constant, whereas the second says that the distribution of $\bar{Y}$ approaches a normal distribution and, therefore, $\bar{Y}$ does not approach a constant. This contradiction is easily resolved by more carefully stating the CLT claim. The CLT claim is that the limiting distribution of $\sqrt{n}(\bar{Y} - \mu)/\sigma$ is a normal distribution, namely $N(0, 1)$. The CLT does not claim that the limiting distribution of $\bar{Y}$ is normal. Nonetheless, the CLT does imply that if n is finite and large, then the distribution of $\bar{Y}$ can be accurately approximated by $N(\mu, \sigma^2/n)$. The limit of this distribution as $n \rightarrow \infty$, however is not a normal distribution. Specifically, $\lim_{n \rightarrow \infty} N(\mu, \sigma^2/n)$ is a distribution that has all of its mass on one point, namely μ . That is,

$$(2) \quad \lim_{n \rightarrow \infty} P(\bar{Y}_n \leq t) = \begin{cases} 0 & \text{if } t < \mu, \\ 1 & \text{if } t > \mu. \end{cases}$$

Equation (2) is merely a restatement of the law of large numbers. Accordingly, the limiting distribution of $\bar{Y}$ is not a normal distribution.

Less restrictive forms of the CLT than that in Theorem 1 apply to independent but not identically distributed random variables (see Theorem 4 in this chapter). The CLT also has been extended to non-independent random variables and vectors. If a sample of size n is randomly selected from a finite population without replacement, then the sample $Y_1, Y_2, \dots, Y_n$ consists of non-independent random variables. This situation is discussed in §5.1 in this chapter. See [Lehmann, 1999, §2.8] for an introduction to other extensions of the CLT to non-independent random variables. These basic and less restrictive theorems are important because many random variables of practical importance (e.g., log likelihood functions) can be represented as the sum of component random variables. Also, in observational studies and experiments, the response variable sometimes can be represented as the sum of more elementary variables. For example, the weight of an organism can be represented as the sum of the weights of the components of the organ-

ism. If these component weights satisfy certain moment conditions (i.e., existence of bounded moments and restrictions on covariances among component weights), then a version of the CLT will apply.

A special case of Theorem 1 pre-dates Laplace's work by nearly 80 years. In the context of gambling problems, De Moivre [1733] found that if $T_n \sim \text{Binomial}(n, \theta)$, then

$$\frac{T_n - n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow{\text{dist}} N(0, 1).$$

This result can be obtained from Theorem 1 by recognizing that T_n has the same distribution as the sum of n identically and independently distributed Bernoulli random variables, each with mean θ and variance $\theta(1-\theta)$.

In practice, sample sizes are finite rather than infinite. The accuracy of the normal approximation when sample size is finite was addressed by Berry [1941] and Esseen [1945]. They verified that the magnitude of the error induced by the normal approximation is bound by a constant of order $O(n^{-\frac{1}{2}})$. Order of magnitude notation O , o , O_p and o_p is defined in Appendix A.2. See [Bishop *et al.*, 1975, ch. 14] for additional details about order of magnitude. The result of Berry and Esseen is the following.

THEOREM 2 Berry-Esseen. *If the conditions of Theorem 1 hold and the third absolute moment, $\eta_3 = E(|Y_i - \mu|^3)$, is finite, then*

$$\sup_{z \in \mathbb{R}} |F_{Z_n}(z) - \Phi(z; 0, 1)| < \frac{c\eta_3}{\sigma^3\sqrt{n}} \text{ for all } n$$

and for all distributions whose third moment exists, where c is a constant. ■

The value of the constant c is known to satisfy $c \in (0.4097, 0.7975)$ (see [van Beek, 1972]).

Many applications in statistical theory and methods require extensions of the CLT beyond the version presented in Theorem 1. First, a multivariate version of the CLT can be obtained simply by replacing the scalar random variable Y_i by the random vector $\mathbf{Y}_i$. The proof of Theorem 1 also holds for the multivariate case. In Theorem 3, the Euclidean norm of a matrix $\mathbf{M}$ is denoted by $\|\mathbf{M}\| \stackrel{\text{def}}{=} \sqrt{\text{trace}(\mathbf{M}'\mathbf{M})}$.

THEOREM 3 Multivariate Central Limit Theorem. *Suppose that $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_\infty$ is a sequence of independent and identically distributed random p -vectors each with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, where $\boldsymbol{\mu}$ is a p -vector of finite constants and $\boldsymbol{\Sigma}$ is a finite non-negative definite $p \times p$ matrix. Define $\bar{\mathbf{Y}}_n$, $\mathbf{U}_n$, and $F_{\mathbf{U}_n}(\mathbf{u})$ as*

$$\bar{\mathbf{Y}}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i, \quad \mathbf{U}_n \stackrel{\text{def}}{=} \sqrt{n} (\bar{\mathbf{Y}}_n - \boldsymbol{\mu}), \quad \text{and } F_{\mathbf{U}_n}(\mathbf{u}) \stackrel{\text{def}}{=} P(\mathbf{U}_n \leq \mathbf{u}),$$

where $\mathbf{u}$ is any finite p -vector of constants. If $0 < \|\boldsymbol{\Sigma}\|$, then

$$\lim_{n \rightarrow \infty} F_{\mathbf{U}_n}(\mathbf{u}) = \Phi(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\Phi(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the cdf of the multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Furthermore, if $\boldsymbol{\Sigma}$ is nonsingular, then

$$\begin{aligned} \boldsymbol{\Sigma}^{-\frac{1}{2}} \sqrt{n}(\bar{\mathbf{Y}}_n - \boldsymbol{\mu}) &\xrightarrow{\text{dist}} N(\mathbf{0}, \mathbf{I}_p) \text{ and} \\ \Phi(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \int_{-\infty}^{u_1} \int_{-\infty}^{u_2} \cdots \int_{-\infty}^{u_p} \varphi(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) dw_1 dw_2 \cdots dw_p, \\ \text{where } \varphi(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{e^{-(\mathbf{w}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{w}-\boldsymbol{\mu})/2}}{|\boldsymbol{\Sigma}|^{\frac{1}{2}} (2\pi)^{\frac{p}{2}}} \text{ and } \boldsymbol{\Sigma}^{-\frac{1}{2}} \text{ satisfies } \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}'} \end{aligned}$$

■

The conclusion of the multivariate CLT also can be written as

$$\sqrt{n}(\bar{\mathbf{Y}}_n - \boldsymbol{\mu}) \xrightarrow{\text{dist}} N(\mathbf{0}, \boldsymbol{\Sigma}).$$

A second extension is required if the random vectors $\{\mathbf{Y}_i\}_{i=1}^n$ are not identically distributed. Extensions in this direction follow from the work of Lindeberg [1922] and Feller [1935].

THEOREM 4 Lindeberg-Feller. *Suppose that for each n , $\{\mathbf{Y}_{ni}\}_{i=1}^n$ are independently distributed random p -vectors with $E(\mathbf{Y}_{ni}) = \boldsymbol{\mu}_{ni}$ and $\text{Var}(\mathbf{Y}_{ni}) = \boldsymbol{\Sigma}_{ni}$. Define $\bar{\mathbf{Y}}_n$, $\bar{\boldsymbol{\mu}}_n$, $\bar{\boldsymbol{\Sigma}}_n$, and Q_{ni} as*

$$\begin{aligned} \bar{\mathbf{Y}}_n &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_{ni}, \quad \bar{\boldsymbol{\mu}}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \boldsymbol{\mu}_{ni}, \quad \bar{\boldsymbol{\Sigma}}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \boldsymbol{\Sigma}_{ni}, \\ \text{and } Q_{ni} &\stackrel{\text{def}}{=} (\mathbf{Y}_{ni} - \boldsymbol{\mu}_{ni})' \bar{\boldsymbol{\Sigma}}_n^{-1} (\mathbf{Y}_{ni} - \boldsymbol{\mu}_{ni}), \end{aligned}$$

where it is assumed that $\bar{\boldsymbol{\Sigma}}_n$ is positive definite. If $\bar{\boldsymbol{\Sigma}}_n \rightarrow \boldsymbol{\Sigma}$ as $n \rightarrow \infty$ and, for every $\varepsilon > 0$,

$$\frac{1}{n} \sum_{i=1}^n E \left[Q_{ni} I \left(\sqrt{Q_{ni}} \geq \varepsilon \sqrt{n} \right) \right] \rightarrow 0 \text{ as } n \rightarrow \infty,$$

$$\text{where } I(\sqrt{Q_{ni}} \geq \varepsilon \sqrt{n}) = \begin{cases} 1 & \text{if } \sqrt{Q_{ni}} \geq \varepsilon \sqrt{n}, \\ 0 & \text{otherwise,} \end{cases}$$

then,

$$\sqrt{n}(\bar{\mathbf{Y}}_n - \bar{\boldsymbol{\mu}}_n) \xrightarrow{\text{dist}} N(\mathbf{0}, \boldsymbol{\Sigma}),$$

■

The condition on $E \left[Q_{ni} I(\sqrt{Q_{ni}} \geq \varepsilon \sqrt{n}) \right]$ ensures that for any fixed p -vector $\mathbf{t}$, each individual variance, $\text{Var}(\mathbf{t}' \mathbf{Y}_{ni}) = \mathbf{t}' \boldsymbol{\Sigma}_{ni} \mathbf{t}$ for $i = 1, \dots, n$, is small compared to their sum, $\sum_{i=1}^n \text{Var}(\mathbf{t}' \mathbf{Y}_{ni})$, (see Feller, 1971, p. 264).

3 THE DELTA METHOD AND SLUTSKY'S THEOREM

In statistical practice, one often seeks the asymptotic distribution of a function of sample means, rather than of the means themselves.

Example. Suppose that the conditions of Theorem 1 are satisfied and, in addition, the third and fourth moments of Y_i are finite. The first two sample moments are $\bar{Y}_n$ and S_n^2 , where $\bar{Y}_n$ is defined in Theorem 1 and

$$S_n^2 \stackrel{\text{def}}{=} \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2.$$

It is readily shown that $E(\bar{Y}_n) = \mu$ and that $E(S_n^2) = \sigma^2$. Suppose that it is of interest to find the asymptotic distribution of

$$(3) \quad \mathbf{U}_n = \begin{pmatrix} U_{n1} \\ U_{n2} \end{pmatrix} = \begin{bmatrix} \sqrt{n}(\bar{Y}_n - \mu) \\ \sqrt{n}(S_n^2 - \sigma^2) \end{bmatrix}.$$

To find this distribution, a first step is to express $(\bar{Y}_n \ S_n^2)'$ as a function of sample means. Define $\mathbf{W}_i$, $\bar{\mathbf{W}}_n$, and $\boldsymbol{\theta}$ as

$$(4) \quad \begin{aligned} \mathbf{W}_i &\stackrel{\text{def}}{=} \begin{pmatrix} Y_i \\ Y_i^2 \end{pmatrix}, \quad \bar{\mathbf{W}}_n = \begin{pmatrix} \bar{W}_{n1} \\ \bar{W}_{n2} \end{pmatrix} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \mathbf{W}_i, \\ \text{and } \boldsymbol{\theta} &\stackrel{\text{def}}{=} E(\mathbf{W}_i) = \begin{pmatrix} \mu \\ \mu^2 + \sigma^2 \end{pmatrix}. \end{aligned}$$

It can be shown that

$$n \text{Var}(\bar{\mathbf{W}}_n) = \begin{pmatrix} \sigma^2 & \kappa_3 + 2\mu\sigma^2 \\ \kappa_3 + 2\mu\sigma^2 & \kappa_4 + 2\sigma^4 + 4\mu(\mu\sigma^2 + \kappa_3) \end{pmatrix},$$

where κ_j is the j^{th} cumulant of the distribution of Y_i . Cumulants and cumulant generating functions are described in Appendix A.5. It follows from Theorem 3 that

$$\sqrt{n}(\bar{\mathbf{W}}_n - \boldsymbol{\theta}) \xrightarrow{\text{dist}} N(\mathbf{0}, \boldsymbol{\Sigma}), \text{ where } \boldsymbol{\Sigma} \stackrel{\text{def}}{=} n \text{Var}(\bar{\mathbf{W}}_n).$$

The random vector $\mathbf{U}_n$ in (3) can be written in terms of $\bar{\mathbf{W}}_n$ as follows:

$$\mathbf{U}_n = \sqrt{n} \left[\begin{pmatrix} \bar{W}_{n1} \\ \frac{n}{n-1} \{ \bar{W}_{n2} - \bar{W}_{n1}^2 \} \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \right].$$

To find the asymptotic distribution of $\mathbf{U}_n$, it is necessary to find the asymptotic distribution of a differentiable function of $\bar{\mathbf{W}}_n$. To accomplish this task, the delta method is useful. Theorem 5 summarizes the delta method. See [Sen and Singer, 1993, pp. 136–137] for an accessible proof.

THEOREM 5 Delta Method. *Suppose that $\{\mathbf{T}_n\}_{n=1}^\infty$ is a sequence of p -variate random vectors with asymptotic distribution $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{\text{dist}} N(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \rightarrow \infty$,*

where $\boldsymbol{\theta}$ is a fixed p -vector of parameters and $\boldsymbol{\Sigma}$ is a $p \times p$ covariance matrix. Further, suppose that $\mathbf{g}(\mathbf{T}_n)$ is a $k \times 1$ function whose first derivative exists for $\mathbf{T}_n$ in an open neighborhood of $\boldsymbol{\theta}$. Denote the $k \times p$ matrix of first partial derivatives of $\mathbf{g}(\mathbf{T}_n)$ evaluated at $\mathbf{T}_n = \boldsymbol{\theta}$ by $\mathbf{G}$. That is

$$\mathbf{G} \stackrel{\text{def}}{=} \left. \frac{\partial \mathbf{g}(\mathbf{T}_n)}{\partial \mathbf{T}_n'} \right|_{\mathbf{T}_n = \boldsymbol{\theta}} = \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \begin{pmatrix} \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \theta_p} \end{pmatrix}.$$

If $0 < \|\mathbf{G}\| < \infty$, then

$$\sqrt{n}[\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})] \xrightarrow{\text{dist}} \mathbf{N}(\mathbf{0}, \mathbf{G}\boldsymbol{\Sigma}\mathbf{G}').$$

■

If $\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}'$ is nonsingular, then the conclusion of Theorem 5 also can be written as

$$(\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}')^{-\frac{1}{2}} \sqrt{n}[\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})] \xrightarrow{\text{dist}} \mathbf{N}(\mathbf{0}, \mathbf{I}_k).$$

In practice, the values of quantities such as $\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}'$ are unknown and must be estimated. If $\mathbf{G}$ is a continuous function of $\boldsymbol{\theta}$, then it can be shown that

$$(5) \quad \left(\widehat{\mathbf{G}} \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{G}}' \right)^{-\frac{1}{2}} \sqrt{n}[\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})] \xrightarrow{\text{dist}} \mathbf{N}(\mathbf{0}, \mathbf{I}_k),$$

where $\widehat{\mathbf{G}} = \mathbf{G}(\mathbf{T}_n)$ and $\widehat{\boldsymbol{\Sigma}}$ is a consistent estimator of $\boldsymbol{\Sigma}$. Slutsky's theorem (Slutsky, 1925) is useful to verify (5). Proofs of Slutsky's theorem can be found in Ferguson [1993, ch. 6] and Sen and Singer [1993, pp. 127–131].

THEOREM 6 Slutsky. Let $\{\mathbf{Y}_n\}_{n=1}^{\infty}$ be a sequence of random p -vectors and let $\{\mathbf{X}_n\}_{n=1}^{\infty}$ be a sequence of random $q \times r$ matrices. Also, let $\mathbf{Y}$ be a random p -vector and let $\mathbf{C}$ be a $q \times r$ matrix of constants. If $\mathbf{Y}_n \xrightarrow{\text{dist}} \mathbf{Y}$, $\mathbf{X}_n \xrightarrow{\text{prob}} \mathbf{C}$, and $\mathbf{h}(\mathbf{X}_n, \mathbf{Y}_n)$ is a continuous function of $\mathbf{X}_n$ and $\mathbf{Y}_n$ (except, possibly, on sets of measure zero), then

$$\mathbf{h}(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{\text{dist}} \mathbf{h}(\mathbf{C}, \mathbf{Y}).$$

■

Special cases that follow from Slutsky's Theorem include the following.

- (a) If $\mathbf{X}_n$ has dimension $p \times 1$, then $\mathbf{X}_n + \mathbf{Y}_n \xrightarrow{\text{dist}} \mathbf{C} + \mathbf{Y}$.
- (b) If $\mathbf{X}_n$ has dimension $q \times p$, then $\mathbf{X}_n \mathbf{Y}_n \xrightarrow{\text{dist}} \mathbf{C} \mathbf{Y}$.
- (c) If $\mathbf{X}_n$ has dimension $1 \times r$, then $\mathbf{Y}_n \mathbf{X}_n \xrightarrow{\text{dist}} \mathbf{Y} \mathbf{C}$.

The importance of Slutsky's Theorem is that it justifies replacing unknown parameters by consistent estimators of those parameters without affecting the asymptotic distribution of a random quantity. For example, let $\mathbf{Z}$ be a random

k -vector with distribution $N(\mathbf{0}, \mathbf{G}\Sigma\mathbf{G}')$. Then, the claim in (5), follows from Slutsky's Theorem because

$$\begin{aligned} & \left(\widehat{\mathbf{G}}\widehat{\Sigma}\widehat{\mathbf{G}}\right)^{-\frac{1}{2}} \xrightarrow{\text{prob}} (\mathbf{G}\Sigma\mathbf{G})^{-\frac{1}{2}} \quad \text{and} \quad \sqrt{n}[\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})] \xrightarrow{\text{dist}} \mathbf{Z} \\ & \implies \left(\widehat{\mathbf{G}}\widehat{\Sigma}\widehat{\mathbf{G}}'\right)^{-\frac{1}{2}} \sqrt{n}[\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})] \xrightarrow{\text{dist}} (\mathbf{G}\Sigma\mathbf{G}')^{-\frac{1}{2}} \mathbf{Z}, \end{aligned}$$

and $(\mathbf{G}\Sigma\mathbf{G}')^{-\frac{1}{2}} \mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_k)$.

Example Revisited. As an illustration of Theorems 5 and 6, re-consider the problem of finding the joint asymptotic distribution of $\bar{Y}_n$ and S_n^2 . Let $\mathbf{T}_n = (T_{n1} \ T_{n2})' = \bar{\mathbf{W}}_n$, where $\bar{\mathbf{W}}_n$ is defined in (4). Define $\mathbf{g}_n(\mathbf{T}_n)$ and $\mathbf{g}(\mathbf{T}_n)$ as

$$\mathbf{g}_n(\mathbf{T}_n) \stackrel{\text{def}}{=} \begin{pmatrix} T_{n1} \\ \frac{n}{n-1} [T_{n2} - T_{n1}^2] \end{pmatrix} \quad \text{and} \quad \mathbf{g}(\mathbf{T}_n) \stackrel{\text{def}}{=} \begin{pmatrix} T_{n1} \\ T_{n2} - T_{n1}^2 \end{pmatrix}.$$

Note that $\mathbf{U}_n$ in (3) can be written as

$$\begin{aligned} \mathbf{U}_n &= \sqrt{n} \begin{pmatrix} \bar{Y}_n - \mu \\ S_n^2 - \sigma^2 \end{pmatrix} = \sqrt{n} [\mathbf{g}_n(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})] = \sqrt{n} [\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})] + \mathbf{X}_n, \\ &\quad \text{where } \mathbf{X}_n = \frac{\sqrt{n}}{n-1} \begin{pmatrix} 0 \\ T_{n2} - T_{n1}^2 \end{pmatrix}, \end{aligned}$$

and $\boldsymbol{\theta}$ is defined in (4). Then, it follows from Theorem 6 and from $\mathbf{X}_n \xrightarrow{\text{prob}} \mathbf{0}$ that $\mathbf{U}_n$ and $\sqrt{n}[\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})]$ have the same asymptotic distribution. Accordingly, it follows from Theorems 5 and 6 that

$$\begin{aligned} & \sqrt{n} \begin{pmatrix} \bar{Y}_n - \mu \\ S_n^2 - \sigma^2 \end{pmatrix} \xrightarrow{\text{dist}} N(\mathbf{0}, \boldsymbol{\Omega}) \quad \text{and} \quad Q = n \begin{pmatrix} \bar{Y}_n - \mu \\ S_n^2 - \sigma^2 \end{pmatrix}' \boldsymbol{\Omega}^{-1} \begin{pmatrix} \bar{Y}_n - \mu \\ S_n^2 - \sigma^2 \end{pmatrix} \xrightarrow{\text{dist}} \chi_2^2, \\ & \text{where } \boldsymbol{\Omega} = \mathbf{G}\Sigma\mathbf{G}' = \begin{pmatrix} \sigma^2 & \kappa_3 \\ \kappa_3 & \kappa_4 + 2\sigma^4 \end{pmatrix}, \quad \text{and} \quad \mathbf{G} = \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \begin{pmatrix} 1 & 0 \\ -2\mu & 1 \end{pmatrix}, \end{aligned}$$

Note that $\bar{Y}_n$ and S_n^2 are asymptotically independent only if the skewness coefficient, κ_3/σ^3 , is zero. Furthermore, it follows from Theorem 6 that $\boldsymbol{\Omega}^{-1}$ can be replaced by $\widehat{\boldsymbol{\Omega}}^{-1}$ without affecting the asymptotic distribution of Q , where $\widehat{\boldsymbol{\Omega}}$ is any consistent estimator of $\boldsymbol{\Omega}$.

4 DISCRETE DISTRIBUTIONS

In this section, normal approximations to discrete distributions are examined. Details are given for the binomial and negative binomial distributions.

4.1 Normal Approximation to the Binomial Distribution

First, consider a sequence of independent and identically distributed Bernoulli random variables, each with success probability θ . Denote the Bernoulli random variables as $Y_1, Y_2, \dots, Y_n$. The value of Y_i is 1 if the i^{th} trial is a success and 0 otherwise. Let $T_n = \sum_{i=1}^n Y_i$. The random variable T_n is the total number of successes in n trials. It follows that T_n has a binomial distribution with parameters n and θ . The probability mass function (pmf) for T_n is

$$f_{T_n}(t) = P(T_n = t) = \binom{n}{t} \theta^t (1 - \theta)^{n-t}, \text{ for } t = 0, 1, \dots, n.$$

It is readily shown that $\mu \stackrel{\text{def}}{=} E(Y_i) = \theta$ and $\sigma^2 \stackrel{\text{def}}{=} \text{Var}(Y_i) = \theta(1 - \theta)$. It follows from the CLT (Theorem 1) that

$$\frac{T_n - n\theta}{\sqrt{n\theta(1 - \theta)}} \xrightarrow{\text{dist}} N(0, 1).$$

Figure 1 displays the exact pmf of T_n for $\theta = 0.2$ and $n = 5, 10, 20$, and 50 . A normal distribution with mean $n\mu$ and variance $n\sigma^2$ also is displayed. It is apparent that the normal approximation improves as n increases.

To approximate individual probabilities, a continuity correction can be used. Specifically, if t is an integer in $\{0, 1, \dots, n\}$, then

$$\begin{aligned} f_{T_n}(t) &= P(T_n = t) = P(t - 0.5 < T_n \leq t + 0.5) \\ (6) \quad &= P(T_n \leq t + 0.5) - P(T_n \leq t - 0.5) \\ &= \Phi\left(\frac{t+0.5-n\mu}{\sqrt{n\sigma^2}}\right) - \Phi\left(\frac{t-0.5-n\mu}{\sqrt{n\sigma^2}}\right) + O\left(n^{-\frac{1}{2}}\right), \end{aligned}$$

where Φ is the cdf of the standard normal distribution. The supremum of the error in the normal approximations to the cdf and to the pmf are listed in Table 1. The last column of Table 1 illustrates that the constant, c , in the Berry-Esseen Theorem (Theorem 2) is less than 0.7975 for this example.

The normal approximation to a binomial random variable often is used to make frequentist inferences about the value of the parameter, θ . Suppose that $T_n \sim \text{Binomial}(n, \theta)$ for fixed n and unknown θ . It follows from Theorems 1 and 2 that

$$P\left[\left|\frac{T_n - n\theta}{\sqrt{n\theta(1 - \theta)}} - 0.5\right| \leq z_{\alpha/2}\right] = 1 - \alpha + O\left(n^{-1/2}\right),$$

where z_α is the $100(1 - \alpha)$ percentile of $N(0, 1)$. A confidence interval with nominal confidence coefficient $1 - \alpha$ can be obtained by inverting the inequality. The result

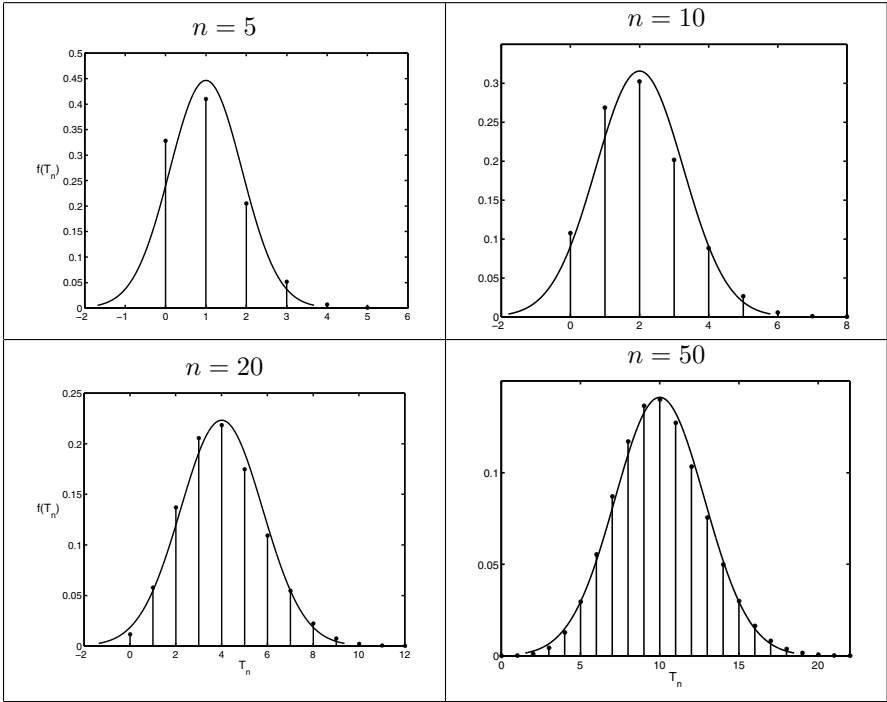


Figure 1. Normal Approximation to the Binomial Distribution

Table 1. Error in Approximating the PMF and CDF

Distribution	n	Supremum Error		Actual c
		PMF	CDF	
Binomial	5	0.0864	0.2373	0.3344
	10	0.0400	0.1778	0.3540
	20	0.0172	0.1296	0.3643
	50	0.0074	0.0836	0.3707
Negative Binomial	5	0.0202	0.0800	0.1005
	10	0.0050	0.0565	0.1058
	20	0.0023	0.0399	0.1087
	50	0.0009	0.0252	0.1105

is

$$P(L_{n,\alpha} \leq \theta \leq U_{n,\alpha}) = 1 - \alpha + O\left(n^{-1/2}\right), \text{ where}$$

$$L_{n,\alpha} = \frac{2T_n^* - 1 + z_{\alpha/2}^2 - z_{\alpha/2} \sqrt{z_{\alpha/2}^2 + 4(T_n^* - 0.5) - 4(T_n^* - 0.5)^2/n}}{2(n + z_{\alpha/2}^2)},$$

$$U_{n,\alpha} = \frac{2T_n^{**} + 1 + z_{\alpha/2}^2 + z_{\alpha/2} \sqrt{z_{\alpha/2}^2 + 4(T_n^{**} + 0.5) - 4(T_n^{**} + 0.5)^2/n}}{2(n + z_{\alpha/2}^2)},$$

$$T_n^* = \begin{cases} 1 & \text{if } T_n = 0, \\ T_n & \text{otherwise,} \end{cases} \text{ and } T_n^{**} = \begin{cases} n - 1 & \text{if } T_n = n, \\ T_n & \text{otherwise.} \end{cases}$$

Note that $(L_{n,\alpha}, U_{n,\alpha})$ is a frequentist confidence interval; the endpoints are random variables whereas the parameter θ is a fixed, but unknown constant. For example, if $n = 20$ and $T_n = 7$ is observed, then the nominal 95% confidence interval is (0.16, 0.59). This interval is correct if θ is captured by the interval and is incorrect otherwise.

The actual confidence coefficient that corresponds to the nominal 95% intervals can be obtained by computing the interval for each value of T_n from 0 to n and then summing the probability of T_n over those intervals that capture θ . For example, if $\theta = 0.2$, then the actual confidence coefficient (i.e., coverage) for nominal 95% intervals is 0.993, 0.967, 0.990, and 0.980 for sample sizes $n = 5$, $n = 10$, $n = 20$, and $n = 50$, respectively.

4.2 Normal Approximation to the Negative Binomial Distribution

For the second application, consider an infinite sequence of independent and identically distributed Bernoulli random variables, each with success probability θ . Again, denote the Bernoulli random variables as $Y_1, Y_1, \dots, Y_\infty$. The value of Y_i is 1 if the i^{th} trial is a success and 0 otherwise. Let W be the number of failures before the first success (i.e., $Y_i = 1$) occurs in the sequence. The random variable W is said to have a geometric distribution and its pmf is

$$P(W = w) = (1 - \theta)^w \theta \text{ for } w = 0, 1, \dots, \infty.$$

Now consider a sequence of independent and identically distributed geometric random variables, $W_1, W_2, \dots, W_n$. Denote the sum of the geometric random variables by T_n . The random variable T_n has the same distribution as the number of failures before the n^{th} success occurs in a sequence of independent and identically distributed Bernoulli trials. The random variable T_n is said to have a negative binomial distribution and its pmf is

$$f_{T_n}(t) = P(T_n = t) = \binom{t + n - 1}{t} \theta^n (1 - \theta)^t, \text{ for } t = 0, 1, \dots, \infty.$$

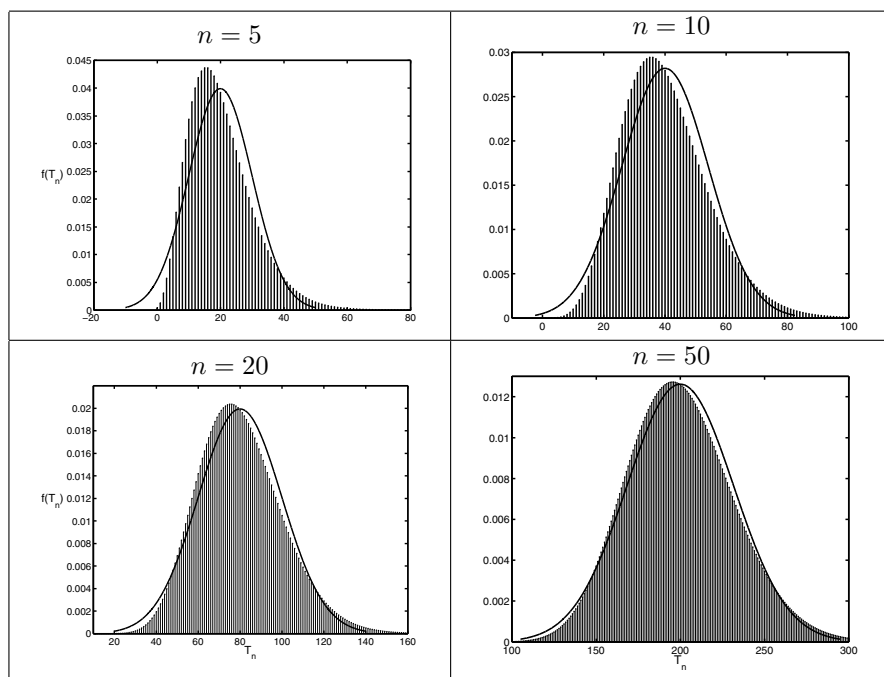


Figure 2. Normal Approximation to the Negative Binomial Distribution, $\theta = 0.2$

It is readily shown that $\mu \stackrel{\text{def}}{=} E(W_i) = (1 - \theta)/\theta$ and $\sigma^2 \stackrel{\text{def}}{=} \text{Var}(W_i) = (1 - \theta)/\theta^2$. Accordingly, it follows from Theorem 1 that

$$\frac{T_n - n(1 - \theta)/\theta}{\sqrt{n(1 - \theta)/\theta^2}} \xrightarrow{\text{dist}} N(0, 1).$$

The exact pmf and the normal approximations to the pmf of T_n for $\theta = 0.2$ and $n = 5, 10, 20$, and 50 are displayed in Figure 2. Table 1 displays the approximation errors. The normal approximation to the pmf was computed using the continuity correction illustrated in (6).

The normal approximation to a negative binomial random variable can be used to make frequentist inferences about the value of the parameter, θ . Suppose that $T_n \sim \text{Negative Binomial}(n, \theta)$ for fixed n and unknown θ . It follows from Theorems 1 and 2 that

$$P \left[\left| \frac{T_n - n(1 - \theta)/\theta - 0.5}{\sqrt{n(1 - \theta)/\theta^2}} \right| \leq z_{\alpha/2} \right] = 1 - \alpha + O(n^{-1/2}),$$

where z_α is the $100(1 - \alpha)$ percentile of $N(0, 1)$. A confidence interval with nominal confidence coefficient $1 - \alpha$ can be obtained by inverting the inequality. The result

is

$$P(L_{n,\alpha} \leq \theta \leq U_{n,\alpha}) = 1 - \alpha + O\left(n^{-1/2}\right), \text{ where}$$

$$L_{n,\alpha} = \frac{n(T_n + 0.5 + n - z_{\alpha/2}^2/2) - z_{\alpha/2} \sqrt{n(T_n + 0.5)(T_n + 0.5 + n) + n^2 z_{\alpha/2}^2/4}}{(T_n + 0.5 + n)^2},$$

$$U_{n,\alpha} = \frac{n(T_n^* - 0.5 + n - z_{\alpha/2}^2/2) + z_{\alpha/2} \sqrt{n(T_n^* - 0.5)(T_n^* - 0.5 + n) + n^2 z_{\alpha/2}^2/4}}{(T_n^* - 0.5 + n)^2},$$

$$\text{and } T_n^* = \begin{cases} 1 & \text{if } T_n = 0, \\ T_n & \text{otherwise.} \end{cases}$$

For example, if $n = 20$ and $T_n = 46$ is observed, then the approximate 95% confidence interval is (0.18, 0.41). This interval is correct when $\theta = 0.2$ because θ is captured by the interval.

The actual confidence coefficient that corresponds to the nominal 95% intervals can be obtained by computing the interval for each value of T_n from 0 to ∞ and then summing the probability of T_n over those intervals that capture θ . For example, if $\theta = 0.2$, then the actual confidence coefficient (i.e., coverage) for nominal 95% intervals is 0.962, 0.960, 0.954, and 0.953 for sample sizes $n = 5$, $n = 10$, $n = 20$, and $n = 50$, respectively.

5 RANDOMIZATION INFERENCE

In many circumstances, frequentist statistical inference can be justified by the act of randomization (i.e., random selection and/or random assignment). In this section, normal approximations are applied to two areas in which randomization justifies statistical inference, namely estimating the mean of a finite population (random selection) and testing hypotheses about equality of population distributions (random assignment). These applications are illustrated in §5.1 and §5.2, respectively.

5.1 Inference about the Mean of a Finite Population

The central limit theorems in §2 require that $\{Y_i\}_{i=1}^{\infty}$ be independently distributed random variables. This requirement is satisfied whenever units are randomly selected from an infinite population or are randomly selected with replacement from a finite population. If n units are selected at random without replacement from a finite population of size $N \geq n$, however, then $\{Y_i\}_{i=1}^n$ are not independently distributed.

Example Suppose that a sample of size n is selected at random and without replacement from a population containing N units. The mean and variance of the

population are

$$\mu \stackrel{\text{def}}{=} E(Y_i) = \frac{1}{N} \sum_{i=1}^N Y_i \text{ and } \sigma^2 \stackrel{\text{def}}{=} \text{Var}(Y_i) = \frac{1}{N} \sum_{i=1}^N (Y_i - \mu)^2.$$

Denote the sample mean based on a sample of size n by $\bar{Y}_n$. Then $\bar{Y}_n = \mu$ with probability 1 and this implies that $\text{Var}(\bar{Y}_n) = 0$. It is readily shown that for any sample size, the joint distribution of $Y_1, Y_2, \dots, Y_n$ is exchangeable. Exchangeability is defined in Appendix A.3. Exchangeability implies that $E(Y_i) = E(Y_1)$ for all i and that $\text{Cov}(Y_i, Y_j) = \text{Cov}(Y_1, Y_2)$ for all $i \neq j$. See Berry and Lindgren [1996, §2.6] for additional details. Accordingly,

$$\begin{aligned} 0 = \text{Var}(\bar{Y}_n) &= \frac{1}{n^2} \left[\sum_{i=1}^n \text{Var}(Y_i) + \sum_{i \neq j}^n \text{Cov}(Y_i, Y_j) \right] \\ &= \frac{1}{n^2} [n\sigma^2 + n(n-1) \text{Cov}(Y_1, Y_2)] \end{aligned}$$

which implies that $\text{Cov}(Y_i, Y_j) = -\sigma^2/(n-1)$ for all $i \neq j$. These nonzero covariances imply that $\{Y_i\}_{i=1}^n$ are not mutually independent.

Fortunately, the central limit theorem can be extended to the case of sampling without replacement from a finite population (David, 1938; Madow, 1948; Erdős and Rényi, 1959; Hajek, 1960). Let $P_1, P_2, \dots, P_\infty$ be a sequence of populations such that the N^{th} population consists of units with values $Y_{N1}, Y_{N2}, \dots, Y_{NN}$. The mean and the r^{th} central moment of P_N are

$$\mu_N = \frac{1}{N} \sum_{j=1}^N Y_{Nj} \text{ and } \mu_{Nr} = \frac{1}{N} \sum_{j=1}^N (Y_{Nj} - \mu_N)^r.$$

Note that $\mu_{N2} = \sigma_N^2$, the variance of P_N . It is assumed that $\sigma_N^2 < \infty$ and that higher-order central moments of P_N are bounded as $N \rightarrow \infty$. Specifically,

$$(7) \quad \mu_{Nr}/\sigma_N^r = O(1) \text{ for } r = 3, 4, \dots \text{ as } N \rightarrow \infty,$$

Let $Y_{N1}, Y_{N2}, \dots, Y_{Nn}$ be a random sample of size n selected from P_N without replacement. Consider the linear function $T_n = \sum_{j=1}^n c_{Nj} Y_{Nj}$, where the coefficients $\{c_{Nj}\}_{j=1}^n$ are fixed numbers that are not all zero. It is readily shown that

$$\begin{aligned} E(T_n) &= n\bar{c}_n\mu_N \text{ and } \text{Var}(T_n) = n(V_{c_n}^2 + \bar{c}_n^2) \sigma_N^2 \left(\frac{N}{N-1} \right) \left[1 - \frac{n\bar{c}_n^2}{N(V_{c_n}^2 + \bar{c}_n^2)} \right], \\ \text{where } \bar{c}_n &= \frac{1}{n} \sum_{j=1}^n c_{Nj}, \text{ and } V_{c_n}^2 = \frac{1}{n} \sum_{j=1}^n (c_{Nj} - \bar{c}_n)^2. \end{aligned}$$

Following Madow [1948], it is assumed that the coefficients satisfy

$$(8) \quad (a) \frac{1}{n} \sum_{j=1}^n c_{Nj}^r / (V_{c_n}^2 + \bar{c}_n^2)^{\frac{r}{2}} = O(1) \text{ for } r = 3, 4, \dots \text{ and (b) } \frac{n\bar{c}_n^2}{N(V_{c_n}^2 + \bar{c}_n^2)} < 1 - \varepsilon,$$

as $n \rightarrow \infty$ and $N \rightarrow \infty$, where ε is a positive constant that does not depend on either n or N . Condition (a) ensures that no single coefficient dominates the rest and condition (b) ensures that $\text{Var}(T_n) > 0$. Under these conditions, the CLT in Theorem 1 can be extended to finite populations. The extension is summarized in Theorem 7.

THEOREM 7 Finite Population Central Limit Theorem. *If (7) and (8) are satisfied then*

$$\frac{T_n - E(T_n)}{\text{Var}(T_n)} \xrightarrow{\text{dist}} N(0, 1)$$

as $n \rightarrow \infty$ and $N \rightarrow \infty$. ■

It is important to note that Theorem 7 requires that both N and n increase. It is not sufficient for only n to increase, regardless of how large N might be (see Plane and Gordon, 1982). To apply Theorem 7 to the sample mean, equate each coefficient c_{Nj} to $1/\sqrt{n}$. Note $n/N < 1$ as $n \rightarrow \infty$ and $N \rightarrow \infty$ is sufficient to satisfy Madow's conditions in (8) for this choice of c_{Nj} . Also $c_{Nj} = 1/\sqrt{n}$ implies that $T_n = \sqrt{n} \bar{Y}_n$, $E(T_n) = \sqrt{n} \mu_N$, and $\text{Var}(T_n) = \sigma_N^2 [1 - (n-1)/(N-1)]$. It follows from Theorem 7 that

$$\frac{\sqrt{n}(\bar{Y}_n - \mu_N)}{\{\sigma_N^2 [1 - (n-1)/(N-1)]\}^{\frac{1}{2}}} \xrightarrow{\text{dist}} N(0, 1).$$

As an illustration, consider the population of size $N = 25$ displayed in Figure 3. The population values were generated by sampling from a Poisson distribution with mean 4. The mean and variance of the finite population are $\mu_N = 4.16$ and $\sigma_N^2 = 4.21$. The sampling distribution of $\bar{Y}_n$ based on all samples of size $n = 5$ selected without replacement is displayed in Figure 4. The $\binom{25}{5} = 53,130$ samples yield only 29 distinct values of $\bar{Y}_n$, namely 1.4 to 7.0 in steps of size 0.2. A normal distribution with mean $E(\bar{Y}_n) = \mu_N = 4.16$ and variance $\text{Var}(\bar{Y}_n) = (\sigma_N^2/n)[1 - (n-1)/(N-1)] = 0.70$ is superimposed on the line graph. In this example, the supremum absolute cdf approximation error is 0.048 and the maximum absolute pmf error is 0.002.

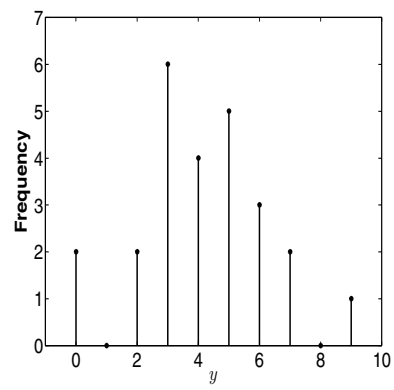
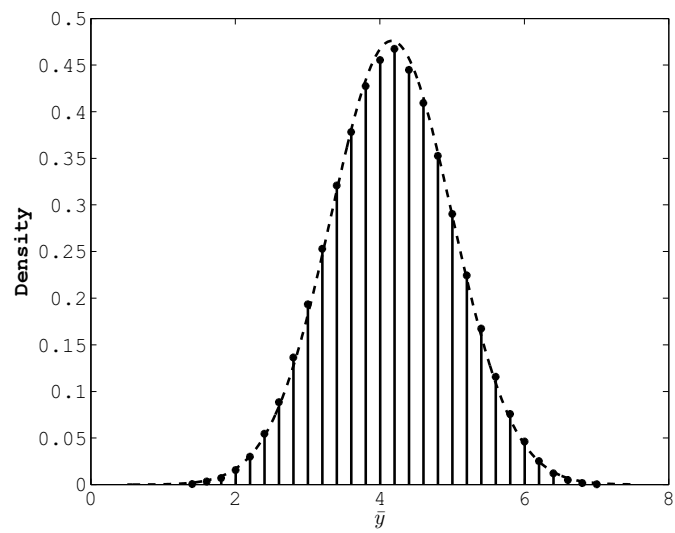
Using $\text{Cov}(X_i, X_j) = -\sigma_N^2/(N-1)$, it is readily shown that

$$\frac{S^2}{n} \left(1 - \frac{n}{N}\right) \text{ is unbiased for } \text{Var}(\bar{Y}_n) = \frac{\sigma_N^2}{n} \left(1 - \frac{(n-1)}{(N-1)}\right),$$

where $S^2 = (n-1)^{-1} \sum_{j=1}^n (Y_{Nj} - \bar{Y}_n)^2$ is the sample variance. Accordingly, the endpoints of two-sided nominal $100(1-\alpha)\%$ confidence intervals, based on Theorem 7, can be constructed as

$$\bar{Y}_n \pm t_{1-\alpha/2, n-1} \left[\frac{S^2}{n} \left(1 - \frac{n}{N}\right) \right]^{\frac{1}{2}},$$

where $t_{1-\alpha/2, n-1}$ is the $100(1-\alpha/2)$ percentile of the t distribution with $n-1$ degrees of freedom. For example, if a sample of size $n = 5$ is randomly selected

Figure 3. Finite Population, $N=25$ Figure 4. Distribution of $\bar{Y}$, Sampling from a Finite Population

and the sample is $(2, 3, 3, 6, 9)$, then $\bar{y} = 4.6$, $s^2 = 8.3$, and the nominal 95% confidence interval is $4.6 \pm 2.776\sqrt{8.3(1 - .2)/5} = (1.40, 7.80)$. This specific interval is correct because it does capture the population mean, $\mu = 4.16$. For the population in Figure 3 and samples of size $n = 5$, the coverage of confidence intervals can be found by computing the proportion of the $\binom{25}{5} = 53,130$ confidence intervals that capture the population mean. For two-sided intervals with nominal confidence coefficients 90%, 95%, and 99%, the coverage is 0.916, 0.958, and 0.995, respectively.

5.2 Permutation Tests

Fisher [1936; 1937] introduced a class of tests that can be used when the form of the population distribution is unknown. The tests are based on the distribution of the sample conditional on a minimal sufficient statistic. Informally, a minimal sufficient statistic is a function of the sample that provides the greatest data reduction while still preserving all information about the unknown parameters that is contained in the sample. More precise definitions of sufficient statistics and minimal sufficient statistics are given in Appendix A.4. Suppose that $Y_1, Y_2, \dots, Y_N$ is a random sample from a continuous distribution and that nothing is known about which or what type of continuous distribution the data have been sampled from. In this case, it can be shown that the vector of order statistics $\mathbf{T} = (Y_{(1)} Y_{(2)} \dots Y_{(N)})'$ is minimal sufficient, where $Y_{(i)}$ is the i^{th} smallest value in the sample. The joint distribution of $Y_1, Y_2, \dots, Y_N$ conditional on the order statistics is

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_N = y_N | \mathbf{T} = \mathbf{t}) \\ = \begin{cases} \frac{1}{N!} & \text{if } y_1, y_2, \dots, y_N \text{ is a permutation of } t_1, t_2, \dots, t_N, \\ 0 & \text{otherwise.} \end{cases}$$

That is, each possible rearrangement (permutation) of the sample is equally likely and has probability $1/N!$ regardless of which continuous distribution the data have been sampled from.

Fisher suggested that this conditional distribution could be used to construct distribution-free hypothesis tests. In this section, a normal approximation to Fisher's permutation test in a one-way analysis of variance is illustrated. Suppose that N experimental units are randomly assigned to k treatments such that n_j units are assigned to treatment j and $N = \sum_{j=1}^k n_j$. Denote the responses of the n_j units under treatment j by Y_{ij} , for $i = 1, \dots, n_j$ and denote their common cdf by $F_j(y)$. The form of F_j is unknown, but it is assumed that its mean (μ_j) and variance are finite and that the treatment effects, if any, merely shift the distribution up or down on the number line. Accordingly, $F_j(y - \mu_j) = F_0(y - \mu_j)$ for $j = 1, \dots, k$, where F_0 is a cdf. The linear model that corresponds to this setting is

$$(9) \quad Y_{ij} = \mu_j + \epsilon_{ij}, \text{ for } i = 1, \dots, n_j \text{ and } j = 1, \dots, k,$$

where $\{\epsilon_{ij}\}$ are iid with mean zero, variance σ^2 , and cdf F_0 . It is of interest to test the null hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_k$. If the k populations are normally distributed, then under H_0 , the conventional F statistic has an F distribution with degrees of freedom $k - 1$ and $N - k$, where

$$F = \frac{SS \text{ Treat} / (k - 1)}{SSE / (N - k)}, \quad N = \sum_{j=1}^k n_j, \quad SS \text{ Treat} = \sum_{j=1}^k n_j (\bar{Y}_j - \bar{Y})^2,$$

$$\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}, \quad \bar{Y} = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} Y_{ij}, \text{ and } SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2.$$

Alternatively, the F statistic can be transformed to a statistic, B that has a Beta distribution with parameters $(k - 1)/2$ and $(N - k)/2$, where

$$B = \frac{(k - 1)F}{(k - 1)F + N - k} = \frac{SS \text{ Treat}}{SS \text{ Total}} \text{ and}$$

$$SS \text{ Total} = SS \text{ Treat} + SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^2.$$

The null hypothesis is rejected for large values of F or, equivalently, for large values of B .

Fisher's permutation test is conducted conditional on the order statistics for the total sample of size N . The conditional sampling distribution of F or B is constructed by computing the value of the test statistic under each of the $N! / \prod_{j=1}^k n_j!$ possible assignments of data to groups. If the null hypothesis is true, then, conditional on the order statistics, each assignment of sample values to the k treatments is equally likely and has probability $1/[N! / \prod_{j=1}^k n_j!]$. Accordingly, an exact test can be constructed by comparing the observed value of F or B to its conditional sampling distribution. If sample sizes are not small, then construction of the exact conditional sampling distribution may not be feasible. For example, if $k = 4$ and $n_1 = n_2 = n_3 = n_4 = 10$, then there are $40! / [(10!)^4] \approx 4.71 \times 10^{21}$ possible assignments of data to groups. Because sample sizes are equal across groups, only $(1/4!)40! / [(10!)^4] \approx 1.96 \times 10^{20}$ assignments of data to groups need be examined. Nonetheless, computation of the test statistic (F or B) under each of these assignments is a formidable task, even with a high speed computer. In this case, using the CLT to approximate the conditional sampling distribution can save time without inducing appreciable loss in accuracy.

This permutation test (exact or normal approximation) must be used with caution, however, even though the form of the distribution, F_0 , is arbitrary. Some investigators have employed the permutation procedure to test the hypothesis of equality of means when model (9) is not satisfied. In particular, the permutation test is sometimes used to test equality of population means when population

variances differ among the k sub-populations. Boik [1987] showed that the permutation test is not a robust alternative to the conventional F test in this situation. Denote the arithmetic mean of the k sample sizes by $\bar{n}$ and denote that harmonic mean of the k sample sizes by $\tilde{n}$. That is,

$$\bar{n} = \frac{1}{k} \sum_{j=1}^k n_j \text{ and } \tilde{n} = k \left(\sum_{j=1}^k \frac{1}{n_j} \right)^{-1}.$$

If population variances are heterogeneous and $\tilde{n}/\bar{n}$ is large (i.e., near 1), then the type I error rate for the permutation test is larger than the type I error rate of the normal-theory F test. If population variances are heterogeneous and $\tilde{n}/\bar{n}$ is small, then the type I error rate for the permutation test is smaller than the type I error rate of the normal-theory F test. In either case, use of the permutation test rather than the normal theory F test does not ensure better control over type I error rate.

Central limit theorems for linear statistics under permutation were obtained by Wald and Wolfowitz [1944], Hoeffding [1951] and others. The finite population version of the CLT that was obtained by Madow [1948] also applies to permutation distributions. In the present application, the finite sample version of the CLT ensures that the joint permutation distribution of the scaled sample means converges in distribution to a multivariate normal distribution.

THEOREM 8 Permutation Central Limit Theorem. *Denote the conditional variance of the $\{Y_{ij}\}$ values by S_N^2 . That is $S_N^2 = SS \text{ Total} / N$. Assume that (a) the higher-order central moments of $\{Y_{ij}\}$ are bounded as in (7) and (b) $n_j/N \rightarrow \lambda_j^2$ as $N \rightarrow \infty$, where $\lambda_j^2 \in (0, 1)$ for $j = 1, \dots, k$. Denote the k -vector with components $\lambda_1, \lambda_2, \dots, \lambda_k$ by $\boldsymbol{\lambda}$. Define Z_{Nj} and $\mathbf{Z}_N$ as*

$$Z_{Nj} = \frac{\sqrt{n_j}(\bar{Y}_j - \bar{Y})}{S_N} \text{ for } j = 1, \dots, k, \text{ and } \mathbf{Z}_N = (Z_{N1} \quad Z_{N2} \quad \cdots \quad Z_{Nk})'.$$

Then, as $N \rightarrow \infty$, the distribution of $\mathbf{Z}_N$, conditional on the order statistics $\mathbf{T}$, converges to the singular multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}_k - \boldsymbol{\lambda}\boldsymbol{\lambda}'$. Note that $\boldsymbol{\Sigma}$ is idempotent with rank $k - 1$. It follows from [Christensen, 1996, Theorem 1.3.6] and Theorem 6 in this chapter that conditionally and unconditionally

$$\frac{SS \text{ Treat}}{SS \text{ Total} / N} = NB = \mathbf{Z}_N' \mathbf{Z}_N \xrightarrow{\text{dist}} \chi_{k-1}^2$$

as $N \rightarrow \infty$. ■

Theorem 8 reveals that the asymptotic permutation distributions of NB and $(k - 1)F$ are identical to their asymptotic distributions when sampling from a normal distribution. In practice, however, the asymptotic χ^2 distribution is rarely employed. A more accurate approximation to the permutation distribution of B

can be obtained by matching the exact permutation moments of B to those of a beta distribution. The first two permutation moments of B are

$$\mu_B = (k-1)/(N-1), \quad \sigma_B^2 = \frac{2(N-k)(k-1)}{(N^2-1)(N+1)} + R,$$

$$\text{where } R = \left(\hat{\kappa}_4 + \frac{6}{N+1} \right) \frac{N(N+1) \sum_{j=1}^k n_j^{-1} - 2(N-k)(k-1) + (N+1)k^2}{N(N-1)(N-2)(N-3)},$$

$$\text{and } \hat{\kappa}_4 = \frac{\frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y})^4}{(SS \text{ Total} / N)^2} - 3.$$

The quantity $\hat{\kappa}_4$ is the kurtosis of the $\{Y_{ij}\}$ values. See [Pitman, 1937; Box and Anderson, 1955; Robinson, 1983] for details on computing permutation moments. Matching the conditional moments of B to those of a Beta(α, β) distribution yields

$$\alpha = \mu_B \left[\frac{\mu_B(1-\mu_B)}{\sigma_B^2} - 1 \right] = \frac{k-1}{2} [1 + O(N^{-1})] \text{ and}$$

$$\beta = (1-\mu_B) \left[\frac{\mu_B(1-\mu_B)}{\sigma_B^2} - 1 \right] = \frac{N-k}{2} [1 + O(N^{-1})].$$

Under normality, the exact values of α and β are $(k-1)/2$ and $(N-k)/2$ respectively.

For example, Table 2 displays a stem and leaf plot of $N = 15$ values sampled from the χ_3^2 distribution and then rounded to 2 decimal places. Suppose that $k = 3$ and sample sizes are $n_1 = 4$, $n_2 = 5$, and $n_3 = 6$. The fitted beta parameters are $\alpha = 0.958$ and $\beta = 5.748$. The exact normal-theory values are 1 and 6, respectively. Figure 5 displays the difference between the exact permutation cdf of B and the beta(0.958, 5.748) cdf approximation. The exact permutation cdf is based on all 630,630 possible assignments of data to groups. The absolute difference in Figure 5 is no larger than 0.006, except when b is near zero. If the beta(0.958, 5.748) distribution is used as the reference distribution for testing H_0 , then the actual test sizes that correspond to nominal test sizes of 0.10, 0.05, and 0.01 are 0.0982, 0.0506, and 0.0116, respectively.

As a specific example, suppose that the $n = 15$ values in Table 2 are randomly assigned to $k = 3$ groups and the results in Table 3 are observed. For this data set, $SS \text{ Total} = 40.516$, $SS \text{ Treat} = 1.8905$, and $b = 0.0467$. The beta approximation to the p -value is $P(B \geq 0.0467) = 0.742$, where $B \sim \text{beta}(0.958, 5.748)$. The exact p -value (conditional on the order statistics) computed from the permutation distribution is $P(B \geq 0.0467 | \mathbf{T}) = 0.746$.

6 LIKELIHOOD-BASED INFERENCE

Suppose that $\mathbf{Y}_i$ for $i = 1, \dots, n$ are mutually independent observable random p -vectors, where the pdf or pmf for $\mathbf{Y}_i$ is $f_i(\mathbf{y}_i; \boldsymbol{\theta})$ and $\boldsymbol{\theta}$ is a k -vector of unknown

Table 2. Stem and Leaf Plot of Sample of size $N = 15$ from χ_3^2 . Stem unit is ones, leaf unit is hundredths

Stems	Leaves		
0	69		
1	16	19	28
2	03	23	60
3	11	51	95
4	00	19	56
5	45		
6	46		

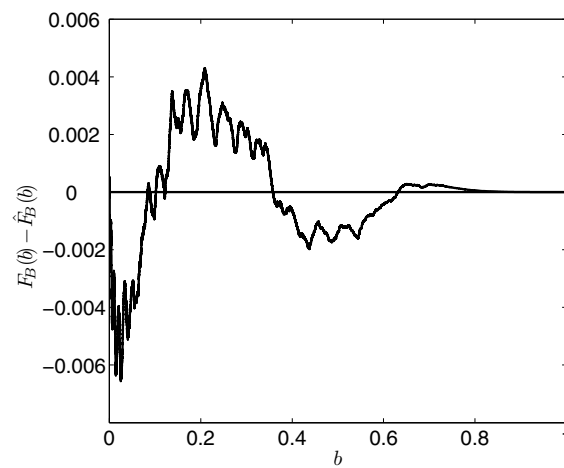


Figure 5. Normal Approximation to the Permutation Sampling Distribution of B

Table 3. Data for $k = 3$ Groups

	Group 1	Group 2	Group 3
	3.95	2.60	4.56
	4.00	2.23	2.03
	1.19	1.16	1.28
	5.45	4.19	3.11
		3.51	6.46
			0.69
Mean	3.65	2.74	3.02

parameters. The function f_i may depend on known covariates. Denote the parameter space for $\boldsymbol{\theta}$ by Θ . The likelihood function of $\boldsymbol{\theta}$ given the data is defined as the joint pdf or pmf of $\{\mathbf{Y}_i\}_{i=1}^n$ given $\boldsymbol{\theta}$. In the present case, the $\mathbf{Y}_i$ vectors are independent so the likelihood and log likelihood functions are

$$(10) \quad L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^n f_i(\mathbf{y}_i; \boldsymbol{\theta}) \text{ and } \ell(\boldsymbol{\theta}; \mathbf{y}) \stackrel{\text{def}}{=} \ln[L(\boldsymbol{\theta}; \mathbf{y})] = \sum_{i=1}^n \ln[f_i(\mathbf{y}_i; \boldsymbol{\theta})],$$

respectively, where $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^n$ and $\mathbf{y}_i$ is a realization of $\mathbf{Y}_i$. Note that $L(\boldsymbol{\theta}; \mathbf{y})$ is a function of the observed values $\mathbf{y}$, whereas $L(\boldsymbol{\theta}; \mathbf{Y})$ is a function of the random variables $\mathbf{Y} = \{\mathbf{Y}_i\}_{i=1}^n$.

The method of maximum likelihood estimation was introduced by Fisher [1912; 1922; 1925]. A maximum likelihood estimate (mle) of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ that maximizes the likelihood function $L(\boldsymbol{\theta}; \mathbf{y})$. The maximum likelihood estimator (MLE) is a random variable and is the maximizer of $L(\boldsymbol{\theta}; \mathbf{Y})$. Both the mle and the MLE can be denoted by $\hat{\boldsymbol{\theta}}$, provided that the context provides a clear distinction. Many properties of the MLE have been established, including consistency [Doob, 1934; Cramér, 1946; Wald, 1949], asymptotic normality [Doob, 1936; Cramér, 1946; Wald, 1943; Le Cam, 1970], and efficiency [Daniels, 1961].

Theorem 9 gives the asymptotic distribution of the MLE under the assumption that $\{\mathbf{Y}_i\}_{i=1}^n$ are mutually independent but not necessarily identically distributed. The regularity conditions listed in Appendix C.1 are sufficient to ensure validity of the theorem. A heuristic proof is given in Appendix B.3. Details and complete proofs were given by Bradley and Gart [1962] and Hoadley [1971]. The theorem can be extended to the case where $\{\mathbf{Y}_i\}_{i=1}^n$ are not independently distributed [Lehmann and Casella, 1998, §6.7].

THEOREM 9 Asymptotic Normality of MLEs. *If the regularity conditions in Appendix C.1 are satisfied, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\text{dist}} N\left(\mathbf{0}, \bar{\mathbf{I}}_{\boldsymbol{\theta}, \infty}^{-1}\right)$$

where $\bar{\mathbf{I}}_{\boldsymbol{\theta}, \infty}$ is the limit of the average information matrix as $n \rightarrow \infty$, $\hat{\boldsymbol{\theta}}$ is a consistent root of the likelihood equations $\mathbf{S}(\boldsymbol{\theta}; \mathbf{Y}) = \mathbf{0}$ and $\mathbf{S}(\boldsymbol{\theta}; \mathbf{Y})$ is the score function. The score function and the average limiting information matrix are defined as

$$\begin{aligned} \mathbf{S}(\boldsymbol{\theta}; \mathbf{Y}) &\stackrel{\text{def}}{=} \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i), \text{ and } \bar{\mathbf{I}}_{\boldsymbol{\theta}, \infty} = \lim_{n \rightarrow \infty} n^{-1} \mathbf{I}_{\boldsymbol{\theta}, n}, \text{ where} \\ \mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i) &\stackrel{\text{def}}{=} \frac{\partial \ln[f_i(\mathbf{Y}_i; \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \text{ and } \mathbf{I}_{\boldsymbol{\theta}, n} \stackrel{\text{def}}{=} E[\mathbf{S}(\boldsymbol{\theta}; \mathbf{Y}) \mathbf{S}(\boldsymbol{\theta}; \mathbf{Y})']. \end{aligned}$$

■

In practice, Theorem 9 is employed by using the approximation

$$\hat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}, \mathbf{I}_{\boldsymbol{\theta}, n}^{-1}\right).$$

For example, consider a dose-response study in which the outcome is binary. Denote the response of the i^{th} experimental unit by Y_i , where $Y_i = 1$ if the drug is effective (i.e., success) and $Y_i = 0$ otherwise. Denote the probability of success for the i^{th} case by τ_i ; i.e., $P(Y_i = 1) = \tau_i$. Suppose that τ_i satisfies the logit function: $\ln[\tau_i/(1 - \tau_i)] = \mathbf{x}_i' \boldsymbol{\theta}$, where $\mathbf{x}_i$ is a $k \times 1$ vector of known covariates and $\boldsymbol{\theta}$ is a $k \times 1$ vector of unknown parameter values. These assumptions yield the familiar logistic regression model for Y_i for $i = 1, \dots, n$. It follows that

$$\tau_i = \frac{e^{\mathbf{x}_i' \boldsymbol{\theta}}}{1 + e^{\mathbf{x}_i' \boldsymbol{\theta}}},$$

$$L(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{i=1}^n \tau_i^{Y_i} (1 - \tau_i)^{1-Y_i}, \quad \ell(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{i=1}^n [Y_i \ln(\tau_i) + (1 - Y_i) \ln(1 - \tau_i)],$$

$$\mathbf{S}(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{i=1}^n \mathbf{x}_i (Y_i - \tau_i), \quad \text{and} \quad \mathbf{I}_{\boldsymbol{\theta}, n} = \sum_{i=1}^n \mathbf{x}_i \tau_i (1 - \tau_i) \mathbf{x}_i'.$$

As an illustration, suppose that $k = 2$, $\boldsymbol{\theta} = (-1 \ 0.2)'$, and $\mathbf{x}_i = (1 \ i)'$ with frequency $n_i = 10$ for $i = 1, \dots, 5$. For this model, the inverse information matrix is

$$\text{Var}(\hat{\boldsymbol{\theta}}) \approx \mathbf{I}_{\boldsymbol{\theta}, n}^{-1} = \begin{pmatrix} 0.492 & -0.132 \\ -0.132 & 0.043 \end{pmatrix}.$$

To examine the accuracy of the normal approximation, 100,000 data sets that followed the logistic regression model were randomly generated. The sample size for each data set was $n = \sum_{i=1}^5 n_i = 50$. The simulation-based means and variances of $\hat{\boldsymbol{\theta}}$ are

$$\widehat{\text{E}}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} -1.058 \\ 0.213 \end{pmatrix} \text{ compared to } \boldsymbol{\theta} = \begin{pmatrix} -1.0 \\ 0.2 \end{pmatrix} \text{ and}$$

$$\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} 0.565 & -0.152 \\ -0.152 & 0.049 \end{pmatrix} \text{ compared to } \mathbf{I}_{\boldsymbol{\theta}, n}^{-1} = \begin{pmatrix} 0.492 & -0.132 \\ -0.132 & 0.043 \end{pmatrix}.$$

It is apparent that the MLEs are slightly biased and that $\mathbf{I}_{\boldsymbol{\theta}, n}^{-1}$ slightly underestimates the covariance matrix of $\hat{\boldsymbol{\theta}}$.

To capture the bivariate nature of the distribution of $\hat{\boldsymbol{\theta}}$, the asymptotic principal components of $\hat{\boldsymbol{\theta}}$ can be examined. Principal components are successive linear functions that have maximum variability subject to being uncorrelated with previous linear functions and subject to a scale constraint on the coefficients of the linear functions. These linear functions are obtained by conducting an eigen-analysis on the asymptotic covariance matrix. The two principal components are

$$PCA = \begin{pmatrix} PCA_1 \\ PCA_2 \end{pmatrix} = \begin{pmatrix} 0.965 & -0.263 \\ 0.263 & 0.965 \end{pmatrix} \hat{\boldsymbol{\theta}} \text{ with mean and variance}$$

$$\text{E}(PCA) = \begin{pmatrix} -1.017 \\ -0.070 \end{pmatrix} \text{ and } \text{Var}(PCA) = \begin{pmatrix} 0.528 & 0.000 \\ 0.000 & 0.007 \end{pmatrix}.$$

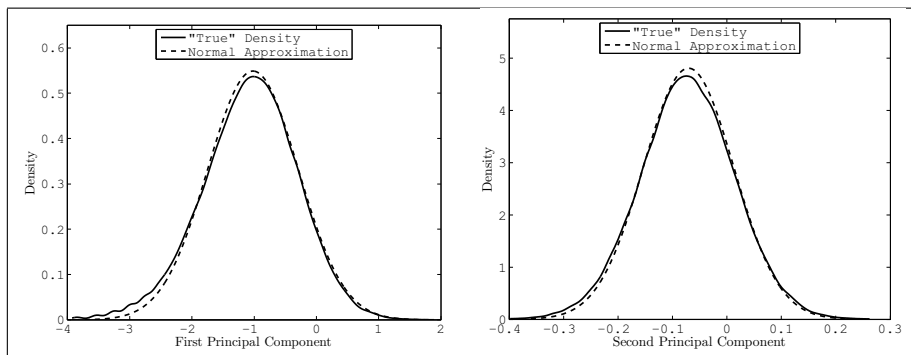


Figure 6. Normal Approximation to the Distribution of MLEs

The simulation-based means and variances of the principal components of $\hat{\boldsymbol{\theta}}$ (using asymptotic coefficients) are

$$\widehat{\mathbf{E}}(PCA) = \begin{pmatrix} -1.076 \\ -0.073 \end{pmatrix}, \text{ and } \widehat{\mathbf{Var}}(PCA) = \begin{pmatrix} 0.607 & 0.000 \\ 0.000 & 0.008 \end{pmatrix}.$$

Figure 6 displays the “true” and the normal approximations to the pdf of each of the two principal components. The “true” pdfs were estimated by applying a kernel density estimator to the simulation results. With minor exceptions, the normal approximations are reasonably accurate.

7 BAYESIAN POSTERIOR DISTRIBUTIONS

Let $\boldsymbol{\theta}$ be a k vector of parameters with parameter space $\boldsymbol{\theta} \in \Theta$. Suppose that, prior to observing data from the study at hand, the investigator can summarize his/her beliefs about $\boldsymbol{\theta}$ as the probability density function $h(\boldsymbol{\theta})$. Let $\mathbf{Y}_i$ for $i = 1, \dots, n$ be independent random p -vectors with pdf or pmf $f_i(\mathbf{y}_i; \boldsymbol{\theta})$, where f_i may depend on known covariates. The likelihood and log likelihood of $\boldsymbol{\theta}$ given the data are given in (10).

The distribution of $\boldsymbol{\theta}$ conditional on $\mathbf{Y} = \mathbf{y}$ is called the posterior distribution of $\boldsymbol{\theta}$. Using Bayes Theorem, the posterior distribution can be expressed as follows:

$$h(\boldsymbol{\theta}; \mathbf{y}) = \frac{h(\boldsymbol{\theta}) \exp \{ \ell_n(\boldsymbol{\theta}; \mathbf{y}) \}}{\int_{\Theta} h(\boldsymbol{\theta}) \exp \{ \ell_n(\boldsymbol{\theta}; \mathbf{y}) \} d\boldsymbol{\theta}}.$$

In practice, the posterior density can be difficult to evaluate and, therefore, approximations can be useful. It was recognized as early as 1774 [Laplace, 1774] that if $\boldsymbol{\theta}$ is suitably scaled and mild regularity conditions are satisfied, then its asymptotic posterior distribution is normal. Proofs of the asymptotic normality of

posterior distributions under various conditions were obtained by Walker [1969], Dawid [1970], Heyde and Johnstone [1979], Chen [1985], and Schervish [1995, §7.4.2]. Extensions to semiparametric and nonparametric likelihoods were given by Shen [2002]. 'indexinvariant prior

The asymptotic normal result applied to independent, but not necessarily identically distributed random variables is summarized in Theorem 10. A heuristic proof is sketched in Appendix B.4. The proof is based on an expansion of the log likelihood function around $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is the mle of $\boldsymbol{\theta}$.

THEOREM 10 Bayesian Central Limit Theorem. *Under the regularity conditions in Appendix C.2, the asymptotic posterior distribution of $\boldsymbol{\theta}$ conditional on the data is the following:*

$$\bar{\mathbf{J}}_{\hat{\boldsymbol{\theta}},n}^{\frac{1}{2}} \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \xrightarrow{\text{dist}} \mathbf{N}(\mathbf{0}, \mathbf{I}_k) \text{ where } \bar{\mathbf{J}}_{\hat{\boldsymbol{\theta}},n} = n^{-1} \mathbf{J}_{\hat{\boldsymbol{\theta}},n}, \quad \mathbf{J}_{\hat{\boldsymbol{\theta}},n} = - \left. \frac{\partial^2 \ell_n(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \otimes \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$. The matrix $\mathbf{J}_{\hat{\boldsymbol{\theta}},n}$ is the observed information matrix evaluated at the mle. ■

In practice, Theorem 10 is used by approximating the posterior distribution of $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} \sim \mathbf{N}(\hat{\boldsymbol{\theta}}, \mathbf{J}_{\hat{\boldsymbol{\theta}},n}^{-1}).$$

Note that the normal asymptotic posterior distribution does not depend on the prior, $h(\boldsymbol{\theta})$. This occurs because as $n \rightarrow \infty$, the information about $\boldsymbol{\theta}$ that is contained in the data dominates the information about $\boldsymbol{\theta}$ that is contained in the prior. If the conditions in Appendix C.2 are not satisfied, then the posterior distribution need not be asymptotically normal. Examples that illustrate the necessity of these regularity conditions were described by Le Cam and Yang [1990, §7.5], and by Gelman *et al.* [1995; §4.3].

An example, in the spirit of Laplace [1774], concerns the posterior distribution of the binomial parameter. Suppose that n independent trials are conducted, where the outcome of each trial is success or failure. Denote the outcome of trial i by Y_i , where $Y_i = 1$ if the trial is a success and $Y_i = 0$ if the trial is a failure. If the probability of success, say θ , is constant over trials, then the total number of successes, $T_n = \sum_{i=1}^n Y_i$, has a binomial distribution with parameters n and θ . Suppose that the investigator's prior beliefs about θ can be summarized as $\theta \sim \text{Beta}(\alpha, \beta)$, where α and β are the two parameters of the beta distribution. Common choices for α and β are (1, 1) which corresponds to the uniform distribution and (0.5, 0.5) which corresponds to Jeffreys's [1961] invariant prior. The,

exact posterior density of θ is

$$\begin{aligned} h(\boldsymbol{\theta}; \mathbf{y}) &= \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1} \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i} d\theta} \\ &= \frac{\theta^{t_n+\alpha-1}(1-\theta)^{n-t_n+\beta-1}}{B(t_n+\alpha, n-t_n+\beta)} I_{(0,1)}(\theta), \end{aligned}$$

where $B(a, b)$ is the beta function, t_n is the total number of observed successes, and $I_{(0,1)}(\theta)$ is the indicator function that takes value 1 if $\theta \in (0, 1)$ and zero otherwise. Accordingly, $\theta|\mathbf{y} \sim \text{Beta}(t_n + \alpha, n - t_n + \beta)$. From Theorem 10, the asymptotic posterior distribution of θ , after suitable scaling, is

$$\frac{\sqrt{n}(\theta - \hat{\theta})}{\sqrt{\hat{\theta}(1 - \hat{\theta})}} \xrightarrow{\text{dist}} N(0, 1),$$

where $\hat{\theta} = t_n/n$. In practice, the Bayesian central limit theorem can be applied by using

$$\theta|\mathbf{y} \sim N\left[\hat{\theta}, n^{-1}\hat{\theta}(1 - \hat{\theta})\right].$$

Figure 7 displays the exact posterior distribution of θ together with the normal approximation for two cases, namely $n = 20$, $t_n = 3$ and $n = 60$, $t_n = 9$. In both cases, $\alpha = \beta = 0.5$. The normal approximation is reasonably accurate for the $n = 60$ case, but not for the $n = 20$ case. Approximate Bayesian $100(1 - \alpha)\%$ credible intervals are given by

$$\hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}},$$

where z_α is the $100(1 - \alpha)$ percentile of $N(0, 1)$. Note that these credible intervals are identical to the conventional large-sample frequentist intervals for $\boldsymbol{\theta}$. Nonetheless, the interpretation is quite different. In the credible interval, the endpoints are fixed and θ is random. The probability that θ falls in the fixed interval is approximately $1 - \alpha$.

8 EXPANSIONS

Convergence of a random variable in distribution to a normal distribution does not guarantee that the normal approximation will be sufficiently accurate when employed with finite sample sizes. Theorem 2 revealed that the cdf approximation error is of magnitude $O(n^{-1/2})$ and, for small sample size this error can be substantial. Various approaches are available for improving the normal approximation.

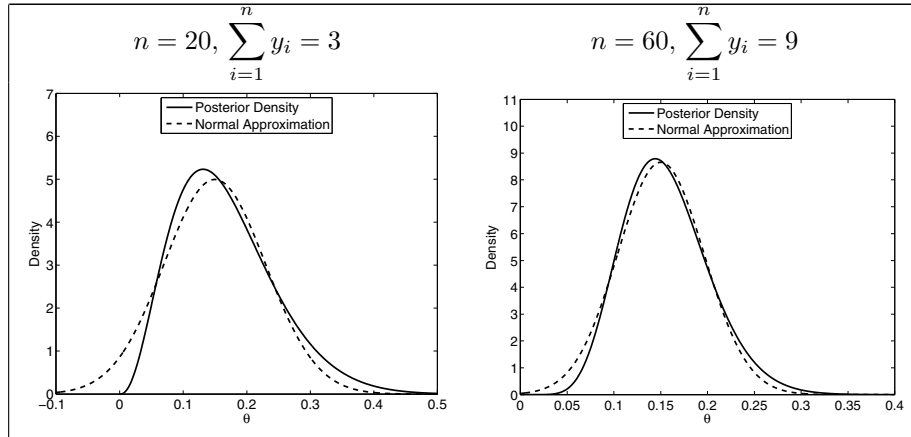


Figure 7. Normal Approximation to the Posterior Distribution

This section describes the Edgeworth expansion for the sampling distribution of a random quantity. An expansion of the Bayesian posterior distribution also is described. These expansions reduce the magnitude of the cdf approximation error to $O(n^{-1})$ or even to $O(n^{-3/2})$, depending on how many terms in the expansion are retained.

8.1 Edgeworth Expansions

The expansion described in this section was pioneered by Chebyshev [1890], Edgeworth [1905; 1907] and Charlier [1905]. See [Cramér, 1972] for additional historical notes.

Suppose that W_n is a continuous scalar random variable that satisfies $W_n \xrightarrow{\text{dist}} N(0, 1)$. In practice, W_n often is of the form

$$W_n = \frac{\sqrt{n}(T_n - \theta)}{\sigma_T} \text{ or } W_n = \frac{\sqrt{n}(T_n - \theta)}{\hat{\sigma}_T},$$

where $n \text{Var}(T_n) \rightarrow \sigma_T^2$ as $n \rightarrow \infty$ and $\hat{\sigma}_T^2$ is a consistent estimator of σ_T^2 . Typically, the random variable T_n is a function of sample means or a function of an estimator obtained as the solution to an estimating equation (e.g., a MLE). Denote the mean and variance of W_n by μ_n and σ_n^2 , respectively, and denote the r^{th} cumulant of W_n by $\kappa_r(W_n)$. In particular,

$$\begin{aligned} \kappa_1(W_n) &= \mu_n = E(W_n), & \kappa_2(W_n) &= \sigma_n^2 \stackrel{\text{def}}{=} \text{Var}(W_n), \\ \kappa_3(W_n) &= E(W_n - \mu_n)^3, & \text{and } \kappa_4(W_n) &= E(W_n - \mu_n)^4 - 3\sigma_n^4. \end{aligned}$$

Typically, the first four cumulants of W_n satisfy

$$\begin{aligned}\kappa_1(W_n) &= \frac{\omega_1}{\sqrt{n}} + O(n^{-3/2}), \quad \kappa_2(W_n) = 1 + \frac{\omega_2}{n} + O(n^{-2}), \\ \kappa_3(W_n) &= \frac{\omega_3}{\sqrt{n}} + O(n^{-3/2}), \quad \text{and } \kappa_4(W_n) = \frac{\omega_4}{n} + O(n^{-2}),\end{aligned}\quad (11)$$

where ω_j has magnitude $O(1)$ for $j = 1, \dots, 4$.

An Edgeworth expansion of the distribution of W_n modifies the standard normal approximation such that the first r cumulants (typically 3 or 4) of the approximating distribution match those of W_n . The terms in the expansion are functions of Hermite polynomials. Let z be a scalar constant. Then, the r^{th} Hermite polynomial, $H_r(z)$, is defined as

$$H_r(z) \stackrel{\text{def}}{=} \frac{(-1)^r}{\varphi(z)} \frac{d^r \varphi(z)}{(dz)^r}, \quad \text{for } r = 1, 2, \dots$$

where $\varphi(z) = \varphi(z, 0, 1)$ is defined in (1). The first seven Hermite polynomials are the following:

$$\begin{aligned}H_0(z) &= 1, \quad H_1(z) = z, \quad H_2(z) = z^2 - 1, \quad H_3(z) = z^3 - 3z, \quad H_4(z) = z^4 - 6z^2 + 3, \\ H_5(z) &= z^5 - 10z^3 + 15z, \quad \text{and } H_6(z) = z^6 - 15z^4 + 45z^2 - 15.\end{aligned}$$

The Edgeworth expansion of the distribution of W_n up to order $O(n^{-1})$ is given in Theorem 11. A proof is sketched in Appendix B.5

THEOREM 11 *Edgeworth. Regularity conditions for Edgeworth expansions are described by Bhattacharya and Rao [1976, §19-20] and Hall [1992b, §2.4]. If the regularity conditions are satisfied, then the pdf and cdf of W_n can be expanded as follows:*

$$f_n(w) = \varphi(w) \left[1 + \frac{\omega_1 H_1(w)}{\sqrt{n}} + \frac{g_2 H_2(w)}{2n} + \frac{\omega_3 H_3(w)}{6\sqrt{n}} + \frac{g_4 H_4(w)}{24n} + \frac{\omega_3^2 H_6(w)}{72n} \right] + O(n^{-3/2}),$$

and

$$F_n(w) = \Phi(w) - \varphi(w) \left[\frac{\omega_1}{\sqrt{n}} + \frac{g_2 H_1(w)}{2n} + \frac{\omega_3 H_2(w)}{6\sqrt{n}} + \frac{g_4 H_3(w)}{24n} + \frac{\omega_3^2 H_5(w)}{72n} \right] + O(n^{-3/2}),$$

where $g_2 = \omega_2 + \omega_1^2$, $g_4 = \omega_4 + 4\omega_1\omega_3$,

φ is the standard normal pdf, Φ is the standard normal cdf, and $H_j(w)$ is the j^{th} Hermite polynomial. ■

As an example, suppose that Y_i for $i = 1, \dots, n$ are iid $\chi_{\nu, \lambda}^2$ random variables; i.e., each Y_i has a noncentral chi-squared distribution with ν degree of freedom and noncentrality parameter λ . Let $W_n = \sqrt{n}(\bar{Y}_n - \mu)/\sigma$, where $\mu = E(Y_i) = \nu + 2\lambda$

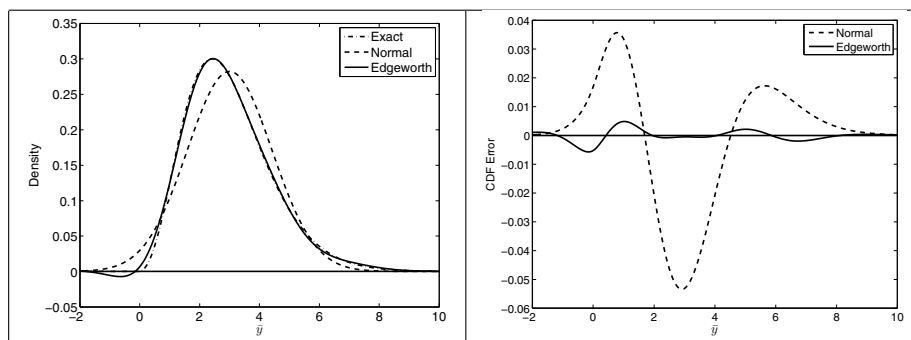


Figure 8. Edgeworth Expansion of the Distribution of $\bar{Y}_n$, $n = 5$ and $Y_i \stackrel{\text{iid}}{\sim} \chi_{1,1}^2$

and $\sigma^2 = \text{Var}(Y_i) = 2\nu + 8\lambda$. The first four cumulants of W_n are readily shown to be

$$\kappa_1(W_n) = 0, \quad \kappa_2(W_n) = 1, \quad \kappa_3(W_n) = \frac{4\nu + 24\lambda}{\sqrt{2n(\nu + 4\lambda)}^{\frac{3}{2}}}, \quad \text{and} \quad \kappa_4(W_n) = \frac{12\nu + 96\lambda}{n(\nu + 4\lambda)^2}.$$

Accordingly, the cumulant functions required in Theorem 11 are

$$\omega_1 = 0, \quad \omega_2 = 1, \quad \omega_3 = \frac{4\nu + 24\lambda}{\sqrt{2}(\nu + 4\lambda)^{\frac{3}{2}}}, \quad \text{and} \quad \omega_4 = \frac{12\nu + 96\lambda}{(\nu + 4\lambda)^2}.$$

Figure 8 displays the normal approximation and the Edgeworth expansion for the pdf of $\bar{Y}_n$ when $n = 5$, and $Y_i \stackrel{\text{iid}}{\sim} \chi_{1,1}^2$. Also shown is the cdf error of the normal approximation and of the Edgeworth expansion. The Edgeworth expansion represents a substantial improvement over the normal approximation. The pdf plot also reveals a problem with the Edgeworth expansion; namely that the approximating function can be negative and therefore it is not a pdf. Hall's [1992a] cubic transformation remedies this problem for Edgeworth expansions up to order $O(n^{-1/2})$.

The Edgeworth and related expansions have been applied to many commonly used test statistics and pivotal quantities. The expansions typically are used to improve control of type I error rate or to improve coverage of confidence intervals. For example, the expansions can be used to improve the coverage of nominal $100(1 - \alpha)\%$ confidence intervals from $1 - \alpha + O(n^{-1/2})$ to $1 - \alpha + O(n^{-1})$ or even to $1 - \alpha + O(n^{-3/2})$. Edgeworth and related expansions have been applied to t statistics [Johnson, 1978], regression coefficients [Boik, 2008a], simple, partial and multiple correlation coefficients [Boik and Haaland, 2006; Ogasawara, 2006a; 2006b], eigenvalues of correlation and covariance matrices [Boik, 2005], estimators of principle components weights and variances [Ogasawara, 2006c], and factor loadings in factor analysis [Ogasawara, 2007]. Boik [2008b] gave a method for

obtaining Edgeworth expansions for parameter estimators when the parameters are subject to constraints. —indexEdgeworth expansion

8.2 Expansion of Posterior Distributions

Various higher-order expansions for the posterior distribution have been proposed (e.g., [Woodroffe, 1992; Sun, 1994; Ghosal & Samanta, 1997; Weng & Woodroffe, 2000]). In this section, the expansion proposed by Johnson [1970] is described. For this expansion, it is assumed that the likelihood function depends on a scalar parameter, θ . Johnson's expansion is based on a second-order expansion of the log likelihood function evaluated at the mle. To employ Johnson's expansion for the posterior distribution of θ , the following quantities must be evaluated:

$$h^{(j)}(\hat{\theta}) \stackrel{\text{def}}{=} \frac{d^j h(\theta)}{(d\theta)^j} \Big|_{\theta=\hat{\theta}}, \quad a_{nj} \stackrel{\text{def}}{=} \left(\frac{1}{n j!} \right) \frac{\partial^j \ell(\theta; \mathbf{y})}{(\partial \theta)^j} \Big|_{\theta=\hat{\theta}}, \quad \text{and } \sigma_{\hat{\theta}}^2 \stackrel{\text{def}}{=} -\frac{1}{2a_{n2}},$$

where $h(\theta)$ is the prior pdf and $\ell(\theta; \mathbf{y})$ is the log likelihood function. Johnson showed that if the required regularity conditions are satisfied, then the posterior cdf of $Z = \sqrt{n}(\theta - \hat{\theta})/\sigma_{\hat{\theta}}$ is

$$\begin{aligned} h(z; \mathbf{y}) &= P(Z \leq z | \mathbf{y}) = \Phi(z) \\ &- \varphi(z) \left[\frac{c_{31}(z^2 + 2) + c_{32}}{\sqrt{n}} + \frac{c_{41}(z^5 + 5z^3 + 15z) + c_{42}(z^3 + 3z) + c_{43}z}{n} \right] + O\left(n^{-\frac{3}{2}}\right), \\ \text{where } c_{31} &= \sigma_{\hat{\theta}}^3 a_{n3}, \quad c_{32} = \frac{\sigma_{\hat{\theta}} h^{(1)}(\hat{\theta})}{h(\hat{\theta})}, \quad c_{41} = \frac{\sigma_{\hat{\theta}}^6 a_{n3}^2}{2}, \\ c_{42} &= \sigma_{\hat{\theta}}^4 \left[a_{n4} + \frac{a_{n3} h^{(1)}(\hat{\theta})}{h(\hat{\theta})} \right], \quad \text{and } c_{43} = \frac{\sigma_{\hat{\theta}}^2 h^{(2)}(\hat{\theta})}{2h(\hat{\theta})}. \end{aligned}$$

Because of a typographical error, the term c_{02} in equation 2.26 in Johnson needs to be divided by 2 to obtain the above expression for c_{43} . Figure 9 displays the exact posterior as well as the expansions of the posterior pdfs and cdfs to $O(n^{-1/2})$ and to $O(n^{-1})$ for the example in §7. The expansion of the pdf is obtained by differentiating the expansion of the cdf. The expansions yield substantial improvement over the normal approximation, with the expansion to order $O(n^{-1})$ being most accurate.

Higher-order expansions of the posterior density can be used to improve the coverage of nominal $100(1 - \alpha)\%$ credible intervals from $1 - \alpha + O(n^{-1/2})$ to $1 - \alpha + O(n^{-1})$ or even to $1 - \alpha + O(n^{-3/2})$.

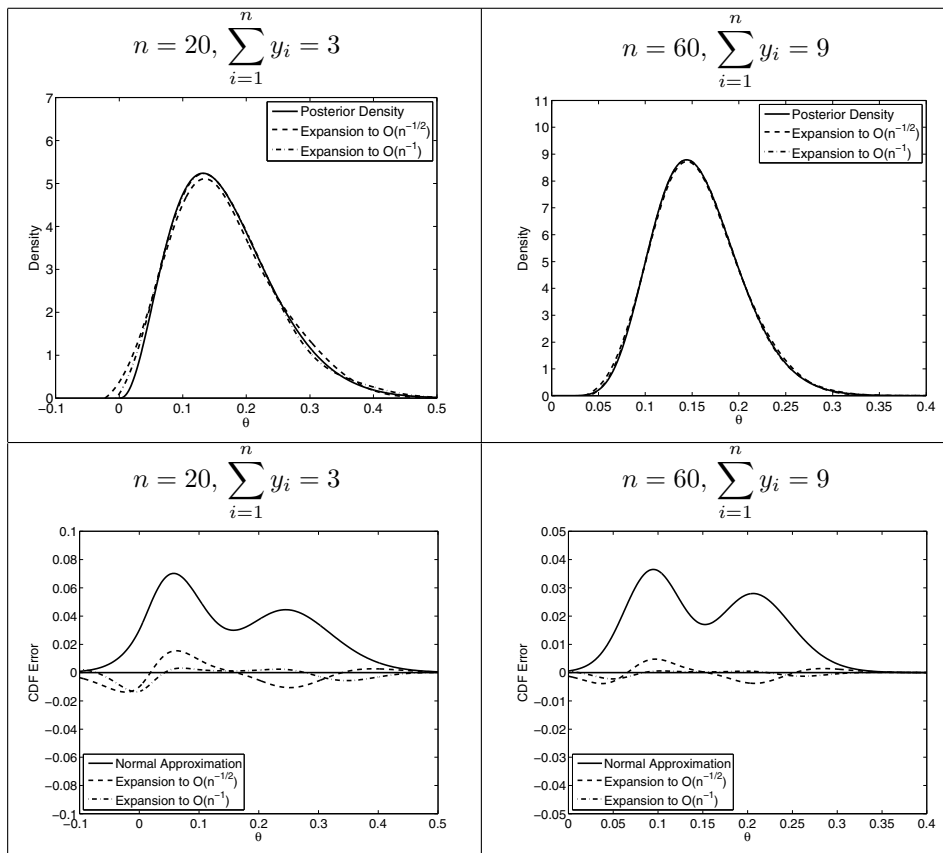


Figure 9. Expansion of Posterior Distribution.

A SELECTED MATHEMATICAL AND STATISTICAL CONCEPTS

A.1 Conditional Distributions

Suppose that $\mathbf{X}$ is a p -vector of random variables, $\mathbf{Y}$ is a q -vector of random variables, and $(\mathbf{X}, \mathbf{Y})$ has a joint probability function, $f_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y})$, with respect to σ -finite measures. Then the probability function of $\mathbf{Y}$, conditional on $\mathbf{X} = \mathbf{x}$, is

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \stackrel{\text{def}}{=} \frac{f_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})},$$

where $f_{\mathbf{X}}(\mathbf{x})$ is the marginal probability function for $\mathbf{X}$. In particular, if $\mathbf{Y}$ is a continuous random vector, then $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ is a probability density function (pdf) and if $\mathbf{Y}$ is a discrete random vector, then $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ is a probability mass function (pmf).

A.2 Convergence Concepts

Definitions for a few of the most important terms and concepts about convergence are given in this section.

1. Convergence in Probability. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables, let X be a random variable, and let c be a constant. The notation $X_n \xrightarrow{\text{prob}} X$ or $X_n \xrightarrow{\text{prob}} c$ is read as X_n converges in probability to X (or to c). Furthermore,

$$X_n \xrightarrow{\text{prob}} X \text{ if } \lim_{n \rightarrow \infty} P(|X_n - X| < \epsilon) = 1 \text{ for every } \epsilon > 0, \text{ and}$$

$$X_n \xrightarrow{\text{prob}} c \text{ if } \lim_{n \rightarrow \infty} P(|X_n - c| < \epsilon) = 1 \text{ for every } \epsilon > 0.$$

2. Convergence in Distribution. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and let X be a random variable. The notation $X_n \xrightarrow{\text{dist}} X$ is read as X_n converges in distribution (or in law) to X . Denote the cumulative distribution functions of X_n and X by $F_{X_n}(x)$ and $F_X(x)$, respectively. Then, $X_n \xrightarrow{\text{dist}} X$ if

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$$

for all x at which $F_X(x)$ is continuous. It can be shown that

$$X_n \xrightarrow{\text{prob}} X \implies X_n \xrightarrow{\text{dist}} X.$$

3. Order of Magnitude. Let $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ be two sequences of real numbers and let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables.

- (a) Little o Notation. The notation $a_n = o(b_n)$ is read as a_n is little o of b_n and it means that the ratio $|a_n/b_n|$ converges to zero as $n \rightarrow \infty$. That is, for any $\varepsilon > 0$, there exists an integer, $n(\varepsilon)$, such that $n > n(\varepsilon) \implies |a_n/b_n| < \varepsilon$.
- (b) Big O Notation. The notation $a_n = O(b_n)$ is read as a_n is big O of b_n and it means that the ratio $|a_n/b_n|$ is bounded for large n . That is, there exists a finite number, K , and an integer, $n(K)$, such that $n > n(K) \implies |a_n/b_n| < K$.
- (c) Little o_p Notation. The notation $X_n = o_p(1)$ is read as X_n is little o_p of 1 and it means that $X_n \xrightarrow{\text{prob}} 0$ as $n \rightarrow \infty$. That is, for every $\varepsilon > 0$, $\lim_{n \rightarrow \infty} \Pr(|X_n| < \varepsilon) = 1$, or, equivalently, for every $\varepsilon > 0$ and for every $\eta > 0$, there exists an integer, $n(\varepsilon, \eta)$, such that if $n > n(\varepsilon, \eta)$ then $\Pr(|X_n| < \varepsilon) \geq 1 - \eta$.
- (d) Big O_p Notation. The notation $X_n = O_p(1)$ is read as X_n is big O_p of 1 and it means that for every $\eta > 0$, there exists a number $K(\eta)$ and an integer $n(\eta)$ such that if $n > n(\eta)$, then $\Pr[|X_n| \leq K(\eta)] \geq 1 - \eta$.
- (e) Operations on Order of Magnitude. The following relationships follow directly from the definitions:

$$\begin{aligned}
 a_n = o(b_n) &\iff a_n = b_n o(1) \iff \frac{a_n}{b_n} = o(1), \\
 a_n = O(b_n) &\iff a_n = b_n O(1) \iff \frac{a_n}{b_n} = O(1), \\
 a_n = o_p(b_n) &\iff a_n = b_n o_p(1) \iff \frac{a_n}{b_n} = o_p(1), \text{ and} \\
 a_n = O_p(b_n) &\iff a_n = b_n O_p(1) \iff \frac{a_n}{b_n} = O_p(1).
 \end{aligned}$$

A.3 Exchangeable Distributions

The random variables $Y_1, Y_2, \dots, Y_k$ are said to have an exchangeable distribution if their joint probability density function or probability mass function satisfies the following condition:

$$f_{Y_1, Y_2, \dots, Y_k}(y_1, y_2, \dots, y_k) = f_{Y_1, Y_2, \dots, Y_k}(y_1^*, y_2^*, \dots, y_k^*), \text{ for all } y_1, y_2, \dots, y_k,$$

where $y_1^*, y_2^*, \dots, y_k^*$ is any permutation of $y_1, y_2, \dots, y_k$. It is readily shown that if the distribution of $Y_1, Y_2, \dots, Y_k$ is exchangeable, then marginal distributions of any order are identical. For example,

$$\begin{aligned}
 f_{Y_1}(u) &= f_{Y_2}(u) = \dots = f_{Y_k}(u) \text{ for all } u, \text{ and} \\
 f_{Y_1, Y_2}(u, v) &= f_{Y_1, Y_3}(u, v) = f_{Y_i, Y_j}(u, v) \text{ for all } u, v \text{ and for all } i \neq j.
 \end{aligned}$$

It follows that $E(Y_1) = E(Y_i)$ and $\text{Var}(Y_1) = \text{Var}(Y_i)$ for $i = 2, 3, \dots, k$ and that $\text{Cov}(Y_1, Y_2) = \text{Cov}(Y_i, Y_j)$ for all $i \neq j$.

A.4 Sufficient Statistics

A well-defined collection of distributions is called a family of distributions and the collection is denoted by $\mathcal{F}$. For example, the collection of all p -dimensional normal distributions is a family. Typically, an investigator obtains a sample of data from some distribution $F_Y(y) \in \mathcal{F}$, where $\mathcal{F}$ is known (or assumed), but $F_Y(y)$ is unknown. In most applications, the investigator uses the sample data in an attempt to determine which member of the family generated the data. Denote the sample by $\mathbf{Y} = (Y_1 \ Y_2 \ \dots \ Y_N)'$ and let $\mathbf{T} = \mathbf{T}(\mathbf{Y})$ be a function of the sample. The statistic $\mathbf{T}$ is said to be sufficient if it contains all the information that the sample contains about which member of $\mathcal{F}$ generated the data. Technically, the statistic $\mathbf{T}$ is said to be sufficient if the distribution of the sample, conditional on $\mathbf{T}$, is identical for every $F_Y(y) \in \mathcal{F}$. If this is satisfied, then after conditioning on $\mathbf{T}$, the sample contains no additional information about which member of $\mathcal{F}$ generated the data. For example, consider the family that consists of all Bernoulli distributions. A Bernoulli distribution assigns a probability, say τ , to a success and assigns a probability $1 - \tau$ to a failure for a binary random variable. It can be shown that if N independent and identically distributed binary trials are observed, then $T =$ the number of successes is a sufficient statistic. The particular sequence of successes and failures that yields T successes contains no additional information about which member of the Bernoulli family (i.e., which value of τ) generated the data.

A minimal sufficient statistic is a sufficient statistic that is a function of every other sufficient statistic. For example, consider the family of all univariate normal distributions. It is trivially true that $\mathbf{T}_1 = \mathbf{Y}$ is sufficient. It also is true that $\mathbf{T}_2 = (\bar{Y} \ S_Y^2 \ \tilde{Y})'$ is sufficient, where $\bar{Y}$ is the sample mean, S_Y^2 is the sample variance, and $\tilde{Y}$ is the sample median. The statistic $\mathbf{T}_2$ is not minimal sufficient, however, even though $\mathbf{T}_2$ is a function of $\mathbf{T}_1$. It can be shown that a minimal sufficient statistic for the normal family is $\mathbf{T}_3 = (\bar{Y} \ S_Y^2)'$.

The family of interest in §5.2 is quite large and consists of all univariate continuous distributions. For this family, the minimal sufficient statistic, $\mathbf{T}$, is the set of order statistics. Fisher's permutation test does not use the sufficient statistic to determine which member of $\mathcal{F}$ generated the data. Instead, it uses the distribution of the sample, conditional on $\mathbf{T}$, precisely because this distribution is the same for all members of the family.

A.5 Characteristic and Cumulant Generating Functions

Let Y be a random variable. The characteristic function (CF) and cumulant generating function (CGF) of Y are defined as

$$\text{CF}_Y(t) \stackrel{\text{def}}{=} E(e^{itY}) \text{ and } \text{CGF}_Y(t) \stackrel{\text{def}}{=} \ln[\text{CF}_Y(t)], \text{ where } i^2 = -1.$$

Let r be a nonnegative integer. If $E(|Y|^r) < \infty$ then it follows from Theorem 3.5 in Severini [2005, p. 78] that the characteristic function can be expanded as

$$(12) \quad \text{CF}_Y(t) = 1 + \sum_{j=1}^r \frac{(it)^j}{j!} \text{E}(Y^j) + o(t^r) \text{ as } t \rightarrow 0, \text{ where } i^2 = -1.$$

It follows that

$$\text{E}(Y^j) = \left(\frac{1}{i^j} \right) \frac{d^j \text{CF}_Y(t)}{(dt)^j} \Big|_{t=0} \text{ for } j = 0, 1, \dots, r.$$

Also, Theorem 4.13 in Severini [2005, p. 115] verified that if $\text{E}(|Y|^r) < \infty$, where r is a positive integer, then the cumulant generating function of Y can be expanded as

$$(13) \quad \text{CGF}_Y(t) = \sum_{j=1}^r \frac{(it)^j}{j!} \kappa_j(Y) + o(|t|^r) \text{ as } t \rightarrow 0,$$

where κ_j is called the j^{th} cumulant of Y . The cumulants of Y are polynomial functions of the moments of Y . For example, the first six cumulants can be written in terms of the moments as follows:

$$\begin{aligned} \kappa_1 &= \text{E}(Y) = \mu, \quad \kappa_2 = \text{E}(Y - \mu)^2 = \sigma^2, \quad \kappa_3 = \text{E}(Y - \mu)^3, \quad \kappa_4 = \text{E}(Y - \mu)^4 - 3\sigma^4, \\ \kappa_5 &= \text{E}(Y - \mu)^5 - 10\sigma^2\kappa_3, \text{ and } \kappa_6 = \text{E}(Y - \mu)^6 - 15\kappa_4\sigma^2 - 10\kappa_3^2 - 15\sigma^6 \end{aligned}$$

The j^{th} standardized cumulant of Y is denoted by ρ_j and is defined as

$$\rho_j \stackrel{\text{def}}{=} \frac{\kappa_j}{\sigma^j} \text{ for } j = 1, \dots, r.$$

The third standardized cumulant, ρ_3 , is a measure of skewness and the fourth standardized cumulant, ρ_4 , is a measure of kurtosis.

B HEURISTIC PROOFS OF SELECTED THEOREMS

B.1 Theorem 1: Central Limit Theorem

Suppose that Y_j for $j = 1, \dots, \infty$ are independent and identically distributed random variables with mean μ and variance σ^2 . Define Z_n as $Z_n \stackrel{\text{def}}{=} \sqrt{n}(\bar{Y} - \mu)/\sigma$,

where $\bar{Y} = n^{-1} \sum_{j=1}^n Y_j$. The characteristic function of Z_n is

$$\begin{aligned}
 \text{CF}_{Z_n}(t) &= \mathbb{E} \left(e^{it \frac{\sqrt{n}(\bar{Y}_n - \mu)}{\sigma}} \right) = e^{\frac{-it\sqrt{n}\mu}{\sigma}} \mathbb{E} \left(e^{\frac{it}{\sqrt{n}\sigma} \sum_{j=1}^n Y_j} \right) \\
 &= e^{\frac{-it\sqrt{n}\mu}{\sigma}} \left[\mathbb{E} \left(e^{\frac{it}{\sqrt{n}\sigma} Y} \right) \right]^n \text{ because } \{Y_j\}_{j=1}^n \text{ are iid} \\
 \implies \text{CGF}_{Z_n}(t) &\stackrel{\text{def}}{=} \ln [\text{CF}_{Z_n}(t)] = \frac{-it\sqrt{n}\mu}{\sigma} + n \ln \left[\text{CF}_Y \left(\frac{t}{\sqrt{n}\sigma} \right) \right] \\
 &= \frac{-it\sqrt{n}\mu}{\sigma} + n \ln \left[1 + \frac{it}{\sqrt{n}\sigma} \mu + \frac{(it)^2}{2n\sigma^2} (\sigma^2 + \mu^2) + o(n^{-1}) \right] \text{ from (12)} \\
 &= \frac{-it\sqrt{n}\mu}{\sigma} + n \left[\frac{it}{\sqrt{n}\sigma} \mu + \frac{(it)^2}{2n\sigma^2} (\sigma^2 + \mu^2) - \left(\frac{1}{2} \right) \frac{(it)^2}{n\sigma^2} \mu^2 + o(n^{-1}) \right] \\
 &\quad \left(\text{using the Taylor expansion } \ln(1 + \varepsilon) = \varepsilon - \frac{\varepsilon^2}{2} + o(\varepsilon^2) \right) \\
 &= -\frac{t^2}{2} + o(1) \implies \lim_{n \rightarrow \infty} \text{CGF}_{Z_n}(t) = -\frac{t^2}{2},
 \end{aligned}$$

which is the CGF of a $N(0, 1)$ random variable. By the uniqueness of the CF and CGF, it follows that the limiting distribution of Z_n is $N(0, 1)$.

B.2 Theorem 5: Delta Method

Expand $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta})$ in a Taylor series around $\mathbf{T}_n = \boldsymbol{\theta}$:

$$\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) = \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}) + \mathbf{G}\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) + o_p(1).$$

It follows from Slutsky's Theorem (Theorem 6) that $\mathbf{G}\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{\text{dist}} N(\mathbf{0}, \mathbf{G}\boldsymbol{\Sigma}\mathbf{G}')$ and that $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{\text{dist}} N(\mathbf{0}, \mathbf{G}\boldsymbol{\Sigma}\mathbf{G}')$ because $o_p(1) \xrightarrow{\text{prob}} 0$ as $n \rightarrow \infty$.

B.3 Theorem 9: Asymptotic Normality of MLEs

Interchanging integration and differentiation reveals that

$$\begin{aligned}
 \mathbb{E}[\mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i)] &= \mathbf{0} \text{ because} \tag{14} \\
 \mathbb{E}[\mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i)] &= \int \frac{\partial \ln[f_i(\mathbf{y}_i; \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} f_i(\mathbf{y}_i; \boldsymbol{\theta}) d\mathbf{y}_i = \int \left(\frac{1}{f_i(\mathbf{y}_i; \boldsymbol{\theta})} \right) \frac{\partial f_i(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f_i(\mathbf{y}_i; \boldsymbol{\theta}) d\mathbf{y}_i \\
 &= \int \frac{\partial f_i(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} d\mathbf{y}_i = \frac{\partial}{\partial \boldsymbol{\theta}} \int f_i(\mathbf{y}_i; \boldsymbol{\theta}) d\mathbf{y}_i = \frac{\partial 1}{\partial \boldsymbol{\theta}} = \mathbf{0}.
 \end{aligned}$$

Similar operations on the second derivative of $\ln[f_i(\boldsymbol{\theta}; \mathbf{Y}_i)]$ reveal that

$$(15) \quad \mathbb{E}[\mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i) \mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i)'] = -\mathbb{E} \left[\frac{\partial^2 \ln[f_i(\boldsymbol{\theta}; \mathbf{Y}_i)]}{\partial \boldsymbol{\theta} \otimes \partial \boldsymbol{\theta}'} \right].$$

Specifically,

$$\begin{aligned}
& -\mathbf{E} \left[\frac{\partial^2 \ln [f_i(\boldsymbol{\theta}; \mathbf{Y}_i)]}{\partial \boldsymbol{\theta} \otimes \partial \boldsymbol{\theta}'} \right] = - \int \frac{\partial^2 \ln [f_i(\boldsymbol{\theta}; \mathbf{y}_i)]}{\partial \boldsymbol{\theta} \otimes \partial \boldsymbol{\theta}'} f_i(\mathbf{y}_i; \boldsymbol{\theta}) d\mathbf{y}_i \\
& = - \int \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \left[\left(\frac{1}{f_i(\mathbf{y}_i; \boldsymbol{\theta})} \right) \frac{\partial f_i(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right] \right\} f_i(\mathbf{y}_i; \boldsymbol{\theta}) d\mathbf{y}_i \\
& = \int \left[\frac{\partial f_i(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left(\frac{1}{f_i(\mathbf{y}_i; \boldsymbol{\theta})^2} \right) \frac{\partial f_i(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} - \left(\frac{1}{f_i(\mathbf{y}_i; \boldsymbol{\theta})} \right) \frac{\partial^2 f_i(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \otimes \partial \boldsymbol{\theta}'} \right] f_i(\mathbf{y}_i; \boldsymbol{\theta}) d\mathbf{y}_i \\
& = \int \left\{ \frac{\partial \ln [f_i(\mathbf{y}_i; \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \frac{\partial \ln [f_i(\mathbf{y}_i; \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}'} \right\} f_i(\mathbf{y}_i; \boldsymbol{\theta}) d\mathbf{y}_i - \int \frac{\partial^2 f_i(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \otimes \partial \boldsymbol{\theta}'} d\mathbf{y}_i \\
& = \mathbf{E} [\mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i) \mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i)'] - \frac{\partial^2}{\partial \boldsymbol{\theta} \otimes \partial \boldsymbol{\theta}'} \int f_i(\mathbf{y}_i; \boldsymbol{\theta}) d\mathbf{y}_i \\
& = \mathbf{E} [\mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i) \mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i)'] - \frac{\partial^2 1}{\partial \boldsymbol{\theta} \otimes \partial \boldsymbol{\theta}'} = \mathbf{E} [\mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i) \mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i)'].
\end{aligned}$$

Define $\bar{\mathbf{S}}(\boldsymbol{\theta}; \mathbf{Y})$ as $n^{-1}\mathbf{S}(\boldsymbol{\theta}; \mathbf{Y})$. Then

$$\mathbf{E} [\sqrt{n} \bar{\mathbf{S}}(\boldsymbol{\theta}; \mathbf{Y})] = \mathbf{0} \text{ and } \text{Var} [\sqrt{n} \bar{\mathbf{S}}(\boldsymbol{\theta}; \mathbf{Y})] = \frac{1}{n} \mathbf{E} [\mathbf{S}(\boldsymbol{\theta}; \mathbf{Y}) \mathbf{S}(\boldsymbol{\theta}; \mathbf{Y})'] = \bar{\mathbf{I}}_{\boldsymbol{\theta}, n},$$

where $\bar{\mathbf{I}}_{\boldsymbol{\theta}, n} = n^{-1} \bar{\mathbf{I}}_{\boldsymbol{\theta}, \infty}$. Furthermore, it follows from Theorem 3 that

$$(16) \quad \sqrt{n} \bar{\mathbf{S}}(\boldsymbol{\theta}; \mathbf{Y}) \xrightarrow{\text{dist}} \mathbf{N}(\mathbf{0}, \bar{\mathbf{I}}_{\boldsymbol{\theta}, \infty}), \text{ where } \bar{\mathbf{I}}_{\boldsymbol{\theta}, \infty} = \lim_{n \rightarrow \infty} \bar{\mathbf{I}}_{\boldsymbol{\theta}, n}.$$

Recall that the MLE, $\hat{\boldsymbol{\theta}}$, is the solution to $\mathbf{S}(\boldsymbol{\theta}; \mathbf{Y}) = \mathbf{0}$ for $\boldsymbol{\theta}$. Expand $\mathbf{S}(\hat{\boldsymbol{\theta}}; \mathbf{Y})$ around $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ to obtain

$$\begin{aligned}
\sqrt{n} \bar{\mathbf{S}}(\hat{\boldsymbol{\theta}}; \mathbf{Y}) = \mathbf{0} &= \sqrt{n} \bar{\mathbf{S}}(\boldsymbol{\theta}; \mathbf{Y}) - \bar{\mathbf{J}}_{\boldsymbol{\theta}, n} \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1), \\
\text{where } \bar{\mathbf{J}}_{\boldsymbol{\theta}, n} &= -\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta} \otimes \partial \boldsymbol{\theta}'}.
\end{aligned}$$

By the law of large numbers, $\bar{\mathbf{J}}_{\boldsymbol{\theta}, n} = \bar{\mathbf{I}}_{\boldsymbol{\theta}, n} + o_p(1)$. Accordingly,

$$[\bar{\mathbf{I}}_{\boldsymbol{\theta}, n} + o_p(1)] \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \sqrt{n} \bar{\mathbf{S}}(\boldsymbol{\theta}; \mathbf{Y}).$$

It follows from (16) and from Theorem 6 that

$$\begin{aligned}
\bar{\mathbf{I}}_{\boldsymbol{\theta}, n} \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \sqrt{n} \bar{\mathbf{S}}(\boldsymbol{\theta}; \mathbf{Y}) &\xrightarrow{\text{prob}} \mathbf{0}, \quad \bar{\mathbf{I}}_{\boldsymbol{\theta}, n} \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\text{dist}} \mathbf{N}(\mathbf{0}, \mathbf{I}_{\boldsymbol{\theta}, \infty}), \\
\text{and } \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &\xrightarrow{\text{dist}} \mathbf{N}(\mathbf{0}, \mathbf{I}_{\boldsymbol{\theta}, \infty}^{-1}).
\end{aligned}$$

B.4 Theorem 10: Bayesian Central Limit Theorem

The proof of Theorem 10 proceeds by expanding the log likelihood function around $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and then performing the integration in the denominator. This approach to approximating exponential integrals is called the Laplace method (see [Severini, 2000, §2.11]).

Under the assumed regularity conditions, the log likelihood function can be expanded as

$$\begin{aligned} \ell_n(\boldsymbol{\theta}; \mathbf{y}) &= \ell_n(\hat{\boldsymbol{\theta}}; \mathbf{y}) + \left[\frac{\partial \ell_n(\boldsymbol{\theta}; \mathbf{y})}{\sqrt{n} \partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right] \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \\ &\quad + \frac{1}{2} \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \bar{\mathbf{J}}_{\hat{\boldsymbol{\theta}},n} \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + O_p(n|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^3) \\ &= \ell_n(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \frac{1}{2} \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \bar{\mathbf{J}}_{\hat{\boldsymbol{\theta}},n} \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + O_p(n|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^3), \\ \text{where } \bar{\mathbf{J}}_{\hat{\boldsymbol{\theta}},n} &\stackrel{\text{def}}{=} - \frac{\partial^2 \ell_n(\boldsymbol{\theta}; \mathbf{y})}{n \partial \boldsymbol{\theta} \otimes \partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \text{ because } \frac{\partial \ell_n(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}. \end{aligned}$$

Substituting this expansion into the posterior distribution yields

$$\begin{aligned} h(\boldsymbol{\theta}; \mathbf{y}) &= \frac{h(\boldsymbol{\theta}) \exp \{ \ell_n(\boldsymbol{\theta}; \mathbf{y}) \}}{\int_{\Theta} h(\boldsymbol{\theta}) \exp \{ \ell_n(\boldsymbol{\theta}; \mathbf{y}) \} d\boldsymbol{\theta}} \\ &= \frac{h(\boldsymbol{\theta}) \exp \left\{ -\frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \bar{\mathbf{J}}_{\hat{\boldsymbol{\theta}},n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + O_p(n|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^3) \right\}}{\int_{\Theta} h(\boldsymbol{\theta}) \exp \left\{ -\frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \bar{\mathbf{J}}_{\hat{\boldsymbol{\theta}},n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + O_p(n|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|^3) \right\} d\boldsymbol{\theta}} \end{aligned}$$

Transforming the densities in the numerator and denominator from $\boldsymbol{\theta}$ to $\mathbf{u}_n = \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, expanding $h(\boldsymbol{\theta})$ around $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and using multivariate Mill's ratio

yields

$$\begin{aligned}
 h(\boldsymbol{\theta}; \mathbf{Y}) &= \frac{\left[h(\widehat{\boldsymbol{\theta}}) + O_p(n^{-1/2}) \right] \exp \left\{ -\frac{1}{2} \mathbf{u}'_n \bar{\mathbf{J}}_{\widehat{\boldsymbol{\theta}},n} \mathbf{u}_n + O_p(n^{-1/2}) \right\}}{\int_{\mathbf{R}^k} \left[h(\widehat{\boldsymbol{\theta}}) + O_p(n^{-1/2}) \right] \exp \left\{ -\frac{1}{2} \mathbf{u}' \bar{\mathbf{J}}_{\widehat{\boldsymbol{\theta}},n} \mathbf{u} + O_p(n^{-1/2}) \right\} d\mathbf{u}_n} \\
 &= \frac{\left[h(\widehat{\boldsymbol{\theta}}) + O_p(n^{-1/2}) \right] \exp \left\{ -\frac{1}{2} \mathbf{u}'_n \bar{\mathbf{J}}_{\widehat{\boldsymbol{\theta}},n} \mathbf{u}_n + O_p(n^{-1/2}) \right\}}{h(\widehat{\boldsymbol{\theta}}) (2\pi)^{k/2} \left| \bar{\mathbf{J}}_{\widehat{\boldsymbol{\theta}},n}^{-1} \right|^{1/2} + O_p(n^{-1/2})} \\
 &= \frac{\exp \left\{ -\frac{1}{2} \mathbf{u}'_n \bar{\mathbf{J}}_{\widehat{\boldsymbol{\theta}},n} \mathbf{u}_n \right\}}{(2\pi)^{k/2} \left| \bar{\mathbf{J}}_{\widehat{\boldsymbol{\theta}},n}^{-1} \right|^{1/2}} \left[1 + O_p(n^{-1/2}) \right] \\
 &\quad \text{because } \int_{\mathbf{R}^k} \exp \left\{ -\frac{1}{2} \mathbf{u}' \bar{\mathbf{J}}_{\widehat{\boldsymbol{\theta}},n} \mathbf{u} \right\} d\mathbf{u} = (2\pi)^{k/2} \left| \bar{\mathbf{J}}_{\widehat{\boldsymbol{\theta}},n}^{-1} \right|^{1/2}.
 \end{aligned}$$

Accordingly, $\bar{\mathbf{J}}_{\widehat{\boldsymbol{\theta}},n}^{\frac{1}{2}} \sqrt{n}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \xrightarrow{\text{dist}} N(\mathbf{0}, \mathbf{I}_k)$.

B.5 Theorem 11: Edgeworth Expansion

The Edgeworth expansion is based on properties of Hermite polynomials, defined in §8, and properties of characteristic and cumulant generating functions, defined in §B.1. Severini [2005, Theorem 3.8] verified that if Y is a random variable whose characteristic function, $\text{CF}_Y(t)$ satisfies

$$(17) \quad \int_{-\infty}^{\infty} |\text{CF}(t)| dt < \infty, \text{ then the pdf of } Y \text{ is } f_Y(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ity} \text{CF}_Y(t) dt.$$

Suppose that $Z \sim N(0, 1)$. Then the characteristic function of Z is $\text{CF}_Z(t) = \exp\{-t^2/2\}$. It follows from (17) that

$$\varphi(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itz} e^{-t^2/2} dt,$$

where $\varphi(z) = \varphi(z, 0, 1)$. Differentiating both sides of the above equality and using the definition of Hermite polynomials (see §8) reveals that

$$(18) \quad \frac{d^r \varphi(z)}{(dz)^r} = (-1)^r \varphi(z) H_r(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itz} (-it)^r e^{-\frac{1}{2}t^2} dt.$$

Also, if $r \geq 1$, then

$$\begin{aligned} \int_{-\infty}^z H_r(u) \varphi(u) du &= -H_{r-1}(z) \varphi(z) \text{ because} \\ \int_{-\infty}^z H_r(u) \varphi(u) du &= \int_{-\infty}^z \frac{(-1)^r}{\varphi(u)} \frac{d^r \varphi(u)}{(du)^r} \varphi(u) du = (-1)^r \int_{-\infty}^z \frac{d^r \varphi(u)}{(du)^r} du \\ &= (-1)^r \frac{d^{r-1} \varphi(u)}{(du)^{r-1}} \Big|_{-\infty}^z = (-1)^r (-1)^{r-1} H_{r-1}(u) \varphi(u) \Big|_{-\infty}^z = -H_{r-1}(z) \varphi(z). \end{aligned} \quad (19)$$

To justify Theorem 11, first use (11) and (13) to expand the cumulant generating function of W_n . The result is

$$\text{CGF}_{W_n}(t) = \frac{it}{\sqrt{n}} \omega_1 + \frac{(it)^2}{2} \left(1 + \frac{\omega_2}{n}\right) + \frac{(it)^3}{6\sqrt{n}} \omega_3 + \frac{(it)^4}{24n} \omega_4 + o(n^{-1}).$$

Second, use the inversion formula (17) and the above expansion to obtain

$$\begin{aligned} f_n(w) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itw} \text{CF}_{W_n}(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itw} \exp\{\text{CGF}_{W_n}(t)\} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itw} \exp\left\{\frac{it}{\sqrt{n}} \omega_1 + \frac{(it)^2}{2} \left(1 + \frac{\omega_2}{n}\right) + \frac{(it)^3}{6\sqrt{n}} \omega_3 + \frac{(it)^4}{24n} \omega_4 + o(n^{-1})\right\} dt. \end{aligned}$$

Using $(it)^2 = -t^2$ and expanding the exponential function yields

$$\begin{aligned} f_n(w) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itw} e^{-\frac{1}{2}t^2} \exp\left\{\frac{it}{\sqrt{n}} \omega_1 + \frac{(it)^2}{2n} \omega_2 + \frac{(it)^3}{6\sqrt{n}} \omega_3 + \frac{(it)^4}{24n} \omega_4 + o(n^{-1})\right\} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itw} e^{-\frac{1}{2}t^2} \left[1 + \frac{it}{\sqrt{n}} \omega_1 + \frac{(it)^2}{2n} \omega_2 + \frac{(it)^3}{6\sqrt{n}} \omega_3 + \frac{(it)^4}{24n} \omega_4 \right. \\ &\quad \left. + \frac{(it)^2}{2n} \omega_1^2 + \frac{(it)^4}{6n} \omega_1 \omega_3 + \frac{(it)^6}{72n} \omega_3^2 + o(n^{-1})\right] dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itw} e^{-\frac{1}{2}t^2} \left[1 + \frac{it}{\sqrt{n}} \omega_1 + \frac{(it)^2}{2n} (\omega_2 + \omega_1^2) + \frac{(it)^3}{6\sqrt{n}} \omega_3 + \frac{(it)^4}{24n} (\omega_4 + 4\omega_1 \omega_3) \right. \\ &\quad \left. + \frac{(it)^6}{72n} \omega_3^2 + o(n^{-1})\right] dt. \end{aligned}$$

Lastly, use (18) to integrate term by term. The result is as follows:

$$\begin{aligned} f_n(w) &= \varphi(w) \left[1 + \frac{H_1(z)}{\sqrt{n}} \omega_1 + \frac{H_2(z)}{2n} (\omega_2 + \omega_1^2) + \frac{H_3(z)}{6\sqrt{n}} \omega_3 + \frac{H_4(z)}{24n} (\omega_4 + 4\omega_1 \omega_3) \right. \\ &\quad \left. + \frac{H_6(z)}{72n} \omega_3^2 + o(n^{-1})\right]. \end{aligned}$$

The Edgeworth expansion for the cdf, $F_n(w)$ is obtained by using (19) to integrate the pdf expansion term by term.

C REGULARITY CONDITIONS

C.1 Asymptotic Normality of MLEs

1. The support of $\mathbf{Y}_i$, namely $\mathcal{Y}_i$, does not depend on $\boldsymbol{\theta}$.
2. The parameter $\boldsymbol{\theta}$ is identifiable. That is,

$$f(\mathbf{Y}_i; \boldsymbol{\theta}_1) = f(\mathbf{Y}_i; \boldsymbol{\theta}_2) \forall \mathbf{Y}_i \in \mathcal{Y}_i \implies \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2.$$

3. The parameter space, namely Θ , is an open subset in $\mathbf{R}^k$ where $k = \dim(\boldsymbol{\theta})$.
4. The log likelihood, $\ell(\boldsymbol{\theta}; \mathbf{Y})$, is a third order differentiable function of $\boldsymbol{\theta}$ in an open neighborhood in Θ that contains the true value. Furthermore, the order of integration and differentiation can be interchanged.
5. The average information, namely $\bar{\mathbf{I}}_{\boldsymbol{\theta},n}$, is finite and positive definite,

$$\lim_{n \rightarrow \infty} \bar{\mathbf{I}}_{\boldsymbol{\theta},n} = \bar{\mathbf{I}}_{\boldsymbol{\theta},\infty} \quad \text{and} \quad \lim_{n \rightarrow \infty} \mathbf{I}_{\boldsymbol{\theta},n}^{-1} = \mathbf{0}, \quad \text{where} \quad \bar{\mathbf{I}}_{\boldsymbol{\theta},n} \stackrel{\text{def}}{=} n^{-1} \mathbf{I}_{\boldsymbol{\theta},n} \quad \text{and} \\ \mathbf{I}_{\boldsymbol{\theta},n} \stackrel{\text{def}}{=} \mathbf{E} \left[\left(\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}} \right)' \right].$$

6. Denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$. If $\boldsymbol{\theta}$ is in a neighborhood of $\boldsymbol{\theta}_0$, then

$$\left\| \frac{\partial^3 \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \otimes \partial \boldsymbol{\theta}' \otimes \boldsymbol{\theta}'} \right\| < M_i(\mathbf{Y}_i),$$

where $M_i(\mathbf{Y}_i)$ satisfies $\mathbf{E}_{\boldsymbol{\theta}_0}[M_i(\mathbf{Y}_i)] < \infty$.

7. The score function for the i^{th} observation, namely

$$\mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i) \stackrel{\text{def}}{=} \frac{\partial}{\partial \boldsymbol{\theta}} \ln [f_i(\mathbf{Y}_i; \boldsymbol{\theta})],$$

satisfies the Lindeberg-Feller condition (see Theorem 4). Specifically, for every $\varepsilon > 0$,

$$\frac{1}{n} \sum_{i=1}^n \mathbf{E} \left[Q_{ni} I(\sqrt{Q_{ni}} \geq \sqrt{n\varepsilon}) \right] \rightarrow 0 \text{ as } n \rightarrow \infty, \text{ where} \\ Q_{ni} = \mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i)' \bar{\mathbf{I}}_{\boldsymbol{\theta},n}^{-1} \mathbf{S}_i(\boldsymbol{\theta}; \mathbf{Y}_i).$$

C.2 Asymptotically Normal Posterior Distributions

The following conditions are adapted from [Schervish, 1995, §7.4].

1. The parameter space is $\Theta \in \mathbf{R}^k$, where k is finite.

2. The point θ_0 is an interior point of Θ .
3. The prior distribution has a density with respect to Lebesgue measure that is positive and continuous at θ_0 . The prior density is denoted by $h(\theta)$.
4. The log likelihood function, $\ell_n(\theta; \mathbf{y})$, is twice continuously differentiable in a neighborhood of θ_0 .
5. The observed information matrix is defined as

$$\mathbf{J}_{\theta,n} \stackrel{\text{def}}{=} -\frac{\partial^2 \ell_n(\theta; \mathbf{y})}{\partial \theta \otimes \partial \theta'}.$$

The observed information matrix evaluated at the MLE is denoted by $\mathbf{J}_{\hat{\theta},n}$.

Denote the smallest eigenvalue of $\mathbf{J}_{\hat{\theta},n}$ by $\lambda_{n,1}$. Then $\lambda_{n,1}^{-1} \xrightarrow{\text{prob}} 0$ as $n \rightarrow \infty$.

6. Let $\lambda_{n,k}$ be the largest eigenvalue of $\mathbf{J}_{\hat{\theta},n}$ and denote an open ball of radius δ around θ_0 by $N_0(\delta)$. Denote the part of Θ that does not include $N_0(\delta)$ by Θ_0^c . If $N_0(\delta) \subseteq \Theta$, then there exists $K(\delta) > 0$ such that

$$\lim_{n \rightarrow \infty} P_{\theta_0} \left\{ \sup_{\theta \in \Theta_0^c} \lambda_{n,k}^{-1} [\ell_n(\theta; \mathbf{y}) - \ell_n(\theta_0; \mathbf{y})] < -K(\delta) \right\} = 1.$$

7. Denote the smallest eigenvalue of $\mathbf{J}_{\hat{\theta},n}^{-1} \mathbf{J}_{\theta,n}$ by $\lambda_{n,1}^*$. For each $\epsilon > 0$, there exists $\delta(\epsilon) > 0$ such that

$$\lim_{n \rightarrow \infty} P_{\theta_0} \left(\sup_{\theta \in N_0(\delta(\epsilon))} |1 - \lambda_{n,1}^*| < \epsilon \right) = 1.$$

ACKNOWLEDGEMENTS

I thank an anonymous referee for comments on earlier versions of this chapter. The comments were very helpful in improving the chapter.

BIBLIOGRAPHY

- [Adams, 1974] W. J. Adams. *The Life and Times of the Central Limit Theorem*, New York: Kaedmon Publishing Company, 1974.
- [Bhattacharya and Rao, 1976] R. N. Bhattacharya and R. R. Rao. *Normal Approximation and Asymptotic Expansions*, New York: John Wiley & Sons, 1976.
- [Berry, 1941] A. C. Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, **49**, 122–136, 1941.
- [Berry and Lindgren, 1996] D. A. Berry and B. W. Lindgren. *Statistics Theory and Methods*, second edition, Belmont CA: Duxbury Press, 1996.
- [Bishop et al., 1975] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*, Cambridge MA: MIT Press, 1975.

- [Boik, 1987] R. J. Boik. The Fisher-Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology*, **40**, 26–42, 1987.
- [Boik, 2005] R. J. Boik. Second-order accurate inference on eigenvalues of covariance and correlation matrices. *Journal of Multivariate Analysis*, **96**, 136–171, 2005.
- [Boik, 2008a] R. J. Boik. Accurate confidence intervals in regression analysis of non-normal data. *Annals of the Institute of Statistical Mathematics*, **60**, 61–83, 2008.
- [Boik, 2008b] R. J. Boik. An implicit function approach to constrained optimization with applications to asymptotics. *Journal of Multivariate Analysis*, **99**, 465–489, 2008.
- [Boik and Haaland, 2006] R. J. Boik and B. Haaland. Second-order accurate inference on simple, partial, and multiple correlations. *Journal of Modern Applied Statistical Methods*, **5**, 283–308, 2006.
- [Box and Anderson, 1955] G. E. P. Box and S. L. Anderson. Permutation theory in the derivation of robust criteria and the study of departures from assumptions. *Journal of the Royal Statistical Society, Series B*, **17**, 1–26, 1955.
- [Bradley and Gart, 1962] R. A. Bradley and J. J. Gart. The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika*, **49**, 205–214, 1962.
- [Charlier, 1905] C. V. L. Charlier. Über das Fehlergesetz. *Arkiv för Matematik, Astronomi och Fysik*, **2**, No. 8, 1905.
- [Chebyshev, 1890] P. L. Chebyshev. Sur deux théorèmes relatifs aux probabilités. *Acta Mathematica*, **14**, 305–315, 1890.
- [Chen, 1985] C. F. Chen. On asymptotic normality of limiting density functions with Bayesian implication. *Journal of the Royal Statistical Society, Series B*, **47**, 540–546, 1985.
- [Christensen, 1996] R. Christensen. *Plane Answers to Complex Questions*, New York: Springer-Verlag, 1996.
- [Cramér, 1946] H. Cramér. *Mathematical Methods of Statistics*. Princeton New Jersey: Princeton University Press, 1946.
- [Cramér, 1972] H. Cramér. Studies in the history of probability and statistics. XXVIII. On the history of certain expansions used in mathematical statistics. *Biometrika*, **59**, 205–207, 1972.
- [Daniels, 1961] H. E. Daniels. The asymptotic efficiency of a maximum likelihood estimator. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, 1961.
- [David, 1938] F. N. David. Limiting distributions connected with certain methods of sampling human populations. *Statistical Research Memoirs*, **2**, 69–90, 1938.
- [David, 1995] H. A. David. First (?) occurrence of common terms in mathematical statistics. *American Statistician*, **49**, 121–133, 1995.
- [Dawid, 1970] A. P. Dawid. On the limiting normality of posterior distributions. *Proceedings of the Cambridge Philosophical Society*, **67**, 625–633, 1970.
- [de Moivre, 1733] A. de Moivre. *Approximatio ad Summam terminorum Binomii $a + b\sqrt[n]{}$ in Seriem expansi*, privately printed and dated November 12, 1733.
- [Doob, 1934a] J. L. Doob. Probability and statistics. *Transactions of the American Mathematical Society*, **36**, 759–775, 1934.
- [Doob, 1934b] J. L. Doob. Statistical estimation. *Transactions of the American Mathematical Society*, **39**, 410–421, 1934.
- [Edgeworth, 1905] F. Y. Edgeworth. The law of error. *Transactions of the Cambridge Philosophical Society*, **20**, 36–65, 113–141, 1905.
- [Edgeworth, 1907] F. Y. Edgeworth. On the representation of statistical frequency by a series. *Journal of the Royal Statistical Society*, **70**, 102–106, 1907.
- [Erdős and Rényi, 1959] P. Erdős and A. Rényi. On the central limit theorem for samples from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, **4**, 49–57, 1959.
- [Esseen, 1945] G. Esseen. Fourier analysis of distribution functions. A mathematical study of the Gauss-Laplacian law. *Acta Mathematica*, **77**, 1–125, 1945.
- [Feller, 1935] W. Feller. Über den Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, **40**, 512–559, 1935.
- [Feller, 1971] W. Feller. *Introduction to Probability Theory and its Applications, Vol II, 2nd ed.* New York: John Wiley & Sons, 1971.
- [Fisher, 1912] R. A. Fisher. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, **41**, 155–160, 1912.

- [Fisher, 1922] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, Ser A, **222**, 309–368, 1922.
- [Fisher, 1925] R. A. Fisher. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, **22**, 700–725, 1925.
- [Fisher, 1936] R. A. Fisher. *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd, 1936.
- [Fisher, 1937] R. A. Fisher. *The Design of Experiments*, Edinburgh: Oliver & Boyd 1937.
- [Galton, 1877] F. Galton. Typical laws of heredity. *Nature*, **15**, 492–495, 512–514, 532–533, 1877.
- [Galton, 1895] F. Galton. *Natural Inheritance*, London: MacMillan, 1895.
- [Gauss, 1809] K. F. Gauss. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Arbitrariis*, Hamburg: Perthes and Besser 1809.
- [Gelman *et al.*, 1995] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*, London: Chapman & Hall, 1995.
- [Ghosal and Samanta, 1997] S. Ghosal and T. Samanta. Asymptotic expansions of posterior distributions in nonregular cases. *Annals of the Institute of Statistical Mathematics*, **49**, 181–197, 1997.
- [Hajek, 1960] J. Hajek. Limiting distribution in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, **5**, 361–374, 1960.
- [Hall, 1992a] P. Hall. On the removal of skewness by transformation. *Journal of the Royal Statistical Society, Series B*, **54**, 221–228, 1992.
- [Hall, 1992b] P. Hall. *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag, 1992.
- [Heyde and Johnstone, 1979] C. C. Heyde and I. M. Johnstone. On asymptotic posterior normality for stochastic processes. *Journal of the Royal Statistical Society, Series B*, **41**, 184–189, 1979.
- [Hoadley, 1971] B. Hoadley. Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of Mathematical Statistics*, **42**, 1977–1991.
- [Jeffreys, 1961] H. Jeffreys. *Theory of Probability*, third edition. New York: Oxford University Press, 1961.
- [Johnson, 1978] N. J. Johnson. Modified *t* tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*, **73**, 536–544, 1978.
- [Johnson, 1970] R. A. Johnson. Asymptotic expansions associated with posterior distributions. *Annals of Mathematical Statistics*, **41**, 851–864, 1970.
- [Kruskal and Stigler, 1977] W. H. Kruskal and S. M. Stigler. Normative terminology ‘normal’ in statistics and elsewhere. In B.D. Spencer (Ed.), *Statistics and Public Policy*, pp. 85–111, Oxford: Oxford University Press, 1977.
- [Laplace, 1774] P. S. Laplace. Mémoire sur la Probabilité de causes par les évènements. *Mémoires de Mathématique et de Physique, Présentés à l’Académie Royal des Sciences par divers Savans & lus dans ses Assemblées*, **6**, 621–656. (English translation by S.M. Stigler printed in *Statistical Science*, 1986, **1**, 364–378, 1774.
- [Laplace, 1810] P. S. Laplace. Mémoire sur les approximations des formules qui sont fonctions de très grands nombres et sur leur application aux probabilités. *Mémoires de l’Académie des Sciences de Paris*, 1809: 353–415, 1810.
- [Le Cam, 1970] L. Le Cam. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, **41**, 802–828, 1970.
- [Le Cam, 1986] L. Le Cam. The central limit theorem around 1935, *Statistical Science*, **1**, 78–96, 1986.
- [Le Cam and Yang, 1990] L. Le Cam and G. L. Yang. *Asymptotics in Statistics Some Basic Concepts*, New York: Springer-Verlag, 1990.
- [Lehmann, 1999] E. L. Lehmann. *Elements of Large-Sample Theory*, New York: Springer-Verlag, 1999.
- [Lexis, 1877] W. Lexis. *Theorie der Massenerscheinungen in der menschlichen Gesellschaft*, Freiburg: Wagner, 1877.
- [Lindeberg, 1922] J. W. Lindeberg. Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung, *Mathematische Zeitschrift*, **15**, 211–225, 1922.

- [Madow, 1948] W. G. Madow. On the limiting distribution of estimates based on samples from finite universes. *Annals of Mathematical Statistics*, **19**, 535–545, 1948.
- [Ogasawara, 2006a] H. Ogasawara. Asymptotic expansion and conditional robustness for the sample multiple correlation coefficient under nonnormality. *Communications in Statistics—Simulation and Computation*, **35**, 177–199, 2006.
- [Ogasawara, 2006b] H. Ogasawara. Asymptotic expansion of the sample correlation coefficient under nonnormality. *Computational Statistics and Data Analysis*, **50**, 891–910, 2006.
- [Ogasawara, 2006c] H. Ogasawara. Higher-order asymptotic standard error and asymptotic expansion in principal component analysis. *Communications in Statistics—Simulation and Computation*, **35**, 201–223, 2006.
- [Ogasawara, 2007] H. Ogasawara. Asymptotic expansion of the distributions of the estimators in factor analysis under normality. *British Journal of Mathematical and Statistical Psychology*, 2007.
- [Peirce, 1873] C. S. Peirce. On the theory of errors of observations. Of the *Report of the Superintendent of the U.S. Coast Survey for the year ending June 1870*, appendix No. 21, pp. 200–224 and plate 27. Reprinted in S.M. Stigler (Ed.), 1980, *American Contributions to Mathematical Statistics in the Nineteenth Century*, New York: Arno Press, 1973.
- [Pitman, 1937] E. J. G. Pitman. Significance tests that may be applied to samples from any populations III: The analysis of variance. *Biometrika*, **29**, 322–335, 1937.
- [Plane and Gordon, 1982] D. R. Plane and K. R. Gordon. A simple proof of the nonapplicability of the central limit theorem to finite populations. *The American Statistician*, **36**, 175–176, 1982.
- [Polya, 1920] G. Polya. Über den Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Moment problem. *Mathematische Zeitschrift*, **8**, 171–178, 1920.
- [Robinson, 1983] J. Robinson. Approximations to some test statistics for permutation tests in a completely randomized design. *Australian Journal of Statistics*, **25**, 358–369, 1983.
- [Schervish, 1995] M. Schervish. *Theory of Statistics*, New York: Springer-Verlag, 1995.
- [Sen and Singer, 1993] P. K. Sen and J. M. Singer. *Large Sample Methods in Statistics An Introduction with Applications*, London: Chapman & Hall, 1993.
- [Severini, 2005] T. Severini. *Elements of Distribution Theory*, New York: Cambridge University Press, 2005.
- [Shen, 2002] X. Shen. Asymptotic normality of semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association*, **97**, 222–235, 2002.
- [Slutsky, 1925] E. E. Slutsky. Über atochastische Asymptoten und Grenzwerte. *Metron*, **5**, 3–89, 1925.
- [Sun, 1994] D. Sun. Integrable expansions for posterior distributions for a two-parameter exponential family. *Annals of Statistics*, **22**, 1808–1830, 1994.
- [van Beek, 1972] P. van Beek. An application of Fourier methods to the problem of sharpening the Berry-Esseen inequality. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **23**, 187–196, 1972.
- [Wald, 1943] A. Wald. Tests of hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**, 426–482, 1943.
- [Wald, 1949] A. Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, **20**, 595–601, 1949.
- [Wald and Wolfowitz, 1944] A. Wald and J. Wolfowitz. Statistical tests based on permutations of the observations. *Annals of Mathematical Statistics*, **15**, 358–372, 1944.
- [Walker, 1969] A. M. Walker. On the asymptotic behavior of posterior distributions. *Journal of the Royal Statistical Society, Series B*, **31**, 80–88, 1969.
- [Weng and Woodroffe, 2000] R. C. Weng and M. Woodroffe. Integrable expansions for posterior distributions for multiparameter exponential families with applications to sequential confidence levels. *Statistica Sinica*, **10**, 693–713, 2000.
- [Woodroffe, 1992] M. Woodroffe. Integrable expansions for posterior distributions for one-parameter exponential families. *Statistica Sinica*, **2**, 91–111, 1992.

This page intentionally left blank

STEIN'S PHENOMENON

Richard Charnigo and Cidambi Srinivasan

1 INTRODUCTION

One of the first inferential principles students learn in a beginning statistics course is that they can estimate the mean of a population using the mean of a random sample drawn from that population. In the special case that the sample size is one, the population mean may be estimated by the value of the single observation drawn from the population.

Suppose, however, that we wish to estimate not just the mean of one population but rather the means of several distinct populations. Let p denote the number of population means that we wish to estimate. Let these means be denoted by $\mu_1, \mu_2, \dots, \mu_p$ and the corresponding population standard deviations by $\sigma_1, \sigma_2, \dots, \sigma_p$. Let X_1 denote the mean of the sample of size n_1 drawn from the first population (or, if the sample is of size one, the value of the single observation), X_2 the mean of the sample of size n_2 drawn from the second population, and so forth.

Our intuition tells us that we should estimate μ_1 by X_1 , μ_2 by X_2 , and so forth. In particular, our intuition suggests that X_2 through X_p should be of no use in estimating μ_1 because the populations are distinct.

And, had we been asked to estimate only μ_1 , our intuition to use only X_1 — in other words, to disregard X_2 through X_p — would have been correct. Indeed, ample statistical justifications have been offered to support the use of a sample mean to estimate a single unknown population mean (or to predict the next observation from a population).

Gauss [1821; 1823; 1826] showed that if the observations drawn from a population follow a normal (or bell-shaped) distribution, the sample mean of the observations — which is then also normally distributed — is the best estimator of the population mean in the sense of minimizing the sum of squared deviations between the individual observations and the estimator. Satisfaction of this least squares criterion implies that the sample mean is, in the case of normally distributed observations, what statisticians refer to as the maximum likelihood estimator of the population mean.

Subsequently, several eminent statisticians (for example, [Fisher, 1925; Neyman, 1937; Rao, 1945; 1967; Wald, 1949; Hodges and Lehmann, 1950; Blyth, 1951; LeCam, 1953]) rigorously demonstrated by various criteria the optimality

of the sample mean or, more generally, of the maximum likelihood estimator and advanced statistical theory or methodology accordingly.

Thus, for much of the 20th century, there was a general belief that sample means should also be optimal when one wanted to estimate several population means simultaneously. The insight of Charles Stein changed all that.

Before proceeding to Stein's insight, we provide a few details about maximum likelihood to make this article self-contained. The likelihood function evaluated at $a_1, a_2, \dots, a_p$ is roughly proportional to the probability that $X_1, X_2, \dots, X_p$ (here assumed to be continuous random variables) fall within a small neighborhood of $a_1, a_2, \dots, a_p$,

$$\mathbb{L}(a_1, a_2, \dots, a_p) := \lim_{\epsilon \rightarrow 0} \mathbb{P}(a_j - \epsilon/2 \leq X_j \leq a_j + \epsilon/2 \text{ for } 1 \leq j \leq p) / \epsilon^p,$$

where $\mathbb{L}$ represents likelihood and $\mathbb{P}$ represents probability.

If $X_1, X_2, \dots, X_p$ are normally distributed with means $\mu_1, \mu_2, \dots, \mu_p$ and a common standard deviation of 1, then the likelihood function has the form

$$(1) \quad \mathbb{L}(a_1, a_2, \dots, a_p) = \prod_{j=1}^p (2\pi)^{-1/2} \exp[-0.5(\mu_j - a_j)^2].$$

Clearly, maximizing (1) is the same as minimizing $\sum_{j=1}^p (\mu_j - a_j)^2$, which is accomplished when $\mu_j = a_j$ for $1 \leq j \leq p$. This suggests that, if $X_1, X_2, \dots, X_p$ assume the numerical values $a_1, a_2, \dots, a_p$ and we don't know $\mu_1, \mu_2, \dots, \mu_p$, then the most plausible guesses for $\mu_1, \mu_2, \dots, \mu_p$ based on the data are $a_1, a_2, \dots, a_p$. We refer to the numbers $a_1, a_2, \dots, a_p$ as maximum likelihood estimates and to the underlying random variables $X_1, X_2, \dots, X_p$ as maximum likelihood estimators. For brevity, we also refer to $X_1, X_2, \dots, X_p$ as maximum likelihood estimates when we have in mind that they are assuming numerical values, even if we do not specify those numerical values explicitly.

Besides having an evidentiary interpretation [Royall, 1997], maximum likelihood estimation enjoys several desirable properties in statistical inference, including asymptotic consistency, normality, and efficiency. Readers interested in learning more about maximum likelihood are directed to other articles in this book and to texts such as [Cox and Hinkley, 1974].

2 STEIN'S INSIGHT

In his 1956 paper, Stein established an amazing result contradicting the general belief that sample means should be optimal when estimating several population means simultaneously, even with normally distributed observations. He showed that, when $p \geq 3$, better estimators of $\mu_1, \mu_2, \dots, \mu_p$ were available than $X_1, X_2, \dots, X_p$ in the sense of minimizing the expected squared Euclidean distance between the estimators and their targets.

Let us assume that $\sigma_1/\sqrt{n_1} = \sigma_2/\sqrt{n_2} = \cdots = \sigma_p/\sqrt{n_p} = 1$ and that $X_1, X_2, \dots, X_p$ are normally distributed. The first assumption, which is mainly for convenience, says that the random variables $X_1, X_2, \dots, X_p$ have a common standard deviation of 1. Note that the standard deviations of the random variables $X_1, X_2, \dots, X_p$ are smaller than the standard deviations of the corresponding populations (except when the sample size is one). Indeed, this reduction of variation provided a reason to advocate the use of sample means.

Yet, Stein [1956] showed that the estimators

$$(2) \quad \begin{aligned} \hat{\mu}_1 &:= \left(1 - \frac{b}{a + \sum_{j=1}^p X_j^2}\right) X_1, \\ \hat{\mu}_2 &:= \left(1 - \frac{b}{a + \sum_{j=1}^p X_j^2}\right) X_2, \dots, \\ \hat{\mu}_p &:= \left(1 - \frac{b}{a + \sum_{j=1}^p X_j^2}\right) X_p \end{aligned}$$

satisfied

$$(3) \quad \mathbb{E} \left[\sum_{j=1}^p (\hat{\mu}_j - \mu_j)^2 \right] < \mathbb{E} \left[\sum_{j=1}^p (X_j - \mu_j)^2 \right]$$

for suitably chosen nonnegative constants a and b depending on $p \geq 3$. Above, the $\mathbb{E}$ symbol represents expected value, which can be conceptualized as a long-term average if the processes of sampling and constructing estimators could be repeated an unlimited number of times. The $\hat{}$ symbol denotes an estimator.

Inequality (3) is true regardless of the numerical values for $\mu_1, \mu_2, \dots, \mu_p$, which implies a sort of uniform superiority of the estimators in (2). This inequality is referred to as Stein's phenomenon or, sometimes, Stein's paradox. The paradox is that X_2 through X_p have no apparent connection to μ_1 , but nonetheless X_2 through X_p can enhance the estimation of μ_1 — as long as the estimation of μ_1 is part of the broader goal of estimating $\mu_1, \mu_2, \dots, \mu_p$ simultaneously.

This capacity to borrow strength from unrelated data is all the more remarkable because the unrelated data need not have a common subject matter interpretation. An entertaining quotation from Robbins [1951] drives home the message: “ X_1 could be an observation on a butterfly in Ecuador, X_2 on an oyster in Maryland, X_3 the temperature of a star, and so on.” Indeed, Robbins anticipated the possibility of improving upon maximum likelihood estimators. He did so in a study on compound decision problems that focused on inferences about binary-valued $\mu_1, \mu_2, \dots, \mu_p$.

Specifically, Robbins assumed that each μ_j , $1 \leq j \leq p$, was known to have one of the two values $+1$ or -1 and that X_j , $1 \leq j \leq p$, was normally distributed with mean μ_j and standard deviation 1. He considered the statistical problem of

deciding whether the value of μ_j was $+1$ or -1 based on the data $X_1, X_2, \dots, X_p$, with the proportion of incorrect decisions as the loss (a measure of “badness” of the decisions). Denoting the signum of x by $\text{sgn}(x)$, the maximum likelihood estimators of $\mu_1, \mu_2, \dots, \mu_p$ are $\text{sgn}(X_1), \text{sgn}(X_2), \dots, \text{sgn}(X_p)$. They provide natural decision rules asserting that the value of μ_j is $\text{sgn}(X_j)$ for $1 \leq j \leq p$. The expected loss, or risk, for each of these decision rules is 0.1587. Note that the maximum likelihood decision rule for μ_j depends solely on X_j and not on $X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_p$.

Using an elegant argument, Robbins developed competing decision rules asserting that the value of μ_j is $\text{sgn}(X_j - X^*)$ if $|\sum_{j=1}^p X_j/p| < 1$, that it is $+1$ if $\sum_{j=1}^p X_j/p \geq 1$, and that it is -1 if $\sum_{j=1}^p X_j/p \leq -1$, where $X^* := 0.5 \log[(1 + \sum_{j=1}^p X_j/p)/(1 - \sum_{j=1}^p X_j/p)]$. The decision rule for μ_j now depends not only on X_j but also on $X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_p$. These decision rules outperform the maximum likelihood decision rules as $p \rightarrow \infty$, in that the expected loss tends to a number strictly less than 0.1587, unless the proportion of μ_j equaling $+1$ tends to $1/2$ as $p \rightarrow \infty$, in which case the expected loss tends to 0.1587 exactly.

On the other hand, Stein was the first author to establish that the maximum likelihood estimators $X_1, X_2, \dots, X_p$ could be surpassed in problems involving real-valued $\mu_1, \mu_2, \dots, \mu_p$ and a finite p . Stein’s subsequent collaboration with James [1961] yielded the following concrete prescriptions for estimators of real-valued $\mu_1, \mu_2, \dots, \mu_p$:

$$(4) \quad \hat{\mu}_1 := \left(1 - \frac{(p-2)}{\sum_{j=1}^p X_j^2}\right) X_1,$$

$$\hat{\mu}_2 := \left(1 - \frac{(p-2)}{\sum_{j=1}^p X_j^2}\right) X_2, \dots,$$

$$\hat{\mu}_p := \left(1 - \frac{(p-2)}{\sum_{j=1}^p X_j^2}\right) X_p$$

and, as discussed by Efron and Morris [1973],

$$(5) \quad \hat{\mu}_1 := \left(\frac{(p-3)}{\sum_{j=1}^p (X_j - \bar{X})^2}\right) \bar{X} + \left(1 - \frac{(p-3)}{\sum_{j=1}^p (X_j - \bar{X})^2}\right) X_1,$$

$$\hat{\mu}_2 := \left(\frac{(p-3)}{\sum_{j=1}^p (X_j - \bar{X})^2}\right) \bar{X} + \left(1 - \frac{(p-3)}{\sum_{j=1}^p (X_j - \bar{X})^2}\right) X_2, \dots,$$

$$\hat{\mu}_p := \left(\frac{(p-3)}{\sum_{j=1}^p (X_j - \bar{X})^2}\right) \bar{X} + \left(1 - \frac{(p-3)}{\sum_{j=1}^p (X_j - \bar{X})^2}\right) X_p,$$

where the so-called grand mean $\bar{X}$ is the arithmetic average of all of the sample means, $\sum_{j=1}^p X_j/p$. The grand mean can be viewed as an estimator of the arithmetic average of all of the population means, $\sum_{j=1}^p \mu_j/p$.

We note that prescriptions (4) and (5) can be generalized to accommodate the situation in which $\sigma_1/\sqrt{n_1}, \sigma_2/\sqrt{n_2}, \dots, \sigma_p/\sqrt{n_p}$ have a common value $\tau \neq 1$. The generalization is that each instance of $(p-2)$ or $(p-3)$ gets multiplied by τ^2 , which represents a common variance for the random variables $X_1, X_2, \dots, X_p$.

Comparing prescriptions (4) and (5), we see that the former entails shrinking the maximum likelihood estimators $X_1, X_2, \dots, X_p$ toward 0, while the latter entails drawing the maximum likelihood estimators $X_1, X_2, \dots, X_p$ toward the grand mean $\bar{X}$. In effect, prescription (5) represents a compromise between a maximum likelihood estimator and the grand mean.

3 A DATA ANALYSIS EXAMPLE

To illustrate Stein's phenomenon, we consider data on bladder cancer mortality among males in 13 southeastern states of the U.S. (AR, TN, NC, LA, MS, AL, GA, SC, MO, KY, WV, VA, FL). The data were acquired from the website <http://statecancerprofiles.cancer.gov/historicaltrend/joinpoint.noimage.html>, which is maintained by the National Cancer Institute and the Centers for Disease Control and Prevention.

We have $p = 13$, $\mu_1 =$ expected annual deaths from bladder cancer per 100,000 males in Arkansas (computed as the average observed number of deaths from bladder cancer per 100,000 males in Arkansas over the years 1991 to 2000), $X_1 =$ observed number of deaths from bladder cancer per 100,000 males in Arkansas during the single year 1990, $\mu_2 =$ expected annual deaths from bladder cancer per 100,000 males in Tennessee (computed as the average observed number of deaths from bladder cancer per 100,000 males in Tennessee over the years 1991 to 2000), $X_2 =$ observed number of deaths from bladder cancer per 100,000 males in Tennessee during the single year 1990, and so forth. These quantities are listed in Table 1.

We view $\mu_1, \mu_2, \dots, \mu_{13}$ as targets to be estimated. The intuitive, maximum likelihood estimate of μ_1 is X_1 , while the James-Stein estimate based on prescription (5) is $0.199X_1 + 0.801\bar{X}$. The maximum likelihood and James-Stein estimates of μ_2 through μ_{13} are defined analogously. The technical details of how we acquired the 0.801 for the James-Stein estimates are presented at the end of this section.

So, which are better, the maximum likelihood estimates or the James-Stein estimates? Figure 1 compares both the maximum likelihood estimates and the James-Stein estimates to the targets.

Compared to the maximum likelihood estimates, the James-Stein estimates are less than half the distance to the targets for Tennessee, South Carolina, Missouri, and Virginia. More specifically, the James-Stein estimates reduce the absolute

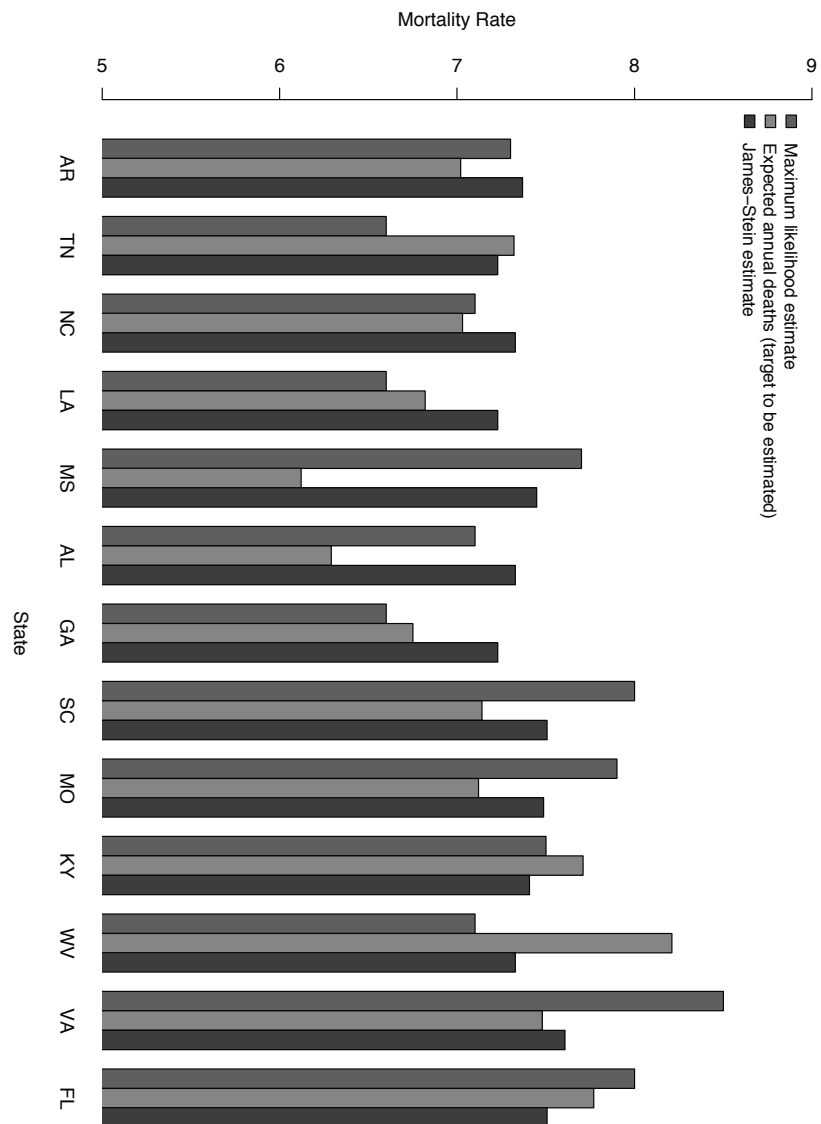


Figure 1. Data analysis exemplifying Stein's phenomenon

Table 1. Bladder cancer data

State	j	X_j	μ_j
Arkansas	1	7.3	7.02
Tennessee	2	6.6	7.32
North Carolina	3	7.1	7.03
Louisiana	4	6.6	6.82
Mississippi	5	7.7	6.12
Alabama	6	7.1	6.29
Georgia	7	6.6	6.75
South Carolina	8	8.0	7.14
Missouri	9	7.9	7.12
Kentucky	10	7.5	7.71
West Virginia	11	7.1	8.21
Virginia	12	8.5	7.48
Florida	13	8.0	7.77

estimation error by 87.3% for Tennessee, 57.3% for South Carolina, 52.9% for Missouri, and 87.6% for Virginia.

On the other hand, the maximum likelihood estimates are much closer to the targets than are the James-Stein estimates for North Carolina and Georgia.

For the other seven states individually, neither the James-Stein estimates nor the maximum likelihood estimates are dramatically better.

However, when we consider all 13 states simultaneously, the James-Stein estimates emerge as superior. The sum of squared estimation errors — called the loss by statisticians — is

$$\sum_{j=1}^{13} (\text{target } j - \text{estimate } j)^2 = \sum_{j=1}^{13} (\mu_j - \{0.199X_j + 0.801\bar{X}\})^2 = 4.677$$

for the James-Stein estimates, whereas the loss is

$$\sum_{j=1}^{13} (\text{target } j - \text{estimate } j)^2 = \sum_{j=1}^{13} (\mu_j - X_j)^2 = 7.543$$

for the maximum likelihood estimates. Hence, the James-Stein estimates reduce the loss by 38%.

Besides perceiving whether the maximum likelihood estimates or the James-Stein estimates are closer to the targets, another key to viewing Figure 1 is noticing that the James-Stein estimates for different states are more similar to each other than are the maximum likelihood estimates for different states. This is because the James-Stein estimate for a given state is a compromise between the maximum

likelihood estimate for that state and the average of the maximum likelihood estimates for all of the states:

$$\begin{aligned} & \text{James-Stein estimate for state } j \\ = & \quad 0.199 \text{ maximum likelihood estimate for state } j \\ & + 0.801 \sum_{k=1}^{13} \text{maximum likelihood estimate for state } k / 13. \end{aligned}$$

Thus, the James-Stein estimate takes the corresponding maximum likelihood estimate and moves it 80.1% of the way to the average of the maximum likelihood estimates for all of the states.

At this juncture, perhaps some readers may think that it is natural to use data from neighboring states to estimate the expected number of deaths from bladder cancer in a given state. Such readers may, without yet being aware of it, already have in mind an empirical Bayes perspective. We will have more to say about that in Section 5. In any event, we conclude this data analysis with three remarks.

First, there is some heterogeneity in the expected numbers of deaths, ranging from 6.12 per 100,000 in Mississippi to 8.21 per 100,000 in West Virginia. Thus, whatever may explain the superior performance of the James-Stein estimates, it is emphatically not that $\mu_1 = \mu_2 = \cdots = \mu_{13}$. Indeed, if it were true that $\mu_1 = \mu_2 = \cdots = \mu_{13}$, then all of the targets would be the same and consideration of the grand mean $\bar{X}$ would seem obvious. Hence, a more subtle resolution to the paradox is required.

Second, the expected values in inequality (3) should not be overlooked. In particular, inequality (3) is not a promise that the loss will be reduced every time we use the James-Stein estimators. Rather, the long-term average of the loss – if somehow if it were possible to repeat the processes of sampling and constructing estimators an unlimited number of times — is promised to be less with the James-Stein estimators than with the maximum likelihood estimators.

Third, even when the loss is reduced dramatically, some of the individual James-Stein estimates may not be as good as the corresponding individual maximum likelihood estimates. We saw this in our data analysis: the individual James-Stein estimates for North Carolina and Georgia were not as good as the individual maximum likelihood estimates for those states.

We now close this section with a few technical details for readers interested in the probability modeling underlying our data analysis. Other readers may proceed to Section 4 without loss of continuity.

For our data analysis, we calculated the 0.801 used in the James-Stein estimates as

$$\frac{(p-3)\tau^2}{\sum_{j=1}^p (X_j - \bar{X})^2},$$

in accord with prescription (5) and the generalization described thereafter. We

approximated τ^2 , the presumed-common variance of $X_1, X_2, \dots, X_{13}$, as

$$\sum_{j=1}^{13} X_j(100,000/m_j)/13,$$

where m_j is the number of males living in state j during the year 1990. Websites with addresses $\{\text{http://www.census.gov/population/cencounts/zz190090.txt}\}$, maintained by the U.S. Census Bureau, provided $m_1, m_2, \dots, m_{13}$. Note that “zz” is a placeholder for the usual state abbreviation in lower case letters.

The above approximation for τ^2 is based on a Poisson probability model for the observed number of bladder cancer deaths in each state. If $Y_j := X_j(m_j/100,000)$ — that is, the observed number of bladder cancer deaths in state j — has a Poisson distribution with expected value $\mu_j(m_j/100,000)$, then the variance of Y_j is also $\mu_j(m_j/100,000)$. Hence, the variance of $X_j = Y_j(100,000/m_j)$ must be $\mu_j(m_j/100,000)(100,000/m_j)^2 = \mu_j(100,000/m_j)$. We can roughly estimate this as $X_j(100,000/m_j)$, and averaging the latter quantity over all 13 states yields the above approximation for τ^2 .

4 A GEOMETRIC HEURISTIC

There is an elegant and simple heuristic reasoning underlying Stein's phenomenon, which we now describe. As before, suppose that $X_1, X_2, \dots, X_p$ are sample means from distinct populations with unknown means $\mu_1, \mu_2, \dots, \mu_p$. Moreover, assume that $\sigma_1/\sqrt{n_1} = \sigma_2/\sqrt{n_2} = \dots = \sigma_p/\sqrt{n_p} = 1$ and that $X_1, X_2, \dots, X_p$ are normally distributed.

Using elementary algebra, one obtains the identity

$$(6) \quad \sum_{j=1}^p X_j^2 = \sum_{j=1}^p (X_j - \mu_j)^2 + \sum_{j=1}^p \mu_j^2 + 2\sqrt{\sum_{j=1}^p \mu_j^2} Z,$$

where $Z := \sum_{j=1}^p \mu_j(X_j - \mu_j)/\sqrt{\sum_{j=1}^p \mu_j^2}$. Since $X_1, X_2, \dots, X_p$ are normally distributed, so is Z , with mean 0 and variance 1. Hence, $2\sqrt{\sum_{j=1}^p \mu_j^2} Z$ is negligible in relation to the larger of p and $\sum_{j=1}^p \mu_j^2$ as p becomes large. Moreover, $\sum_{j=1}^p (X_j - \mu_j)^2$ equals p plus a remainder that is negligible in relation to p as p becomes large. Thus, identity (6) simplifies to

$$(7) \quad \sum_{j=1}^p X_j^2 \approx p + \sum_{j=1}^p \mu_j^2.$$

Approximation (7) holds for all numerical values of $\mu_1, \mu_2, \dots, \mu_p$.

We see from approximation (7) that, for large p , a good estimator of $\sum_{j=1}^p \mu_j^2$ is $\sum_{j=1}^p X_j^2 - p$, assuming that the latter turns out to be nonnegative. A negative

value of $\sum_{j=1}^p X_j^2 - p$ is possible, especially when $\sum_{j=1}^p \mu_j^2$ is small. If such a contingency arises, we may sensibly replace $\sum_{j=1}^p X_j^2 - p$ by 0. However, for simplicity in what follows, we do not consider such a contingency.

The maximum likelihood estimator of $\sum_{j=1}^p \mu_j^2$ is $\sum_{j=1}^p X_j^2$, since the maximum likelihood estimator of μ_j , $1 \leq j \leq p$, is X_j . Thus, the maximum likelihood estimator of $\sum_{j=1}^p \mu_j^2$ tends to err on the high side, suggesting that a better estimator of μ_j , $1 \leq j \leq p$, is available. In particular, we may consider multiplicatively adjusting X_j , $1 \leq j \leq p$, by

$$(8) \quad \sqrt{\frac{\sum_{k=1}^p X_k^2 - p}{\sum_{k=1}^p X_k^2}} = \sqrt{1 - \frac{p}{\sum_{k=1}^p X_k^2}} \approx 1 - \frac{p/2}{\sum_{k=1}^p X_k^2},$$

the last approximation in (8) following from the calculus result $\sqrt{1+x} \approx 1 + x/2$ for small $|x|$. Such multiplication shrinks the vector $\mathbf{X} := (X_1, X_2, \dots, X_p)'$ toward the origin in p -dimensional space while leaving its direction $\mathbf{X}/\sqrt{\sum_{j=1}^p X_j^2}$ undisturbed.

Figure 2 illustrates the geometry for $p = 2$. The joint probability density function for X_1 and X_2 is greater everywhere inside the dark-and-light circle than anywhere outside it. Roughly speaking, this means that the most likely values of X_1 and X_2 are those within the dark-and-light circle. However, there is an asymmetry. The region in which $X_1^2 + X_2^2 < \mu_1^2 + \mu_2^2$, colored light, is smaller than the region in which $X_1^2 + X_2^2 > \mu_1^2 + \mu_2^2$ colored dark. This asymmetry arises because the circle centered at the origin with radius $\sqrt{\mu_1^2 + \mu_2^2}$ is a convex set in 2-dimensional space. Hence, $X_1^2 + X_2^2 > \mu_1^2 + \mu_2^2$ is a more probable outcome than $X_1^2 + X_2^2 < \mu_1^2 + \mu_2^2$. Shrinkage of $(X_1, X_2)'$ toward the origin works in our favor when $(X_1, X_2)'$ falls in the dark region and works against us when $(X_1, X_2)'$ falls in the light region, but the former is more likely than the latter. A similar geometry applies for $p > 2$ because the hypersphere centered at the origin with radius $\sqrt{\sum_{j=1}^p \mu_j^2}$ is a convex set in p -dimensional space.

Figure 3 provides a numerical example in which $\mu_1 = \mu_2 = 5$ and various values of $(X_1, X_2)'$ are identified as belonging either to the dark region in which $X_1^2 + X_2^2 > 50$ or to the light region in which $X_1^2 + X_2^2 < 50$. Of particular interest is that there exist points symmetric about $(5, 5)'$ that land inside the dark region. For instance, $(2.5, 7)'$ and $(7.5, 3)'$ both fall inside the dark region. On the other hand, there do not exist points symmetric about $(5, 5)'$ that fall inside the light region. This indicates that $(X_1, X_2)'$ is more likely to land inside the dark region, which is to say that $X_1^2 + X_2^2 > 50$ is more probable than $X_1^2 + X_2^2 < 50$.

Based on the above geometric heuristic, Stein [1956] provided a rigorous mathematical justification for the existence of better estimators of $\mu_1, \mu_2, \dots, \mu_p$ than the maximum likelihood estimators $X_1, X_2, \dots, X_p$. The technical meaning of “better” entails the decision theoretic concepts of loss and risk.

For any estimators $D_1, D_2, \dots, D_p$ of $\mu_1, \mu_2, \dots, \mu_p$, we define the loss to be the sum of squared estimation errors,

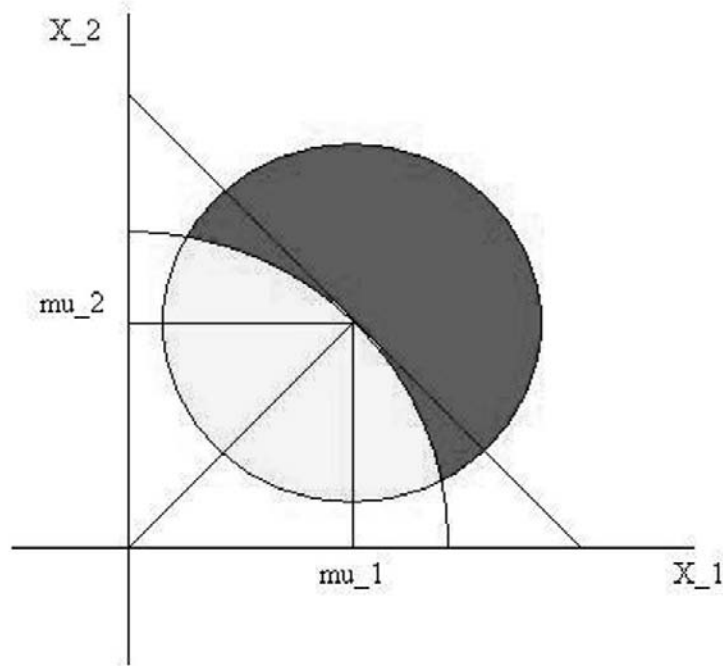


Figure 2. Geometric Heuristic for Shrinkage

$$(9) \quad \sum_{j=1}^p (\mu_j - D_j)^2.$$

(Other definitions of the loss are possible, as will be noted toward the end of this section.) The risk is defined as the expected value of the loss,

$$(10) \quad \mathbb{E} \left[\sum_{j=1}^p (\mu_j - D_j)^2 \right].$$

Assembling $\mu_1, \mu_2, \dots, \mu_p$ into a vector $\boldsymbol{\mu}$ and $D_1, D_2, \dots, D_p$ into a vector $\mathbf{D}$, we may use the shorthand $L(\boldsymbol{\mu}, \mathbf{D})$ for the loss (9) and the shorthand $R(\boldsymbol{\mu}, \mathbf{D})$ for the risk (10).

Stein [1956] showed that, for any $p \geq 3$, there exists a $\mathbf{D}$ such that

$$(11) \quad R(\boldsymbol{\mu}, \mathbf{D}) < R(\boldsymbol{\mu}, \mathbf{X}),$$

where, as before, $\mathbf{X}$ is a vector containing the maximum likelihood estimators $X_1, X_2, \dots, X_p$. Importantly, relation (11) holds regardless of what $\boldsymbol{\mu}$ may be.

James and Stein [1961] subsequently established relation (11) for the specific choice

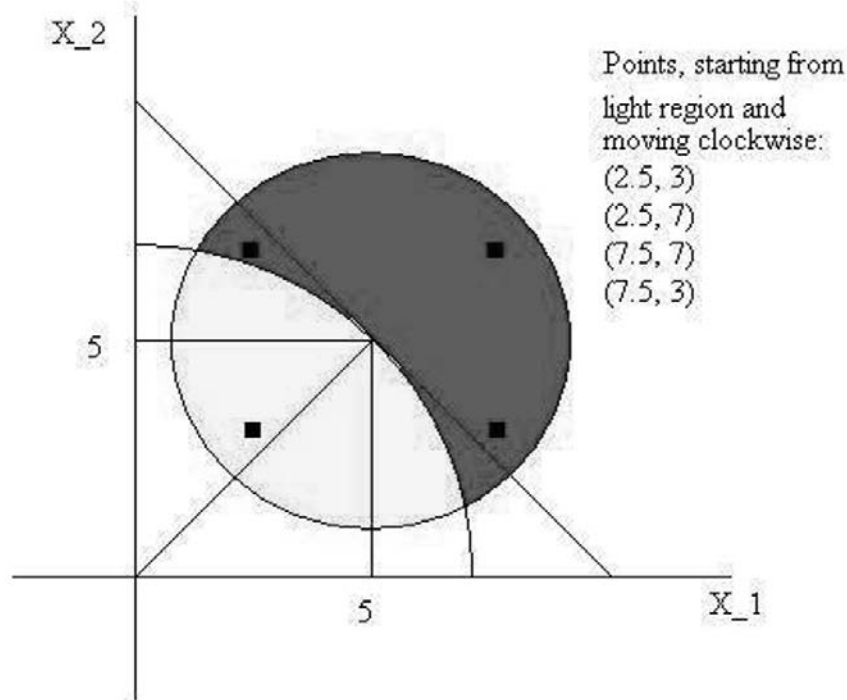


Figure 3. Geometric Heuristic for Shrinkage (including representative points)

$$(12) \quad D_j := \left(1 - \frac{p-2}{\sum_{k=1}^p X_k^2}\right) X_j, \quad 1 \leq j \leq p,$$

which entails the kind of shrinkage displayed in (8) except that $p/2$ has been replaced by $p-2$.

Figure 4 illustrates the reduction in risk associated with the estimators (12) for a fixed $p \geq 3$. While $R(\boldsymbol{\mu}, \mathbf{X}) = p$ regardless of $\boldsymbol{\mu}$, $R(\boldsymbol{\mu}, \mathbf{D})$ with $\mathbf{D}$ as in (12) has the form

$$p - (p-2) b(\boldsymbol{\mu}),$$

where $b(\boldsymbol{\mu})$ is bounded above by 1, equal to 1 when $\boldsymbol{\mu}$ is the zero vector, and vanishingly small as $\sqrt{\sum_{j=1}^p \mu_j^2}$ becomes large. Thus, the estimators (12) are dramatically better than the maximum likelihood estimators when $\sqrt{\sum_{j=1}^p \mu_j^2}$ is small. Intuitively, shrinking the vector of maximum likelihood estimators toward the origin must work in our favor if $\boldsymbol{\mu}$ itself is at the origin!

The estimators (12) can be modified to pull the vector of maximum likelihood estimators toward any point of the form $(C, C, \dots, C)'$, where C is a real constant or a real-valued random variable. A case of particular interest, already considered in the data analysis of Section 3, is when C equals the grand mean $\bar{X}$. In this case, (12) is changed to

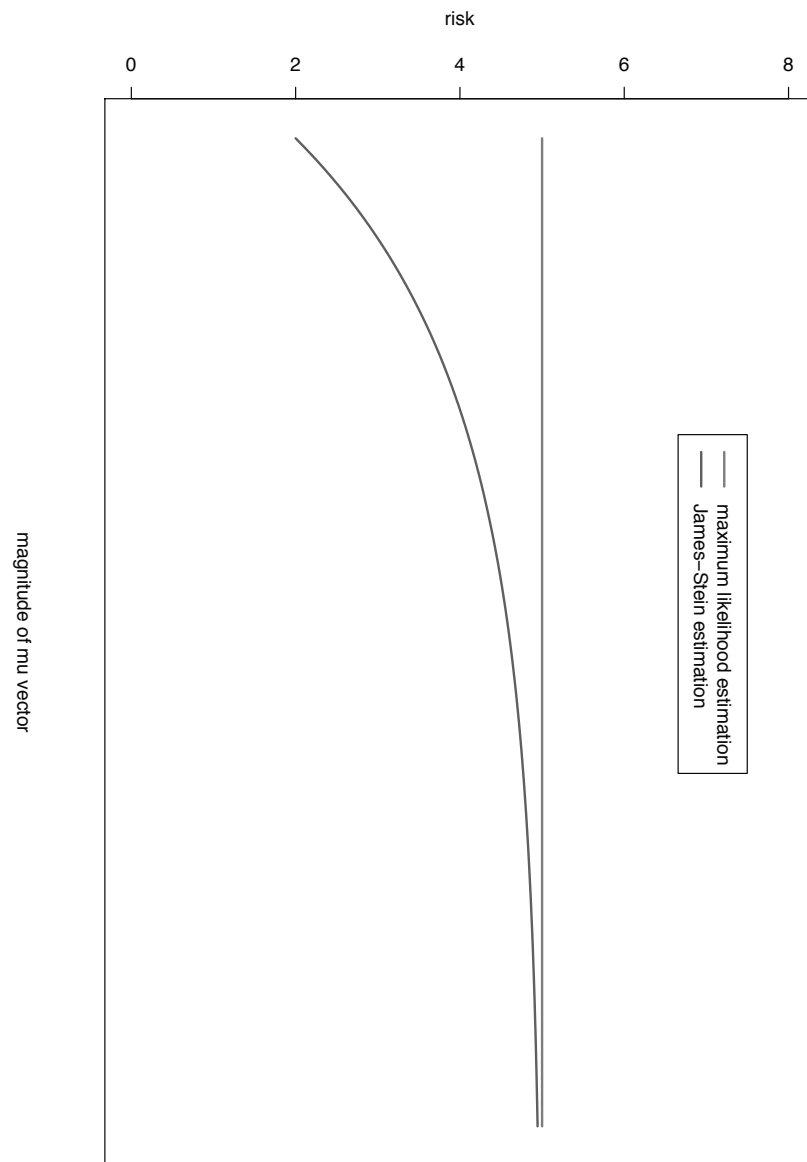


Figure 4. Comparison of risks with and without shrinkage, $p = 5$

$$(13) \quad D_j := \bar{X} + \left(1 - \frac{p-3}{\sum_{k=1}^p (X_k - \bar{X})^2}\right) (X_j - \bar{X}), \quad 1 \leq j \leq p.$$

Also, as was suggested in Section 2, a further modification is possible if $\sigma_1/\sqrt{n_1}, \sigma_2/\sqrt{n_2}, \dots, \sigma_p/\sqrt{n_p}$ share a common value other than 1.

We close this section by noting that Stein's phenomenon holds under far more general conditions than indicated above. Hudson [1974], Efron and Morris [1976], and Berger [1976; 1985] are among those who have exhibited Stein-type estimators superior to $X_1, X_2, \dots, X_p$ under more general conditions. One generalization allows $X_1, X_2, \dots, X_p$ to have unequal variances and/or to be correlated. Another generalization entails replacing the loss (9) and risk (10) by

$$(14) \quad L_{\mathbf{Q}}(\boldsymbol{\mu}, \mathbf{D}) := (\boldsymbol{\mu} - \mathbf{D})' \mathbf{Q} (\boldsymbol{\mu} - \mathbf{D}) = \sum_{j=1}^p \sum_{k=1}^p q_{jk} (\mu_j - D_j)(\mu_k - D_k)$$

and

$$R_{\mathbf{Q}}(\boldsymbol{\mu}, \mathbf{D}) := \mathbb{E}[L_{\mathbf{Q}}(\boldsymbol{\mu}, \mathbf{D})],$$

respectively, where $\mathbf{Q}$ is a given $p \times p$ positive definite matrix whose (j, k) element is q_{jk} , $1 \leq j, k \leq p$.

As an example, suppose that $X_1, X_2, \dots, X_p$ have known positive variances $\tau_1^2, \tau_2^2, \dots, \tau_p^2$ (but are uncorrelated) and that $\mathbf{Q}$ is a diagonal matrix (i.e., $q_{jk} = 0$ for $j \neq k$). If $p \geq 3$, then better estimators of $\mu_1, \mu_2, \dots, \mu_p$ than $X_1, X_2, \dots, X_p$ are given by

$$D_j := \left(1 - \frac{p-2}{\tau_j^2 q_{jj} \sum_{k=1}^p X_k^2 / (\tau_k^4 q_{kk})}\right) X_j, \quad 1 \leq j \leq p.$$

They simplify to the James-Stein estimators in (12) when $\tau_1^2 = \tau_2^2 = \dots = \tau_p^2 = 1$ and $q_{11} = q_{22} = \dots = q_{pp} = 1$.

5 AN EMPIRICAL BAYES PERSPECTIVE

The fundamental ideas underlying Stein's phenomenon are by no means restricted to a classical frequentist perspective, in which $\mu_1, \mu_2, \dots, \mu_p$ are treated as unknown constants. Indeed, the logic behind Stein's phenomenon transcends different schools of statistical thought.

Efron and Morris [1971; 1972; 1973; 1975; 1976] and Berger [1982; 1985] are among those who have examined Stein's phenomenon from a Bayesian or empirical Bayesian perspective. Such a perspective entails treating $\mu_1, \mu_2, \dots, \mu_p$ as random quantities with their own distributions. A further distinction must be made between prior distributions, which pertain before $X_1, X_2, \dots, X_p$ have been observed, and posterior distributions, which pertain after $X_1, X_2, \dots, X_p$ have been observed. An elegant and intuitive justification of Stein's phenomenon emerges, as described next.

Suppose that each μ_j , $1 \leq j \leq p$, has the same normal prior distribution with mean θ and variance δ^2 . The values of θ and δ^2 reflect an individual's subjective view — before $X_1, X_2, \dots, X_p$ have been observed — about what numerical values are likely for $\mu_1, \mu_2, \dots, \mu_p$. Suppose also that the distribution of X_j given μ_j , $1 \leq j \leq p$, is normal with mean μ_j and variance τ^2 . Then, once $X_1, X_2, \dots, X_p$ have been observed, a famous result called Bayes' Theorem may be applied to derive the posterior distributions of $\mu_1, \mu_2, \dots, \mu_p$.

In fact, the posterior distribution of μ_j , $1 \leq j \leq p$, is [Gelman *et al.*, 1995] normal with mean

$$(15) \quad \theta + \left(1 - \frac{1}{1 + \delta^2/\tau^2}\right) (X_j - \theta)$$

and variance

$$\frac{1}{1/\tau^2 + 1/\delta^2}.$$

If we have to make a single number guess for μ_j after observing $X_1, X_2, \dots, X_p$, we may be inclined to use (15). The resemblance to (13) is unmistakable, but the connection to Stein's phenomenon can be made even more transparent.

Suppose that, instead of specifying θ and δ^2 a priori, we choose them after observing $X_1, X_2, \dots, X_p$. In essence, we are now treating θ and δ^2 as unknown constants that must be estimated. One may argue that a good choice for θ is $\bar{X}$ and that a good choice for δ^2 is $\sum_{k=1}^p (X_k - \bar{X})^2 / (p - 3) - \tau^2$, assuming that the latter is positive. Substituting these choices for θ and δ^2 into (15) yields

$$(16) \quad \bar{X} + \left(1 - \frac{(p-3)\tau^2}{\sum_{k=1}^p (X_k - \bar{X})^2}\right) (X_j - \bar{X}),$$

which simplifies to (13) when $\tau^2 = 1$. To distinguish (15) from (16), we refer to the former as a Bayesian estimator and to the latter as an empirical Bayesian estimator. The empirical Bayesian estimator reflects the idea, also exemplified by Robbins' [1951] quotation, of borrowing strength from unrelated data. Here the data $X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_p$ are apparently unrelated to X_j , yet they provide strength in the form of a reduced risk in estimating μ_j . —idexcancer

We conclude this section with two remarks. First, by adopting an empirical Bayesian perspective, Efron and Morris [1973] were able to give a short and easily accessible proof for the superiority of the James-Stein estimators (12) over maximum likelihood estimators. Second, some further interpretation is now possible for the data analysis example of Section 3, as explained below.

If $\mu_1, \mu_2, \dots, \mu_{13}$ share a common prior mean θ that we treat as an unknown constant, then every observation $X_1, X_2, \dots, X_{13}$ contains information about the value of θ . Hence, although X_1 contains the most information about μ_1 , $X_2, \dots, X_{13}$ also contain some information about μ_1 . Analogous statements apply to $\mu_2, \dots, \mu_{13}$. In essence, θ unites what otherwise may appear to be 13 distinct estimation problems.

From a subject matter viewpoint, we may conceptualize θ as the expected annual deaths from bladder cancer per 100,000 males over the whole 13-state region.

This makes highly plausible the idea that using every observation to estimate μ_1 should be better than using only X_1 , using every observation to estimate μ_2 should be better than using only X_2 , and so forth. However, while conceptualizing θ from a subject matter viewpoint makes Stein's phenomenon highly plausible in this example, in general the presence of Stein's phenomenon does not depend on the existence of — let alone our recognition of — such a conceptualization.

6 A REGRESSION PERSPECTIVE

The classical idea of regression as described by Galton [1890] provides another intuitive explanation for Stein's phenomenon. The excellent article by Stigler [1990] discusses Stein's phenomenon from this perspective, rendering transparent the justification for estimators like (13). We follow Stigler [1990] closely in the presentation below.

As in Section 2, we have pairs $(X_1, \mu_1), (X_2, \mu_2), \dots, (X_p, \mu_p)$ such that $X_1, X_2, \dots, X_p$ are observable but $\mu_1, \mu_2, \dots, \mu_p$ are not. We assume that $X_1, X_2, \dots, X_p$ are normally distributed with respective means $\mu_1, \mu_2, \dots, \mu_p$ and common variance 1. As such, we may view $X_1, X_2, \dots, X_p$ as generated from the model

$$(17) \quad X_j = \mu_j + \epsilon_j$$

for $1 \leq j \leq p$, where $\epsilon_1, \epsilon_2, \dots, \epsilon_p$ are normally distributed “error” terms with common mean 0 and common variance 1.

Even though $\mu_1, \mu_2, \dots, \mu_p$ are unknown, we can conceptualize a scatterplot of $\mu_1, \mu_2, \dots, \mu_p$ against $X_1, X_2, \dots, X_p$. This is illustrated for simulated data in Figure 5. Now imagine fitting two regressions to the points in the scatterplot: a regression of $X_1, X_2, \dots, X_p$ on $\mu_1, \mu_2, \dots, \mu_p$ symbolized as $R_1(X|\mu) = \alpha_1 + \beta_1\mu$, and a regression of $\mu_1, \mu_2, \dots, \mu_p$ on $X_1, X_2, \dots, X_p$ symbolized as $R_2(\mu|X) = \alpha_2 + \beta_2 X$.

Not surprisingly given (17), $R_1(X|\mu)$ turns out to be the 45-degree line: $\alpha_1 = 0$ and $\beta_1 = 1$. Figure 5 shows $R_1(X|\mu)$ in red, from which we see that $R_1(X|\mu)$ avoids large horizontal deviations by the points in the scatterplot. Note that maximum likelihood estimation of $\mu_1, \mu_2, \dots, \mu_p$ is, in effect, using $R_1(X|\mu)$ to predict μ from X — even though the form of $R_1(X|\mu)$ indicates prediction of X from μ !

To predict μ from X , a more appropriate option than using $R_1(X|\mu)$ is to employ $R_2(\mu|X)$. Geometrically, we want to avoid large vertical deviations by the points in the scatterplot. Appealing to least squares, we may estimate β_2 and α_2 by

$$(18) \quad \hat{\beta}_2 := \frac{\sum_{j=1}^p (X_j - \bar{X})(\mu_j - \bar{\mu})}{\sum_{j=1}^p (X_j - \bar{X})^2} \quad \text{and} \quad \hat{\alpha}_2 := \bar{\mu} - \hat{\beta}_2 \bar{X},$$

where $\bar{\mu} := \sum_{j=1}^p \mu_j / p$. Figure 5 shows $R_2(\mu|X)$ in blue, with α_2 and β_2 estimated from the points in the scatterplot per (18). The blue line in Figure 5 illustrates Galton's [1890] phenomenon of regression to the mean:

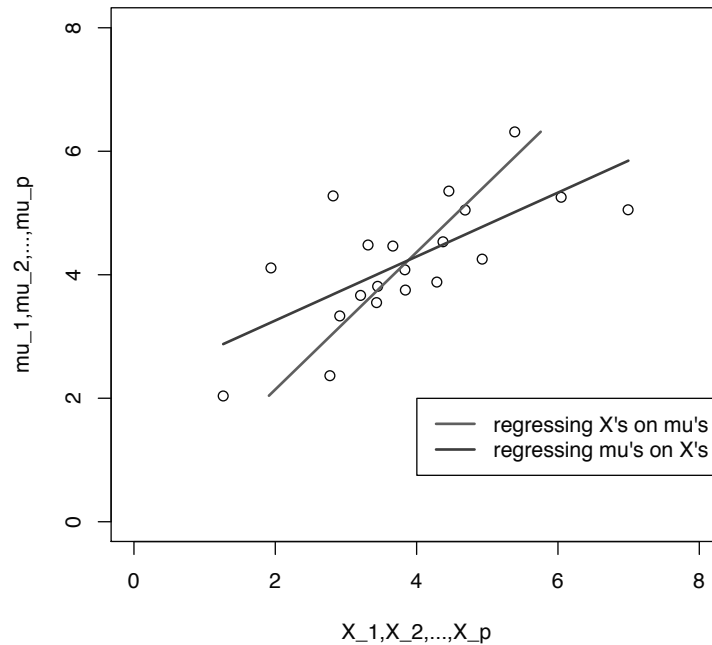


Figure 5. Stein estimation from a regression perspective

extremely low values of X suggest low — but not as extremely low — values of μ , while extremely high values of X suggest high — but not as extremely high — values of μ .

This idea is further illustrated in Figure 6, which pertains to the bladder cancer data from Table 1. With these real data, we find that $R_1(X|\mu)$ has greater than unit slope. So, in fact, predicting μ from X based on $R_1(X|\mu)$ is not merely counterintuitive but also outlandish. For instance, a one-year mortality rate of 8.5 (per 100,000) would lead to a predicted ten-year mortality rate of 12.2, whereas a one-year mortality rate of 7.5 would lead to a predicted ten-year mortality rate of 7.7. On the other hand, predicting μ from X based on $R_2(\mu|X)$ seems quite reasonable.

Of course, we took liberties here for the sake of illustration: the practitioner cannot use $\mu_1, \mu_2, \dots, \mu_p$ to estimate α_2 and β_2 because the practitioner does not know $\mu_1, \mu_2, \dots, \mu_p$. Stigler [1990] argued that $\hat{\beta}_2$ and $\hat{\alpha}_2$ in (18) may be reasonably replaced by

$$\tilde{\beta}_2 := 1 - \frac{p-1}{\sum_{k=1}^p (X_k - \bar{X})^2} \quad \text{and} \quad \tilde{\alpha}_2 := \bar{X} - \tilde{\beta}_2 \bar{X}.$$

Then, if we estimate μ_j for $1 \leq j \leq p$ by $\tilde{\alpha}_2 + \tilde{\beta}_2 X_j$, we obtain

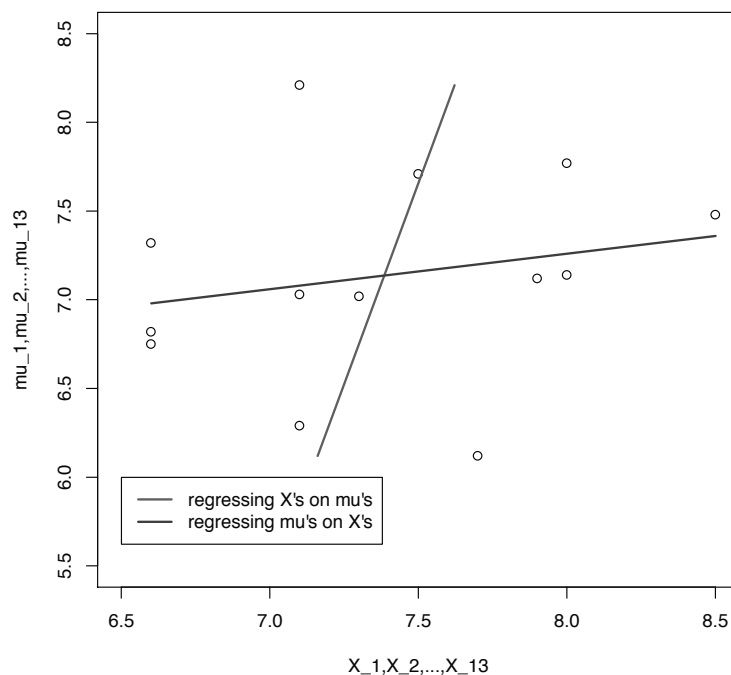


Figure 6. Stein estimation from a regression perspective (bladder cancer data)

$$(19) \quad \bar{X} + \left(1 - \frac{p-1}{\sum_{k=1}^p (X_k - \bar{X})^2}\right) (X_j - \bar{X}).$$

Expression (19) is identical to expression (13), except that $p-3$ has been replaced by $p-1$. The estimators in (19) are superior to maximum likelihood estimators, in that they have a lower risk (10), as long as $p \geq 6$.

We remark that Meng [2005] recovered the estimators (13), with the $p-3$ intact, by modifying Stigler's [1990] derivation. Meng's [2005] approach employed a special formula relating the expected value of a ratio of random variables to a statistical entity called the moment generating function.

7 RELATED METHODOLOGICAL DEVELOPMENTS

The impact of Stein's phenomenon on statistical science for multidimensional problems has been striking. Indeed, Stein's phenomenon has motivated numerous methodological and theoretical advances, some of which are taking place even now — more than a half-century after Stein's crucial insight!

This section briefly highlights some such advances, mostly in the context of a normal probability model like that described at the beginning of Section 4. We

emphasize that this section is by no means an exhaustive account of the advances that have been inspired by Stein's phenomenon. In particular, much work has also been done outside the context of a normal probability model.

To clarify the impetus for methodological and theoretical developments following the revelation of Stein's phenomenon, we reiterate that a James-Stein estimator such as (4) or (5) performs admirably in reducing the risk (10) in the simultaneous estimation of $\mu_1, \mu_2, \dots, \mu_p$. Yet, an individual mean μ_j ($1 \leq j \leq p$) may be grossly misestimated by (4) or (5), particularly if μ_j is not close to most of the other means $\mu_1, \mu_2, \dots, \mu_{j-1}, \mu_{j+1}, \dots, \mu_p$.

Considering again the data analysis example in Section 3, the expected annual deaths per 100,000 males for Georgia was only 6.75. This was smaller than for all but two of the other states, so the maximum likelihood estimate of 6.60 for Georgia was much better than the James-Stein estimate of 7.23.

To emphasize the distinction between reducing $\mathbb{E} \left[\sum_{j=1}^p (\mu_j - D_j)^2 \right]$ and controlling $\mathbb{E} [(\mu_j - D_j)^2]$ for a single specific j ($1 \leq j \leq p$), statisticians sometimes refer to the former as the “ensemble risk” and the latter as a “component risk”. Unlike the ensemble risk, a component risk can be higher with a James-Stein estimator than with the maximum likelihood estimator. This might have discouraged some data analysts from using a James-Stein estimator in practical applications. Recognizing this, Efron and Morris [1971; 1972; 1973; 1975; 1976] made significant methodological and theoretical advances that rendered Stein-type estimators more attractive and accessible to practitioners:

1. They proposed “limited translation estimators” expressible in the form

$$\begin{aligned} & \left[1 - \frac{(p-2)}{\sum_{j=1}^p X_j^2} \rho_D \left((p-2)X_1^2 / \sum_{j=1}^p X_j^2 \right) \right] X_1, \\ & \left[1 - \frac{(p-2)}{\sum_{j=1}^p X_j^2} \rho_D \left((p-2)X_2^2 / \sum_{j=1}^p X_j^2 \right) \right] X_2, \dots, \\ & \left[1 - \frac{(p-2)}{\sum_{j=1}^p X_j^2} \rho_D \left((p-2)X_p^2 / \sum_{j=1}^p X_j^2 \right) \right] X_p \end{aligned}$$

with $\rho_D(u) := \min\{1, D/u^{1/2}\}$ for a specified $D > 0$. The practitioner can choose D to avoid deviating “too much” from the maximum likelihood estimator, and thereby avoid having too large a component risk, while still gaining some reduction of the ensemble risk. In particular, taking D smaller exercises more control over the component risks, while taking D larger achieves greater reduction of the ensemble risk. If desired, these limited translation estimators can be modified to shrink $X_1, X_2, \dots, X_p$ toward $\bar{X}$ (or to any specific nonzero constant) rather than to 0.

2. They described how to accommodate the more realistic scenario in which $X_1, X_2, \dots, X_p$ have unknown rather than known variances.
3. They articulated Bayesian and empirical Bayesian justifications for Stein-type estimators, providing convincing arguments for the use of Stein-type estimators in real life data analysis problems.

Following Efron and Morris, other authors developed numerous variants of Stein-type estimators to provide further appeal to practitioners and to accommodate loss functions of the form (14) rather than (9) exclusively. Berger [1980; 1985] presents an excellent account of these developments.

An important methodological advance based on Stein's phenomenon is the creation of improved confidence sets. A confidence set is a data-dependent region of the parameter space that has a guaranteed probability, called the "confidence level", of containing the true value of the unknown parameter describing how the data are generated. A student in an introductory statistics course learns that a confidence interval — a special kind of confidence set when the unknown parameter is one-dimensional — quantifies uncertainty in the estimation of the unknown parameter and that, moreover, there is a duality between confidence intervals and hypothesis tests. Here duality refers to a general correspondence between a significance level α test of a null hypothesis concerning the unknown parameter and a confidence interval for the same unknown parameter that has confidence level $100(1 - \alpha)\%$. More specifically, the $100(1 - \alpha)\%$ confidence interval consists precisely of those candidate values for the parameter that would not be rejected by the significance level α test.

For example, suppose that we have a single observation X available from a normal distribution with unknown mean μ and known standard deviation 1. Consider testing the null hypothesis that μ equals the candidate value μ_0 . The test rejects the null hypothesis at significance level 0.05 if and only if $|X - \mu_0| > 1.96$, which is to say that the null hypothesis is not rejected if and only if X is in the interval $\mu_0 \pm 1.96$. With the help of a little algebra, we see that X falling in the interval $\mu_0 \pm 1.96$ is the same as μ_0 belonging to the interval $X \pm 1.96$. Hence, $X \pm 1.96$ defines a 95% confidence interval for μ . This process can also be reversed, to obtain a hypothesis test from a confidence interval.

As the above example suggests, duality facilitates the use of confidence intervals in drawing inferences about the unknown parameter. Moreover, a similar duality pertains to hypothesis tests and confidence sets when the unknown parameter is multi-dimensional, so that developing better confidence sets is indeed a step forward in statistical science.

Regarding the normal probability model with which we began Section 4, a conventional confidence set for the vector $\boldsymbol{\mu} := (\mu_1, \mu_2, \dots, \mu_p)'$ is a p -dimensional sphere centered at the maximum likelihood estimator $\mathbf{X} := (X_1, X_2, \dots, X_p)'$,

$$C_0(\mathbf{X}) := \left\{ \boldsymbol{\mu} : \sum_{j=1}^p (\mu_j - X_j)^2 \leq m^2 \right\}.$$

The radius $m > 0$ of the p -dimensional sphere is chosen to achieve a desired confidence level. In the special case that $p = 1$, the sphere is just an interval and $C_0(\mathbf{X})$ has the form

(20) estimate $\pm$ margin of error,

where the radius is identified with the margin of error. Although formula (20) does not apply when $p > 1$, the radius is still referred to as the margin of error even when $p > 1$.

Motivated by his work leading to the James-Stein estimator, Stein [1962] initiated the search for improved confidence sets. Based on approximate calculations, he conjectured that there should exist a confidence set with higher confidence level than $C_0(\mathbf{X})$ but with equal or lesser margin of error when $p \geq 3$. This conjecture was settled affirmatively by Brown [1966] and Joshi [1967], although their arguments did not lend themselves to the explicit construction of the better confidence set. Employing Bayesian reasoning, Faith [1976] and Berger [1980] derived confidence sets that improved upon $C_0(\mathbf{X})$. However, these confidence sets were not directly related to a Stein-type estimator.

On the other hand, Hwang and Casella [1982] explicitly appealed to a Stein-type estimator in deriving a confidence set superior to $C_0(\mathbf{X})$. The improved confidence set is

$$C_{HC}(\mathbf{X}) := \left\{ \boldsymbol{\mu} : \sum_{j=1}^p (\mu_j - \tilde{\mu}_j)^2 \leq m^2 \right\},$$

where

$$\begin{aligned} \tilde{\mu}_1 &:= \left[1 - \frac{a}{\sum_{j=1}^p X_j^2} \right]^+ X_1, \\ \tilde{\mu}_2 &:= \left[1 - \frac{a}{\sum_{j=1}^p X_j^2} \right]^+ X_2, \dots, \\ \tilde{\mu}_p &:= \left[1 - \frac{a}{\sum_{j=1}^p X_j^2} \right]^+ X_p \end{aligned}$$

for an appropriately chosen positive constant a . Above, $u^+ := \max\{0, u\}$ denotes the “positive part” of u . The confidence set $C_{HC}(\mathbf{X})$ achieves a higher confidence level than $C_0(\mathbf{X})$ when $p \geq 4$.

Tseng and Brown [1997] and Samworth [2005] subsequently constructed confidence sets with the same confidence level as $C_0(\mathbf{X})$ but strictly smaller margin of error. These confidence sets are potentially more appealing to data analysts, as many practitioners prefer to fix the confidence level and seek a smaller margin of error rather than fix the margin of error and seek a larger confidence level. Efron [2006]d) recently presented a very elegant and general method of constructing confidence sets with smaller margin of error than $C_0(\mathbf{X})$ and quantified the extent to

which the margin of error could be reduced. The aforementioned duality between hypothesis tests and confidence sets was part of Efron's derivation.

Another key methodological advance related to Stein's phenomenon involves estimation of the "covariance matrix" Σ or the corresponding "precision matrix" Σ^{-1} , where the superscript $^{-1}$ denotes the operation of matrix inversion. We refer to Σ as the covariance matrix because its (j, k) element ($1 \leq j, k \leq p$) is the covariance between X_j and X_k . (If $j = k$, then this is just the variance of X_j .) The conventional estimator of Σ is

$$\widehat{\Sigma}_0 := \mathbf{S},$$

whose (j, k) element ($1 \leq j, k \leq p$) is the sample analogue to the (j, k) element of Σ in much the same way that X_j ($1 \leq j \leq p$) is the sample analogue to μ_j . The corresponding conventional estimator of Σ^{-1} is

$$\widehat{\Sigma}_0^{-1} := K\mathbf{S}^{-1},$$

where K is a positive constant depending on p and the sample sizes $n_1, n_2, \dots, n_p$.

Surprisingly, a Stein-type phenomenon holds in the estimation of Σ and Σ^{-1} , even though the quantities being estimated are matrices.

Regarding Σ^{-1} , Efron and Morris [1976] derived a Stein-type estimator that expands, rather than shrinks, the conventional estimator $K\mathbf{S}^{-1}$. This estimator is

$$\widehat{\Sigma}_{EM}^{-1} := K\mathbf{S}^{-1} + \frac{p^2 + p - 2}{tr(\mathbf{S})} \mathbf{I}_{p \times p},$$

where tr represents the trace of a matrix (sum of the diagonal elements) and $\mathbf{I}_{p \times p}$ is the identity matrix (diagonal elements equal to 1, non-diagonal elements equal to 0) with p rows and p columns. Efron and Morris [1976] showed that $\widehat{\Sigma}_{EM}^{-1}$ has smaller risk than $\widehat{\Sigma}_0^{-1}$. In this setting, the risk of an estimator $\widehat{\Sigma}^{-1}$ is defined, up to a constant of proportionality, as

$$(21) \quad \mathbb{E} \left[\frac{tr((\widehat{\Sigma}^{-1} - \Sigma^{-1})^2 \mathbf{S})}{tr(\Sigma^{-1})} \right].$$

The loss underlying (21) is not as transparent as (9). However, the loss underlying (21) has the quadratic-like feature of depending on the squared difference between the estimator $\widehat{\Sigma}^{-1}$ and the target Σ^{-1} .

On the other hand, Stein-type estimators for Σ do entail a sort of shrinkage. In particular, quantities called the "eigenvalues" of $\mathbf{S}$ are pulled toward each other, in much the same way that the James-Stein estimator (5) pulls $X_1, X_2, \dots, X_p$ toward each other. Dey and Srinivasan [1985] constructed a wide class of Stein-type estimators with smaller risk than the conventional estimator $\mathbf{S}$. In this setting, the risk of an estimator $\widehat{\Sigma}$ is defined as

$$(22) \quad \mathbb{E} \left[\text{tr}(\widehat{\Sigma}\Sigma^{-1}) - \log\{\det(\widehat{\Sigma}\Sigma^{-1})\} - p \right],$$

where $\det$ is the determinant of a matrix. The loss underlying (22) is referred to as “entropy loss”.

Finally, a major theoretical development associated with Stein's phenomenon is the characterization of estimators of μ that can be improved upon when quadratic loss underlies the risk. Brown [1971] and Srinivasan [1981] provided a complete characterization of such estimators.

8 STEIN'S PHENOMENON AND MODERN SCIENCE

We conclude this article with a brief discussion of how Stein's phenomenon is helping to shape modern science. As with the last section, we emphasize that the present discussion does not provide an exhaustive account. We offer some general remarks and then consider one specific illustration in detail.

Many of the challenging statistical problems in modern science relate to the dimensionality of the data, specifically a large ensemble size p accompanied by comparatively small sample sizes $n_1, n_2, \dots, n_p$. These are often referred to as “large p , small n ” scenarios. Such scenarios arise in numerous fields, including molecular biology and genetics, finance, econometrics, computer science, and high energy physics. A prototypical example is microarray data analysis, in which expression data are obtained for each of p genes on $n = n_1 = n_2 = \dots = n_p$ subjects with $p \gg n$. Classical statistical methods fare poorly in this “large p , small n ” scenario, which has inspired a massive body of literature on microarray data analysis; see Dai and Charnigo [2008] and references therein for just a few of the many works. Some of the more recent literature on microarray data analysis employs Stein's phenomenon. For instance, Schafer and Strimmer [2005] developed a Stein-type covariance matrix estimator to map gene associations. Their application reinforces the idea that Stein's phenomenon is not limited to measures of central tendency.

An interesting illustration of Stein's phenomenon in modern science, which we consider in some detail, arises in molecular systems biology. Here a major goal is to characterize the interplay among genes so that molecular mechanisms of diseases and associated cellular functions can be identified and described. Fulfillment of this goal requires estimating associations between pairs of genes and mapping the interplay among genes as a network. A general model for the association between two genes with expression levels X and Y entails a quantity $M(X, Y)$ called “mutual information”. The mutual information is related to the statistical concept of “entropy”. When X and Y are jointly normally distributed, the mutual information is a measure of linear association that is closely related to the familiar Pearson correlation.

If we discretize the expression levels into p categories and model X and Y with

multinomial distributions, then the mutual information is given by

$$M(X, Y) := H(X) + H(Y) - H(X, Y),$$

where $H(X) := -\sum_{j=1}^p \theta_j \log \theta_j$ and $\theta_1, \theta_2, \dots, \theta_p$ are the probabilities that X falls within the respective categories. We may define $H(Y)$ and $H(X, Y)$ analogously, noting that there are p probabilities underlying $H(Y)$ and $p \times p$ probabilities underlying $H(X, Y)$. Estimating $M(X, Y)$ is a simple matter if the underlying probabilities can be recovered. Let $\boldsymbol{\theta}$ be a vector containing these probabilities. The conventional maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{ML}$ does not work well when p is large relative to the sample sizes.

Hausser and Strimmer [2008] therefore developed a Stein-type estimator of the form

$$\hat{\boldsymbol{\theta}}^* := \lambda \hat{\boldsymbol{\theta}}^{Target} + (1 - \lambda) \hat{\boldsymbol{\theta}}_{ML},$$

where $\hat{\boldsymbol{\theta}}^{Target}$ is a “target estimator” obtained by reducing the dimension of the data and λ is determined adaptively from the data. Hausser and Strimmer [2008] provided empirical evidence for the superiority of $\hat{\boldsymbol{\theta}}^*$ in recovering $M(X, Y)$. Furthermore, when they applied $\hat{\boldsymbol{\theta}}^*$ in the analysis of *E. Coli* stress response data from 102 genes, they found three new significant associations that had not been detected with other methods.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. DMS-0706857. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Prasanta Bandhyopadhyay and an anonymous Reviewer for their comments on an earlier version.

BIBLIOGRAPHY

- [Berger, 1976] J. Berger. Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Annals of Statistics*, **4**, 223-226, 1976.
- [Berger, 1980] J. Berger. A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Annals of Statistics*, **8**, 716-751, 1980.
- [Berger, 1982] J. Berger. Bayesian robustness and the Stein effect. *Journal of the American Statistical Association*, **77**, 358-368, 1982.
- [Berger, 1985] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 1985.
- [Blyth, 1951] C. Blyth. On minimax statistical decision problems and their admissibility. *Annals of Mathematical Statistics*, **22**, 22-42, 1951.
- [Brown, 1966] L. D. Brown. On the admissibility of invariant estimators of one or more location parameters. *Annals of Mathematical Statistics*, **37**, 1087-1136, 1966.
- [Brown, 1971] L. D. Brown. Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Annals of Mathematical Statistics*, **42**, 855-903, 1971.

- [Cox and Hinkley, 1974] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman & Hall/CRC, Boca Raton, 1974.
- [Dai and Charnigo, 2008] H. Dai and R. Charnigo. Omnibus testing and gene filtration in microarray data analysis. *Journal of Applied Statistics*, **35**, 31–47, 2008.
- [Dey and Srinivasan, 1985] D. Dey and C. Srinivasan. Estimation of covariance matrix under Steins loss. *Annals of Statistics*, **13**, 1581–1591, 1985.
- [Efron, 2006] B. Efron. Minimum volume confidence regions for a multivariate normal mean vector. *Journal of the Royal Statistical Society Series B*, **68**, 655–670, 2006.
- [Efron and Morris, 1971] B. Efron and C. Morris. Limiting the risk of Bayes and empirical Bayes estimators — part I: the Bayes case. *Journal of the American Statistical Association*, **66**, 807–815. 1971.
- [Efron and Morris, 1972] B. Efron and C. Morris. Limiting the risk of Bayes and empirical Bayes estimators — part II: the empirical Bayes case. *Journal of the American Statistical Association*, **67**, 130–139. 1972.
- [Efron and Morris, 1973] B. Efron and C. Morris. Stein's estimation rule and its competitors — an empirical Bayes approach. *Journal of the American Statistical Association*, **68**, 117–130, 1973.
- [Efron and Morris, 1975] B. Efron and C. Morris. Data analysis using Stein's estimator and its competitors. *Journal of the American Statistical Association*, **70**, 311–319, 1975.
- [Efron and Morris, 1976] B. Efron and C. Morris. Multivariate empirical Bayes and estimation of covariance matrices. *Annals of Statistics*, **4**, 22–32, 1976.
- [Faith, 1976] R. E. Faith. Minimax Bayes set and point estimators of a multivariate normal mean. Technical Report 66, University of Michigan, 1976.
- [Fisher, 1925] R. Fisher. Theory of statistical estimation. In *Proceedings of the Cambridge Philosophical Society*, **22**, 700–725, 1925.
- [Galton, 1890] F. Galton. Kinship and correlation. *North American Review*, **150**, 419–431, 1890.
- [Gauss, 1821] G. K. Gauss. Theory on the combination of observations least subject to errors (Part One), 1821. Translation in same-titled book by G. Stewart (1995), Society for Industrial and Applied Mathematics.
- [Gauss, 1823] K. Gauss. Theory on the combination of observations least subject to errors (Part Two), 1823. Translation in same-titled book by G. Stewart (1995), Society for Industrial and Applied Mathematics.
- [Gauss, 1826] K. Gauss. Theory on the combination of observations least subject to errors (Supplement), 1826. Translation in same-titled book by G. Stewart (1995), Society for Industrial and Applied Mathematics.
- [Gelman *et al.*, 1995] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, 1995.
- [Hausser and Strimmer, 2008] J. Hausser and K. Strimmer. Entropy inference and the James-Stein estimator, with applications to nonlinear gene association networks. Technical report, arXiv:0811.3579v2 [stat.ML] 31 Dec 2008.
- [Hodges and Lehmann, 1950] J. Hodges and E. Lehmann. Some problems in minimax point estimation. *Annals of Mathematical Statistics*, **21**, 182–197, 1950.
- [Hudson, 1974] M. Hudson. Empirical Bayes estimation. Technical Report 58, Stanford University, 1974.
- [Hwang and Casella, 1982] J. Hwang and G. Casella. Minimax confidence sets for the mean of a multivariate normal distribution. *Annals of Statistics*, **10**, 868–881, 1982.
- [James and Stein, 1961] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman, ed.), **1**, 361–379. University of California Press, 1961.
- [Joshi, 1967] V. Joshi. Confidence intervals for the mean of a finite population. *Annals of Mathematical Statistics*, **38**, 1180–1207, 1961.
- [LeCam, 1953] L. LeCam. On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *University of California Publications in Statistics*, **1**, 277–329, 1953.
- [Meng, 2005] X. Meng. From unit root to Stein's estimator to Fisher's k statistics: if you have a moment, I can tell you more. *Statistical Science*, **20**, 141–162, 2005.
- [Neyman, 1937] J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*. **236**, 333–380, 1937.

- [Rao, 1945] C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, **37**, 81-91, 1945.
- [Rao, 1967] C. R. Rao. Least squares theory using an estimated dispersion matrix and its application to measurement of signals. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (L. LeCam and J. Neyman, ed.), **1**, 355-372. University of California Press, 1967.
- [Robbins, 1951] H. Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman, ed.), 131-148. University of California Press, 1951.
- [Royall, 1997] R. Royall. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall/CRC, London, 1997.
- [Samworth, 2005] R. Samworth. Small confidence sets for the mean of a spherically symmetric distribution. *Journal of the Royal Statistical Society Series B*, **67**, 343-361, 2005.
- [Schafer and Strimmer, 2005] J. Schafer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**, Issue 1, Article 32, 2005.
- [Srinivasan, 1981] C. Srinivasan. Admissible generalized Bayes estimators and exterior boundary value problems. *Sanhky Series A*, **43**, 1-25, 1981.
- [Stein, 1956] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman, ed.), **1**, 197-206. University of California Press, 1956.
- [Stein, 1962] C. Stein. Confidence sets for the mean of a multivariate normal distribution (with discussion). *Journal of the Royal Statistical Society Series B*, **24**, 265-296, 1962.
- [Stigler, 1990] S. Stigler. The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science*, **5**, 147-155, 1990.
- [Tseng and Brown, 1997] Y. L. Tseng and L. D. Brown. Good exact confidence sets for a multivariate normal mean. *Annals of Statistics*, **25**, 2228-2258, 1997.
- [Wald, 1949] A. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20**, 595-601.

DATA, DATA, EVERYWHERE: STATISTICAL ISSUES IN DATA MINING

Choh Man Teng

1 OCEANS OF DATA

Data used to be hard to come by and even harder to analyze in any large scale fashion. Advances in collection and storage capabilities have made it relatively convenient to produce and accumulate large volumes of data, often automatically archiving them in data repositories. NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) onboard the Terra spacecraft, for instance, generates 850 gigabytes of data per day [MODIS, 1999]. The number of nucleotide bases in GenBank, National Center for Biotechnology Information's genetic sequence database, has been doubling approximately every 18 months; a recent release contains more than 110 billion bases [GenBank, 2010].

On the other hand, data are still scarce in some domains. Microarray data, for example, are typically high dimensional and have a high variable-to-case ratio. It is not uncommon to have more variables than cases in a data sample. The problem space, however, is still large, as the space is determined not by the sample size but by the number of variables. This situation further complicates the analysis process as the amount of data often does not provide sufficient statistical support for many analysis methods.

The amount of data available combined with the number of variables that need to be considered are of a scale far beyond what is amenable to manual inspection. Automated and semi-automated data analysis is thus essential to sieve through the data for meaningful conclusions. This process is variously called data mining, knowledge discovery, machine learning, inductive inference and others, depending on the discipline. We will use all these terms depending on the aspect of the analysis process we would like to emphasize.

Below we will examine automated data analysis and some of its associated issues. Many of these issues are inherent in traditional statistical analysis before automation as well, but the necessary reliance on computerized methods applied to large databases exacerbates their import. Overviews and foundational issues regarding statistics in data mining and knowledge discovery are discussed in for example [Elder and Pregibon, 1996; Fayyad *et al.*, 1996; Glymour *et al.*, 1997; Hand, 2000; Breiman, 2001].

2 KNOWLEDGE DISCOVERY FROM DATA

Knowledge discovery from data (KDD) aims to extract useful information from large amounts of data. The “knowledge” to be discovered can be broadly grouped into three, not mutually exclusive, veins, depending on the use the extracted information is to be put to.

Description

For description, the goal is to provide an intensional summary of the extensional data given in the form of individual database records. For descriptive tasks, intelligibility is an important consideration. The target set of data points are to be summarized in a succinct and easily understandable way, such that its characteristic features can be easily grasped. Association rule mining [Agrawal *et al.*, 1993] is an example of a descriptive task. Description is not in itself inductive, however. There is no uncertainty involved in the results, and good descriptions of a database do not necessarily possess inferential power for generalizing to data outside of the described target database.

Prediction

For prediction, the goal is to produce a mechanism that will reliably foretell the value of a target feature from the values of other features of a new unseen instance. In contrast to description, prediction is concerned with induction and inference, and less emphasis is placed on providing a comprehensible description of the resulting model or method. In predictive tasks, accuracy of the predictions for new data instances is paramount. Black boxes that do not reveal the inner workings of the mechanism are quite acceptable as long as they provide accurate predictions of new data. For instance, neural networks are powerful but notoriously obscure predictors. They are universal approximators but it is not always easy to decipher the underlying predictive mechanisms.

Explanation

For explanation, the goal is to recover the underlying mechanism by which the data is generated. Explanatory tasks aim at uncovering the causal relationships between variables. Variables that are highly correlated with the target variable are valuable as predictors. However, they are not necessarily direct or indirect causes of the target variable. This distinction is crucial in some situations, as while we can predict the value of a target variable based on other closely correlated variables, for instance, an effect variable of the target, only causes of the target can influence its value. Causal pathways allow one to determine how a desirable outcome may be achieved by strategically manipulating its direct or indirect causes.

These three tasks are inter-related, but they differ in the characteristics required of the discovered knowledge and the prime criteria for evaluating whether the results are satisfactory. In practice, we desire some of each kind of characteristics. Results that are extremely accurate but entirely incomprehensible, or vice versa, are deemed wanting. In particular, we tend to bestow causal interpretations on the discovered relationships and provide and sometimes invent an explanation for the connections. These explanations may be warranted with background knowledge provided by subject matter experts, but sometimes they are contrived even when causation is not warranted.

We also need to distinguish between making predictions from passively observing the characteristics of the data, and making predictions when external perturbations are introduced. In the former, (combinations of) features that are highly correlated with the target, for example, are valuable predictors. In the latter, when a variable is held or changed to a particular value through external intervention, the way this manipulation affects the target variable depends on the structural relationships between the variables.

For instance, in an observational setting, observing a change in the value of an effect variable of the target will influence the prediction of the target value, whereas in a perturbation setting, a change in the value of an effect variable by external means has no bearing on the value of the target variable. This is in contrast to when the change is imposed on a variable that is a cause of the target variable, in which case the target value will vary accordingly. In addition, associations between the target and other variables may be changed or broken by the perturbation. The difference between these two prediction settings is a crucial consideration when we aim to bring about a particular state of the system through the enactment of some intervention action or policy.

We will mostly concentrate on statistical prediction as related to automated knowledge discovery, but will touch on other aspects here and there.

3 MONKEYS AND TYPEWRITERS; BANGLADESHI BUTTER AND THE S&P 500

Friends of automated knowledge discovery advocate exploratory data analysis [Tukey, 1977]. Foes have another term for this activity: “data dredging” has often been used derogatorily to describe the unconstrained blanket exploration of data [Selvin and Stuart, 1966; Smith and Ebrahim, 2002, for example]. Chance associations and regularities are bound to turn up in the data if we are allowed to test for an unlimited number of patterns. Some of these patterns will be spurious, but if the analysis procedure has not been corrected for such effects, we could end up with dubious but seemingly valid “discoveries” that may not be justified on the basis of the given data.

This is akin to the inverse scenario of the monkey pounding on a typewriter: not only that if the monkey types long enough, the complete works of Shakespeare will probably be produced, but if one looks long and hard enough, intelligible patterns

can be found in almost anything, including ink blots and any collection of pages randomly generated by the monkey.

In a slightly less fantastical setting, in [Leinweber, 2007] a search was initiated for predictors of the S&P 500 index annual closing values over a ten year period. Using regression analysis, an R^2 of 0.75 was obtained using a single predictor variable:

1: Butter production in Bangladesh.

The fit of the model was improved to $R^2 = 0.95$ by using instead the following two slightly extended variables:

1': Butter production in Bangladesh and the United States;

2 : Cheese production in the United States.

A further increase of the fitness to $R^2 = 0.99$ was achieved by adding a third variable into the mix:

3: Sheep production in Bangladesh and the United States.

Thus, to strike it rich it appears that we should pay great attention to dairy and farm products, especially those in Bangladesh. This case study was of course tongue-in-cheek:

The butter fit was the result of a lucky fishing expedition. The rest comes from throwing in a few other series that were uncorrelated to the first one. Pretty much anything would have worked, but we like sheep. [Leinweber, 2007]

The study went on to reveal that the regression model obtained was wildly inaccurate in predicting the stock index values outside of the given period used for the initial analysis, even though the regression model fitted the training data almost perfectly. In other words, the regression model obtained was purely, highly, descriptive but not predictive at all.

This case study illustrates amusingly how easy it is to leverage chance associations and overfit the data. Models so obtained are good descriptors of the given data and only the given data. They have little predictive power beyond the data to which they have been specifically fitted. In less outrageous cases, it might not be as easy to spot the incongruence, particularly when much of the data analysis is automated.

4 UNDERFITTING AND OVERFITTING

Models learned from the given data may underfit or overfit. These phenomena relate to the bias-variance dichotomy (together with the noise component) in a way that is not completely well understood, although there are general rules of thumb on the tradeoffs between the two.

4.1 *Bias, Variance and Noise*

The discrepancy between the value predicted for a variable and its actual value is attributable to three components: bias, variance, and noise. Bias refers to the average difference between the predicted value and the actual value. Variance refers to the variation arising from predictions obtained using different data samples drawn from the same population. Noise refers to the random variations inherent in the data. Bias and variance are related to the prediction model, whereas noise is a characteristic of the data.

In regression analysis, the expected squared error of the estimated value can be decomposed additively into functions of the three components. For the prediction or classification of discrete variables, using for example a zero-one loss function instead of a squared loss function, the bias-variance-noise decomposition is not as neatly delineated. The decomposition of the error involves multiplicative terms whose interaction may explain the non-trivial behavior of some of the machine learning methods [Friedman, 1997; Domingos, 2000a; Domingos, 2000b].

The Bias-Variance Spectrum

The noise component is a property of the data and cannot be directly influenced by the learning method. The bias-variance tradeoff however can be adjusted to some extent to achieve different prediction behaviors. In general models that underfit the data are of high bias, whereas models that overfit the data are of high variance. A decrease in one component tends to increase the other component, although not necessarily by the same amount.

Neither underfitting nor overfitting is easy to detect on the basis of the data used to learn the model. However, underfitting is less of a problem in practice since overfitting tends to increase the *descriptive* accuracy of the model, which would seem desirable *prima facie* and thus is a more tempting pitfall. An increase in the descriptive accuracy however does not necessarily translate into an increase in the predictive accuracy, and in many cases leads to a decrease instead, as the variance in the sample data that has been minutely fitted by the learned model does not generalize to other samples or the population from which the sample was drawn.

With enough finesse and models of sufficient degrees of freedom, we can always construct a model that describes the input data perfectly. For instance, suppose we use a polynomial curve fitting method to construct a predictive model from a set of sample data points with two variables. A linear model would underfit in many cases. The fit can be improved by increasing the order of the polynomial curve. In the extreme case, for $n + 1$ data points, there always exists an n -degree polynomial that exactly fits all the given data points.

Such a curve is perfectly descriptive of the data points at hand. However, if these data points are merely a random sample of a much larger noisy population, or if the true model is not really polynomial in the nominated independent variable, the n -degree polynomial is most likely overfitting and would predict poorly for

data points outside of the given sample. This problem is magnified when the data sample is noisy and the perfect curve that passes through all the data points is fitting the noise rather than the signal.

4.2 *The Reference Class*

Finding the optimal balance along the bias-variance spectrum can be thought of as a variant of the reference class problem:

If we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class. An individual thing or event may be incorporated in many reference classes... [Reichenbach, 1949]

A single case is a member of many reference classes. A data point with n known binary attributes belongs to 2^n reference classes, for example. Each of these reference classes has its own distributional properties, many of which are not compatible with those derived from other reference classes. In addition, when these reference statistics are compiled from sample data, not all of them are available and not all of them are equally well supported by the data.

The challenge then is to identify the appropriate reference class or reference classes among the competing classes to arrive at a justifiable inference for the target case. This is analogous to finding the model encompassing the appropriate bias-variance split. Too broad a reference class and we will underfit; too narrow and we will overfit.

Reichenbach's recommendation for the reference class problem was to

... proceed by considering the narrowest reference class for which reliable statistics can be compiled. [Reichenbach, 1949]

whereas Salmon proposed

... the single case should be referred to the broadest homogeneous reference class of which it is a member. [Salmon, 1967]

The two criteria are not as contradictory as they seem when *not* quoted out of context (guilty here!). The intuition is that among all the reference classes to which the single case in question belongs, the appropriate reference class should be as narrow as possible, but no narrower than what the accompanying statistics can demonstrate to be relevant. In other words, any narrower the reference class will be unsuitable as the basis of inference with respect to the target case. There are two ways a reference class can be too narrow to be suitable: the lack of statistical support and the lack of relevance.

It would appear that the more specific the reference class, the better matched it is to the target case in question, and thus we should always prefer the narrowest reference class. However, as the reference class becomes narrower, the amount of

data belonging to this reference class becomes smaller. In the extreme case we may arrive at a reference class that matches the target case very closely but contains no known data points, rendering total ignorance with regard to the target prediction.

The second way a reference class can be too narrow is by including features that are irrelevant to the target prediction. Again as we add more irrelevant features, the corresponding reference class maps to fewer known data points, but the underlying probability distribution of the target property does not differ substantively from the distribution in the reference class without the irrelevant features. Thus, even if the narrower reference class still contains enough data points, the inclusion of irrelevant features effectively increases the variance of the prediction without increasing its accuracy.

Reference classes are not always nested and sometimes there is no single narrowest suitable reference class. The reference class formulation was further extended in [Kyburg, 1974; Kyburg and Teng, 2001].

4.3 *Competing Hypotheses*

The reference class problem surfaces in a number of different settings. In nonmonotonic reasoning, multiple inheritance leads to multiple extensions in, for example, default logic [Reiter, 1980]. Taxonomic hierarchies are used to resolve inheritance conflicts between rules that give competing conclusions [Etherington and Reiter, 1983; Touretzky, 1984].

In machine learning, a model or classifier is abstracted from a data set and predictions about new data instances are made according to the model. Choosing among competing models may be considered a generalized reference class problem. Some classifiers can be thought of as algorithmic blueprints for picking out a reference class to be used for the prediction of each incoming new instance. For example, a decision tree [Quinlan, 1993] guides a new instance down successive branches of the tree at each level according to specific features possessed by the new instance. The path from the root to a leaf in the decision tree encapsulates the set of feature criteria that determine the appropriate reference class for the test instance.

Occam's razor is often used in the KDD literature to justify preferring simpler hypotheses to more complicated ones. For example, the pruning of decision trees aims at reducing the size of the tree by removing branches at the bottom that are deemed to be overfitting the data. Stepwise logistic regression and Markov blanket procedures are developed to identify the features relevant for the prediction task at hand from among all the input features.

At first glance the Occam's razor principle appears to run counter to the directive for preferring the more specific reference class. A simpler model translates to preferring the less complicated, more general, models that take into account fewer attributes. However, simplicity does not automatically translate into relevance. Selecting for a simpler model is justified only when irrelevant variables are included in the more complicated models to overfit the noise and sampling variance.

There have been theoretical arguments and supporting experimental evidence indicating that with the same training error, the simpler model is not always the better [Domingos, 1998, for example].

5 TESTING AND EVALUATION

Many data mining procedures involve administering statistical tests repeatedly. For example, in stepwise regression, variables are selected by their statistical significance in successive groupings. For decision trees, bottom leaves and branches are pruned iteratively according to a statistical criterion. In addition, hypotheses abstracted from the data need to be evaluated for predictive accuracy, and at times many hypotheses are tested together or serially in search of the “nugget” in data mining. We will examine below some issues pertaining to the statistical significance of a test result.

5.1 Hypothesis Testing

In hypothesis testing, typically a single, simple, null hypothesis is tested against difference in all directions. The null hypothesis is rejected when the chance of false rejection is less than a specified level α . The value of the parameter α is meant to be calibrated with respect to the test setting. For example, the larger the sample size, the higher the power of the test and the easier the null hypothesis is rejected. This problem can be counteracted by lowering the value of α . However, either because the guideline is vague or because the quantification is complex, in practice many studies adopt a standard parameter value of 0.05 or 0.01 for α . The use of these conventional values is seldom questioned, but adopting any value other than these two defaults necessitates a thorough defense of the choice, which may additionally discourage efforts to calibrate the parameter.

Multiple hypothesis testing further complicates the determination of a suitable value for the parameter α . Each hypothesis, independently tested at level α , stands a chance of α of being falsely rejected. Even if all the null hypotheses are true, at a level of 0.05 it is to be expected that 5 out of every 100 hypotheses tested would be falsely rejected simply by chance. The chance of obtaining at least one false rejection from multiple hypothesis tests is thus higher, and sometimes *much* higher, than the stipulated threshold α . For 100 independent hypotheses each tested at a level of 0.05, the probability of at least one false rejection is 99.4%: we can be practically certain of having committed the error.

Multiple hypothesis testing relates to the problem of overfitting discussed in the previous section. Test enough hypotheses and sooner or later we are bound to hit upon one that by sheer chance fits the training data well.

5.2 Corrections and Approximations

Various methods can be adopted to more reasonably bound the overall chance of error for multiple testing. This is often accomplished by lowering the acceptable chance of error with which each individual hypothesis in a multiple testing setting can be rejected. Instead of utilizing an analytic bound, it is also common in machine learning to validate a hypothesis experimentally with additional data not used for constructing the hypothesis in the first place.

Statistical Corrections

The level of significance parameter α can be corrected to reflect the testing of multiple hypotheses. One commonly used statistical correction is the Bonferroni method:

For n hypotheses tested at the level α , each hypothesis is rejected if its chance of error is less than α/n .

The Bonferroni method is a rather conservative method for handling the estimation of the chance of error in multiple testing. This method essentially provides a provably correct upper bound on the overall chance of error of falsely rejecting at least one hypothesis in the whole set of hypotheses considered. This upper bound holds regardless of the relationships between the hypotheses, but in many cases the bound may be higher than the true bound.

In areas such as genomics, it is not uncommon to test thousands or even millions of hypotheses at a time. To achieve an overall bound of α on the chance of committing a false rejection error, according to the Bonferroni method the individual hypotheses need to be tested at a level of a millionth of α , which may be all but unsatisfiable.

Other procedures have been proposed which impose less stringent individual significance levels, for instance the Holm-Bonferroni method [Holm, 1979] and the False Discovery Rate method [Benjamini and Hochberg, 1995], with alternative characterizations of the control of the chance of error.

Experimental Methods

For computationally expensive error estimation procedures such as permutation methods, the chance of error can be approximated by experimental methods. For example, Monte Carlo resampling can be used when the number of possible permutations is too large for exact computation. Random permutations are repeatedly composed from the data and the test statistics are compiled from these sampled permutations rather than from the whole enumerated set of permutations.

The predictive accuracy of an hypothesis learned from a data set can be evaluated alternatively by testing the hypothesis against unseen data. Ideally, the evaluation is performed on data collected in repeated trials, or, a portion of the data in a single trial is reserved for validation purposes only. Resource constraints

often do not allow either, in which case cross-validation procedures are engineered to extend the use of the available data.

In an n -fold cross validation, the data set is partitioned into n (almost) identical parts. In each of the n trials, an hypothesis is learned from $n - 1$ parts of the data and tested on the remaining one part. n trials are performed and n results are obtained from such an arrangement. The individual trials are correlated to some extent, since the input data are reused among them, but the predictive error rate can be estimated from the cross validation test results.

Even though in each cross validation trial the training and testing data are separate, meta-level overfitting can occur when a large number of models are trained and tested on the supposedly unseen test sets. Again, perform enough trials and one of the models would fit the “unseen” test data well. This concern led to an approach that splits the data into three parts, one part for training, a second part for testing, and a third part for validation. Models vetted by cross validation are evaluated against the still more unseen validation data set. While the use of a validation set further lowers the effects of data snooping, overfitting may still occur at the meta-meta level.

5.3 *Adjunction*

The problem of multiple testing can be restated as the problem of adjunctive inference, as exemplified by the lottery paradox [Kyburg, 1961]. Briefly, the lottery paradox can be paraphrased as follows.¹

Consider a fair lottery with a million tickets. One of the tickets will be drawn as the winner. The chance for a particular ticket, for example ticket number 37841, to win is one in a million, a number so small that we can be practically certain that it will lose. The same line of reasoning can be applied to all one million of the tickets available in the lottery. At the same time one of these tickets is guaranteed to be drawn as the winner.

These accepted statements taken conjunctively are inconsistent: every individual ticket will lose, yet one of these tickets will surely win. Worse, because of the symmetry it is not clear whether any one of these accepted statements should be rejected, even though they are jointly inconsistent. Removing any one of the statements will restore consistency; yet there is no one statement that seems to be less justified to be accepted than the others², and any choice we make will seem arbitrary and irrational.

Analogous problems arise qualitatively in the paradox of the preface [Makinson, 1965] and in default reasoning [Poole, 1989]. The problem encountered in multiple testing follows the same structure. Each hypothesis considered by itself has an

¹The lottery may perhaps more properly be described as a raffle.

²Not counting the statement stating that one of the tickets will be the winning ticket, unless we also consider the possibility of fraud in the lottery setup.

acceptable chance of false rejection, yet we can be practically certain that among a reasonably large repertoire of hypotheses at least one has been falsely rejected, even though there may not be any one particular hypothesis that can be reasonably singled out as the culprit.

The statistical corrections to the level of significance parameter α for multiple testing are in response to the failure of adjunction. When multiple conclusions are conjoined, the resulting compound conclusion is not warranted to the same degree as the individual conclusions. For the conjoined conclusions to be warranted at the original desired level α , the member conclusions need to be subject to more stringent thresholds, for example α/n for n individual hypotheses according to the Bonferroni method.

Similar problems are encountered by models learned inductively from sample data. Each statistically sound predictive rule incurs a small, acceptable, chance of error. Among a large number of these predictions accumulated from repeatedly applying individual predictive rules, we can be practically certain that there is at least one false prediction.

So far there does not appear to be a ready made solution to this problem. Adjunction is a very persuasive form of inference. Rather than giving it up completely to avoid inconsistent beliefs, perhaps it is more pragmatic to correct for the error bound whenever we can, and be prepared to retract some of the conclusions in any case, when these conclusions turn out to be false even though they appeared to be warranted initially.

6 ASSUMPTIONS AND VIOLATIONS

Theoretical analysis provides bounds and optimality results based on certain assumptions about the nature and distribution of data. These assumptions are more often than not violated in practice. Although such violations should prompt us to rethink and re-qualify the assumptions where possible, they should not discourage us from applying the methods with suitable precautions and a fair understanding of the robustness of the methods.

6.1 *Random Sampling*

Given a set of data, a model is constructed and used to predict future or unseen events. For a model to be predictive, we need the future to be just like the past, and we need the data used to construct the model to be a representative sample of the parent population. This notion is usually captured as a random sampling assumption: the data observations are independent and identically distributed (i.i.d.) with respect to the parent population.

Data collection is often opportunistic. Data are collected whenever and wherever they can be collected. They may consist of Psychology 101 students, or birds that were spotted on particular days, or households who have access to the internet (so they may participate in online polls). These convenience samples are hardly

i.i.d. samples. Not taking into account the provenance of the data and instead treating the data as an i.i.d. sample will distort the inferences drawn from the data.

The problem goes deeper than that. Even in the best of circumstances (other than artificial formulations), i.i.d. samples simply cannot be had. Consider the case of Psychology 101 students. The parent population is usually not the restrictive set of students in Psychology 101 classes, or all college students, or even the far more expansive “people here and now”. The inferences drawn from the data on the behavior of Psychology 101 students in many cases are intended to implicate the behavior of the all inclusive set of human beings, geographically and temporally dispersed.

Even if we can sample members independently and with equal probability from all corners of the world, future members of the human population cannot be available for i.i.d. sampling at the present time at any cost. Yet prediction often are concerned with future events (although we can also aim to predict retrospective or current events that are not included in the sample data). Future events are then members of the parent population that are inaccessible for sampling, random or otherwise.

Thus the i.i.d. sampling requirement is at best only an idealized requirement. Not only do we not have good reasons for considering a data sample to have been drawn i.i.d. over the entire parent population, but we usually have strong evidence against such a claim.

6.2 *Representative Sampling*

Our goal in sampling is not to obtain a random sample per se, but to obtain a sample representative of the parent population, which will allow us to infer from the characteristics of the sample to the characteristics of the population, based on the argument that the sample is “similar” to the rest of the population. Other than in the most trivial cases, it is not possible to establish the representativeness of a sample, as that would require knowing the composition of the parent population in the first place, in which case we do not need to go to the trouble of drawing a sample and performing an inference from the sample to the population.

A random sample is often used as a stand in for a representative sample. Random sampling however is neither necessary nor sufficient to ensure a representative sample. A random sample might still be skewed, as it must happen every now and then as a mathematical fact, and a non-random sample might have the appropriate proportion of elements, either by careful crafting of the sample or just by chance. Nonetheless, random sampling allows us to relate to the representativeness of a data sample in a mathematically rigorous way. The degree of similarity between the sample and the population can be quantified by the characteristic variations inherent in i.i.d. sampling.

Still, the i.i.d. sampling assumption is provably false in many situations. Instead of insisting on an untenable assumption (or effectively ignoring the implications

of accepting a false statement), we may consider reformulating the requirement as a weaker, but not as crisp, requirement on the data sample. The data sample does not need to be provably random or even highly probably random. Rather we require the absence of strong evidence against the representativeness of the sample. While we cannot prove representativeness, where appropriate we can cast very reasonable doubt upon it and sometimes even falsify it.

With such a requirement, good sampling practice then is followed not to ensure the impossible i.i.d. sampling, but to serve as a safeguard against some obvious sampling biases. However, even if we adhere to all good sampling practice, if there are reasons to cast doubt on a sample's representativeness, the statistical inference from the sample to the population should justifiably be undermined.

6.3 Nonmonotonic Inference

The above formulation is in the tradition of nonmonotonic reasoning. In default logic [Reiter, 1980], for example, a distinction is made between *prerequisites*, which need to be provably true, and *justifications*, which merely need to be plausible. A default rule is triggered if its prerequisite is provably true and its justifications are each not provably false. If later one of the justifications is found to be false, the default conclusion has to be retracted.

Random sampling, traditionally cast in the form of a default prerequisite, is not satisfiable. A more reasonable formulation would be to require as a prerequisite the sample be drawn by following “good sampling practice”. Good sampling practice does not ensure random or representative samples, but its protocols are relatively well defined and falsifiable, and adherence to such protocols can serve as positive evidence supporting the fitness of the sample.

The representativeness of a sample cannot be verified and thus cannot take the form of a prerequisite. Rather, we can require as a default justification that we *not* know that the sample drawn is *not* representative. This requirement can be embodied in a number of conditions that should not be known to be false, even though we may not know them to be true either. For example, to determine the average IQ of people, if it comes to our attention that all the members of the sample, drawn carefully following good sampling practice, happen to be members of mensa, we should not proceed with the inference. The inference would be blocked as the default justification that the sample be representative becomes suspect. Note that this can happen even with truly i.i.d. sampling, as samples comprising of only mensa members are equally likely to be drawn as samples of other compositions (unless the sample is so large that there are fewer mensa members than sample members). On the other hand, it is not necessary to establish positively that the sample is representative, a condition which is all but impossible to prove in most cases.

With these prerequisites and justifications, the conclusion of the default rule is then the permissible statistical inference from the sample to the population.

Similarly, these conditions for statistical inference can be captured in a modal logic framework.

7 THE CASE OF ASSOCIATIONS IN ASSOCIATION RULES

One of the predominant subjects in data mining is the extraction of association rules, first introduced in [Agrawal *et al.*, 1993]. A well-known application is the analysis of market basket data to discover customer buying patterns based on register transactions. An association rule is a rule of the form $X \Rightarrow Y$, intended to capture dependencies between, for example, items that are likely to be purchased together in a single transaction. A well known and well cited, but unattributed, discovered rule is $\text{diapers} \Rightarrow \text{beer}$, which suggests, loosely, that among customers surveyed, those who bought diapers also tend to buy beer at the same time. The interpretation of this association rule is left as an exercise for the reader.

More formally, the rule $X \Rightarrow Y$ says that if event X occurs in the database, so does event Y with a certain level of frequency, as measured by two values, *support* and *confidence*:

support: baskets in which X and Y appear together (as the percentage over all baskets in the database);

confidence: occurrences of Y among those baskets containing X (as the percentage over all baskets that contain X in the database).

Each basket consists of a set of items from a given domain, for example, a set of items purchased together in a single register transaction among all items that are available for purchase at this store. The database is given as a set of baskets, or a set of sets of items. The objective of association rule mining is to find all the rules with support and confidence values above some user-defined thresholds.

7.1 Description

Association rules are sometimes advanced as rules of inference and used in a predictive setting. For example, rules concerning the associations found between items in a market basket analysis are supposed to reflect consumer behavior in general, and it was suggested that the discovered rules can guide business decisions such as the running of store promotions and new product placement in the store. [Agrawal *et al.*, 1993]

Statistically speaking, this stance is problematic. Association rules are purely descriptive: they present a summary of the relationship between the attributes, as manifested in the instances in the existing data set. A rule $X \Rightarrow Y$, with 1% support and 90% confidence for example, indicates that at least 90% of X 's found in the given data set are also Y 's, and that X 's and Y 's co-occur in the data set at least 1% of the time. These rules however say nothing conclusive about the relationships between attributes in the instances that are not part of the given

data set. They merely report verbatim patterns found in the static data set at hand.

Mining association rules is a clearly defined task. The objective there is to generate all rules of the form $X \Rightarrow Y$ which are above some given support and confidence thresholds. The problem of evaluation and validation is thus reduced to one of correctness and efficiency. Correctness in this case is unambiguous. Any algorithm is required to return *the* set of rules meeting the given support and confidence criteria. Since there is no difference between the set of rules returned by one algorithm and the next in the standard association rule framework, much of the research effort in this area has been understandably focused on efficiency issues, in terms of time and storage. High performance association rule mining aims at overcoming the challenges imposed by the tremendous size of the data sets involved and the potential number of rules that satisfy the mining criteria [Savasere *et al.*, 1995; Agrawal *et al.*, 1996, for example].

Another line of research in this area is to find from among the complete set of association rules a subset of interesting, novel, surprisingly, anomalous or otherwise more noteworthy rules [Bayardo and Agrawal, 1999; Klemettinen *et al.*, 1999, for example]. Still, this amounts to selecting a subset of descriptive rules which may be smaller but with no more predictive power than the full set of rules.

7.2 Prediction

In prediction, the goal is to infer about the characteristics of instances we have not seen, instances that are outside of the original data set, either because these instances have not been sampled for our data set, or because these instances are in the future. We would like the rules we derive from the given data set to be indicative of the patterns that can be found in a much larger domain. The target domain may be the set of all potential transactions, for example, including both those performed by current customers who did not shop on the day we collected the data, as well as those that will be performed by future customers who have not shopped here yet but one day soon will. (We note also that the data collected are usually a convenience sample, with all its attendant problems. For example the data set may consist of all transactions that occurred on a particular day in a particular store, among all days and all stores in a chain.)

The target domain thus is not only much larger than the given data set, but also is typically of infinite cardinality and extending into the future. Examining the whole population is out of the question in many cases of interest, both because of the size of the population, and also because many members of the population, such as those that exist in the future, cannot be available for examination here and now.

In the association rule framework, the minimum support requirement is expressed as a relative proportion of the given data set. This does not make statistical sense in the predictive setting, when the data set is only a sample of our target population, which in addition may be infinite. Rather, statistical support is

based upon the absolute number of instances that satisfy a certain condition, for example the antecedent of a proposed rule. Furthermore, this support varies from rule to rule, according to the content and absolute number of occurrences of the rule antecedent. This is in contrast to the support value for standard association rules, which is always specified relative to the size of the whole data set.

This is not a matter of tweaking the support and confidence thresholds to arrive at a different set of association rules. The problem lies in distinguishing between a rule that is justifiably acceptable and one whose supporting evidence is inconclusive. It is not the case that

While confidence is a measure of the rule's strength, support corresponds to statistical significance. [Agrawal *et al.*, 1993]

The support criterion as construed in the standard framework is inapplicable for statistical inference.

In practice, however, the standard approach has been widely used in many situations, for both description and inference, with little appreciable difficulty. Depending on the utility and the sensitivity of the application, the descriptive support and confidence of the discovered rules may suffice as approximate indicators of their predictive capability in the larger intended population. This is especially true when the sample size in question is large, which is bolstered in the standard association rule framework by the combined effect of huge data sets and a reasonably high support threshold.

8 REPLACING STATISTICIANS WITH COMPUTERS

Traditionally experienced statisticians and data analysts are the custodians of the etiquette of statistical analysis. We rely on expert judgement and common sense to ensure analysis protocols and inference procedures are properly executed, and to detect potential methodological and experimental trouble spots. Common sense stops us from deriving the eating habits of the general public from polling people in a Buddhist temple, or using the butter production in Bangladesh to predict the S&P 500. Experts can in many cases separate spurious associations from relevant ones and extract meaningful patterns with insight into the data.

We may consider replacing statisticians with computers in automated data analysis and knowledge discovery. There are pros and cons in such an arrangement, as discussed below.

8.1 The Flip Side

The voluminous amount of data available for analysis these days poses a cognitive barrier for applying human judgement. Analysis is increasingly carried out by automated or semi-automated processes. The human analyst is increasingly removed from the nitty-gritty of the data. None of the issues discussed above are very new, but with automated data analysis and knowledge discovery these issues

have come to the forefront as computers for the most part lack both common sense and expert judgement.

There are efforts to formalize common sense and background knowledge, for example the project Cyc [Lenat, 1995], as well as efforts to develop objective and quantifiable measures of intuitive but non-trivial concepts such as what knowledge discoveries are considered “interesting” or “surprising” [Silberschatz and Tuzhilin, 1996; Chakrabarti *et al.*, 1998, for example]. It is hoped that computerized background knowledge can substitute for human common sense and expert judgement to some extent, or at least help cut down on the induced search space of the analysis task.

8.2 *The Upside*

Even though the state of the art is still far from capturing human insight and judgement, and a considerable amount of background knowledge would be needed for general reasoning and inference, in the near term usable knowledge bases can be constructed for relatively restricted domains. For scientific inference and knowledge discovery, at least initially we typically do not need to worry about emotions, metaphorical implications, and the like. The analysis of scientific data is considerably simplified when it is confined to a specialized and well defined domain.

On the upside, the use of computers to automate at least part of the data analysis process has made feasible a large repertoire of methods that were too computationally intensive when calculations had to be done entirely by hand. For example, factor analysis was preferred to tetrad methods because tetrad constraints are difficult to compute manually even for problems of moderate size [Guilford, 1954]. In addition, complex, high dimensional problems, frequently encountered in for example bioinformatics, do not easily yield an analytic solution, but they can be tackled numerically by Monte Carlo methods assisted by computers.

8.3 *On Balance*

Data are often too much or too complex to allow human insight. The automation of knowledge discovery allows us to discover more patterns more rapidly than ever. Computers however do not possess the expert judgement and common sense to choose the appropriate analysis methods and to interpret the significance and relevance of the findings. Without expert oversight, bizarre and nonsensical findings from the institutionalized process can easily pass as sober findings.

To realize the potential of automated knowledge discovery, we need to install explicit safeguards against misuse of methods and data. The data characteristics and conditions necessary for an analysis method need to be articulated formally, such that the analyst can be alerted when assumptions are violated, data are suspect, and matters are not as they should be.

Analysis methods developed for application under one set of conditions are routinely used even when the assumptions are violated. For instance, data distri-

butions are often not Gaussian, yet normality are often assumed in the methods applied. Similarly, the construction of naive Bayes classifiers depends on the independence between attributes, which is often patently false. We need to understand how robust an analysis method is in terms of the amount of deviation from the assumed conditions that can be tolerated, and the magnitude and nature of the effect of such deviations.

On the meta-level, there is the challenge of learning how to learn. Expert analysts select the methods to apply based on the problem and data at hand. This process may involve multiple steps and different methods at each step. Automated knowledge discovery systems could learn this selection from a repository of data and problem characteristics together with the analysis methods applied, provided that these meta-data are properly articulated and formalized. The issues discussed here and elsewhere will be of paramount relevance at this level as well.

BIBLIOGRAPHY

- [Agrawal *et al.*, 1993] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on the Management of Data*, pages 207–216, 1993.
- [Agrawal *et al.*, 1996] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast discovery of association rules. In U. Fayad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press, Menlo Park, CA, 1996.
- [Bayardo and Agrawal, 1999] R. Bayardo and R. Agrawal. Mining the most interesting rules. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 145–154, 1999.
- [Benjamini and Hochberg, 1995] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300, 1995.
- [Breiman, 2001] Leo Breiman. Statistical modeling: The two cultures (with comments and rejoinder). *Statistical Science*, 16(3):199–231, 2001.
- [Chakrabarti *et al.*, 1998] Soumen Chakrabarti, Sunita Sarawagi, and Byron Dom. Mining surprising patterns using temporal description length. In *Proceedings of the 24th International Conference on Very Large Data Bases*, pages 606–617. Morgan Kaufmann, 1998.
- [Domingos, 1998] Pedro Domingos. Occam’s two razors: The sharp and the blunt. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 37–43, 1998.
- [Domingos, 2000a] Pedro Domingos. A unified bias-variance decomposition and its applications. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 231–238, 2000.
- [Domingos, 2000b] Pedro Domingos. A unified bias-variance decomposition for zero-one and squared loss. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 564–569, 2000.
- [Elder and Pregibon, 1996] John F. Elder, IV and Daryl Pregibon. A statistical perspective on knowledge discovery in databases. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 83–113. AAAI Press, 1996.
- [Etherington and Reiter, 1983] David W. Etherington and Raymond Reiter. On inheritance hierarchies with exceptions. In *Proceedings of the Third National Conference on Artificial Intelligence*, pages 104–108, 1983.

- [Fayyad *et al.*, 1996] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery: An overview. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthrusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press, 1996.
- [Friedman, 1997] Jerome H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77, 1997.
- [GenBank, 2010] National Center for Biotechnology Information. *NCBI-GenBank Flat File Release 179.0. Distribution Release Notes*, 2010.
- [Glymour *et al.*, 1997] Clark Glymour, David Madigan, Daryl Pregibon, and Padhraic Smyth. Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery*, 1:11–28, 1997.
- [Guilford, 1954] J. P. Guilford. *Psychometric Methods*. McGraw-Hill, second edition, 1954.
- [Hand, 2000] David J. Hand. Methodological issues in data mining. In *Proceedings in Computational Statistics (COMPSTAT)*, pages 77–85, 2000.
- [Holm, 1979] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.
- [Klemettinen *et al.*, 1999] Mika Klemettinen, Heikki Mannila, and A. Inkeri Verkamo. Association rule selection in a data mining environment. In *Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, pages 372–377, 1999.
- [Kyburg and Teng, 2001] Henry E. Kyburg, Jr. and Choh Man Teng. *Uncertain Inference*. Cambridge University Press, 2001.
- [Kyburg, 1961] Henry E. Kyburg, Jr. *Probability and the Logic of Rational Belief*. Wesleyan University Press, 1961.
- [Kyburg, 1974] Henry E. Kyburg, Jr. *The Logical Foundations of Statistical Inference*. Reidel, Dordrecht, 1974.
- [Leinweber, 2007] David J. Leinweber. Stupid data miner tricks: Overfitting the S&P 500. *Journal of Investing*, 16(1):15–22, 2007.
- [Lenat, 1995] Douglas B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.
- [Makinson, 1965] David Makinson. The paradox of the preface. *Analysis*, 25:205–207, 1965.
- [MODIS, 1999] National Aeronautics and Space Administration, Goddard Space Flight Center. *TERRA: The Earth Observing System (EOS) AM-1*, 1999.
- [Poole, 1989] David Poole. What the lottery paradox tells us about default reasoning. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 333–340, 1989.
- [Quinlan, 1993] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
- [Reichenbach, 1949] Hans Reichenbach. *The Theory of Probability. An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability*. University of California Press, second edition, 1949.
- [Reiter, 1980] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.
- [Salmon, 1967] Wesley C. Salmon. *The Foundations Of Scientific Inference*. University of Pittsburgh Press, 1967.
- [Savasere *et al.*, 1995] Ashoka Savasere, Edward Omiecinski, and Shamkant Navathe. An efficient algorithm for mining association rules in large databases. In *Proceedings of the 21st Conference on Very Large Databases*, pages 432–444, 1995.
- [Selvin and Stuart, 1966] Hanan C. Selvin and Alan Stuart. Data-dredging procedures in survey analysis. *The American Statistician*, 20(3):20–23, 1966.
- [Silberschatz and Tuzhilin, 1996] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, 1996.
- [Smith and Ebrahim, 2002] George Davey Smith and Shah Ebrahim. Data dredging, bias, or confounding. *British Medical Journal*, 325:1437–1438, 2002.
- [Touretzky, 1984] David S. Touretzky. Implicit ordering of defaults in inheritance systems. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 322–325, 1984.
- [Tukey, 1977] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

This page intentionally left blank

Part XIII

An Application of Statistics to Climate Change

This page intentionally left blank

AN APPLICATION OF STATISTICS IN CLIMATE CHANGE: DETECTION OF NONLINEAR CHANGES IN A STREAMFLOW TIMING MEASURE IN THE COLUMBIA AND MISSOURI HEADWATERS

Mark C. Greenwood, Joel Harper and Johnnie Moore

Recent analyses of regional snow and river data for the late 20th century have suggested that western North American snow pack has decreased [Mote *et al.*, 2003] and the spring snow melt pulse has arrived earlier [Stewart *et al.*, 2005]. These papers investigate linear change over this time period allowing the results to differ across locations by using distinct linear regression models at each location. Moore *et al.* [2007] found less conclusive evidence for linear change when adjustments for multiple testing were considered but used similar statistical models otherwise. A class of nonparametric spatial-temporal models for the same data used in Moore *et al.* [2007] are used to estimate a common regional level trend while accounting for spatial and temporal dependencies, allowing evidence for linear and nonlinear regional level trends to be compared. We find evidence for a nonlinear regional level trend between 1951 and 2005, generally demonstrating earlier arrival over the study period as well as a more complex pattern of change than previously reported.

1 INTRODUCTION

Detecting the impact of climate change in natural systems is typically hampered by limitations caused by short time series, especially for instrumental records. In some systems, high levels of natural variability make the task particularly difficult due to low signal to noise ratios. To obtain more precise estimates of the average change, information can be aggregated across many locations to provide an average change in the system. By borrowing information across locations we can often find a less variable estimate of the average. When information is aggregated across locations, correlations between locations need to be considered to properly estimate the precision in the estimate. Many researchers trying to assess change have assumed that the change is linear over time. In some natural systems over moderate time scales that may be reasonable, but over long enough time frames that assumption certainly becomes problematic. The models used in this work

incorporate correlations over space and time and compare models that assume linearity to those that estimate trends nonlinearly using a nonparametric regression technique.

We propose a method to assess the evidence for nonlinear trends in the timing of streamflow in the headwaters of the Missouri and Columbia Rivers. Streamflow timing refers to the arrival date of characteristics of the yearly distribution of streamflow, most often the mean or median of the flow. We consider the day of the median or when half the yearly flow has occurred and look for changes in that timing from 1951 to 2005. Streamflow timing in gage locations such as ours, that are high elevation and not disturbed by irrigation practices, should primarily record climate change induced trends in snowmelt. The conjecture is that earlier timing measures should be observed over the study period due to warmer temperatures causing earlier snowmelt. The details of this trend and statistical evidence regarding it are the main concerns of this work, which extends Moore *et al.* [2007] into nonparametric spatial-temporal models.

2 PREVIOUS RESEARCH

Estimating the amount of change that has occurred due to global warming is a challenging problem due to relatively short physical records and relatively high natural variability in the systems. One approach that has often been employed is to estimate the amount of change at each location and test for statistical evidence for the presence of that change. This provides many localized estimates of change and can lead to problems with compounding of statistical error from using many hypothesis tests. Numerous researchers have searched for evidence of climate change in the Western US, focusing on different measures that might record that change. One area of this research involves analyzing streamflow discharge in high elevation stream gages, searching for changes in the shape or timing of the discharge over time. Stewart *et al.* [2005] and Moore *et al.* [2007] consider different streamflow metrics and use simple linear regression to assess evidence of a linear change over the last half century in Western U.S. rivers. The results of Stewart *et al.* [2005] are often cited as evidence of earlier snowmelt in this region. Moore *et al.* [2007] also found some evidence of earlier streamflow timing but found less convincing evidence of that change after adjusting for the number of tests considered. There are two issues with using site specific analysis techniques, one is the number of hypothesis tests to consider and the other is the difficulty in synthesizing those results to provide a regional level result.

In analyzing streamflow timing measures for evidence of climate change there are a number of issues to consider in the choice of analysis technique, especially related to the site specific analyses. The first is the choice of metric used to measure timing, the second is to consider the number of hypothesis tests considered (which scales with the number of locations and response variables considered), third is whether linearity is a reasonable assumption for the trends, and fourth is whether local analyses can provide the overall inference that is really of interest. These

issues are considered in order along with solutions to each considered in this work.

First, there is the question of the timing measure(s) to use. Stewart *et al.* [2005] use a peak detection algorithm and what they describe as the center of mass of timing (CT). CT is defined as $\sum_i t_i q_i / \sum_i q_i$ where t_i is the day of the water year for $i=1, \dots, 365$, q_i is the daily discharge. If we let $\text{total}Q = \sum_i q_i$ be the total annual discharge, CT can be re-written to show that it is the average of the day of the water year weighted by the proportion of the annual flow on that day (p_i) as follows: $\text{CT} = \sum_i t_i q_i / \text{total}Q = \sum_i t_i \frac{q_i}{\text{total}Q} = \sum_i t_i p_i$. This provides a day of the average flow for the water year. Hydrographs (plots of daily flow) are typically defined by the water year, starting Oct. 1, which is typically during the lower flow season in the areas in this study. Flows increase rapidly when the snow begins to melt in late winter and then flows taper off as the snow is exhausted and precipitation is low in the later portion of the summer. This tends to result in a highly skewed distribution as displayed in Figure 1 due to the choice of the starting date in the water year. It is well known that the mean, to retain its center of mass interpretation, moves in the direction of the long tail of the distribution (see, for example, [DeVeaux *et al.*, 2008]). The median is a better measure of the center in skewed distributions. We denote the day of the median as D50, calculated as $\sum_{i=1}^{D50} p_i = 0.5$. Additionally, outliers have more impact on the mean than on the median. Outliers could easily be generated in the low flow portions of the year due to precipitation events, dragging the mean in the direction of those unusual events. A final note on CT relates to the arbitrariness of the beginning of the water year. If, for some reason, this were to be re-defined, then the CT could be dramatically impacted whereas the median, which we use here, would be more robust to these changes.

To illustrate this, the flow records from USGS Gage 13186000, South Fork of the Boise River, near Featherville, ID at an elevation of 4218 ft, are shown in Figure 1. The plot contains the daily hydrographs and cumulative discharge functions [Greenwood *et al.*, 2007] for 1951 and 2005 at this location. The CT for 1951 was day 218.1 and D50 was 229 and for 2005 the CT was 215 and D50 was 232. The sensitivity of the mean to the length of the tail of the distribution is demonstrated by the difference between the measures of center in these two hydrographs.

Inspection of the two hydrographs in Figure 1 suggests the difficulty of using the date of the first large spring peak. It is especially difficult to visually choose a “peak” in the 2005 hydrograph that represents spring onset. For this reason, we focus on measures of the center of discharge. Moore *et al.* [2007] explore some other reasons for the preference of the median over timing of peak discharge or CT, including the stronger relationship between total yearly discharge and CT than total yearly discharge and the day of the median. The median seems to record more information that is independent of the total yearly discharge than CT and does not suffer the other drawbacks noted above; we will use the D50 in this analysis.

The second issue with site specific analyses involves repeated use of hypothesis tests where each test is subject to errors. Hypothesis testing can be criticized

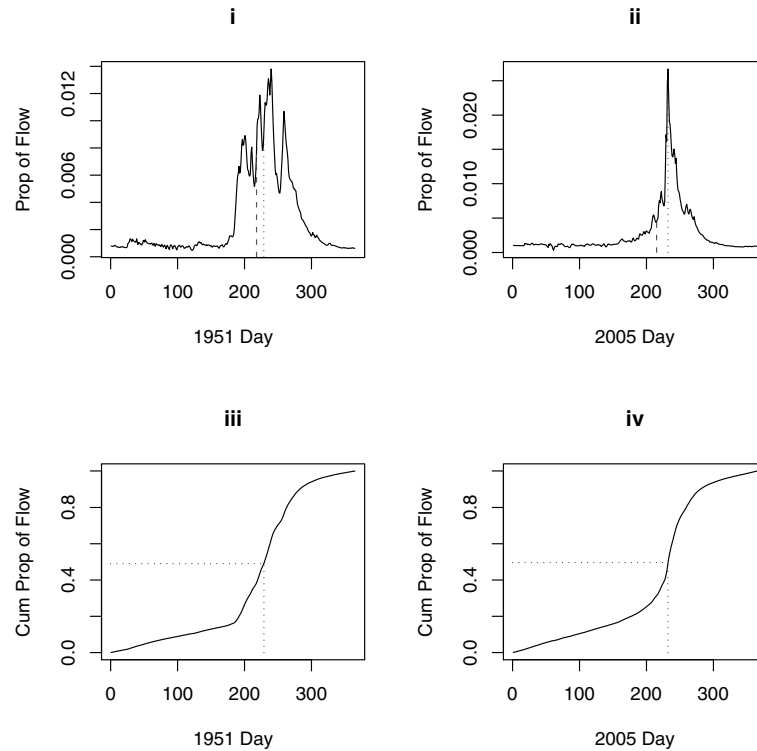


Figure 1. Gage on the South Fork of the Boise River, 1951 and 2005. Panel i and ii contain hydrographs, iii and iv contain cumulative yearly discharge, with CT displayed as the bold dashed line and D50 as the light dashed line.

merely for its inability to assess practical importance [Armstrong, 2007], the arbitrariness of the choice of significance level, the impacts of choice of direction in tests, and because repeated tests lead to increased probability of spuriously detected significant results. For example, Stewart *et al.* [2005] consider 294 different tests for significant linear trends in CT. The tests used are 2-sided, meaning that they attempt to detect change either to earlier or later timing, which is a conservative choice if researchers are only looking for earlier timing and that is all that is present. However, it is liberal in that it allows for significant results in either direction, with significant results of later timing providing a strong contradiction to the warming driven earlier timing. Additionally, Stewart *et al.* [2005] use a 10% significance level for most of their tests, which is generally considered a liberal choice, except as the previous discussion suggests, these 10% 2-sided tests are equivalent to 5% 1-sided tests if all the resulting estimates are in the direction

of earlier timing. Of their 294 tests, 105 are reported to be significant based on simple linear regression. This is a higher rate of significant results than would be expected just based on chance, which is 29.4 based on their assumed error rate, number of tests, and assuming that there is no trend present in any location. The problem is then to consider which 30 of the 105 significant results might have been due to chance alone. More generally in the same paper, around 3200 hypothesis tests were performed (the exact number of tests and significant results found is unclear), which is around 10 tests per study site. It is possible to control the overall error rate in testing by using an adjustment for multiple testing, leading to increased belief in the locations where significant changes are detected. The most well known adjustment for multiple testing is the Bonferroni method [1936], but it is known to be a conservative approach and with thousands of tests makes it nearly impossible to find any significant results. Newer approaches generated from the Biostatistics literature would seem to be more useful here, such as Benjamini and Hochberg [1995] and Storey and Tibshirani's q -value [2003]. These methods are less conservative, providing higher power to detect changes, while still providing useful protection for multiple testing by controlling other measures of error instead of the experiment-wise errors rates that Bonferroni controls. Moore *et al.* [2007] used the methods of Benjamini and Hochberg [1995] and found significant linear trends for D50 for 8 of 21 locations after adjusting for 126 tests considered in that paper (6 tests per study site). It is difficult to aggregate a set of significant and nonsignificant tests across locations in a region to be able to quantify the overall magnitude of change in the region; the judgement about the magnitude of change is often non-statistical.

A further concern in climate change detection involves the assumption of linearity of trends. A more detailed discussion of different methods for trend detection is given below, but this considers the assumption that climate change can and should be measured with linear or even monotonic assessments. While warming of around 0.2 degrees C per decade has been reported in Hansen *et al.* [2006], that paper highlights the nonlinearity of the global temperature trend by using 5-year running means to describe the trend in observed temperatures. This result alone suggests that linear assessments of climate change are crude approximations to the actual signals that temperature changes might generate and that extending linear trend detection to longer time scales can produce un-representative trend estimates. Over shorter time scales, more interesting local variation will be ignored if all trends are assumed to be linear. The best arguments for linear trend detection methods are that many curves can be reasonably approximated by straight lines locally, the results are simple to interpret, the statistical tests are well known, and that the linear trend could be a component of a more complicated polynomial but the only interest is in that component of the model. Other trend tests are based on monotonic trend detection, but these can also be problematic as local decreases may be found in the presence of overall increases and some methods would fail to detect the overall change. Additionally, these methods typically do not provide estimates of the amount of change, just whether there is evidence of its presence.

We take a different approach, comparing support for models that contain linear or nonlinear components, and then using the most supported model(s) to describe the form and magnitude of any trends.

In some, but certainly not all, climate change detection papers, the measurements were assumed to be independent over time. This is always the case when Pearson correlation and simple linear regression tests are used. When the autocorrelation is negligible, those procedures are not impacted and when the autocorrelation is negative, the procedures are actually conservative [Cryer and Chan, 2008]. The latter situation was encountered in Moore *et al.* [2007] so simple linear regression was used without adjustment there. In general, it is dangerous to use statistical procedures that assume independent measurements on time series without at least assessing the magnitude of the dependency. Similar concerns arise when attempting to aggregate effects spatially. All models considered here incorporate spatial-temporal correlations but tend to demonstrate minimal dependencies on either aspect for D50.

A major benefit of attempting to estimate a regional level trend is that the display of the trend can be simplified into a single plot as opposed to plots of significant trends by location. For example, in Stewart *et al.* [2005], the reader is left to visually assess the size and magnitude of trend tests and somehow visually aggregate the results. Based on the many trends explored in that paper, the International Panel on Climate Change [Lemke *et al.*, 2007] summarizes the observed trends as:

Date of maximum mountain SWE [snow water equivalent of the mountain snowpack] appears to have shifted earlier by about two weeks since 1950, as inferred from streamflow measurements [Stewart *et al.*, 2005]. That these reductions are predominantly due to warming is shown by regression analysis of streamflow. [Stewart *et al.*, 2005]

This summary result of about a 2 week shift is not accompanied by any measure of precision and is not really a statistical estimate as it is based on a non-statistical aggregation of many observed slope coefficients, some that are significant and some that are not. In Moore *et al.* [2007], confidence intervals for each question of interest are displayed, which allows assessment of the practical significance of the change at each location but fails to provide a direct regional level generalization. While it is more difficult statistically to estimate regional level trends, it is simpler to then interpret the results, as the average change in the region.

3 DATA

We use, as in [Moore *et al.*, 2007], a set of 21 high elevation streamflow gages. These gages were selected after considerable screening of locations to eliminate locations that have been impacted by irrigation withdrawal and small dams. They are a small subset of the gages used in [Stewart *et al.*, 2005] that are in the Columbia and Missouri Headwaters. A map of the gage locations is provided in Figure 2.

These should provide records that are not affected by human modification and are recording primarily snowmelt signals due to their high elevation locations, leading to an analysis that is focused on the change in timing in streamflow records that should correspond to snowmelt.

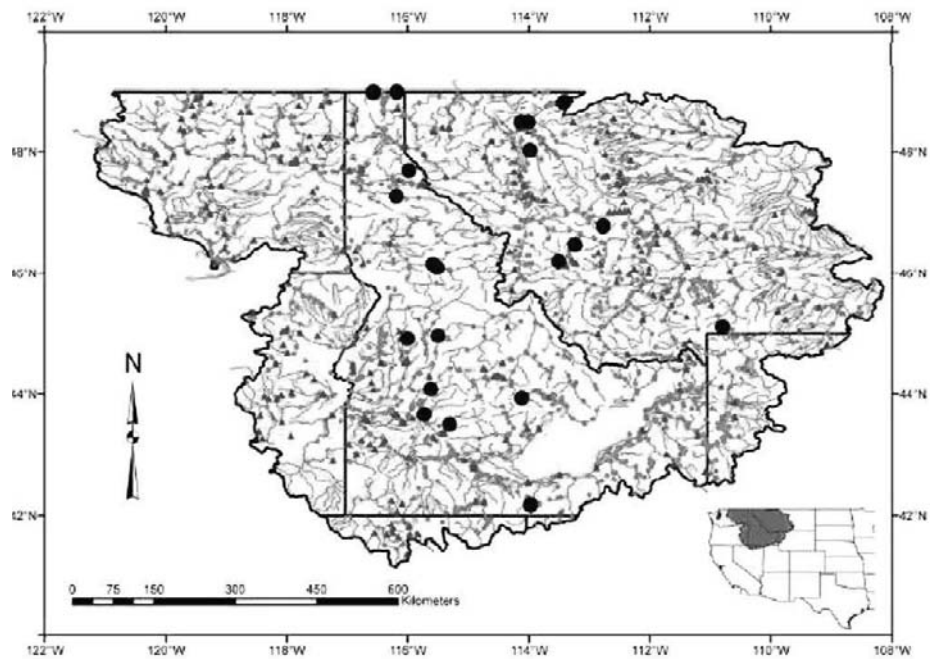


Figure 2. Map of gage locations, with black dots indicating the locations used.

Streamflow timing measures are driven by precipitation, temperature, solar radiation, and basin specific characteristics. Basin specific characteristics are incorporated into our models by allowing for different average timing for each gage through a flexible spatial intercept in the models explained below. Weather is both a local and regional phenomenon. On a regional scale, differences in weather between years can be explained, to some degree, using regional climate forcing functions such as El Niño/Southern Oscillation (ENSO, [Wolter *et al.*, 1993; 1998]) and Pacific Decadal Oscillation (PDO, [Zhang *et al.*, 1997]). ENSO has a period of around 7 years with positive ENSO values (El Niño Winters) related to drier winters and warmer springs in the Northwest U.S. This should lead to earlier spring runoff in this region. Positive PDO values are also suggested to lead to drier winters and warmer springs in the Northwest U.S., which should also be associated with earlier spring. The period for PDO is thought to be around 20 to 30 years. It has also been suggested that positive values of both PDO and ENSO in the same year can have a synergistic effect. To address the potential importance of these

two variables, we incorporate yearly averages of each measure individually and together in the models as well as considering an interaction between them to assess that potential synergistic effect. Figure 3 displays both functions with estimated trend lines for easier visualization of the periodicity and trend in each measure. PDO is more correlated with year than ENSO ($r = 0.51$ vs. 0.36) and the measures are moderately correlated with each other over this time period ($r = 0.72$).

Total yearly discharge (totalQ) at a gage is a function of the precipitation in the basin. Moore *et al.* [2007] demonstrated the impact of larger total yearly discharge on delaying timing measures both theoretically and using actual measurements. The variability in timing measures were shown to be usefully explained using totalQ but visual diagnostics also suggested that this effect may be nonlinear, with upper and lower thresholds within each gage for the effects of changing amounts of totalQ. These values are transformed first by detrending the results for each gage to remove any collinearity with year. This allows the time trends estimated in the models to contain all the linear change over time. To allow for models that incorporate total discharge across the study area in single effect, the detrended total discharge is also standardized for each gage. Any further references to totalQ are to this detrended, standardized measure. These transformations had minimal impacts on the results but make the trend component in the models interpretable as the change over time even though this makes the totalQ effect more difficult to interpret. Figure 4 displays D50 over time and versus the detrended, standardized totalQ for two gage locations. The modest negative linear time trend is visible along with the indication of a nonlinear relationship between totalQ and D50.

4 TREND DETECTION METHODS

There are a wide variety of techniques that have been developed for detecting trends over time. The simplest and most commonly used involves estimating a linear trend using simple linear regression, which provides a single estimate of the amount of linear change over time and often a hypothesis test is used to assess the evidence against a slope of 0. This technique carries a wide variety of assumptions including a constant linear relationship between the measure of interest and time and errors that are independent, normally distributed and that have constant variance over time (homoskedasticity). Many different approaches are available to retain its simplicity while making the technique more robust to violations of some or all of those assumptions.

The linearity assumption can be addressed via nonparametric trends tests such as the Cox-Stuart [1955] trend test or Mann's [1945] trend test (also called the Mann-Kendall test), both tests assessing the patterns in increases or decreases and assessing how unusual those patterns might be as evidence of a trend. These methods are robust to the linearity assumption and to varying degrees to the other assumptions, although the robustness of these techniques to heteroskedasticity are less clear. These methods do not directly lead to estimates of the amount of change or the potential to consider trends adjusted for other effects. The Mann-Kendall

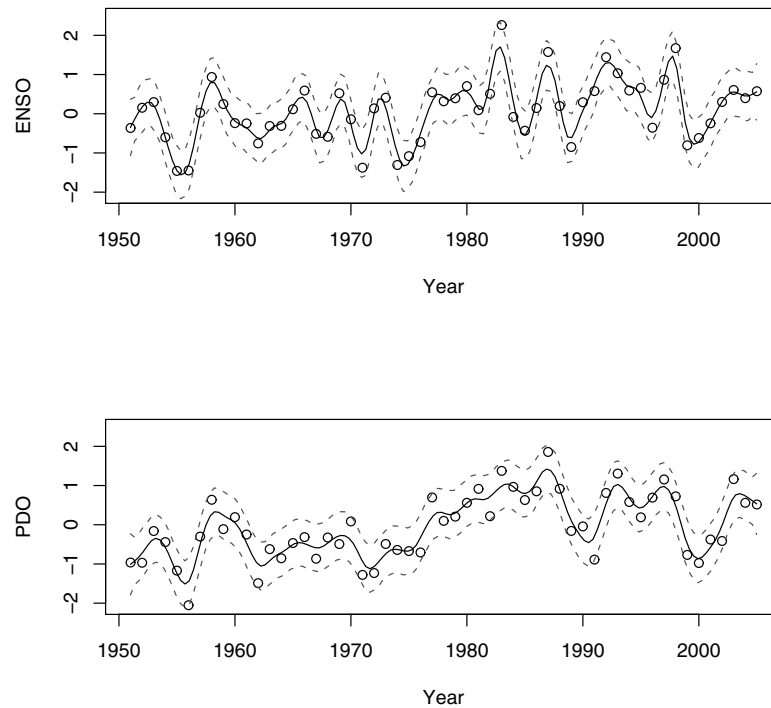


Figure 3. Plots of Water Year averaged PDO [Zhang *et al.*, 1997] and ENSO [Wolter *et al.*, 1993; 1998] effects.

procedure has been adapted to consider regional level effects based on multiple sites for change detection in Renard *et al.* [2008]. Another alternative involves a resistant slope estimate such as in Theil-Sen regression ([Theil, 1950; Sen, 1968], which provide resistance to the effect of some unusual observations. Fully robust techniques based on Huber's M-estimation [Huber, 1981] provide robustness to the normality assumption by using estimation techniques not based on least squares as in typical regression models. These robust regression models, like the Theil-Sen resistant methods, provide slope coefficients that have relatively standard interpretations, avoid least squares estimation which corresponds to the normality assumption, and can incorporate other variables in the models. Houseman [2005] has proposed a version of robust regression that incorporates an estimate of autocorrelation between responses providing a technique that is robust to violations of the independence and normality assumptions. While these techniques are robust to unusual observations, they do not directly address non-constant variance which

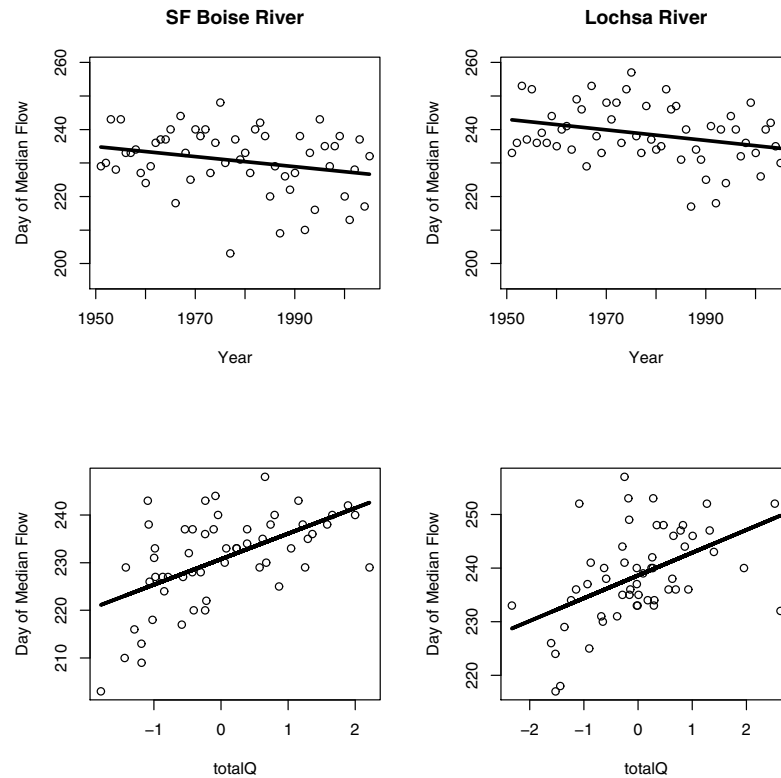


Figure 4. Linear regression effects for two gages.

may be present in climate change situations where the variability is also changing over time. Generalized least squares (described in [Pinheiro and Bates, 2003]) techniques can be used to allow the variance to change either as a function of the mean or as a function of time and also can be used to incorporate correlation between responses to potentially address violations of independence and constance variance assumptions.

Another option is to relax the linearity assumption for the trend but retain independence, normality and equal variance assumptions, at least initially. This is possible using a nonparametric trend estimate such as those possible using local polynomial regression, kernel smoothing or spline based methods. These methods assume that the trend is varying slowly over time but estimate both the level of smoothness and the trend simultaneously. These models can be compared with either models with a linear trend or no trend, with that comparison providing information about the form of the trend (nonlinear, linear, or not present). Incorporation of additional variables is most easily accomplished using what is called a

generalized additive model (GAM, [Hastie and Tibshirani, 1990]) which incorporates nonparametric components. Accommodation of certain types of violations of normality are possible using the generalized component of GAMs where response distributions other than normal are possible. While GAMs are not technically robust to outliers, if a skewed response distribution is needed, then a gamma distribution can be used. Adjustments for lack of independence and heteroskedasticity are possible using generalized least squares modifications of GAMs, leading to a generalized additive mixed models (GAMM, see [Wood, 2006] for a general discussion).

Across the different methods, there are different levels at which the analysis can be considered, the site specific analyses that amount to single time series analysis and different levels of regional aggregation, seeking an overall change on a larger scale based on aggregating that information. Much recent research has followed the detection and attribution methods as in [Berliner *et al.*, 2000] or recently in [Barnett *et al.*, 2008]. The detection part of the analysis, whether Bayesian or frequentist in perspective, amounts to assessing evidence of change, often after incorporating an adjustment for temporal and spatial dependency in the measurements. Researchers such as McCabe and Clark [2005] and Stewart *et al.* [2005] focus on site specific versions of these methods using simple linear regression or Theil-Sen versions of simple linear regression in contrast to Berliner *et al.* [2000] who perform this type of “detection” analysis on the regional scale, accounting for spatial-temporal correlations. After evidence of change has been established, then attempts are made to attribute the change to anthropogenic factors [Barnett *et al.*, 2008].

We focus on the first step in these types of analyses, using a “detection” type model, focusing on estimation of the trend adjusting for spatial-temporal correlations. We do not consider “attribution” in this work, working on the evidence for and estimation of the trends over time. To do this, three types of models are compared, a model with a nonparametric trend, one with a linear trend and one with no time trend. We also incorporate the effect of totalQ in each model to adjust for the total yearly discharge and search amongst models that incorporate effects of PDO and ENSO to see if the support for the trend remains in presence of those regional forcing functions. This provides further information on the form of the trend adjusted for other factors that are thought to be important and not directly related to warming. Whenever an effect is considered to be nonlinear in our models, a similar model constraining it to be linear is also considered, to allow support for the nonlinearity of that component to be assessed.

In contrast to other approaches in this area, our model comparison uses a different approach than methods based on hypothesis testing. This avoids some of the arbitrariness of p -value usage and some methodological problems that have yet to be resolved with hypothesis testing in additive mixed models [Wood, 2006]. Instead, we take an information theoretic, model selection approach using Akaike's Information Criterion (AIC, [Akaike, 1973]) to compare support for different potential models. The $AIC = -2 \times \log(Likelihood) + 2p$, where p is based on the degrees

of freedom for the model, is an estimator of the expected Kullback-Liebler discrepancy. It provides a method of comparing models on their predictive ability and models with smaller AIC values are considered to be better. Burnham and Anderson [2002] suggest a rule of thumb that models within 2 AIC units should be considered to be similarly supported. They note that the performance of model selection criteria like AIC improve if a set of reasonable candidate models are defined *a priori*, instead of “data-dredging” in the selection process. Incorporating models in the candidate set with different versions of the trend (none, linear, non-parametric) allows evidence to be compared for each of the forms of the trends in the presence of other variables. Our approach follows this two step process, identifying a set of candidate models and then ranking those models based on their AIC values relative to the top performing model in the candidate set. This represents a departure from the typical “detection” methods that rely on hypothesis testing, but evidential support for different models can be just as convincing.

Any statistical procedure needs to be assessed relative to the assumptions contained in it, even robust or nonparametric procedures. Our models assume normally distributed errors and the aspects of the models resemble the real system in terms of the form of the trend and other explanatory variables as well as the correlation structure. These assumptions can be evaluated using residual diagnostics, which were generally reasonable. Residual temporal autocorrelation was negligible, although it was in the original data set as well. The distribution of the residuals did not suggest mis-specification of the variables in the favored model or a particularly skewed or outlier prone response distribution. However, the tails are slightly heavier than expected based on a normal distribution. A relatively symmetric but heavier tailed error distribution does not suggest that there is a problem with the slope estimates but could lead to an underestimate of the standard error of the effects. However, the sample size is rather large with $n = 1155$, which allows the central limit theorem to provide protection on the overall model inference that assumes normality for the errors. If the residuals demonstrated an extremely skewed distribution with outliers in a single direction, the central limit theorem justification for the results might be more debatable.

5 ADDITIVE MIXED MODELS

Additive mixed models (or semi-parametric mixed models) provide a general framework for estimating potentially nonlinear regression components as well as accommodating spatial-temporal correlations. This description is not explicit as it is often meant to encompass models with random components such as random intercepts and random slopes but also can refer to models that incorporate variance-covariance structures between observations that are more complicated than the identity matrix. That second use of the terminology is employed here. We did consider random intercept and other related models initially but found much better overall performance from the correlated errors additive models discussed here.

When considering GAM or GAMM models, there are different ways to con-

struct the nonparametric, smooth components. We use penalized natural cubic regression splines, which are piecewise cubic polynomials that are patched together at “knots” with conditions requiring continuous second derivatives at the knots, linearity conditions at the edges, and estimated roughness penalties on the form of the effects. Natural splines are used to reduce edge effects by imposing linearity constraints at the edges of the effect. By using cubic regression splines, we mean that there is some initial smoothing producing via knot choice, with fewer knots than observation locations. This contrasts with “smoothing splines” where each observation location defines a knot [Ramsay and Silverman, 2005]. Initial screening was actually performed with smoothing splines but regression splines were used once a typical range of complexity was identified to improve the convergence of the estimation algorithm. The results were similar using either technique and the cubic regression splines are much faster to estimate.

In the estimation of the spline-based component, the knots define an initial restriction on the component, with k knots leading to $k-1$ degrees of freedom. Using generalized cross-validation (GCV, [Craven and Wahba, 1979]), a smoothing parameter is chosen that can further reduce the roughness in the spline component. Based on the original knots and the estimated smoothing parameter, each additive effect can be described by its effective degrees of freedom (EDF). As the EDF increases, the freedom in that spline effect increases, starting with an EDF of 1 corresponding to a linear effect. It is important to note that when the smoothing parameter is being estimated, it is possible for the estimation algorithm to choose to make the effect linear, thus a parametric entry for that effect in the model is nested in the GCV estimation process. Wood [2006] describes the multiple generalized cross validation methods used to estimate more than one nonparametric component in a GAM. This algorithm differs from the original back-fitting method developed for GAMs in Hastie and Tibshirani [1990].

An additive mixed model as we employ here can be written as $y_{it} = \alpha + x_{1it}\beta_1 + \dots + x_{kit}\beta_k + s_1(x_{(k+1)it}) + s_2(x_{(k+2)it}) + \dots + v_{it}$ for $i = 1, \dots, 21$ locations, $t = 1951, \dots, 2005$ and $\mathbf{v} \sim N(0, \sigma^2 \Lambda)$. $s()$ means that the term is estimated nonparametrically. Many different correlation structures including independence between observations can be specified in Λ . Cressie and Huang [1999] define general classes of separable and non-separable spatial-temporal error structures, which are written as a function of the distance apart in space, d , and time, t . The non-separable error structures allow an interaction between space and time, meaning that the spatial dependency can change over time (or vice-versa). One such structure is the non-separable exponential where the correlations needed to define Λ are a function of d and t as $cor(d, t) = \exp(-d/r_d - t/r_t + \lambda * d/r_d * t/r_t)$. This structure becomes separable if $\lambda=0$. Separable structures multiply the spatial correlation by the temporal correlation, providing the same spatial correlation within all years and the same temporal correlation within all locations. It is simpler to estimate than the non-separable structures which has led to the frequent use of separable structures, as in [Berliner *et al.*, 2000].

Especially when considering the separable spatial-temporal errors, there are a

variety of different candidate models for both the spatial and temporal components in the correlation as any spatial and temporal method can easily be combined. We found the linear spatial correlation structure to be most numerically stable in this application. It is one of the simplest correlation structures with a linear decrease up to the spatial range, R_d , and then correlations of 0 beyond that point, calculated as $cor(d) = [1 - (d/R_d)]I_{d < R_d}$ with I being an indicator function. For the temporal autocorrelation, we consider the commonly used first order autoregressive (AR(1)) error structure. This provides a pattern of exponentially dampening correlations as a function of years apart, ϕ^t , with ϕ being the AR(1) coefficient. This structure is also useful for identifying non-stationary processes, as $\phi = 1$ corresponds to a random walk process. Early uses of AR(1) residual temporal autocorrelation regression models are due to Cochrane and Orcutt [1949] although our estimation algorithm differs from theirs. It is interesting to note that if $\phi > 0$, the AR(1) error structure is equivalent to the exponential structure discussed above with $\phi = e^{-1/r_t}$. However, we routinely encountered negative temporal autocorrelation so this constraint would be problematic in this application.

Putting the linear spatial and AR(1) temporal correlation together as a separable structure leads to the linear-AR(1) spatial-temporal structure with $cor(d, t) = [1 - (d/R_d)]I_{d < R_d} \times \phi^t$. We use numerical maximum likelihood methods to identify estimates of R_d and ϕ to estimate Λ . While this structure is simple, it can account for strong dependencies between neighboring sites, allows spatially distant sites to have no relationship, and allows for either positive or negative temporal autocorrelation. An additional advantage of the linear correlation structure is that the correlation matrix can contain many zeros for sites that are further than R_d apart, providing a matrix that is more quickly and accurately inverted, which aids the model estimation process. The error structures we implemented were inspired by the ramps package [Yan *et al.*, 2007] in R [R Development Core Team, 2008].

Once the effects have been estimated, we want to provide estimates of the precision of the terms in the models. These results are not used in the model selection process, which is based on AIC values, but are helpful for understanding the effects and interpreting the selected models. For GAM's, we can consider frequentist or Bayesian inference for the effects. The typical frequentist variance-covariance for the model components that is used for confidence interval generation is $V_e = (X^T W X + S)^{-1} X^T W X (X^T W X + S)^{-1} \sigma^2$, where X is the design matrix, W is a weight matrix based on the inverse of the covariance matrix of the responses (can allow for correlated responses), S is the penalty matrix resulting from the bases used to define the spline components in the model, and σ^2 is the residual variance. The Bayesian posterior variance-covariance matrix can be shown to be $V_B = (X^T W X + S)^{-1} \sigma^2$ [Wood, 2006]. Once the parameters of the model have been estimated, either $\hat{V}_B$ or $\hat{V}_e$ can be calculated, with the square root of the diagonals leading to standard errors for confidence interval generation. The use of $\hat{V}_B$ has become the default in place of $\hat{V}_e$ in the mgcv package [Wood, 2008] in R. Nychka [1988] studied both types of intervals and found that the average coverage probability (a frequentist idea) is closer to nominal levels with

the Bayesian versions of these intervals than with the frequentist ones, suggesting the potential for their use in a frequentist analysis. We chose to use these Bayesian intervals and the reader can choose how they prefer to interpret them, either from a frequentist or Bayesian perspective. We use the intervals as a reference to aid in the interpretation of the estimated effects. Note that plots presented below of nonparametric effects are “partial effect” plots that show only one effect but that these plots also incorporate the uncertainty at the average of the other terms in the model, providing an interpretation of the uncertainty of the effect conditional on the other effects in the model being held constant at their means.

There are a couple of motivations for incorporating correlated errors into GAMs. One is to get more accurate standard errors for any inference that is of interest. Positive spatial or temporal correlation leads to underestimates of precision in estimates (see [Cryer and Chan, 2008] for this in the time series context). One way to think about this is that typical least squares estimates that assume independence essentially overstate the amount of information available in the calculation of the SE, with positive correlations being associated with less independent information than would have been collected with independent observations. Negative correlation actually makes the SEs conservative since it is easier to estimate the parameters, since the observations move from one side of the mean to the other frequently. Negative autocorrelation can actually look like a bit like noise since neighboring observations are not similar to one another. It is extremely dangerous to ignore strong positive autocorrelation in time series as it can also lead to spuriously significant regression coefficients as discussed in [Granger and Newbold, 1974].

An additional issue specific to using GAMs is in estimating the smoothing parameter in the presence of correlation. Wood [2006] demonstrates that additive components tend to undersmooth if positive correlation is present and not incorporated. This result would apply for spatial or temporal smoothers and correlations. In our models, the linear spatial correlation protects us from undersmoothing the spatial intercept component and the AR(1) temporal correlation structure does the same for our trend and totalQ nonparametric effects.

In the models that follow, the “magic” function from the mgcv package [Wood, 2008] is used to allow for correlated error models and multiple GCV is used for selecting the optimal amount of smoothing for each nonparametric term in the presence of a given correlation structure. A search over correlation parameters is used to find the maximum likelihood estimate of the correlation parameter along with the optimal smoothing parameters. The maximized likelihood and total EDF for the model are used to calculate the AIC to allow comparison of the different models.

An additional feature of the models considered below involves the use of bivariate smooths that essentially allow for interactions between two spline components. Bivariate smooths are only considered for a spatial intercept, with the best performing version of these with our sparse spatial grid using the tensor products of the previously described natural cubic regression splines and penalized estimation

conducted similarly to the method described above. This term is denoted $te(x,y)$ and has EDF defined for the entire component. Complete details of these methods are available in Wood [2006].

We considered a suite of candidate models using the linear-AR(1), spatial-temporal errors that incorporate a spatial intercept, a nonparametric, linear or no time trend, a linear or nonparametric totalQ term, and linear PDO and ENSO effects (and an interaction between them). The models without PDO and ENSO could be generally classified as “detection” models if the focus is on the time trend (either linear or nonparametric or none). Incorporating PDO and ENSO effects provide more insight into the system and the presence of trends after incorporating those effects presents even stronger evidence for the presence of those trends (especially that they are not just due to PDO and/or ENSO effects being detected). The spatial intercept in the models allows for different mean timing in different locations, which would account for some of the differences in the basins that are constant over time. This term could estimate similar values for neighboring sites, providing a spatially smooth surface, or a rough surface that provides effectively independent estimates for each location.

6 RESULTS

Initially, we describe the results from our base “detection” model, $D50_{it} = \alpha + te(x_i, y_i)_{20} + s(Year)_{17.05} + s(totalQ)_{5.95} + v_{it}$. This model estimates a nonparametric spatial intercept denoted by $te(x_i, y_i)$ with (x_i, y_i) being the Universal Transfer Mercator (UTM) coordinates of the gages, $s(Year)$ and $s(totalQ)$ denoting the nonparametric entry of trend and totalQ effects. The subscripts indicate the EDF used for each term, providing 20 EDF for the spatial intercept, 17.05 for the trend, and 5.95 for the totalQ effect. For this model, $\hat{R}_d = 60.3$ and $\hat{\phi} = -0.122$, indicating a relatively limited spatial dependency and a modest negative temporal dependency. The focus of this model is on the form of the trend, which is displayed in Figure 5. Also included in this figure is the linear trend from the same model with the trend constrained to linearity, $D50_{it} = \alpha + te(x_i, y_i)_{20} - 7.74 \times Year + s(totalQ)_{5.87} + v_{it}$. In that model, $\hat{R}_d = 76.5$ km's and $\hat{\phi} = -0.03$. In interpreting that model, note that Year is transformed to extend from 0 to 1, so the slope coefficient is an estimate of the change over the 55 years of the study. A 95% confidence interval for the linear change over 55 years from this model is -10.1 to -5.4 days/ 55 years. This result in itself contradicts the IPCC quote referenced above at least as it applies to our subset of gages. Providing the linear slope in Figure 5 provides insight into how to interpret the nonparametric trend. It can be viewed as a localized trend that is comparable in scale to the linear trend effect, without the linearity constraint imposed.

Comparing AIC values, found in Table 1, the evidence regarding the trend can be first observed by comparing the three models that do contain three different versions of the trend and do not contain PDO or ENSO but do contain a non-

parametric totalQ effect. The nonparametric trend model has an AIC of 8325.4, the linear trend has an AIC of 8413.5 and the model with no time trend has an AIC of 8449.3. Each of these models is clearly different on the AIC scale, with much more support for the nonparametric trend but also much more support for a linear trend than no trend. This comparison provides the support to conclude that there is clear support for a trend of some kind and that the nonparametric trend is much better than the linear version of the trend. Inspecting the trends displayed in Figure 5 leads to suspicion about the reasonableness of the assumption of a linear trend. It appears that there is some variation around a constant mean up to around 1980. Then a dramatic decrease seems to occur, followed by a slow increase. The steep drop in timing and subsequent lower values provide the impetus for the linear trend having a significant negative slope, but the nonparametric estimate seems to describe a more interesting trajectory for the change over time. Inspecting the plots of two locations in Figure 4 after viewing this plot, the dip in the late 1980s becomes more obvious and is being detected on average across the locations. It is also interesting that the estimates at the beginning and end of the study period are similar but these could be edge effects on the spline estimates but should be interpreted with some caution.

The inclusion of PDO and ENSO effects provides more information about the trend and attempts to eliminate those two regional forcing functions as the source of the detected trend. The comparison of the models with ENSO and PDO and nonparametric totalQ effects provides a useful comparison of their impacts across the models with no trend, a linear trend, and a nonparametric trend, with AICs of 8409.1, 8398.9, and 8322.0 respectively. Even when we incorporate ENSO and PDO effects we still find noticeable improvement from adding a linear and then nonparametric trend, suggesting that the detected trend is not an artifact of those regional forcing functions.

The top model contains *nonparametric* spatial intercept, trend and *totalQ* effects as well as a linear ENSO effect with an AIC of 8320.2. The next ranked model adds a PDO effect and has an AIC of 8322.0. The presence of an ENSO effect in the top model is found throughout the results when a nonparametric trend is considered, where otherwise similar models tend to rank based on inclusion of ENSO, then the ENSO and PDO, and then the model that contains their interaction. This suggests little to no support for an interaction between ENSO and PDO and that ENSO seems more important than PDO. The top AIC model was estimated to be $\widehat{D50}_{it} = 232.5 + te(x_i, y_i)_{20} - 1.4ENSO_t + s(Year)_{16.97} + s(totalQ_{it})_{5.87}$ with the estimated trend from this model in Figure 6. The same linearly constrained trend line is provided on this plot for a reference. The change in the trend estimate is negligible compared to the results in Figure 5, suggesting that while ENSO is useful in the model it is not explaining the trend over time that was previously discussed. The ENSO effect (95% CI: -2.5 to -0.3) is in the expected direction, with higher ENSO values being related to earlier timing of D50.

Exploring this model further, we can inspect the estimate of the spatial intercept component in Figure 7, where the surface is only displayed near observation

locations. This component consistently used 20 EDF across all the models, with only 21 different locations, leading to it essentially estimating a unique value for each gage. We might have expected some spatial similarity in timing which would have led to a smoother surface and the use of fewer EDF for this effect. Since spatial similarity was not observed, the spatial plot of this effect demonstrates, as much as anything else, the sparseness of the spatial locations that were used in this analysis. The distribution of estimated intercepts for the 21 locations is found in Figure 8 showing the variation in the intercept in the timing across the gages, with two gages that have much earlier timing than the rest and a spread of around 40 days for the other locations.

The effect for totalQ for the top model is found in Figure 10, which shows that the model is apparently picking up on the nonlinearity in the totalQ effects discussed above. It is reasonable to ignore the edges of the estimated effect due to sparsity of observations at those values. Focusing on the results between -1.5 and 2 for totalQ, a clear increase can be observed followed by a thresholding type of effect where similar values are predicted for totalQ values over 0.5. Again, this conforms to the results that were observed in an initial visual exploration of the data set. This suggests that slightly lower flow years than average have the most impact on the timing measures but that the effect is muted for higher flow years for each location.

The spatial range for the correlation matrix was estimated to be $\hat{R}_d = 62$ km's. At this spatial range, only 12 of the 210 pairs of locations maintained a nonzero spatial correlation. The distribution of the distances between observation locations is found in Figure 9, showing that most of the spatial correlations are being set to 0. This suggests a lack of spatial coherency in this measure, as was observed in the spatial intercept.

The estimated AR(1) coefficient was $\hat{\phi} = -0.12$ for the top model, which provides correlations effectively of 0 for observations more than 5 years apart, suggesting that the temporal dependency was weak. Additionally, the detection of negative autocorrelation corresponds to the visual assessment of these records as being “noisy”. As noted above, our precision estimates benefit from the negative temporal dependency as the series bounces back and forth around the mean providing a more precise estimate of the mean. This result for the autocorrelation is consistent with the gage-by-gage analysis performed in [Moore *et al.*, 2007].

7 CONCLUSIONS

Using AIC model selection across the additive mixed models, we found evidence that suggests there is a regional level trend and that the trend is nonlinear. The trend persists even in the presence of PDO and ENSO adjustments and when spatial-temporal correlation in the responses are accounted for. While linear trends are easier to interpret, we have found evidence that streamflow timing in the Western US has followed a more complicated trajectory over the 55 years of the study. The form of that estimated trajectory also provides a reason for the detection of

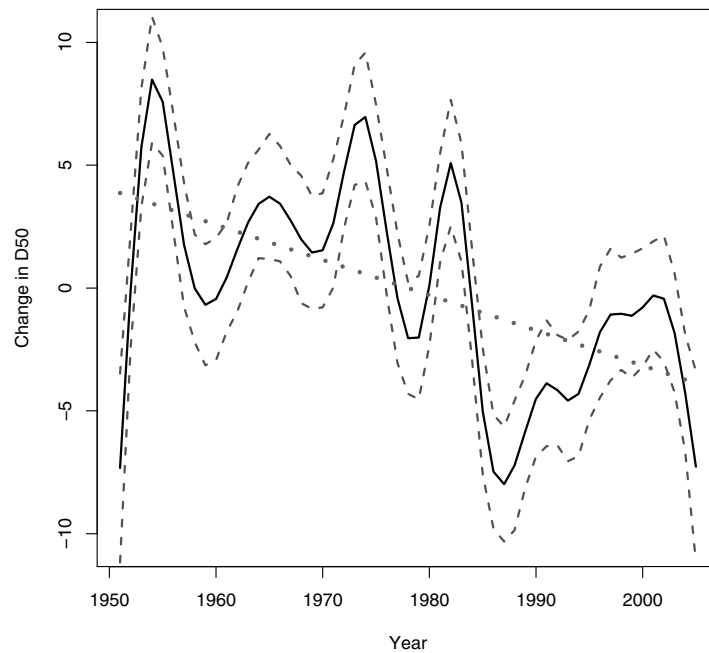


Figure 5. Estimated trend from “detection” model with 95% Bayesian CI for term+mean and linearly constrained estimate for reference.

the linear trend but one that may not be representative of the real trend.

Detection of a trend over time is only one part of the story. Detection of the trend after accounting for other possible sources of trends that are not global warming driven, such as ENSO and PDO, provides an indication that the estimated trend may be attributable to warming. We did not directly assess the relationships between temperature, precipitation and timing measures, but began the process of estimating the trend in this region in the presence of some other sources of variation in the system. Further research intends to address this more directly, this first component of the research was intended to synthesize the work in Stewart *et al.* [2005] and Moore *et al.* [2007] into a single regional level estimate of the change over time, and to assess that trend after some adjustments for known forcing functions. Further research is needed to begin the process of attribution of the change over time in a similar regional level analysis.

We have also proposed a different approach to the detection problem using a different philosophical perspective, comparing support for models that estimate the trend in different ways, including removing it completely. This perspective allows

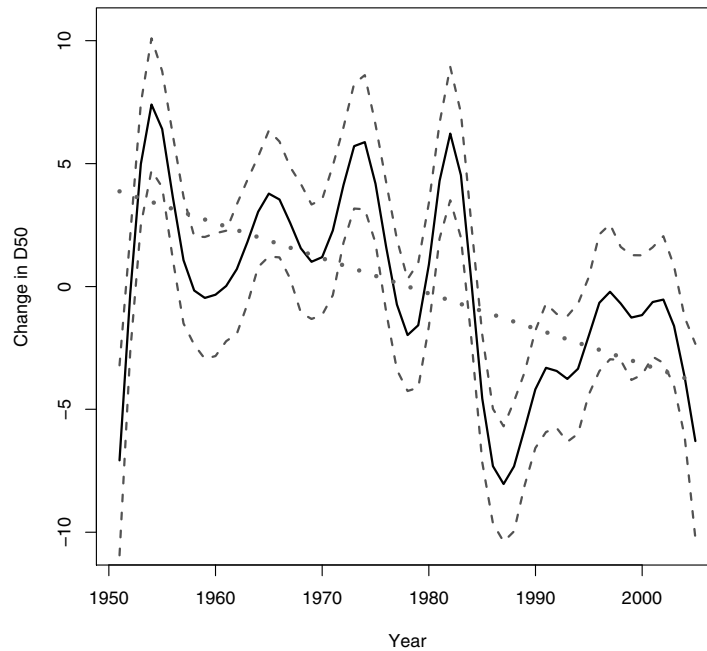


Figure 6. Estimated trend from top model that incorporates ENSO with 95% Bayesian CI for term+mean and linearly constrained estimate for reference.

for comparison of evidence for no trend, linear trend or nonlinear trends. Here we found evidence supporting nonlinear changes over time and an explanation for and criticism of the linearly constrained trend. Models that contain trends clearly outperform models that exclude trends from the models, providing evidence for change over time.

Another aspect of the models used here involved the estimation of a nonlinear relationship between total discharge on the day of median flow. We also find distinct evidence supporting the nonlinearity of this effect as compared to its linearity. This has implications for the understanding of the physical system and how the timing in basins reacts to changing precipitation levels. It suggests that the seasonality of the process controls how early or late in the year D50 can be driven, but that the total amount of precipitation is important until those thresholds are encountered.

There are issues with only using a single quantile of flow from the discharge distribution. We have chosen the day of the median flow, but other percentiles of flow could be considered, as in Moore *et al.* [2007] where two others were also

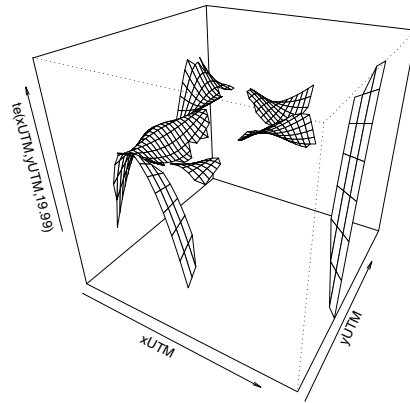


Figure 7. Plot of Spatial Intercept

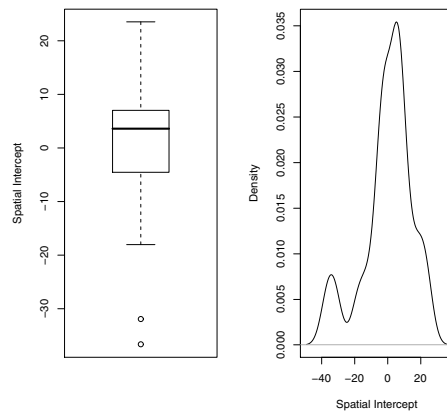


Figure 8. Intercepts from $\alpha + te(x, y) + s(Year) + s(totalQ)$

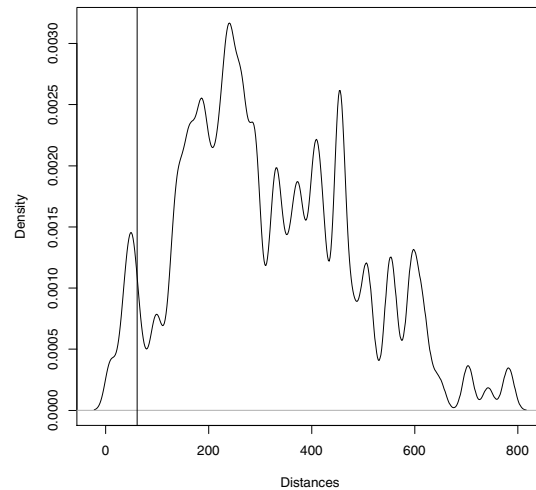


Figure 9. Distance between locations, with vertical line at $\hat{R}_d=62$ from the top model.

explored. The median is a useful measure of timing, but could be missing other changes in distribution over time. For example, the shape of the hydrographs could be changing with little impact on the day of median. Future research should attempt to model the timing of multiple flow percentiles simultaneously either using functional regression models as in Greenwood *et al.* [2007] or via a multivariate response additive mixed model, extending the methods used in this work to multiple percentiles modeled simultaneously.

Streamflow timing is really just a proxy for snowmelt, so these conclusions about changes in timing of yearly discharge are intended to describe changes in snowmelt. Of more interest would be a direct analysis of snow depth records. Due to sparseness of historical snow depth measurements, however, especially into the more distant past, an analysis of snowmelt is more difficult and perhaps no more informative about changes over time. Streamflow records are much more abundant and record a proxy for snowmelt timing, but are also recording other information. We have attempted to synthesize the trend estimates in this region for a timing measure, it remains for further research to see if this trajectory is found in other measures in the region.

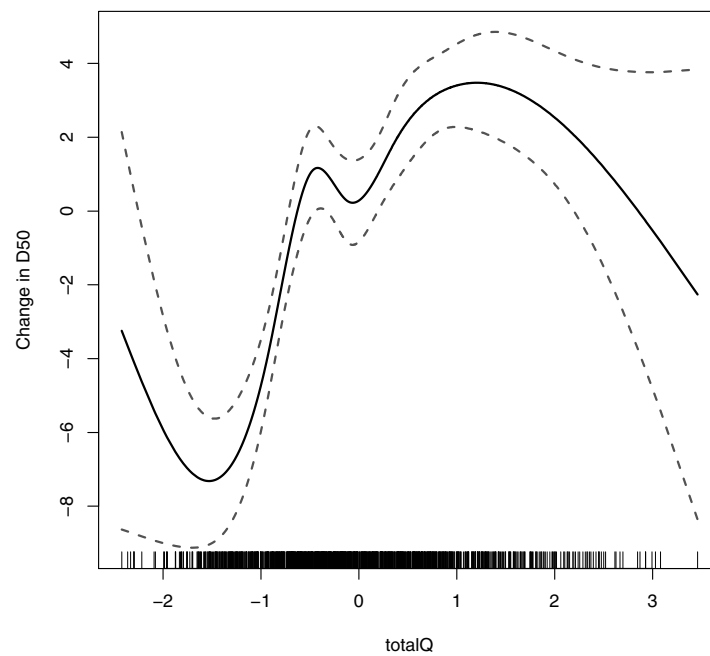


Figure 10. Estimated $s(\text{totalQ})$ from top model with 95% Bayesian CI for term+mean.

Model	Inter- cept (x,y)	Year	totalQ	PDO	ENSO	AIC	∇ AIC
Linear-AR1 Errors	P	s(Year)	s(totalQ)	NP	P	8320.2	0.0
Linear-AR1 Errors	P	s(Year)	s(totalQ)	P	P	8322.0	1.8
Linear-AR1 Errors	P	s(Year)	s(totalQ)	Int	Int	8324.2	4.0
Linear-AR1 Errors	P	s(Year)	s(totalQ)	P	NP	8324.4	4.2
Linear-AR1 Errors	P	s(Year)	s(totalQ)	NP	NP	8325.4	5.3
Linear-AR1 Errors	P	s(Year)	linear	NP	P	8363.0	42.8
Linear-AR1 Errors	P	s(Year)	linear	P	P	8364.4	44.2
Linear-AR1 Errors	P	s(Year)	linear	Int	Int	8366.2	46.0
Linear-AR1 Errors	P	s(Year)	linear	P	NP	8366.7	46.5
Linear-AR1 Errors	P	s(Year)	linear	NP	NP	8371.4	51.2
Linear-AR1 Errors	P	linear	s(totalQ)	P	P	8398.9	78.7
Linear-AR1 Errors	P	linear	s(totalQ)	P	NP	8400.1	79.9
Linear-AR1 Errors	P	linear	s(totalQ)	NP	P	8400.2	80.0
Linear-AR1 Errors	P	linear	s(totalQ)	Int	Int	8400.6	80.5
Linear-AR1 Errors	P	NP	s(totalQ)	P	P	8409.1	89.0
Linear-AR1 Errors	P	NP	s(totalQ)	P	NP	8410.3	90.1
Linear-AR1 Errors	P	NP	s(totalQ)	Int	Int	8410.7	90.5
Linear-AR1 Errors	P	linear	s(totalQ)	NP	NP	8413.5	93.3
Linear-AR1 Errors	P	NP	s(totalQ)	NP	P	8418.2	98.0
Linear-AR1 Errors	P	linear	linear	P	P	8438.0	117.8
Linear-AR1 Errors	P	linear	linear	P	NP	8438.8	118.6
Linear-AR1 Errors	P	linear	linear	Int	Int	8439.9	119.7
Linear-AR1 Errors	P	linear	linear	NP	P	8441.4	121.2
Linear-AR1 Errors	P	NP	s(totalQ)	NP	NP	8449.3	129.1
Linear-AR1 Errors	P	linear	linear	NP	NP	8457.0	136.8

Table 1. Table of Model Selection Results, NP means not present, P means present, and Int means that an interaction term between PDO and ENSO was also included in the model.

BIBLIOGRAPHY

- [Akaike, 1973] H. Akaike. Maximum likelihood identification of Gaussian autoregressive moving average models, *Biometrika*, 60, 255-265, 1973.
- [Armstrong, 2007] J. Armstrong. Significance Tests Harm Progress in Forecasting, *International Journal of Forecasting*, 23, 321-327, 2007.
- [Barnet *et al.*, 2008] T. Barnett, W. Pierce, H. Hidalgo, C. Bonfils, B. Santer, T. Das, G. Bala, A. Wood, T. Nozawa, A. Mirin, D. Cayan, and M. Dettinger. Human-Induced Changes in the Hydrology of the Western United States, *Science*, 319, 1080-1083, 2008.
- [Benjamini and Hochberg, 1995] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, B Met*, 57(1), 289-300, 1995.
- [Berliner *et al.*, 2000] L. Berliner, R. Levine, and D. Shea. Bayesian Climate Change Assessment, *Journal of Climate*, 13, 3805-3820, 2000.
- [Bonferroni, 1936] C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilit , *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3-62, 1936.
- [Burnham and Anderson, 2002] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edition, Springer, New York, 2002.
- [Cochrane and Orcutt, 1949] D. Cochrane and G. H. Orcutt. Application of least squares regression to relationships containing autocorrelated error terms, *JASA*, 44, 32-61, 1949.
- [Craven and Wahba, 1979] P. Craven and G. Wahba. Smoothing noisy data with spline functions, *Numerische Mathematik*, 31, 377-403, 1979.
- [Cressie and Huang, 1999] N. Cressie and H. Huang. Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions. *JASA*, 94(448), 1330-1340, 1999.
- [DeVeaus *et al.*, 2008] D. DeVeaus, P. Velleman, and D. Bock. *Stats: Data and Models*, Pearson, Boston, 2008.
- [Granger, 1974] C. Granger and P. Newbold. Spurious regressions in econometrics, *Journal of Econometrics*, 2(2), 111-120, 1974.
- [Greenwood *et al.*, 2007] M. Greenwood, J. Harper, and J. Moore. Functional linear models for streamflow discharge. 2007 Proceedings of the American Statistical Association, Section on Statistics and the Environment [CD-ROM], Alexandria, VA: American Statistical Association, 2384-2390, 2007.
- [Hansen *et al.*, 2006] J. Hansen, M. Sato, R. Ruedy, K. Lo, D. W. Lea, and M. Medina-Elizade. Global temperature change. *Proc. Natl. Acad. Sci.*, 103, 14288-14293, doi:10.1073/pnas.0606291103, 2006.
- [Hastie and Tibshirani, 1990] T. Hastie and R. Tibshirani. *Generalized Additive Models*, Chapman and Hall, Boca Raton, 1990.
- [Houseman, 2005] E. A. Houseman. A robust regression model for a first-order autoregressive time series with unequal spacing: application to water monitoring. *JRRS-C: Applied Statistics*, 54(4), 769-780, 2005.
- [Huber, 1981] P. Huber. *Robust Statistics*. Wiley, New York, 1981.
- [Lemke *et al.*, 2007] P. Lemke, J. Ren, R. Alley, I. Allison, J. Carrasco, G. Flato, Y. Fujii, G. Kaser, P. Mote, R. Thomas, and T. Zhang. Observations: Changes in Snow, Ice and Frozen Ground. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.
- [Mann, 1945] H. Mann. Nonparametric tests against trend. *Econometrica*, 13(3), 245-259, 1945.
- [McCabe and Clark, 2005] G. McCabe and M. Clark. Trends and Variability in Snowmelt Runoff in the Western United States, *Journal of Hydrometeorology*, 6(4), 476-482, 2005.
- [Moore *et al.*, 2007] J. N. Moore, J. T. Harper, and M. C. Greenwood. Significance of trends toward earlier snowmelt runoff, Columbia and Missouri Basin headwaters, western United States, *Geophysical Research Letters*, 34, doi:10.1029/2007GL031022, 2007.
- [Mote, 2003] P. W. Mote. Trends in snow water equivalent in the Pacific Northwest and their climatic causes, *Geophys. Res. Lett.*, 30(12), 1601, doi:10.1029/2003GL017258, 2003.

- [Nychka, 1988] D. Nychka. Bayesian ‘Confidence’ Intervals for Smoothing Splines, *JASA*, 83, 1134-1143, 1988.
- [Pinheiro and Bates, 2002] J. Pinheiro and D. Bates. *Mixed Effects Models in S and S-Plus*, Springer, New York, 2002.
- [Pinheiro *et al.*, 2008] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and the R Core team. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-89, 2008.
- [R Development Core Team, 2008] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org>, 2008.
- [Ramsay and Silverman, 2005] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis, 2nd edition*, Springer, New York, 2005.
- [Ruppert *et al.*, 2003] D. Ruppert, P. Wand, and R. Carroll. *Semiparametric Regression*. Cambridge University Press, 2003.
- [Sen, 1968] P. K. Sen. Estimates of Regression Coefficient Based on Kendall’s tau. *J. Am. Stat. Ass.* 63, 324, 1379-1389, 1968.
- [Stewart *et al.*, 2005] I. T. Stewart *et al.* Changes toward earlier streamflow timing across western North America, *Journal of Climate*, 18(8), 1136-1155, 2005.
- [Storey and Tibshirani, 2003] J. Storey and R. Tibshirani. Statistical significance for genomewide studies, *Proc. Natl. Acad. Sci.*, 2003, 100(16), 9440-9445, 2003.
- [Theil, 1950] H. Theil. A rank invariant method for linear and polynomial regression analysis. *Nederl. Akad. Wetensch. Proc. Ser. A* 53, 386-392 (Part I), 521-525 (Part II), 1397-1412 (Part III), 1950.
- [Wood, 2006] S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2006.
- [Wood, 2008] S. N. Wood. Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society (B)*, 70(3) 495-518, 2008.
- [Wolter and Timlin, 1993] K. Wolter and M. S. Timlin. Monitoring ENSO in COADS with a seasonally adjusted principal component index. *Proc. of the 17th Climate Diagnostics Workshop*, Norman, OK, NOAA/N MC/CAC, NSSL, Oklahoma Clim. Survey, CIMMS and the School of Meteor., Univ. of Oklahoma, 52-57, 1993.
- [Wolter and Timlin, 1998] K. Wolter and M. S. Timlin. Measuring the strength of ENSO events - how does 1997/98 rank? *Weather*, 53, 315-324, 1998.
- [Yan *et al.*, 2007] J. Yan, M. K. Cowles, S. Wang, and M. P. Armstrong. Parallelizing MCMC for Bayesian Spatiotemporal Geostatistical Models. *Statistics and Computing*, 17(4), 323-355, 2007.
- [Zhang *et al.*, 1997] Y. Zhang, J. M. Wallace, and D. S. Battisti. ENSO-like interdecadal variability: 1900-93. *J. Climate*, 10, 1004-1020, 1997.

Part XIV

Historical Approaches to Probability/Statistics

This page intentionally left blank

THE SUBJECTIVE AND THE OBJECTIVE

Sandy L. Zabell

The history of probability theory is one of a constant battle between two viewpoints, which we shall call briefly the subjective and the objective, although the meaning of these terms is already somewhat historically colored.

Othmar Sterzinger [1911, 1]

In philosophy it is often good policy, where one member of a putative pair falls under suspicion, to view the more innocent-seeming party suspiciously as well.

J. L. Austin [1962, 4]

This paper discusses the origins of the subjective-objective distinction in probability, and the reasons why the resulting debate over the nature of probability later took the form it did in the twentieth century. In brief, we will argue that the explicit distinction between subjective and objective arose when it did — in the middle of the 19th century — because of a tension between two distinct scientific currents. On the one hand, the work of James and Daniel Bernoulli, Condorcet, Laplace, Poisson, and Quetelet dramatically extended the applications of mathematical probability to areas of pressing social concern, and thus for the first time made the nature and scope of probability an issue of interest and importance.

But at approximately the same time, a rising tide of empiricism in European science and philosophy created an antipathy on the part of those most likely to avail themselves of these new methods towards the reigning “idealistic” Laplacian formulation of probability; the view that probability, as Laplace put it in a celebrated passage, “is relative, in part to our ignorance, and in part to our knowledge” [Laplace, 1814, p. viii]. The result was a movement to fashion an alternative approach to probability; what Venn was to term “the thorough working out of the Material view of Logic as applied to Probability” [Venn, 1888, p. x]. The debate about the nature of probability becomes much more understandable when viewed within this context; as part, that is, of a much broader debate about the philosophy and methodology of science.

But to begin, we need to understand the relatively amicable state of affairs that reigned in probability during its youthful first century.

1 THE ERA OF GOOD FEELINGS

Although the word “probability” has a wide spectrum of meanings, nearly all of these may be categorized as being either *epistemic* or *aleatory*; that is, those that view probability as a measure of belief, versus those that view it as a physical attribute of the world. These two categories may in turn be subdivided into several others. Epistemic probabilities may be thought of as *credibilities* or *logical probabilities* (Keynes, Jeffreys, Carnap, Russell), that is, unique, rational degrees of belief dictated by the information at our disposal; or they may be viewed as *subjective* and *personal* (Ramsey, de Finetti, Savage), that is, internally consistent sets of beliefs or preferences that are not, however, uniquely determined; or they may be regarded as purely *psychological*, descriptive of actual belief, but free of any normative constraint. Similarly, physical probabilities admit a number of different interpretations: they may be thought of as either *frequencies* (Venn, von Mises, Reichenbach), finite or infinite, descriptive of outcomes observed over a series of “identical” trials; or *propensities* (Popper, Mellor), dispositional properties manifested in a single trial.¹

These interpretive distinctions, however, are largely absent from the early literature of mathematical probability. The different possible usages just noted can be found in the literature of the time, but there seems to be no clear recognition of the differences between them, or their differing domains of application. They are used interchangeably, depending on context and purpose. One can find propensity in Cardano, frequency in Galileo, logical probability in Leibniz, and personal probability in Bernoulli.²

This facile sliding from one meaning to another is understandable, however, when one remembers the initial uses of the “doctrine of chances”. The primary examples and applications at that time were largely to games of chance where the elementary outcomes — be they the result of tossing a coin, choosing a card, or spinning a roulette wheel — were equiprobable, the disparate meanings effectively merging in consequence.³ If one tosses a “fair” coin, then the probability of heads is $1/2$: one says this whether one means that in the long-run there will be as many heads as tails; or that on a single toss the physical constitution of the coin equally predisposes it to come up heads or tails; or that there is no more warrant for believing it will come up heads than tails; or that it is an even bet either way. No matter what you mean by “probability,” in the classical setting of equipossible cases the numerical value of that probability will be the same.

It is only when not every meaning is applicable (for example, when trials are not repeatable) or where more than one meaning is relevant but the corresponding

¹See generally [Good, 1959; 1965, Chapter 2; Fine, 1973]. The subjective/objective distinction is sometimes used to distinguish between physical and epistemic probabilities; sometimes between credibilities and personal probabilities.

²See generally the classic treatments of Hacking [1975, especially Chapters 6, 10, 16, and 17] and Daston [1988].

³The use of the doctrine of changes in the pricing of annuities is only an apparent exception; see [Hacking, 1975, Chapter 13; Daston, 1988, pp. 27-29, 133-138].

numerical values may differ, that these distinctions acquire force or interest. Thus it is hardly surprising that the earliest treatments of mathematical probability do not explicitly distinguish between its different possible interpretations: as long as the domain of application of the doctrine of chances was limited to games of chance or other applications where a so-called “fundamental probability set” is readily identified, such subtleties were neither useful nor necessary; whatever sense of probability one had in mind, all received the same numerical assessment.⁴

2 LIBERTÉ. EGALITÉ, PROBABILITÉ

Unlike the 17th and early 18th centuries, the latter half of the 18th century saw an increasing number of attempts to apply mathematical probability to other areas such as law, politics, medicine, astronomy, and inductive inference. The result was that such basic epistemological distinctions could no longer be ignored.

2.1 *James Bernoulli’s Ars conjectandi*

The first major attempt to broaden the scope of the “doctrine of chances”, freeing it from its humble beginnings at the gambling table, was James Bernoulli’s seminal book, the *Ars conjectandi* (“The art of conjecture”), posthumously published in 1713. The first three parts of the *Ars conjectandi* deal with purely mathematical questions, but in its unfinished fourth part Bernoulli discusses the possible judicial, medical, and meteorological applications of probability. These are settings in which the resort to and invocation of equipossible cases often seems artificial and forced, but this was the only method then available to Bernoulli. Bernoulli’s celebrated limit theorem gave a method of determining unknown probabilities *a posteriori*:

⁴Daston [1988, Chapter 4] makes the interesting suggestion that “it was the rise and fall of a particular version of associationist psychology that made this characteristic blurring between subjective and objective probabilities tolerable for classical probabilists [from Bernoulli to Laplace]” (p. 189). This version originated with Locke’s comments on belief, perception, and probability in his *Essay on Human Understanding* (1689), and was further developed in Hume’s *Treatise on Human Nature* (1739), Hartley’s *Observations on Man* (1749), and Condillac’s *Traite des sensations* (1754). As Daston observes, elements of this tradition are clearly identifiable in both Condorcet and Laplace. (Note: the Truscott and Emory translation of Laplace’s *Essai* omits his long discussion of psychology on pp. cxxiii–cxxxviii.)

It seems unlikely, however, that associationist psychology can be entirely credited with this conceptual blurring. Instances of it can be identified earlier than Locke, going back to the beginnings of the subject with Pascal, and its presence remains a constant thread throughout its subsequent development. Indeed it is unclear who, prior to Buffon, Condorcet, and Laplace, were thus influenced; associationist psychology does not appear prominently in any of the other “classical probabilists” of the period such as Bernoulli, Montmort, De Moivre, and Bayes. As Daston demonstrates, there were important links between associationist psychology and the probabilists of the French Enlightenment; the influence of associationist psychology on mathematical probability, however, seems to have been largely confined to this one period and country, and its ultimate impact must be assigned a more modest role. The difficulties for the Laplacean position that arose in the 19th century were not the result of trends in psychology, but, as discussed below, the attempt to apply probability to a wider domain in an increasingly empiricist environment.

empirical frequencies express unknown underlying chances. But the justification for forcing such chances into the Procrustean bed of equipossible cases was far from clear.

These were deep questions, and Bernoulli pondered them for nearly twenty years, withholding his book from publication. When in 1703 Leibniz inquired in a letter about the still-unfinished work, Bernoulli expressed surprise that Leibniz had even heard of it, but nevertheless sent Leibniz a copy of the manuscript. In the ensuing correspondence Leibniz challenged the aptness of Bernoulli's implicit urn model, and Bernoulli rallied to its defense. The correspondence is well worth reading, even today, and it illustrates how serious attempts to apply the calculus of probability outside its initial and narrow ambit could immediately provoke serious questions about its nature and scope, questions of a kind rare when the analysis had been confined to games of chance.⁵

There were initially few attempts to further Bernoulli's program, and this aspect of the *Ars conjectandi* — as opposed to its purely mathematical part — remained largely unexplored. This may have been partly because of its clearly fragmentary and incomplete nature, but undoubtedly also because it made no serious attempt to grapple with a specific, concrete problem.

2.2 Daniel Bernoulli vs. Jean d'Alembert

This could not be said of Daniel Bernoulli's analysis of the efficacy and utility of smallpox vaccination, nearly half a century later [Bernoulli, 1760], which immediately attracted the fire of the distinguished French *philosophe* and mathematician Jean Le Rond D'Alembert. Here was a problem of unquestioned importance, one on which reasonable men could (and did) differ, one for which the calculus of probabilities was being invoked to give a definitive answer.

The dispute between D'Alembert and Bernoulli has two-fold significance. First, it again illustrates how readily controversy could arise as soon as the applications of probability shifted from the merely recreational, hypothetical, or mathematical to the domain of science and public policy. And second, the confused nature of D'Alembert's criticisms was only possible because of the lurking epistemological ambiguity in the concept of probability. (For example, D'Alembert argued that in tossing a coin, the chances of seeing at least one head in two tosses were 2:1; and that the probability of heads increased each time a head comes up instead of a tail — or in the coinage of his day, *croix* vs. *pile*). D'Alembert's critics were united in declaring him to be wrong, but it is revealing that none of them were able to state in any clear way why.⁶

⁵For a summary and discussion of the correspondence, see [Hacking, 1975, pp. 163-165; Daston, 1988, pp. 237-241].

⁶Todhunter [1865, pp. 224-228, 265-286, 423-426, 601-602] summarizes the controversy. D'Alembert's criticisms were part of a more general attack on the mathematical theory of probability; see [Baker, 1975, pp. 171-180; Pearson, 1978, pp. 535-573; Daston, 1988, pp. 76-91]. Daston [1979] and Swijtink [1986] are interesting but unconvincing attempts at rehabilitation.

2.3 *Voltaire, Condorcet, and Laplace*

Despite D'Alembert's criticisms, enthusiasm in France for the social uses of probability grew. In part this reflected the interests of the *philosophes* such as Voltaire and Turgot in the reform of the French legal system. In the aftermath of the celebrated Calas case, Voltaire wrote his "Essai sur les probabilités en fait de justice", with its explicit avowal of judicial probabilism; and later, in 1773, Voltaire sought Condorcet's opinion in another case because it "should only be judged by philosophers who know how to weigh probabilities" [Baker, 1975, p. 235].⁷ Condorcet's initial interest in the legal applications of probability can be traced in part to Voltaire's encouragement, as well as that of the French minister Turgot [Gillispie, 1972; Baker, 1975, pp. 231-235; Daston, 1988, Chapter 6].

In the hands of Condorcet and Laplace an elaborate mathematical theory of judicial decision-making was developed. During Laplace's lifetime criticism of this work was relatively muted, but after his death the response was not so restrained. When the mathematician Siméon-Denis Poisson (1781-1840) published his *Recherches sur la probabilité des jugements* in 1837, harsh criticism was voiced.

2.4 *Poisson*

Although Poisson only wrote the occasional paper on probability prior to the 1830s, in the last decade of his life he became interested in the problems of legal decision-making addressed earlier by Condorcet and Laplace. Poisson presented a summary of his results at two sessions of the Académie des Sciences on December 14, 1835 and April 11, 1836, and was attacked after his second presentation by the mathematicians Louis Poinsot and Charles Dupin. While Dupin's criticisms were primarily technical, Poinsot's were more fundamental in nature, questioning the whole enterprise as an inappropriate application of mathematics, it being absurd "to represent by a number the truth of a witness; to thus assimilate men to so many dice, each of which has several sides, some for error, others for truth."⁸

Poinsot was careful, however, not to criticize the legitimate applications of the calculus of probabilities, ones where it was possible to identify a universe of equally likely outcomes. This careful distinction was not always made; a case in point is the distinguished French sociologist Auguste Comte (1798-1857). Comte was a harsh critic of probability, most notably in his *Cours de philosophie positive* (six volumes, 1830-1842).⁹ The philosophical conception on which it rested, he believed, was "radically false and susceptible of leading to the most absurd consequences" (Lesson 27). It was a "vain pretension of many mathematicians to render social

⁷Calas was a French Huguenot wrongfully executed for the supposed murder of his son; see [Gay, 1959, Chapter 6; 1969, pp. 433-437; Baker, 1975, pp. 233-234].

⁸Quoted by Daston [1988, p. 368]. The criticism was later voiced by many others.

⁹The primary discussions appear in Lessons 27, 40, and 49 of the *Cours de philosophie positive*: see [Baker, 1975, pp. 377-382; Porter, 1986, pp. 41, 155; Stigler, 1986, pp. 194-195]. As I. Bernard Cohen notes, at least one English translation of Comte sanitizes his intemperate language.

studies positive by a chimerical subordination to the illusory mathematical theory of chances”; such a “grossiere” might be excusable in the illustrious James Bernoulli, who lived in times unenlightened by the positivist sun, but it was far less so in Condorcet, and truly inexcusable that Laplace should reproduce such a “philosophical aberration” (Lesson 49).

In all this Comte was only the most extreme representative of a growing tradition of criticism and scepticism in France, from Jean-Francois La Harpe in the 18th century, to A. L. C. Destutt de Tracy, Pierre Paul Royer-Collard, Risueno d’Amador, Victor Cousin, and Charles Gouraud in the 19th.¹⁰

3 ANTOINE AUGUSTIN COURNOT

Nothing is more important than to carefully distinguish the double meaning of the term probability, sometimes taken in an objective sense, and sometimes in a subjective sense, if one wishes to avoid confusion and error, as much in the exposition of the theory as in the applications to which it is put.

Antoine Augustin Cournot [1843, p. 438]

It is apparent that the peaceful if somewhat unambitious existence that probability had enjoyed at the beginning of the 18th century was now a thing of the past. Attempts to apply the calculus of probabilities had become widespread, if controversial, thanks to the work of Laplace, which gave such efforts a cachet they had not previously enjoyed. It is precisely at this point that the first careful attempts can be found to distinguish the various senses of probability being employed.

Of course some distinctions had been made earlier. David Hume, in his *Treatise*, wrote that probability could be divided into two kinds, “that which is founded on *chance*, and that which arises from *cases*” [Hume, 1739, pp. 124-125]. By the former Hume meant probabilities calculated using equipossible cases; by the later, probabilities estimated from empirical frequencies.¹¹ Condorcet, looking at the matter somewhat differently, noted that the word probability is used in two senses: one mathematical (“sens abstrait du mot probabilité”), the ratio of the number of favorable cases to all cases; another epistemic, the “motif de croire”.¹² Laplace, although his view of probability is usually uncompromisingly epistemic, occasionally (when discussing what might today be described as the subjective probability of a propensity) refers to the “true possibility” of an event [Hacking, 1975, p. 133].

The distinction between the subjective and objective senses of probability is first made explicit in Poisson’s *Recherches sur la probabilité des jugements*. Here,

¹⁰See generally [Porter, 1986; Daston, 1988, pp. 364-369, 376-378; Schneider, 1988, pp. 195-200].

¹¹Hume’s views on probability are discussed at length by Stove [1973].

¹²See [Rashed, 1974, pp. 58-63; Hacking, 1975, pp. 130-131].

apparently for the first time, the difference is viewed as sufficiently important to merit using different words to distinguish between the different senses:

In this work, the word chance will refer to events in themselves, independent of our knowledge of them, and we will retain the word probability ... for the reason we have to believe. [Poisson, 1837, p. 31]

This contrast was implicit in Laplace's occasional use in tandem of the terms "probabilité" and "possibilité"; making it explicit was certainly a step forward. Unfortunately, Poisson did not carry the idea much further; as Stigler notes [1986, p. 187], "here as elsewhere in Poisson's work the conceptual distinctions most modern readers ... look for were missing." This could not be said, however, of Poisson's younger contemporary, Antoine Augustin Cournot (1801-1877).

Cournot must rank as one of the most underappreciated figures in the history of probability. He is best known today for his contributions to the early development of mathematical economics, and in France is still highly regarded for his contributions to philosophy. His major contribution to the theory of probability is his *Exposition de la theorie des chances et des probabilités* [Cournot, 1843], although he also discussed the philosophical foundations of the subject in several later works [Cournot, 1851; 1861; 1875]. In contrast to this multi-faceted and prolific output, the external details of Cournot's life were relatively quiet: educated at the École Normale Supérieure, Cournot spent most of his life in university administration and died in Paris, where a plaque to his memory may be found today near the Hotel de l'Odeon on the Rue St. Sulpice.

Let us now turn to Cournot's *Exposition*.¹³ From a modern perspective it is very impressive. As we shall see, it introduces the subjective/objective terminology into mathematical probability; it does this in the course of developing the first serious frequentist theory of probability. This leads in turn to the recognition of the confidence interval property for interval estimates of a binomial proportion. The watchword throughout Cournot's book is clarity — clarity of thought, terminology, and application. The mathematical details of the book are equally impressive.¹⁴

In his *Préface*, Cournot states two aims for his book: one, to make accessible the rules of the calculus of probability to those unacquainted with higher mathematics; the other, to "rectify the errors, clarify the ambiguities, dispel the obscurities from which, it appears to me, the works of the most able mathematicians on this delicate subject are not exempt." Probability has "a double sense which

¹³Although there are comments scattered throughout the *Exposition*, Cournot's principal comments occur in his *Préface*, and Chapter 2 ("Des chances et de la probabilité mathématique"), Chapter 4 ("Du hasard, de la possibilité et de l'impossibilité physiques", esp. p. 82), Chapter 8 ("Des probabilités *a posteriori*", esp. pp. 169 and 180), and Chapter 17 ("De la probabilité de nos connaissances, et des jugements fondés sur la probabilité philosophique").

¹⁴Among the many topics discussed are the relationship between simple and stratified random samples (pp. 135-152); the comparison of sample mean and sample median to test normality (pp. 227-228); the accuracy of the Poisson approximation (pp. 331-332); and Bode's law as an illustration of inductive inference and simplicity (pp. 430-434, this was a topic of controversy in the *Journal of the American Statistical Association* in 1971).

sometimes refers to a certain measure of our knowledge, and sometimes to a measure of the possibility of things, independently of the knowledge which we have." Such a clarification was not of mere academic interest: the resulting distinctions would permit the resolution of "difficulties which have up to the present rendered suspect to thoughtful persons ("bons esprits") the entire theory of mathematical probability" (p. 162). What was this double sense?

Fundamentally, the theory of chance and mathematical probability applies to two kinds of questions that are quite distinct: to questions of *possibility* which have an objective existence, as has been explained, and to questions of *probability* which are relative, in part to our knowledge, in part to our ignorance. (p. 84)

As usual, "the ratio of the number of chances favorable to the event, to the total number of chances," is invoked; this will be called the mathematical probability, or simply, the probability of the event.¹⁵ When one considers a large number of trials of the same "hazard," it is physically impossible for the sample frequency to differ sensibly from the mathematical probability of the event; "in this sense, the mathematical probability can be considered as measuring the *possibility* of the event, or the facility with which it is produced" (p. 438).¹⁶

These ideas were not unrelated to those of Poisson, and so Cournot added in a footnote the following comment:

One will pardon me for transcribing here, not so much in order to conserve my rights to the priority of certain ideas, than as a testimony of a friendship of which the memory to me will always be precious, and also as an account of some opinions of a celebrated man, relevant to the subject with which we are occupied, the letter which M. Poisson wrote to me, in response to another in which I had submitted to him the summary of the book which I publish at this moment:

Paris, 26 January 1836

¹⁵ "If it is necessary to give this ratio a name which frees us from constantly giving its definition: one calls it mathematical probability, or simply the probability of the event" (p. 24).

¹⁶ Daston [1988, p. 223], citing Cournot's statement that "mathematical probability becomes the measure of physical possibility" (*Exposition*, p. 81), argues that "Cournot asserted in his *Exposition* ... mathematical probability was not a measure of belief," and that "Cournot's work marks the advent of a new interpretation of mathematical probability exclusively in terms of objective frequencies" (p. 224). (Compare Jorland's statement, 1987, p. 174, that "Cournot annulled this difference [between probability and chance], both being objective.") The actual state of affairs is somewhat more complex. Although Cournot holds that mathematical probability often has an objective sense, he does not claim that this is always so; in his system the objective sense of probability does not replace the subjective one, but coexists with it. Indeed one of the points Cournot stresses is that while probability has a dual acceptance — both legitimate — the failure to recognize this can cause serious confusion. It is, however, certainly true that Cournot regards the objective sense as being the more important of the two.

Monsieur,

It will be with great pleasure that I will read the work which you propose to publish on the doctrine of chances. That which I have actually completed will not be an obstacle, and I will willingly yield the center-stage to a more complete book. I place between the words *chance* and *probability* the same difference as you, and I very much insisted on this difference. . . .

Poisson.

(Cournot's note is an amusing example of claiming one thing while clearly intending quite another. His disclaimer notwithstanding, Cournot — Poisson's junior by twenty years — is obviously concerned here with establishing, if not priority, at least independent credit for making a clear distinction between the subjective and objective senses of probability.)

Both Poisson and Cournot were familiar with the relevant mathematical literature of their day, and Cournot was also widely read in the relevant philosophical literature as well. That both considered the distinction they made novel and important strongly suggests either that it did not previously exist, or at least was not widely appreciated at the time.¹⁷

4 THE INFLUENCE OF KANT

In making the subjective/objective distinction, Cournot was drawing on the philosophical literature, as he himself tells us: "It is however from the language of the metaphysicians that I have borrowed without scruple the two epithets of the objective and the subjective, which appeared to me necessary in order to distinguish radically the two meanings of the term probability which are applied to the combinations of the calculus; but I was authorized here by the example of Jacques Bernoulli."¹⁸

The subjective-objective terminology in philosophy goes back to the medieval scholastics, but the words were used in a way quite different from their modern meanings. When Descartes speaks of "objective reality", what he has in mind would today be described as "subjective" (see, e.g., the discussion of the Third Meditation in [Nadler, 2006]). The shift in meaning to the modern sense seems to

¹⁷For further discussion of Poisson and Cournot, see [Daston, 1994, pp. 335-336].

¹⁸It would be interesting to identify which "metaphysicians" Cournot had in mind. The passage from Bernoulli alluded to is from the opening lines of Part 4 of the *Ars conjectandi*:

The certainty of any event is observed either *objectively* and in itself, and it does not signify anything other than the very truth of the existence or the future existence of that event; or the certainty is observed *subjectively* and according to ourselves, and it lies in the measure of our knowledge about this truth of present or future existence.

All things which exist or are acted upon under the sun — past, present, or future things — always have the greatest certainty in themselves and objectively.

have taken place in the 18th century. Hegel credits its modern use and meaning to Kant (see, e.g., [Priest, 1987]) and although it can be found in earlier philosophers, it is certainly true that it was through Kant that it gained currency and widespread use.

It is very likely that Cournot inherited the distinction directly from Kant. In the final chapter of his *Essai sur les fondements de nos connaissances* (1851), entitled “Examen de quelques systemes philosophiques,” Cournot discusses the systems of Plato, Aristotle, Bacon, Descartes, and Kant from the viewpoint expressed in his own book; and numerous other references to Kant occur throughout the *Essai*. Cournot has great respect for Kant; if Leibniz was the German Plato, then Kant must surely be its Aristotle (p. 468). A brief passage from Kant’s *Logic* on induction and analogy is quoted, as well as the much better known *Critique of Pure Reason* (Preface, introduction, and sections 1-3).

Kant had attempted to steer a middle course between the extreme rationalism of Leibniz, and the extreme skepticism of Hume; one claimed in effect that all knowledge is objective, the other that all knowledge is subjective. The first claim seemed indefensible; the second, unsatisfactory. By delineating the boundaries of the subjective and the objective with sufficient care, Kant had hoped that a credible case could be made for the existence of objective knowledge.¹⁹ In many ways Cournot was now attempting a similar program in probability. If the applications, utility, and even foundations of the subject were being questioned, it was because the boundaries of its legitimate domain of application were being overstepped.²⁰ Delineate those boundaries, separate the often questionable subjective from the acceptable objective uses and controversy would wither away.

Cournot’s debt to Kant went even further. Both in his 1843 *Exposition* and his 1851 *Essai* Cournot argues for the existence of philosophical probability: a species of probability separate from the mathematical kind and thus not reducible to a calculation of chances, dealing with the simplicity of the laws of nature and the sequence of phenomena. The idea effectively appears to originate with Kant in his *Logic* of 1800.²¹ As von Wright notes [1957, p. 217, n. 9], the idea was a common one in the 19th century, but many of the later references to it can be with some confidence traced to this source.²²

¹⁹Of course the distinction between rationalism and empiricism is very different from that between objectivity and subjectivity, but the two are closely linked. For discussion of Kant’s theory of objectivity, in terms of its relationship to theories of probability and indeterminism, see [Brittan, 1994; 2009].

²⁰“If one wishes to inject into the discussion of natural phenomena or social facts the consequences of similar judgments, one risks falling into errors of a kind to be illustrated later, undermining the confidence due legitimate applications.” (*Exposition*, p. 86).

²¹Strong [1976, p. 207, n. 5] notes its earlier appearance in K. H. Frömmichen’s obscure *Über die Lehre des Wahrscheinlichen* (1773).

²²Thus, of six authors listed by von Wright — Fries, Beneke, Drobisch, Cournot, Apelt, and Grelling — Fries was a follower of Kant, Apelt was Fries’s student, Grelling’s contribution appeared in the *Abhandlungen der Fries’schen Schule* (founded by Apelt), Beneke was with Fries the leading advocate in Germany at that time of “psychoiogism”, Drobisch was a fellow “philosopher-psychologist,” and Cournot’s debt to Kant has already been discussed. This is not to suggest that others outside this circle did not adopt the idea, but simply to note that many

5 1842 AND 1843: THE ANNI MIRABILI

Poisson and Cournot had independently come to the conclusion that it was important to distinguish *chance* and *probability*, Cournot by proposing a frequentist theory of probability. Curiously, at almost the same time three other persons proposed very similar theories: Jakob Friedrich Fries in Germany, Richard Leslie Ellis and John Stuart Mill in England.

Fries was a German Neo-Kantian philosopher, and as such his work lies apart from the French and English currents explored here.²³ Fries has the distinction of being one of the earliest writers to join Kant in distinguishing “philosophical probability” from its numerical counterpart (in a youthful work of 1804), and his *Versuch* of 1843, a counterpoint written at the end of his life, is a work of considerable interest.²⁴ But even within Germany, Fries lay outside the philosophical mainstream, a critic of the more fashionable Fichte, Schelling, and Hegel, and his name today is most frequently associated instead with “psychologism.”²⁵

Ellis likewise was a solitary figure in England. Keynes says of his papers on probability that they are “full of spirit and originality, and not now so well known as they deserve to be” [Keynes, 1921, p. 93].²⁶ There is considerable justice in this; it is certainly true that time has been unkind to Ellis: although Venn’s later (and much better known) *Logic of Chance* may fairly be viewed as an elaboration of Ellis’s original position, it contains only a lone reference to Ellis’s work (and that a paper other than the one referred to earlier).²⁷

Such obscurity can scarcely be said however to attach to John Stuart Mill, one of the most famous philosophers of the 19th century. But widely read as Mill’s *System of Logic* was (and is), his impact on the foundations of probability was sharply limited; in part this was because of deficiencies in Mill’s understanding of the subject, in part because his pronouncements on the subject vacillated from

of the names often cited are much more closely linked than might be realized.

²³There appears to have been remarkably little interaction between English and German contributions to the foundations of probability and statistics during this period: the English ignored Continental developments in statistics [Keynes, 1921, p. 394; Borel, 1924, p. 59]; and the Germans ignored the English [Kamlah, 1987, p. 101]. For general information on Fries see the entry in the *Encyclopedia of Philosophy*. For Fries’s critique of induction and probability, see [Porter, 1986, pp. 85-86; Kruger 1987, pp. 67-68].

²⁴Kamlah states that “with the exception of Fries, Central European philosophers [prior to the second half of the 19th century] made no effort to analyze the concept of probability” [Kamlah, 1987, p. 100]. The contributions of Bolzano would appear to contradict this, but it is certainly true that the subject did not receive much attention.

²⁵The present-day philosophical journal *Ratio* provides a link between Fries and the present, being a direct continuation of the *Abhandlungen der Fries’schen Schule*, founded by Fries’s student Apelt in 1847.

²⁶Keynes gives his fellow countryman Ellis priority for the frequency theory on the somewhat dubious grounds that his paper was read before the Cambridge Philosophical Society in February, 1842, although it was not published in the *Transactions* of the CPS until two years later. (Volume 8, Part 1, pp. 1-6. Because the remainder of the volume did not appear until 1849, this later year is sometimes erroneously given as the year of publication.)

²⁷For further discussion of Ellis, see [Porter, 1986, pp. 78-81; Kruger, 1987, pp. 68-70; Daston, 1994, pp. 337-338; Zabell, 2005, pp. 120-121]

edition to edition of the *Logic*.²⁸

6 DE MORGAN VS. VENN: “AND THERE HAS NOT YET BEEN PHILOSOPHY ENOUGH TO EXPEL IT ...”

After the first edition of Mill’s *Logic* appeared, the distinguished English scientist Sir John Frederick William Herschel corresponded with Mill, criticizing his position; as a result, Mill withdrew many of his most fundamental criticisms of the Laplacean position from later editions of the *Logic* (see [Strong, 1978]). Ellis, in contrast, remained true to his position and later wrote further on the matter, but his contributions remained few and brief. Nevertheless, Mill and Ellis presaged a debate soon to take place in England. After their attacks, De Morgan and others felt moved to defend the subjective nature of probability, and this in turn provoked criticism by Boole, Venn, and others. The focus here is not so much this debate itself, as the clear emergence of the subjective-objective distinction in it.

6.1 The “Subjectivists”: De Morgan, Herschel, and Donkin

The most important of the subjectivists was Augustus De Morgan, a staunch Laplacean who had already written both a popular introduction to probability [De Morgan, 1838] and a lengthy encyclopedia article that was largely a technical synopsis of much of Laplace’s *Théorie analytique* [De Morgan, 1845]. Although there is much of foundational interest in both of these works, it is De Morgan’s *Formal Logic* of 1847 that sets out his views on the nature of probability in the most detail. This was uncompromisingly epistemic:

By degree of probability we really mean, or ought to mean, degree of belief. It is true that we may, if we like, divide probability into ideal and objective, and that we must do so, in order to represent common language.

... [O]ur knowledge is *ideal*, as distinguished from *objective*. And it will be found that, frame what circumstances we may, we cannot invent a case of purely objective probability. [De Morgan, 1847, p. 172]

De Morgan’s reason for asserting this was that in the simple examples given to illustrate the existence of objective probabilities, such as an urn containing white balls and ten black, there were always implicit assumptions being made:

I put ten white balls and ten black ones into an urn, and lock the door of the room. I may feel well assured that, when I unlock the room

²⁸Keynes notes Mill suffered from a “complete failure to grasp with any kind of thoroughness the nature and importance of the theory of probability. The treatment of this topic in the *System* is exceedingly bad. His understanding of the subject was, indeed, markedly inferior to the best thought of his own times” [Keynes, 1921, p. 268]. For a more recent appraisal, see [Daston, 1994, pp. 336–337].

again, and draw a ball, I am justified in saying it is an even chance that it will be a white one. If all the metaphysicians who ever wrote on probability were to witness the trial, they would, each in his own sense and manner, hold me right in my assertion. But how many things there are to be taken for granted! Do my eyes still distinguish colours as before? Some persons never do, and eyes alter with age. Has the black paint melted, and blackened the white balls? Has any one else possessed a key of the room, or got in at the window, and changed the balls? We may be very sure, as those words are commonly used, that none of these things have happened, and it may turn out (and I have no doubt will do so, if the reader try the circumstances) that the ten white and ten black balls will be found, as distinguishable as ever, and unchanged. But for all that, there is much to be assumed in reckoning upon such a result, which is not so objective (in the sense in which I have used the word) as the knowledge of what the balls were when they were put into the urn. [De Morgan, 1847, pp. 172-173]

As a result, De Morgan concluded,

... I now premise that I throw away objective *probability* altogether, and consider the word as meaning the state of the mind with respect to an assertion, a coming event, or any other matter on which absolute knowledge does not exist.

This was a fairly forthright statement of the subjectivist position. Particularly revealing of De Morgan's thoughts on the matter is a letter from De Morgan to Herschel in 1849, quoted by Daston [1988, p. 374, n. 11]: Herschel, De Morgan thought, was

apt to escape from the subjective into the objective in viewing a question in probabilities. When you look at my judge — the ego of the problem — you are no ways bound by his opinion of himself — To you he is what he is in the mind of another ego. My question is not whether he is right in his preconception — but what the theory of probabilities binds him to if he has — and cannot help having — preconceptions.

This is a remarkable comment, perhaps the clearest statement of the subjectivist position prior to the 20th century. (Nevertheless, as will be discussed somewhat later, even De Morgan betrays the occasional objectivist sentiment.)²⁹

Herschel was himself a forceful advocate of the Laplacean position. In the 1849 edition of his *Outlines of Astronomy* he had outlined Michell's argument that stars that are optically double are also physically so. Immediately after the Scottish scientist James David Forbes wrote a brief note, criticizing Michell's

²⁹There appears to be no reference to the terms subjective and objective in either De Morgan's *Essay on Probabilities* (1838) or his *Theory of Probabilities* (1845); it would be interesting to trace De Morgan's use of the new terms in his *Formal Logic* and his letter to Herschel.

analysis on several grounds; and Herschel, who shortly after wrote a lengthy review of Quetelet's *Letters on Probability* in the *Edinburgh Review*, took the occasion to reply to Forbes's criticisms. Within a short period Forbes, the logician George Boole, Bishop Charles Terrot of Edinburgh, and William Fishburn Donkin, the Savilian Professor of Astronomy at Oxford, all joined in the fray.³⁰

The resulting debate was not a model of clarity. But during its course Forbes wrote to several colleagues soliciting their opinion about the issues that had arisen. Such correspondence is often more revealing than the judicious language of a published paper. Ellis, not surprisingly, was sympathetic:

I am greatly pleased with your argument in general; it expresses what I have long thought. The foundation of all the confusion is the notion that the numerical expression of a chance expresses the force of expectation, whereas it only expresses the proportion of frequency with which such and such an event occurs on the long run. From this notion that chances express something mental or subjective, is derived the assumption that the force of belief touching past events admits of numerical evaluation as well as the force of expectation touching future. If this were true, it would be a legitimate inquiry to try to assign numerical values to the force of belief in any given case. All this folly, for one cannot give it any other name, grows out of such statements as "certainty is equal to unity," and the like. It belongs to the school of Condillac and the sensationalists — they were in the ascendant when the theory of probabilities *received its present form, and there has not yet been* philosophy enough to expel it [Shairst et al., 1873, p. 481; emphasis added.]

"Its present form" — this was certainly a concession by Ellis that the frequentist view he favored was still very much a minority position in 1850.³¹ Thus Donkin could confidently state: "It will, I suppose, be generally admitted, and has often been more or less explicitly stated, that the subject-matter of calculation in the mathematical theory of Probabilities is *quantity of belief*."

6.2 The "Objectivists": Ellis, Venn, and Peirce

Within 15 years of Donkin's dismissal of Ellis's position, however, a truly formidable opponent of the subjective position would arise in the person of John Venn, the author of the *Logic of Chance*.³²

³⁰The debate is summarized in [Gower, 1982], and briefly discussed by Porter [1986, p. 79].

³¹One other of Forbes's correspondents, however, also espoused the frequentist position. Sir George Airy, commenting on a draft of Forbes's reply to Herschel, remarked on a passage: "This is the only place in which you have adverted to the long run of many trials as entering into a chance-problem: and I think the want of more energetic reference to it is a defect in your paper. I think that that consideration is the foundation of all calculations of probabilities" [Forbes, 1873, p. 476].

³²First published in 1866 (2nd ed. 1876, 3rd ed. 1888). All quotations are from the third edition. Although Venn made major changes in each of the later editions, I have not attempted

Venn's *Logic of Chance* must rank as the first systematic philosophical exploration of a frequentist theory of probability [Salmon, 1980]. What is of interest to us here is not Venn's contribution to the frequentist theory, impressive as it is, but how he viewed his contribution, and the context in which he saw it. Looking back, Venn thought virtually the only person "to have expressed a just view of the nature and foundation of the rules of Probability" was Mill, in his *System of Logic*: the only originality claimed for the "essay" Venn now put before the public was its "thorough working out of the Material view of Logic as applied to Probability."

With what may be called the Material view of Logic as opposed to the Formal or Conceptualist, with that which regards it as taking cognisance of laws of things and not of the laws of our own minds in thinking about things, I am in entire accordance. [Venn, 1888, p. x]

The "material" vs. "formal" view of logic; here was a theory of probability that arose out of a broader philosophical orientation reminiscent of Cournot's remark that in using the terms subjective and objective he was employing "la langue technique de l'école" (i. e., technical academic jargon).³³ Ellis too had viewed matters in this light:

But in what relation, it may be asked, do these conceptions stand to outward realities? How can they be made the foundation of a real science, that is of a science relating to things as they really exist? We are by such questions led back to what was long the great controversy of philosophy; I mean the contest between the realists and the nominalists. ... To assert therefore that the theory of probabilities has for its foundation a statement touching genera and their species, and is at the same time a real science, it to take a realistic view of its nature. [Ellis, 1854]

Closely allied in position to Ellis and Venn was the American philosopher Charles Sanders Peirce, although his position defies ready summarization; see [Porter, 1986, pp. 219-227].

Thus in England the issue had been squarely joined in the two decades from 1843 to 1866. What was happening in France?

7 BERTRAND: OBJECTIVITY AS CONSENSUS

In France, matters moved more slowly, perhaps because of the prestige which Laplace's name lent to his position (see generally [Schneider, 1987]). Between 1843, the year of publication of Cournot's *Exposition*, and 1889, the year of publication of

to trace these as they relate to the subjective/objective issue.

³³Actually more than reminiscent: in philosophical terminology "material" is often used in a sense equivalent to objective, and "formal" in a sense equivalent to subjective. For further discussion of Venn, see [Zabell, 2005, pp. 121–123].

Bertrand's *Calcul de probabilités*. only one major textbook appeared in France, M. H. H. Laurent's *Traite du calcul des probabilités* (published in 1873).³⁴ Although Cournot had criticized elements of the Laplacean position, his comments were often qualified and he retained, albeit with caveats, the classical analysis of judicial and testimonial probabilities. Laurent dispenses with these, but his text remains firmly in the Laplacean tradition.

The first real critic of the Laplacean position in the French textbook tradition is Bertrand.³⁵ He belongs firmly in the tradition of those earlier critics who thought the calculus had outreached its grasp. Alluding to Michell's calculation, he says "the application of the calculus to questions of this genre is an illusion and an abuse"; Mill's strictures are quoted with approval: the application to judicial decisions is "the scandal of mathematics" (p. xliii).³⁶

A century earlier Condorcet had asserted that the grounds for believing (*motifs de croire*) the sun will rise tomorrow are of the same nature as the ground for believing that a ball drawn from a urn, known to contain 10^7 white balls and a single black ball, will be white [Condorcet, 1785, p. xi]. Debating Condorcet across the expanse of a century, Bertrand ridiculed this assertion:

The assimilation is impermissible: one of the probabilities is objective, the other subjective. The probability of drawing a black ball on the first attempt is $1/10000000$, no more, no less. Whoever evaluates it otherwise is mistaken. The probability that the sun will rise varies from person to person. A philosopher can, without being insane, assert on the basis of a false science that the sun will soon extinguish itself; he is within his rights just as Condorcet is in his. [Bertrand, 1889, p. xix]

(As noted earlier, the presumably sane Augustus De Morgan had in fact questioned the objectivity of the first type of probability in his *Formal Logic*.)

To Cournot, the difference between objective and subjective probability was that of physical versus epistemic probability; the difference, that is, between the external world and our knowledge of it. In consequence, subjective probabilities vary from person to person depending on their knowledge; "it will assume a purely subjective character, susceptible to varying from one individual to another depending on the extent of their knowledge" [Cournot, 1843, p. 438].

To Bertrand, in contrast, it is precisely such variation that characterizes subjectivity (rather than being a consequence of it). The probability of rain does not exist,

³⁴The only other comparable work in French, due to the Anton Meyer, was published in Brussels in 1874 [Mansion, 1903; cited in Schneider, 1987, p. 211, n. 7].

³⁵Bertrand [1889]. The data is sometimes given as 1888 or 1889, but the copy I consulted has the 1889 date. The second edition of 1907 appears to be essentially a reprint of the first; the pagination is identical and the book does not appear to have been reset for the second edition.

³⁶Actually Mill employed the stodgier "real opprobrium of mathematics." Bertrand has a distinctively acerbic, staccato French prose which is often lost in translation.

not because it varies from one day to another because of the state of the heavens and the direction of the winds; but because in no circumstances does it have an *objective* value, the same for all who evaluate it without error. [Bertrand, 1889, p.90]³⁷

Bertrand next turns to the question of whether it makes sense to speak of the probability that a specific man who is 40 years old will live another 10: “Neither the event nor its contrary, at the present moment, have the certitude equal to that of an eclipse, but the probability does not have, despite this, an objective value, independent of the known information and good judgement of those who employ it.” Bertrand illustrates this with an example: the king of Siam is forty; what is the probability he will live another decade? It has one value for those who have questioned his physician; yet another for the physician himself; a very different one for those conspirators who have undertaken to strangle him the next day. “All these probabilities are subjective, they do not exist objectively” [Bertrand, 1889, pp. 90-91].

There are really several different strands present here; let us attempt to disentangle them:

- SENSE I:** an epistemic probability is subjective because it is an attribute of the subject perceiving rather than the object perceived.
- SENSE II:** an epistemic probability is subjective because it can vary from one individual to another because of the differing information they possess.
- SENSE III:** an epistemic probability is subjective because it is can vary from one individual to another *even if* they possess the same information.

The first sense is the one employed by Cournot and Venn; the second appears to be the one intended by Bertrand. The third sense goes to the heart of the distinction between epistemic probabilities which are logical or “objective” versus personal or “subjective.” Epistemic probabilities vary from one person to another based on the information they possess, but if they possess the same information then, some argue, they should assign the same probability to an event.

How is it possible for two different persons possessing the same information to have differing personal probabilities? One, less extreme school of subjectivists would argue that differences between individuals ultimately arise because the totality of information available to any two persons necessarily differ in important

³⁷Unfortunately, Bertrand goes on to obscure matters by immediately adding: “It will rain or it will not rain, one of the events is certain, now [“dès à présent”], and the other impossible. The physical forces on which rain depends are moreover determined, subordinate to laws as precise as those which direct the planets”. This is certainly a legitimate point (if one accepts determinism ...), but it is very different from the one before. It does not, however, seem to be the crucial point for Bertrand, merely an additional reason why it makes no reason to speak of probability in this setting; “would one ask for the probability that there would be an eclipse of the moon next month?”

respects. On this view, if hypothetically two persons did possess the same information they should agree — although as a practical matter this hypothetical never occurs. I. J. Good presumably has something like this in mind when he says he does not know whether logical probabilities exist, but he thinks it “mentally healthy” to believe they do [Good, 1965, p. 7].

The will to believe in an objective, logical probability is very strong; it can even be found (once) in a curious passage in De Morgan. De Morgan, as we have seen earlier, regards probability as a “state of the mind with respect to an assertion, a coming event, or any other matter on which absolute knowledge does not exist.” But then he immediately goes on to qualify this by adding, it is the state of mind a person “ought” to have,

for it may happen that the state of mind which is, is not the state of mind which should be. D’Alembert believed that it was two to one that the first head which the throw of a halfpenny was to give would occur before the third throw: a juster view of the mode of applying the theory would have taught him it was *three* to *one*. But he *believed* it, and thought he could show reason for his belief: to him the probability *was* two to one. But I shall say, for all that, that the probability *is* three to one. [De Morgan, 1847, p. 173]

This is a very clear statement of the difference between a psychological probability (the intensity of belief actually held) and an objective probability. But what is the criterion for the later? Why “is” the probability three to one? De Morgan gives the surprising answer:

meaning, that in the universal opinion of those who examine the subject, the state of mind to which a person *ought* to be able to bring himself is to look three times as confidently upon the arrival as upon the non-arrival.

This is a very curious criterion, but it is by no means the only time it has been proposed. In his book *Scientific inference*, Sir Harold Jeffreys copes with the problem of arriving at a so-called ignorance prior by suggesting that a scientific committee be appointed to resolve the matter! [Jeffreys, 1973, p. 41].³⁸ Thus, for De Morgan and Jeffreys, objective probabilities are objective because they represent consensus.

Such notions may seem curious given a long tradition in the past century of epistemic objective probability being viewed as a logical relation between propositions. But it is less surprising — even natural — if one recalls the Kantian tradition from which the subjective/objective distinction was borrowed. In the Critique of *Pure Reason*. Kant says that the “touchstone” which enables us to determine whether

³⁸Various proposals have been made to objectify the choice of prior, including invariance [Jeffreys, 1946], group invariance [Jaynes, 1968], and maximum posterior expected entropy (the “reference priors” of [Bernardo, 1979]). They are sometimes referred to as “uninformative priors”. The principle of indifference may be regarded as the first proposal of this sort.

we believe something on grounds which are objectively sufficient (“conviction”) or not (mere “persuasion”) is “the possibility of communicating it and of finding it valid for all human reason.”³⁹

For there is then at least a presumption that the ground of the agreement of all judgments with each other, notwithstanding the differing characters of individuals, rests upon the common ground, namely, upon the object, and it is for this reason that they are all in agreement with the object — the truth of the judgment being thereby proved.

Indeed, so long as we keep a judgement to ourselves, it is impossible to separate conviction from persuasion; it is the “experiment” of testing that judgment “upon the understanding of others” to see whether “those grounds of the judgment which are valid for us have the same effect on the reason of others as on our own,” which enables us to (subjectively!) detect “any merely private validity in the judgment,” i.e., mere crass persuasion.

Reviewing Keynes’s *Treatise* in 1924, Émile Borel perhaps put the matter best:

there are cases where it is legitimate to speak of the probability of an event: these are the cases where one refers to the probability which is common to the judgments of all the best informed persons

One could, in order to abridge our language, while at the same time not attaching an absolute sense to this expression, call these probabilities objective probabilities. On the other hand, subjective probabilities, those which particularly interest Mr. Keynes, can effectively have different values for different individuals. This is the case, for example, for habits of the race course, where each person attempts to calculate the chances of the entries in a race by taking account of the differing information that he can obtain about the entries. It is frequently said of two bettors that one is a thief and that the other is an imbecile; this is true in certain cases, where one of the two bettors is better informed than the other, and knows it; *but it can also happen that two men in good faith in complex situations where they possess exactly the same elements of information will arrive at different conclusions on the probabilities of an event* [emphasis added] and that betting together, each one figures, to use a picturesque but exaggerated expression, that he is the thief and the other the imbecile.⁴⁰

Precisely so. Objectivity at its base refers to elements of agreement, but this terminology should not be taken too far; subjective disagreement is possible even among those sharing the same information.

³⁹Kant [1781, A 820/ B 848; p. 645 of the 1929 edition].

⁴⁰Borel [1924]; Kyburg and Smokier [1965, pp. 50-51]. For further discussion of Borel, see [Knobloch, 1987, pp. 215-233]. Knobloch notes, in addition to Bertrand, the influence of Poincaré, Keynes, and Reichenbach on Borel’s thought. I do not attempt to trace Borel’s further thoughts on objective and subjective probabilities; see the comments and further references of Jorland [1987, p. 177 and notes 105-106].

How is such a position defensible? In brief, two persons sharing identical information but having different probabilities may have different world views; it is possible to integrate all available information in radically different, yet internally consistent ways.⁴¹ When President Carter's Secretary of State and National Security Advisor differed on the likely success of the Iran rescue mission, presumably that difference of opinion did not arise because of information possessed by one but not the other.⁴²

Despite the pungency of his language, Bertrand's criticisms ultimately lack force. The existence of objective probabilities need not preclude the existence of subjective ones. Nor is the boundary between objective and subjective probabilities at all clear: even objective probabilities will vary depending on our information, and it is far from clear how such instances differ in kind from those discussed by Bertrand. Nevertheless, Bertrand points to an important problem for the subjective theory of probability. If the ultimate nature of probability is subjective, why is it that in many cases apparently objective probabilities exist? This question did not go long unanswered.

8 POINCARÉ: THE METHOD OF ARBITRARY FUNCTIONS

The distinguished mathematician Henri Poincaré firmly allied himself with the Laplacean camp. Although Poincaré's research was largely in areas other than mathematical probability, he did write a textbook on the subject several years after Bertrand's [Poincaré, 1896], and he also discussed the foundations of probability in several of his philosophical writings (e.g., [Poincaré, 1902]). Poincaré saw as the chief difficulty for the subjective position the apparent existence of objective chance: even if one does believe in subjective probability, the fact that most dice come up aces one in six times on average appears to be an objective fact about the world:

A gambler wants to try a *coup*, and he asks my advice. If I give it to him, I use the calculus of probabilities; but I shall not guarantee success. That is what I shall call *subjective probability*. ... But assume that an observer is present at the play, that he knows of the coup, and that play goes on for a long time, and that he makes a summary of his notes. He will find that events have taken place in conformity with the laws of the calculus of probabilities. That is what I shall call *objective probability*, and it is this phenomenon which has to be explained. There are numerous Insurance Societies which apply the rules of the calculus

⁴¹Jacob Burckhardt, in the preface to his *Die Zeit Constantins des Grossen* [Burckhardt, 1853], observes that "In works of general history there is room for differences of opinion on fundamental premises and aims, so that the same fact may seem essential and important to one writer, for example, and to another mere rubbish utterly without interest. Hence the writer is well aware that his treatment may be impugned as being subjective".

⁴²The difficulty with any real example of this type, of course, is that the *utilities* of the two might well differ.

of probabilities, and they distribute to their shareholders dividends, the objective reality of which cannot be contested. In order to explain them, we must do more than invoke our ignorance and the necessity of action. Thus, absolute scepticism is not admissible. We may distrust, but we cannot condemn in *bloc*. Discussion is necessary. [Poincaré, 1902, pp. 187-88.]

Poincaré's explanation of this phenomenon invoked his theory of *arbitrary functions*: for a wide class of microscopic probability distributions for the selection of initial conditions (e. g., in spinning the roulette wheel by imparting an angular momentum, or tossing a die), the probabilities of the macroscopically observed outcomes (1 through 36, heads or tails) are, to an excellent approximation, uniform (see generally [Engel, 1992]). (Poincaré's discovery deeply impressed de Finetti, who drew on a similar idea in his explanation of induction: independent of a person's prior probability distribution, after a large number of trials his posterior probability will be highly peaked about the observed frequency of trials.)

Both Poincaré's explanation of the existence of objective chance, and de Finetti's explanation of induction, belong to a class of results that embed objectivistic phenomena within a subjectivistic framework. Modern philosophers such as Skyrms and Lewis, for example, explain chance as a type of subjective probability that is resilient to change when information of a certain type is received [Lewis, 1980]. In Bayesian statistics it is possible to prove that under conditions of considerable generality the posterior distributions of two "Bayesians" will merge as more and more information is received. Thus beliefs that conform to the constraints of subjective probability will act in ways that are rational even if not *objectively* determinate at any given time.

9 CONCLUSION

In the Middle Ages both sides of human consciousness — that which was turned within as that which was turned without — lay dreaming or half awake beneath a common veil. ... In Italy this veil first melted into air; an *objective* treatment and consideration of the state and of all the things of this world became possible. The *subjective* side at the same time asserted itself with corresponding emphasis; man became a spiritual *individual*, and recognised himself as such.

Jacob Burckhardt, *The Civilization of the Renaissance in Italy*, 1860.

In *Sense and Sensibilia* the philosopher J. L. Austin cautions that certain terms "live by taking in each other's washing — what is spurious is not one term of the pair, but the antithesis itself" [Austin, 1962, p. 4]. Nowhere is this more strikingly illustrated than in the contrast between subjective and objective.

In medieval scholastic terminology, subjective and objective were employed to make much the same distinction as at present, but the terms had meanings *exactly*

opposite to those they possess today. As the *Oxford English Dictionary* notes, the distinction was between

what belongs to things subjectively [subjectivé], or as they are ‘in themselves’, and what belongs to them objectively [objectivé], as they are presented to consciousness.

For example, for William of Ockham “the universal is not something real which has subjective being (that is, of its own right) either in the soul or outside the soul; it has only an objective being (that is, as the object of thought), and it is a kind of mental artefact (*et est quoddam fictivum*) having being as an object of thought corresponding to the being which the thing itself has of its own right” [Knowles, 1962, p. 322].

This reversed sense of the subjective still survives in Descartes, but the Oxford English Dictionary provides examples going as far back as the early 18th century where the shift has already taken place, and as we have seen earlier, it also occurs in this modern sense in Bernoulli’s *Ars conjectandi*. As the OED notes,

In later times the custom of considering the perceiving or thinking consciousness as pre-eminently the ‘subject’ brought about a different use of these words, which now prevails in philosophical language. According to this, what is considered as belonging to the perceiving or thinking self is called subjective, and what is considered as independent of the perceiving or thinking self is called in contrast *objective*.

...

[T]his transition of use ... primarily concerns the word subjective, and affects *objective* as its antithesis

Here we truly see Austin’s phenomenon in action: as “subjective” shifts in its meaning, so does “objective”, always maintaining the antithesis.⁴³ Given this lability of meaning, some authors (e.g., [Jeffreys, 1973, p. 210; Hacking, 1975, pp. 146-147]) prefer when possible to avoid the terminology altogether.

De Finetti, who was always uncompromising, was quite comfortable with the label “subjective,” while L. J. Savage, perhaps fearing the negative connotation that “subjective” had acquired in the 20th century, preferred the term “personal” probability. R. A. Fisher appears to have had an ambivalent attitude. In his book *Statistical Methods and Scientific Inference* [1956; 2nd ed. 1959; 3rd ed. 1973], Fisher initially wrote that the populations used to determine probability levels in tests of significance “have no objective reality, being exclusively the product of the statistician’s imagination” [1956, p. 77; 1973, p. 81], and initially referred to “the role of subjective ignorance, as well as that of objective knowledge in a typical probability statement” (p. 33). This embrace of the subjective was somewhat

⁴³In addition to the discussion and examples in the OED (under “objective” and “subjective”), see generally [Daston, 1994, pp. 332–335]. Even as late as 1888 one can find the English historian James Bryce identifying the newer usage of objective as German in origin, saying he is “stating in a purely positive, or as the Germans say, ‘objective’ way” [Bryce, 1888, vol. 1, p 203].

radical, however, for someone who had once branded Sir Harold Jeffreys's system as "subjective and psychological" [Fisher, 1934, p. 3], and by the third edition of Fisher's book, the statement in the first edition had been altered to read "the role both of well specified ignorance and of specific knowledge in a typical probability statement".⁴⁴

Such quotations may seem surprising to those familiar with some of Fisher's earlier writings, but by the 1950s Fisher had begun to move much closer to the Bayesian position he had so forcefully attacked in the decades before World War II. After a century the balance had begun to tilt, even if only temporarily, and Neyman had emerged as a far more dangerous enemy than the "objective Bayesianism" of a Harold Jeffreys. The subjective-objective debate had begun to have a serious impact on the practice of statistics.

That debate, as we have seen, had been ongoing for more than a century: Condorcet, Laplace, Poisson, Cournot, Bertrand, Poincaré, and Borel in France; De Morgan, Ellis, Mill, Herschel, Donkin, Boole, Venn, Edgeworth, and Peirce in England and America; J. F. Fries, Wundt, Fick, von Kries, Stumpf, and Meinong in Germany; each an element in a larger stream. By 1914 no consensus had emerged, but what had was a clearer picture of the distinctions, limitations, complexities, and paradoxes that surround probability.

Keynes's 1921 *Treatise on Probability* gives a unique snapshot of this ferment. Although Keynes himself has a clear viewpoint about the role and nature of probability, his *Treatise* makes it apparent just how much of a philosophical quagmire the subject must have seemed at the time he wrote. Within two decades however Ramsey and de Finetti were to provide dramatic and fundamental clarifications of the nature of subjective probability, while at the same time Fisher and Neyman advanced new frameworks for statistics entirely independent of effete appeals to inverse probability.

The complexity of probability remains with us; the twentieth century debate over its meaning and domain continues today. But the origins of that debate lie in the 19th century (and even earlier); its terms were in large part framed by that earlier debate; and we can better understand the argument in our own century by listening in on the one that took place earlier.

BIBLIOGRAPHY

- [Austin, 1962] J. L. Austin. *Sense and Sensibilia*. Oxford, University Press, Oxford, 1962.
- [Baker, 1975] K. M. Baker. *Condorcet*. University of Chicago Press, Chicago, 1975.
- [Bernardo, 1979] J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society*, Series B, 41, 113–147, 1979.
- [Bernoulli, 1713] J. Bernoulli. *Ars conjectandi*. Thurnisiorum, Basel, 1713. Reprinted 1975, *Die Werke von Jakob Bernoulli*, vol. 3, Birkhauser, Basel, pp. 107–286.
- [Bertrand, 1889] J. Bertrand. *Calcul des probabilités*. Gauthier-Villars, Paris (2nd ed. 1907). Reprinted 1972, Chelsea, New York, 1889.

⁴⁴Fisher died in 1962. The third edition of [Fisher, 1956] is based on "much new material which the author had entered in his interleaved copy of the book for this purpose."

- [Borel, 1924] E. Borel. A propos d'un traite de probabilités. *Revue Philosophique* 98, 321-326. English translation in Kyburg and Smokier 1964, pp. 47-60, 1924.
- [Brittan, 1994] G. Brittan, Jr. Kant and the quantum theory. In *Kant and Contemporary Epistemology* (The University of Western Ontario Series in Philosophy of Science, 54, Paolo Parrini, ed.), Kluwer Academic, Dordrecht, pp. 131-155, 1994.
- [Brittan, 2009] G. Brittan, Jr. Determinism, determination, and objectivity in modern physics. In *Constituting Objectivity* (Western Ontario Series In Philosophy of Science, Michel Bitbol, Pierre Kerszberg, and Jean Petitot eds.), Springer, New York, pp. 159-168, 2009.
- [Bryce, 1888] J. Bryce. *The American Commonwealth*. Macmillan, London and New York, 1888.
- [Burckhardt, 1953] J. Burckhardt. *Die Zeit Constantins des Grossen*. Basel 1853; second edition, Leipzig, 1880. English translation, *The Age of Constantine the Great*, University of California Press, 1983.
- [Cournot, 1843] A. A. Cournot. *Exposition de la théorie des chances et des probabilités*. Librairie de L. Hachette, Paris, 1843.
- [Cournot, 1851] A. A. Cournot. *Essai sur les fondements de nos connaissances et sur les caractères de la critique philosophique*. Librairie de L. Hachette, Paris, 1851. Reprinted 1975, Librairie Philosophique J. Vrin, Paris; page references are to this edition.
- [Cournot, 1861] A. A. Cournot. *Traite de l'enchaînement des idées fondamentales dans les sciences et dans l'histoire*. Paris, 1861.
- [Cournot, 1875] A. A. Cournot. *Matérialisme, vitalisme, rationalisme: Étude sur l'emploi des données de la science en philosophie*. Paris, 1875.
- [Daston, 1979] L. J. Daston. D'Alembert's critique of probability theory. *Historia mathematica* 6, 259-279, 1979.
- [Daston, 1994] L. J. Daston. How probabilities came to be objective and subjective. *Historia Mathematica* 21, 330-344, 1994.
- [Daston, 1988] L. J. Daston. *Classical Probability in the Enlightenment*. Princeton University Press, Princeton, 1988.
- [De Morgan, 1847] A. De Morgan. *Formal Logic: Or the Calculus of Inference Necessary and Probable*. Taylor and Walton, London, 1847.
- [Donkin, 1851] W. F. Donkin. On certain questions relating to the theory of probabilities. *Philosophical Magazine* 1, 353-368 and 458-466; 2, 55-60, 1851.
- [Ellis, 1843] R. L. Ellis. On the foundations of the theory of probabilities. *Transactions of the Cambridge Philosophical Society* 8, 1-6, 1843.
- [Ellis, 1854] R. L. Ellis. Remarks on the fundamental principle of the theory of probabilities. *Transactions of the Cambridge Philosophical Society* 9, 605-607, 1854.
- [Engel, 1992] E. Engel. *A Road to Randomness in Physical Systems*. Springer, *Lecture Notes in Statistics*, v. 71, 1992.
- [Fine, 1973] T. Fine. *Theories of Probability*. Academic Press, New York, 1973.
- [Fisher, 1956] R. A. Fisher. *Statistical Methods and Scientific Inference*. Hafner, New York, 1956, (2nd ed., 1959; 3rd ed., 1973).
- [Gay, 1959] P. Gay. *Voltaire's Politics: The Poet as Realist*. Princeton University Press, Princeton, 1959.
- [Gay, 1969] P. Gay. *The Enlightenment: An Interpretation*. Volume II: *The Science of Freedom*. Alfred A. Knopf, New York, 1969. Reprinted 1977 by W. W. Norton, New York.
- [Gillispie, 1972] C. C. Gillispie. Probability and politics: Laplace, Condorcet, and Turgot. *Proceedings of the American Philosophical Society* 116, 1-20, 1972.
- [Good, 1965] I. J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, Cambridge, 1965.
- [Gower, 1982] B. Gower. Astronomy and probability: Forbes vs. Michell on the distribution of stars. *Annals of Science* 39, 145-160, 1982.
- [Hacking, 1975] I. Hacking. *The Emergence of Probability*. Cambridge University Press, Cambridge, 1975.
- [Jaynes, 1968] E. T. Jaynes. Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics* 4, 227-241, 1968.
- [Jeffreys, 1946] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A (Mathematical and Physical Sciences)* 186: 453-461, 1946.

- [Jeffreys, 1973] H. Jeffreys. *Scientific Inference*. Cambridge University Press, Cambridge (1st ed. 1931, 2nd ed. 1957, 3rd ed. 1973).
- [Jorland, 1987] G. Jorland. The Saint Petersburg Paradox 1713-1937. In Kruger et al., pp. 157-190, 1987.
- [Kamlah, 1987] A. Kamlah. The decline of the Laplacian theory of probability: a study of Stumpf, von Kries, and Meinong. In Kruger et al., pp. 91-116, 1987.
- [Kant, 1781] I. Kant. *Kritik der Reinen Vernunft*. Johann Friedrich Hartknoch, Riga, 1781. Translated 1929 as Immanuel Kant's *Critique of Pure Reason* (Norman Kemp Smith, trans.), Macmillan, New York. Page references are to the 1929 translation.
- [Keynes, 1921] J. M. Keynes. *A Treatise on Probability*. Macmillan, London, 1921.
- [Knobloch, 1987] E. Knobloch. Emile Borel as probabilist. In Kruger et al., pp. 215-233, 1987.
- [Knowles, 1962] D. Knowles. *The Evolution of Medieval Thought*. Vintage Books, New York, 1962.
- [Kruger et al., 1987] L. Kruger, L. J. Daston, and M. Heidelberger. *The Probabilistic Revolution*. Volume 1: *Ideas in History*. MIT Press, Cambridge, MA, 1987.
- [Kyburg and Smokier, 1964] H. E. Kyburg, Jr. and H. E. Smokier. *Studies in Subjective Probability*. Wiley, New York, 1964.
- [Laplace, 1814] P. S. Laplace, Marquis de. *Essai philosophique sur les probabilités*. Courcier, Paris, 1814. Page references are to *Oeuvres complètes de Laplace*, vol. 7, Gauthier-Villars, Paris, 1886.
- [Lewis, 1980] D. Lewis. A subjectivist's guide to objective chance. In *Studies in Inductive Logic and Probability*, volume II (Richard C. Jeffrey, ed.), University of California Press, Berkeley and Los Angeles, pp. 7-155, 1980.
- [Maistrov, 1974] L. E. Maistrov. *Probability Theory: A Historical Sketch*. Academic Press, New York, 1974.
- [Mansion, 1903] P. Mansion. Sur la portée objective du calcul des probabilités. *Bulletin de l'Académie Royale de Belgique (Classe des sciences)* 4, 1235-1294, 1903.
- [Nadler, 2006] S. Nadler. The doctrine of ideas. In *The Blackwell guide to Descartes' Meditations* (Stephen Gaukroger, ed.), Blackwell, Oxford, pp. 86-103, 2006.
- [Poincaré, 1896] H. Poincaré. *Calcul des probabilités*. Gauthier-Villars, Paris (2nd ed. 1912), 1896.
- [Poincaré, 1902] H. Poincaré. *La science et l'hypothèse*. Paris. Translated 1907 as *Science and Hypothesis* (J. W. Greenstreet, trans.), Walter Scott Publishing Company, London; translation reprinted 1952, Dover, New York. Page references are to the 1902 translation.
- [Poisson, 1837] S. D. Poisson. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. Bachelier, Paris, 1837.
- [Porter, 1986] T. M. Porter. *The Rise of Statistical Thinking: 1820-1900*. Princeton University Press, 1986.
- [Priest, 1987] S. Priest. Subjectivity and objectivity in Kant and Hegel. In *Hegel's Critique of Kant* (S. Priest, ed.), Clarendon Press, Oxford, 103-118, 1987.
- [Rashed, 1974] R. Rashed. *Condorcet: mathématique et société*. Hermann, Paris, 1974.
- [Salmon, 1980] W. C. Salmon. John Venn's Logic of Chance. In *Pisa Conference Proceedings*, vol. 2 (J. Hintikka, D. Gruender, and E. Agazzi, eds.), D. Reidel, Dordrecht, 1980.
- [Schneider, 1987] I. Schneider. Laplace and thereafter: the status of probability theory in the nineteenth century. In Kruger et al., pp. 191-214, 1987.
- [Shairp et al., 1873] J. C. Shairp, P. G. Tait, and A. Adams-Reilly. *Life and Letters of James David Forbes. F. R. S.* London, 1873.
- [Stigler, 1986] S. M. Stigler. *The History of Statistics*. Harvard University Press, Cambridge, 1986.
- [Sterzinger, 1911] O. H. Sterzinger. *Zur Logik und Naturphilosophie der Wahrscheinlichkeitslehre*. Xenien-Verlag, Leipzig, 1911.
- [Strong, 1978] J. V. Strong. John Stuart Mill, John Herschel, and the 'probability of causes.' *PSA 1978: Proceedings of the Philosophy of Science Association* 1. 31-41, 1978.
- [Stove, 1973] D. C. Stove. *Probability and Hume's Inductive Scepticism*. Oxford University Press, Oxford, 1973.
- [Swijtink, 1986] Z. G. Swijtink. D'Alembert and the maturity of chances. *Studies in History and Philosophy of Science* 17, 327-349, 1986.

- [von Wright, 1957] G. H. von Wright. *The Logical Problem of Induction* (2nd revised edition). Macmillan, New York, 1957.
- [Zabell, 2005] S. L. Zabell. *Symmetry and its Discontents*. Cambridge University Press, 2005.

PROBABILITY IN ANCIENT INDIA

C. K. Raju

INTRODUCTION: MATHEMATICAL PRE-REQUISITES

Permutations and Combinations

A first course on probability (at the high-school level) typically begins with an account of the theory of permutations and combinations needed for calculating probabilities in games of chance, such as dice or cards.

Like many other aspects of mathematics, this theory of permutations and combinations first developed in India, although an account of its history is usually missing in stock presentations of combinatorics.

In fact, the theory of permutations and combinations was basic to the Indian understanding of metre and music. The Vedic and post-Vedic composers depended on combinations of two syllables called *guru* (deep, long) and *laghu* (short). The earliest written account of this theory of metre is in Piṅgala's *Chandaḥsūtra* (–3rd c. CE), a book of aphorisms (*sūtra*-s) on the theory of metre (*chanda*). To calculate all possible combinations of these two syllables in a metre containing n syllables, Piṅgala gives the following rule¹ (which explicitly makes use of the symbol for zero). “(Place) two when halved;” “when unity is subtracted then (place) zero;” “multiply by two when zero;” “square when halved.” In a worked example, Dutta and Singh² show how for the Gāyatrī metre with 6 syllables this rule leads to the correct figure of 2^6 possibilities.

That this rule basically involves the binomial expansion is made clear by Piṅgala's commentator the 10th c. CE Halayudha. Thus, in a 3-syllabic metre with two underlying syllables, *guru* and *laghu*, 3 *guru* sounds will occur once, 2 *gurus* and 1 *laghu* will occur twice, as will 1 *guru* and 2 *laghus*, while 3 *laghus* will occur once. Symbolically $(g + l)^3 = g^3 + 3g^2l + 3gl^2 + l^3$. To generalize this to the case of n underlying syllables, Halayudha explains the *meru-prastāra* (pyramidal expansion) scheme for calculation,³ which is identical to “Pascal's” triangle which first appeared in Europe about a century before Pascal (on the title page of the

¹ *Chandaḥsūtra* viii.28. Ed. Sri Sitanath, Calcutta, 1840. Cited by Dutta and Singh, vol. 1, p. 76.

² B. B. Dutta and A. N. Singh, *History of Hindu Mathematics*, Asia Publishing House, 1962, vol. 1, pp. 75–76.

³ A. K. Bag, *Mathematics in Ancient and Medieval India*, Chaukhambha Orientalia, Varanasi, 1979, pp. 189–93.

Arithmetic of Apianus) and in China in the 14th c.⁴ An example, using the *Gāyatrī* meter is also found in Bhaskara's *Līlāvati*.⁵ The accounts found in stock Western histories of mathematics (such as that by Smith⁶) incorrectly state that no attention was paid in India to the theory of permutations and combinations before Bhaskara II (12th c. CE).

Although this theory is built into the Vedic metre, the earliest known written account relating to permutations and combinations actually comes from even before Piṅgala, and is found in the 4th c. Jain *Bhagwatī Sūtra*. Permutations were called *vikalpa-gaṇita* (the calculus of alternatives), and combinations *bhanga*. The text works out the number of combinations of n categories taken 2, 3 etc. at a time.

Incidentally, this throws up large numbers of the sort that cannot easily be written in Greek (Attic) and Roman numerals. It should be noted that while the Yajurveda⁷ already used a place value system, and gives names for numbers up to 10^{12} , the Jain literature typically runs into very large numbers, such as 10^{60} , and the Buddha when challenged (perhaps by a Jain opponent) names numbers up to 10^{53} . Large numbers have an intimate connection with the philosophy of probability, as examined in more detail, later on.

From the earliest Vedic tradition, there is a continuous tradition linking the first accounts of permutations and combinations with those of Bhaskara II (12th c.), and later commentaries on his work, up to the 16th c. CE, such as the *Kriyākramakārī*.⁸ Thus,⁹ the surgeon Suśruta (2nd c. CE) in his compendium (*Suśruta-saṃhitā*) lists the total number of flavours derived from 6 flavours taken 1 at a time, 2 at a time, and so on. Likewise, Varāhamihira (6th c.) who reputedly wrote the first Indian text on astrology (*Brhat-Jātaka*) states in it the number of perfumes that can be made from 16 substances mixed in 1, 2, 3, and 4 proportions.

Similar examples are found in the *Pāṭiḡaṇita* (Slate Arithmetic) of Śrīdhar (10th c.), a widely used elementary school-text, as its name suggests, Mahavira's (8th c.)

⁴Joseph Needham, *The Shorter Science and Civilisation in China*, vol. 2 (abridgement by C. A. Ronan). Cambridge University Press, 1981, p. 55.

⁵Bhaskara, *Līlāvati*, trans. K. S. Patwardhan, S. A. Naimpally, and S. L. Singh, p. 102. The verse is numbered differently in different manuscripts. K. V. Sarma in his critical edition of the 16th c. southern commentary *Kriyākramakārī* (VVRI, Hoshiarpur, 1975) on the *Līlāvati*, gives this as verse number 133, while the other cited source has given it as verse number 121.

⁶D. E. Smith, *History of Mathematics*, Dover Publications, 1958, vol. 2, p. 502.

⁷Yajurveda xvii.2 gives the names for the first 12 powers of 10, the first five being more or less similar to what they are today.

⁸This has an interesting connection with the history of the calculus. Fermat's challenge problem is identical with a solved exercise in Bhaskara II. So this is one of the texts that travelled from Cochin to Rome, and Bhaskara was probably Pascal's source. *Bījagaṇita of Śrī Bhāskarācārya*, ed. Sudhakara Dvivedi, Benares, 1927 (Benares Sanskrit Series, No. 159), chapter on *cakravāla*, p. 40. An account of Bhaskara's *cakravāla* method may be found, for instance, in Bag, cited above (pp. 217–228). For a formalised account of Bhaskara's *cakravāla* method, see I. S. Bhanu Murthy, *A Modern Introduction to Ancient Indian Mathematics*, Wiley Eastern, New Delhi, 1992, pp. 114–21. (Bhanu Murthy's book has a typo here.) For Fermat's challenge problem and "Pell's equation", see D. Struik, *A Source Book in Mathematics 1200–1800*, Harvard University Press, Cambridge, Mass., 1969, pp. 29–30.

⁹A. Bag, *Mathematics in Ancient and Medieval India*, cited above, p. 188.

Gaṇita Sāra Saṁgraha, and Bhaskara II (*Līlāvātī*) etc. Bhaskara mentions that this formula has applications to the theory of metre, to architecture, medicine, and *khaṇḍameru* (“Pascal’s triangle”). In these later texts, one finds explicitly stated formulae for permutations and combinations.

For example, to calculate $\binom{n}{r}$ values, Śridhar, in his text on slate-arithmetic¹⁰ (*Pāṭigaṇita*), gives the following rule.

एकाद्युत्तरविधिना रस विन्यासे विलोमतो गुण्येत्
पूर्वेण परं क्रमशो रूपादिचयैर्हरिर्विभजेत्

This translates as follows (*Pāṭigaṇita*, 72, Eng. p. 58)

Writing down the numbers beginning with 1 and increasing by 1 up to the (given) number of savours in the inverse order, divide them by the numbers beginning with 1 and increasing by 1 in the regular order, and then multiply successively by the preceding (quotient) the succeeding one. (This will give the number of combinations of the savours taken 1, 2, 3, ..., all at a time respectively.)¹¹

Thus, in the case of 6 savours, one writes down the numbers 1 to 6 in reverse order

6, 5, 4, 3, 2, 1

These are divided by the numbers in the usual order, to get the quotients

$$\frac{6}{1}, \frac{5}{2}, \frac{4}{3}, \frac{3}{4}, \frac{2}{5}, \frac{1}{6}.$$

Then, according to the rule, the number of combinations of savours taken 1 at a time, 2 at a time, etc., up to all at a time are respectively

$$\frac{6}{1}, \frac{6}{1} \times \frac{5}{2}, \frac{6}{1} \times \frac{5}{2} \times \frac{4}{3}, \text{ etc.}$$

Although, the formulae are mostly stated in identical terms, they are applied most flamboyantly by Bhaskara II. For example, to illustrate one of his formulae, Bhaskara asks for the total number of 5 digit numbers whose digits sum to 13.¹²

पञ्चस्थानस्थितैरङ्कैर्यद्यद्योगस्त्रयोदश
कतिभेदा भवेत्संख्या यदि वेत्सि निगद्यताम्

He then adds in the next verse that although this question involves “no multiplication or division, no squaring or cubing, it is sure to humble the egotistical and evil lads of astronomers”.

¹⁰Śridhar, *Pāṭigaṇita*, 72, ed. & trans. K. S. Shukla, Dept. of Mathematics and Astronomy, Lucknow University, 1959, Sanskrit, p. 97.

¹¹*Pāṭigaṇita* of Śridhar, trans. K. S. Shukla. As he points out, similar articulations are found in the *Gaṇita Sāra Saṁgrah* of Mahavira, vi.218, *MahāSiddhānta* of Āryabhaṭa 2, xv, 45–46 etc.

¹²*Līlāvātī* of Bhāskarācārya, trans. Patwardhan et al., p. 181. They give the number of this verse as 276, whereas, in K. V. Sarma’s critical edition of *Kriyākramkarī*, a commentary on the *Līlāvātī*, this is at 269, p. 464.

Weighted averages

The notion of simple average was routinely used in Indian planetary models, where each planet had a mean motion, and a deviation from it. Unlike Western planetary models, there was no belief in any divine harmony nor any faith in divine “laws” of any sort involved here, just an average motion and deviations from it in a down-to-earth empirical sense. While the deviations from the mean were not regarded as necessarily *mechanically* explicable, neither were they regarded as quite “random”, for it was believed that the deviation could be calculated in principle, at least to a good degree of approximation (required for the Indian calendar, which identified the rainy season, and hence was a critical input for monsoon-driven agriculture in India).

In various elementary mathematical texts, in the context of computing the density of mixtures or alloys, one also finds the usual formula for weighted averages, which is so closely related to the notion of “mathematical expectation” in probability theory. For example, we find in verse 52(ii) of the *Pāṭīgaṇita* (Eng. p. 36) the following:

“The sum of the products of weight and *varṇa* of the several pieces of gold, being divided by the sum of the weights of the pieces of gold, give the *varṇa* (of the alloy).

That is, if there are n pieces of gold of weights $w_1, w_2, \dots, w_n$, and *varṇas* $v_1, v_2, \dots, v_n$, the *varṇa* v of the alloy is given by $v = \frac{w_1 v_1 + w_2 v_2 + \dots + w_n v_n}{w_1 + w_2 + \dots + w_n}$.

(The term *varṇa* is analogous to the term “carat”, with pure gold consisting of 16 *varṇas*.) Understandably, this topic of “mixtures” is given special emphasis in Jain texts like those of Mahavira.

The relevance of these weighted averages to gambling was understood. It is in this very context of “mixtures” that the *Gaṇita Sāra Saṃgraha*¹³ (268.5, 273.5) gives the example of “Dutch bets” mentioned by Hacking.¹⁴ This is “a rule to ensure profit (in gambling) regardless of victory or loss”, a method of riskless arbitrage, in short. The text illustrates the rule with an example, where “a great man knowing mantra and medicine sees a cockfight in progress. He talks to the owners of the birds separately in a mysterious way. He tells one that ‘if your bird wins, you give me the amount you bet, and if it loses, I will give you $\frac{2}{3}$ of that amount’. Then he goes to the owner of the other bird where on those same conditions he promised to pay $\frac{3}{4}$ of the amount. In either case, he earned a profit of only 12 pieces of gold. O mathematician, blessed with speech, tell me how much money did the owner of each bird bet.”

¹³ *Gaṇitasāra-Saṃgraha*, (Hindi trans. L. C. Jain), Jain Samskṛti Samraksha Sangh, Sholapur, 1963, pp. 159–60.

¹⁴ Ian Hacking, *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*, Cambridge University Press, 1975, pp. 6–9. Hacking uses the English translation of *Gaṇita Sāra Saṃgraha*, trans. M. Rangacharya (1912), pp. 162–3.

Precise fractions

Apart from the ability to work with large numbers, and to calculate permutations and combinations, and weighted averages, there is also needed the ability to work with fractions having large numerators and denominators. Such ability, indicative of greater precision, is not automatic. Such precise fractions with large numerators and denominators are certainly found in Indian mathematical texts from the time of Āryabhata.¹⁵ Their use is also reflected in the Indian calendar.

By way of comparison, the Romans had only a few stock fractions to base 12 (each of which had a separate name). Hence they had a wrong duration of the year as $365\frac{1}{4}$ days, just because it involved such an easy-to-state fraction. And they retained this wrong duration even after the repeated calendar reforms of the 4th to 6th c., and, for over a thousand years, until 1582. The Greek Attic numerals were similar, and the Greeks did not work with such fractions in any text coming actually (and not notionally) from before the 9th c. Baghdad House of Wisdom.¹⁶ The significance of the Baghdad House of Wisdom is that Indian arithmetic texts travelled there, were translated into Arabic, and some of these were further translated from Arabic into Greek¹⁷ like the Indian story book, the *Pañcatantra*.¹⁸ (All known Arabic and Greek manuscripts of “Ptolemy’s” *Almagest*, for example, post-date the Baghdad House of Wisdom—most also post-date the Crusades—and are decidedly post-9th c. accretive texts as is clear, for example, from their star lists. Therefore, it would be anachronistic to attribute, uncritically and ahistori-

¹⁵ Āryabhaṭīya, *Gīṭikā* 3-4, trans. K. S. Shukla and K. V. Sarma, INSA, New Delhi, 1976, p. 6. This verse gives revolution numbers for various planets, and requires us to calculate fractions such as 1582237500/4320000. Up to a hundred years ago, Western historians, who subscribed to the view that the world was created a mere 6000 years ago, invariably described these figures as fantastic cosmological speculations, and meaningless by implication. However, the above fraction leads to a surprisingly accurate figure of 365.25868 days for the length of the *sidereal* year (compared to its modern value of 365.25636 days). In contrast, the value attributed to Ptolemy (365.24666 days) is significantly less accurate, and we know that the length of the (tropical) year on the Roman calendar was less accurate than the figure attributed to “Ptolemy”. This led to a slip in the date of Easter, and hence necessitated the repeated attempts at Roman calendar reform in the 5th and 6th centuries.

¹⁶ The difference between actual and notional dates is important. The only clear way to check a notional date is through the non-textual evidence, and it is hard to believe that the crude Greek and Roman calendars, despite repeated attempts by the state and church to reform them, could have co-existed for centuries with relatively sophisticated astronomy texts, such as the *Almagest* attributed to Ptolemy (which attribution makes its notional date 2nd c., although the actual manuscripts of it come from a thousand years later, and are accretive).

¹⁷ C. K. Raju, *Cultural Foundations of Mathematics*, Pearson Longman, 2007. This text examines in detail the points made in note 16 above. Also, C. K. Raju, *Is Science Western in Origin?*, Citizens International and Multiversity, Penang, 2009, contains a brief account.

¹⁸ Edward Gibbon, *The Decline and Fall of the Roman Empire*, Great Books of the Western World, vols 37–38, Encyclopaedia Britannica, Chicago, 1996, vol. 2, note 55 to chp. 52, p. 608. Others have assigned the date of 1080 to Simon Seth’s Greek translation of the *Pañcatantra* from Arabic. The Arabic translation *Kalilah va Dimnah* by Ibn al Muqaffa (d. 750), was long before the formation of the House of Wisdom, and the movement called the Brethren of Purity (Ikhwan al-Safa) derives inspiration from this text. This translation was from the Pahlavi translation which was from the Sanskrit, and done by Burzoe himself, the vazir of Khusrow I, Noshirvan, according to the *Shahnama* of Firdausi.

cally, the use of sexagesimal fractions in such a late text to a mythical Claudius Ptolemy of the 2nd c.¹⁹)

The game of dice in India

Thus, Indian tradition had all the arithmetic tools needed for the calculation of discrete probabilities. But was there a concept of probability? Stock mathematics texts draw their examples from a range of sources, and we also occasionally find some examples related to games of chance such as dice. The same slate-arithmetic text (*Pāṭiganita*, 99–101, p. 145) gives a long and complex rule for calculating whether one has won or lost in a game of dice. However, this is given as an application of the formulae for arithmetic progression. The substance of these formulae is explained as follows.

Suppose that two persons A and B gamble with dice, and that they alternately win p_1, p_2, p_3, p_4 casts. If the stake-moneys of the casts be in the arithmetic progression $a, a + d, a + 2d, \dots$ then the amount won by $A = [a + (a + d) + (a + 2d) + \dots (p_1 \text{ terms})] + [a + (p_1 + p_2)d + a(p_1 + p_2 + 1)d + \dots (p_3 \text{ terms})]$. Likewise, the amount won by $B = [a + p_1d + a + (p_1 + 1)d + \dots (p_2 \text{ terms})] + [a + (p_1 + p_2 + p_3)d + (p_1 + p_2 + p_3 + 1)d + \dots (p_4 \text{ terms})]$ and this leads to the enunciated rule.

This suggests that the game of dice might not have been played the same way in India, as it is played today, and also that a common strategy followed was (somewhat like martingale bets) to go on increasing the stake as the game went on. But this still does not give us enough information. (Accounts of the game of dice are *not* found in the *Gaṇita Sāra Saṃgraha*.²⁰ Presumably, the Jains did not want their children to start thinking about such things!) The text naturally takes it for granted that the readers are familiar with this game of dice, but *we* do not seem to have adequate sources for that at the moment, or at least such sources are not known to this author.

The hymn on dice in the Rgveda

Games of chance, such as dice, certainly existed in Indian tradition from the earliest times. We find an extraordinary *akṣa sūkta* or hymn on dice in the Rgveda (10.2.34). The long hymn begins by comparing the pleasure of gambling with the pleasure of drinking soma!

“There is enjoyment like the soma in those dice”.²¹ (सोमस्येव मौजवतस्यमच्चो)

It goes on to describe how everyone avoids a gambler, like an old man avoids horses, even his mother and father feign not to recognize him, and he is separated from his loving wife. Many times the gambler resolves to stay away, but each time the fatal attraction of the dice pulls him back. With great enthusiasm he

¹⁹This is argued in more detail in C. K. Raju, *Cultural Foundations of Mathematics*, cited above. See also notes 15 and 16 above.

²⁰Mahavir, *Gaṇita Sāra Saṃgraha*, Hindi trans. L. C. Jain, Jain Samskriti Samrakshaka Sangha, Sholapur, 1963.

²¹Rgveda, 10.2.34.1

reaches the gambling place, hoping to win, but sometimes he wins and sometimes his opponent does. The dice do not obey the wishes of the gambler, they revolt. They pierce the heart of the gambler, as easily as an arrow or a knife cuts through the skin, and they goad him on like the ankus, and pierce him like hot irons. When he wins he is as happy as if a son is born, and when he loses, he is as if dead. The 53 dice dance like the sun playing with its rays, they cannot be controlled by the bravest of the brave, and even the king bows before them. They have no hands, but they rise and fall, and men with hands lose to them. The gambler's wife remains frustrated and his son becomes a vagabond. He always spends his night in other places. Anyone who lends him money doubts that he will get it back. The gambler who arrives in the morning on a steed leaves at night without clothes on his back. [Such is the power of dice!] O Dice, I join the ten fingers of my hands and bow to the leader among you!

THE NOTION OF A FAIR GAME AND THE FREQUENTIST INTERPRETATION OF PROBABILITY

Fair and deceitful gambling in the Mahabharata

This tells us a great deal about the social consequences of gambling, but still very little about how exactly the game was played. (This is naturally assumed to be known to all.) But we do find something interesting from a philosophical perspective on probability. For, from the earliest times, there was also a notion of what constitutes a “fair game”, a notion which is today inextricably linked to the notion of probability.

This can be illustrated by a story from the Mahabharata epic. A key part of the story, and the origin of the Mahabharata war, relates to the way the heroes (Pandava-s) are robbed of their kingdom by means of a game of dice (द्यूत क्रीडा). They cannot very well refuse the invitation to play dice because the game involves risk, and a Kshatriya is dishonoured by refusing to partake in an enterprise involving risk. However, at the start of the game, Yudhisthira, the leader of the Pandavas, makes clear that he knows the dice are loaded, or that the game will involve deceit. He says: “निक्रितिर्देवनं पापं न चात्रोत्र पराक्रमः” (Mbh, Sabhā Parva, 59.5). That is, “Deceitful gambling is sinful, there is no Kshatriya valor in it.”

Philosophically speaking, there clearly are two concepts here: “deceitful gambling” as opposed to “fair gambling”. The notion of a fair game, though not explicitly defined, is critical to the story, and involves some notion of probability implicit or explicit. The Mahabharata narrative itself brings out the unfairness of the game both through Yudhishtira's statement, at the beginning of the game, and by telling us about a long series of throws in each one of which Yudhishtira loses, without ever winning once. Of course, there is room to argue, as the devil's advocate might, and as Duryodhana's uncle Shakuni does, that it is a case of gambler's ruin or a long run of bad luck with a finite amount of capital howsoever large. Apart from the Yudhisthira-Shakuni dialogue, the Mahabharata narrative

itself carries the depiction of this unfairness to the utmost. Yudhisthira goes on losing, and in desperation, he even stakes his wife, Draupadi, who is then deemed to have been won, and publicly stripped and insulted by Duryodhana.²²

Interestingly, a little later in the same epic, Yudhisthira, now banished to the forest, recounts his woes: how he was cheated out of a kingdom and his wife insulted. To console him, a sage recounts to him the celebrated love story of Nala and Damayanti. King Nala too lost his kingdom—once again through deceitful gambling. Having lost also his last remaining clothes in the forest, he covers himself with half of Damayanti's sari, and abandons her in the forest. He takes up a job as a charioteer with king R̥tuparṇa of Ayodhya. His aim is to learn the secret of dice from R̥tuparṇa, for he has been advised by a Naga prince that that knowledge will help him win back his kingdom. He gets an opportunity, at a tense moment, while rushing from Ayodhya to Vidarbha for the announced second marriage (*swayamvara*) of Damayanti whom R̥tuparṇa too wants to wed. Nala stops near a Vibhīṭaka tree²³—it is significant that the nuts of this very tree, with five faces, were used in the ancient Indian game of dice—and R̥tuparṇa can't resist showing off his knowledge of mathematics (*gaṇita vidyā*) by saying: the number of fruits in the two branches of the tree is 2095, count them if you like.

Nala decides to stop and count them. R̥tuparṇa, who is apprehensive of being delayed and knows that he cannot reach in time without Nala's charioteering skills, suggests that Nala should be satisfied with counting a portion of one branch.

Then the king reluctantly told him, 'Count. And on counting the leaves and fruits of a portion of this branch, thou wilt be satisfied of the truth of my assertion.'²⁴

Nala finds the estimate accurate and wants to know how it was done, offering to exchange this knowledge with his knowledge of horses. "And king R̥tuparṇa, having regard to the importance of the act that depended upon Vahuka's [Nala's] good-will, and tempted also by the horse-lore (that his charioteer possessed), said, 'So be it.' As solicited by thee, receive this science of dice from me..." (*ibid*). (The story has a happy ending, and Nala does get back his wife and his kingdom.)

From our immediate point of view, the interesting thing is the way the Mahabharata text conflates sampling theory with the "science of dice". (It is legitimate to call this "science", for there is a clear relation to a process of empirical verification, by cutting down the tree and counting.) This connection of sampling theory to the game of dice is mentioned by Hacking,²⁵ who attributes it to V. P. Godambe. They have, however, overlooked the other aspect of the story, which is that this knowledge was regarded as secret, and R̥tuparṇa parts with it only under

²²Eventually, it is decreed that Yudhisthira incorrectly regarded her as his property to be staked, however, one of the heroes, Bhim, swears that he will break Duryodhana's thigh, on which he seats Draupadi, and also drink the blood of his brother Dussasana who forcibly dragged Draupadi to the court, both of which promises he fulfills years later in the battlefield.

²³Mahabharata, *van parva*, 72, trans. K. M. Ganguly, 1883–1896, Book 3, pp. 150–51.

²⁴*ibid*, p. 151.

²⁵Ian Hacking, *The Emergence of Probability*, cited earlier, pp. 6–9.

extraordinary conditions; therefore, one should not expect to find this knowledge readily in common mathematical texts.

“Fair gambling”, the law of large numbers

There are, however, repeated references to “deceitful gambling”. So, one can ask: what exactly did “fair gambling” mean? As the notion is understood today, in an ultimate sense, the fairness of the game cannot be established merely by the law of large numbers, for it is well known that the frequentist interpretation of probability fails because relative frequency converges to probability only in a probabilistic sense. Shakuni is right that a long streak of bad luck is always possible (and must almost surely occur in the long run). Therefore, also, the randomness of a set of random numbers can *ultimately* be established only by reference to the process which generated those random numbers and not merely by *post-facto* statistical tests of randomness. Note the emphasis on “ultimate”. Given a series of random numbers even a simple χ^2 test could show,²⁶ for most practical purposes, whether the series is concocted; the question is whether it can settle “all” doubt, for there is always some residual probability that the test gave a wrong result, and this problem of what to do with very small numbers takes us back to the old problem of the law of large numbers. One cannot rule out the occurrence even of probability-zero cases, and the best one can say is that they will “almost surely” not occur.

Law of large numbers and the notion of convergence as a supertask

At this point, it is probably a good idea to understand two key differences between the Indian philosophy of mathematics and the Western philosophy of mathematics. The contemporary notion of a fair game already involves some notion of the law of large numbers, hence a notion of *convergence* in some sense (such as convergence in probability). It does not matter how “weak” this notion of convergence is (i.e. whether L^p convergence or convergence in measure): the point is that, as conceptualised in the Western philosophy of mathematics, *any* notion of convergence involves a supertask—an infinite series of tasks. Such supertasks can *only* be performed *metaphysically*, within set theory; however, barring special cases, a supertask is not something that can be empirically performed (in any finite period of time).

Supertasks and the clash of mathematical epistemologies

Historically speaking, the transmission of the calculus²⁷ from India to Europe in the 16th c. led to a clash of epistemologies. The situation is analogous to the

²⁶E.g., C. R. Rao, *Statistics and Truth: Putting Chance to Work*, CSIR, New Delhi, 1989.

²⁷C. K. Raju, *Cultural Foundations of Mathematics*, Pearson Longman, 2007, chp 9, “Math wars and the epistemological divide in mathematics”.

clash of epistemologies which occurred when the Indian arithmetic techniques, called the *Algorismus* (after al Khwarizmi's latinized name), or "Arabic numerals", were imported into Europe and clashed with the abacus or the native Roman and Greek tradition of arithmetic. With regard to the Indian calculus, Western philosophers did not understand its epistemology, and felt that it involved super-tasks. The practical usefulness of the calculus (in calculating trigonometric values, navigational charts etc.) then created the need to justify such supertasks. In the Western philosophy of mathematics, this mistaken belief about supertasks arose because of two reasons, (1) a belief in the perfection of mathematics, and (2) a distrust of the empirical as means of proof.

Thus, for example, reacting to the use of the new-fangled calculus-techniques by Fermat and Pascal, Descartes²⁸ wrote in his *Geometry* that comparing the lengths of straight and curved lines was "beyond the capacity of the human mind". Galileo's reaction, in his letters to Cavalieri, was similar, and he ultimately left it to his student Cavalieri to take credit or discredit for the calculus.

Descartes' reaction seems idiosyncratic and excessive, because any Indian child knew how to measure the length of a curved line empirically by using a piece of string, or *śulba* or *rajju*, and this was done in India, at least since the days of the *śulba sūtra* [–500 CE]. Perhaps, Descartes thought that the European custom of using a rigid ruler was the only way of doing things. (European navigators certainly had a serious problem for just this reason.) In any case, Descartes proceeded from the metaphysical premise that only the length of a straight line was meaningful. On this metaphysical premise, he thought that the length of a curved line could *only* be understood by *approximating* it by a series of straight-line segments. However large the number of straight line segments one might use, and howsoever fine the resulting approximation might be, it remained "imperfect", hence was not quite mathematics, for mathematics (Descartes believed) is perfect and cannot neglect the smallest quantity.

Descartes was willing to make a concession and allow that infinitesimal quantities could be neglected (a premise shared by many of his contemporaries, including, later Berkeley, in his criticism of Newton, and a process later formalised in non-standard analysis). Therefore, the only situation Descartes was willing to contemplate as acceptable was a situation where the neglected error was infinitesimal. But this required an *infinity* of straight line segments, each of *infinitesimal* length. But that created another difficulty: to compute the length of the curved line, one now had to sum an infinity of infinitesimals. And *this*, thought Descartes, was a supertask only God could perform.²⁹ (This was in the days before formal mathematics. Descartes thought of summing an infinite series by actually carrying out the sum.)

²⁸René Descartes, *The Geometry*, Book 2, trans. David Eugene and Marcia L. Latham, Encyclopaedia Britannica, Chicago, 1990, p. 544.

²⁹For more details, see C. K. Raju, "The Indian Rope Trick", *Bhartiya Samajik Chintan* 7(4) (New Series) (2009) pp. 265–269. <http://ckraju.net/papers/MathEducation2RopeTrick.pdf>. Also, *Cultural Foundations of Mathematics*, cited earlier, for more details.

That is, with the European way of thinking about mathematics as “perfect”, the problem of neglecting small numbers got converted into the problem of supertasks or convergence. This got related to the difficulty of summing an “infinity of infinitesimals”. It is these philosophical problems (supertasks, convergence) which are deemed to have been resolved in the Western tradition of mathematics today, through formal set theory, which enables the supertasks needed for the definition of formal real numbers and limits to be carried out formally, and which are needed for the current notion of a “fair game”.

To put matters in another way, the philosophical issue which blocks the frequentist interpretation of probability is the Western belief that mathematics is “perfect”, and hence cannot neglect even the smallest quantity, such as 10^{-200} . In practice, of course, this frequentist understanding is commonly used in physics and works wonderfully well with a gas in a box where a “large” number may mean only 10^{23} molecules.

In Western thought, however, what is adequate for physics is inadequate for mathematics, because empirical procedures are suspect and regarded as inferior to metaphysical procedures which are regarded as certain and “ultimate”. This attitude is, in fact, at the core of almost all Western philosophy: deduction, a purely metaphysical process, is believed to be surer than induction, which involves an empirical process.

(To simplify the discussion, we are not getting here into another fine issue: the very division of physical and metaphysical in terms of Popper’s criterion of refutability already involves inductive processes, and their relative valuation, for logical refutability does not guarantee empirical refutability, which latter requires an inductive process. Popper claimed that he had solved the problem of induction,³⁰ since probabilities, defined the Kolmogorov way, are not ampliative and are left unaffected by any new experiments. However, Popper overlooked the fact that empirically one can only obtain *estimates* of probabilities, never the probabilities themselves, and estimates of probabilities *may* be ampliative. So, Popper did not succeed in throwing any new light on the problem of induction.)

Suffice to say that these difficulties regarding “the fear of small numbers” and “inferiority of the empirical” are closely related to the understanding of the “law of large numbers” on the one hand, and, on the other hand, to the notion of “convergence” and “limits” that developed when Europeans tried to assimilate the calculus imported from India in the 16th c. CE.

Zeroism: beyond the clash of epistemologies

The hate politics that prevailed for centuries in Europe during the Crusades and Inquisition had a decisive impact not only on the Western history of science, but also on its philosophy: both had to be theologically correct (on pain of death). The way this was exploited by colonialists and racists is made abundantly clear

³⁰K. R. Popper, *Realism and the Aim of Science*, PS to LscD, vol 1, Hutchinson, London, 1982.

in various ways. For example, Frank Thilly's text on philosophy, used as a text in pre-independence India, starts with the dismissal of all non-Western traditions as "non-philosophy". This belittling of anything non-Western as inappropriate and not-quite-science still represents a typical Western attitude today.

However, if we abandon the current Western historical myths about mathematics in Indian tradition, and look at the tradition *per se*, that gives us a fresh angle on the philosophical problems related to small numbers, and supertasks and the alleged unreliability of empirical procedures, which philosophical problems are intimately linked to the notion of convergence, and the related issue of the frequentist interpretation of probability.

First, let us take note of the differences in the cultural context. Basically, in India, unlike Europe, there was never a monopoly or hegemony of any single religion, and in this pluralistic (and largely secular) Indian environment, where many different brands of metaphysics were prevalent, there was no question of treating any particular brand of metaphysics as certain. Different people may (and did) have conflicting beliefs even about logic from the earliest times.³¹ The Lokāyata rejected deduction as an inferior technique which did not lead to sure knowledge. The tradition in India, since the time of the Buddha, was to decide truth by debate, and not by the forcible imposition of any one brand of metaphysics.

Since metaphysics carried no certitude, the empirically manifest (*pratyakṣa*) was the primary basis of discourse for two persons with conflicting metaphysical beliefs (even about logic etc.). This was the one (and only) means of proof which was accepted by *all* philosophical schools of thought in India. All this is in striking contrast to Western tradition where mathematics is regarded as metaphysical, and this metaphysics is simply declared, by social fiat, to be "universal" and certain like a religious belief. This attitude emerged naturally in the West, where mathematics was linked to religion and theology, and was spiritual and anti-empirical since the days of Plato and Neoplatonists like Proclus. (Indeed the very name "mathematics" comes from mathesis, meaning learning, and learning, on the well-known Platonic/Socratic doctrine, is recollection of knowledge obtained in past lives, and so relates to stirring the soul. Mathematics was thought of as being especially good for this purpose, since it involved eternal truths which moved the eternal soul by sympathetic magic.) It is this belief—that mathematics contained eternal truths—which led to the belief that mathematics is perfect and hence cannot neglect the smallest quantity.

We have seen that the philosophical issue which blocks the frequentist interpretation of probability is just this Western belief in the perfection of mathematics. A number, howsoever small, such as 10^{-200} cannot be neglected and set equal to zero. On the other hand, Indian tradition allowed this to be done, somewhat in the manner in which rounding is done today, but with a more sophisticated philosophy known as *nyavāda* which I have called zeroism,³² so as to emphasize the

³¹B. M. Baruah, *A History of Pre-Buddhist Indian Philosophy*, Calcutta, 1921; reprint Motilal Banarsidass, Delhi, 1970.

³²C. K. Raju, "Zeroism and Calculus without Limits", 4th dialogue between

fruitful practical aspects of the philosophy, and avoid sterile controversies about the exact interpretation of Buddhist texts.

Now it so happens that the formula for the sum of an infinite (*anantya*) geometric series was first developed in India (we find this in the *Āryabhaṭīya Bhāṣya* of the 15th-16th c. Nīlakaṇṭha³³). This naturally involved the question of what it *means* to sum an infinite series. A procedure existed to test the convergence of an infinite series (of both numbers and functions), and I have explained it elsewhere in more detail.³⁴ This is similar to what is today called the Cauchy criterion of convergence, except for one aspect. Thus, given an infinite series of numbers, $\sum a_n$, and an $\epsilon > 0$, the process checked whether there was an N beyond which $\|\sum_{n=N}^{N+m} a_n\| < \epsilon$. The supertask of actually summing the residual partial sums for *all* numbers m could obviously not be carried out except in some special cases (such as that of the geometric series). However, if $S(k) = \sum_{n=1}^k a_n$ denotes the partial sums, the usual process was to check whether the partial sums became “constant”, beyond some N .³⁵ Obviously, the partial sums $S(k)$ would never literally become constant, and when successive terms are added, there would always be *some* change (except when all the terms of the series are zero beyond N , so that the infinite sum reduces to a finite sum). So this “constancy” or “no change” was understood to hold only *up to the given level of precision* (ϵ) in use. That is, the sum of an infinite series was regarded as meaningful if the partial sums $S(k)$ became constant, after a stage, *up to a non-representable* (or discardable) quantity: $|S(N+m) - S(N)| < \epsilon$, which is just the criterion stated earlier. What exactly constitutes a “non-representable” or “discardable” quantity (ϵ) is context-dependent, decided by the level of precision required, and there need be no “universal” or mechanical rule for it.

Apart from a question of convergence, a key philosophical issue which has gone unnoticed relates to *representability*. The decimal expansion of a real number, such as π , also corresponds to an infinite series. Regardless of the convergence of this series, it can be *written* only up to a given number of terms (corresponding to a given level of precision): even writing down the terms in the infinite decimal expansion is a supertask: this is obvious enough when there is no rule to predict what the successive terms would be. So a real number such as π can *never* be accurately represented; Indian tradition took note of this difficulty from the earliest times, with the *śulba sūtra*-s (-500 CE or earlier) using the words³⁶ सविशेष (“with something left out”) or³⁷ सानित्य (*sa + anitya* = “impermanent, inexact”), and

Buddhism and Science, Nalanda, 2008. Draft at <http://ckraju.net/papers/Zeroism-and-calculus-without-limits.pdf>.

³³ *Āryabhaṭīya of Āryabhaṭācārya with the Bhāṣya of Nīlakaṇṭhasomasūtvān*, ed. K. Sambasiva Sastri, University of Kerala, Trivandrum, 1930, reprint 1970, commentary on *Gaṇita* 17, p. 142.

³⁴ C. K. Raju, *Cultural Foundations of Mathematics*, cited above, chp. 3.

³⁵ *ibid.* p. 177-78, and e.g., *Kriyākramakarī*, cited above, p. 386.

³⁶ Baudhāyana *śulba sūtra*, 2.12. S. N. Sen and A. K. Bag, *The Śulbasūtras of Baudhāyana, Apastamba, Kātyāyana, and Mānava*, INSA, New Delhi, 1983, p. 169.

³⁷ Apastamba *śulba sūtra*, 3.2. Sen and Bag, cited above, p. 103. The same thing is repeated in other *śulba sūtra*-s.

early Jain works (such as the *Sūrya prajñāpati*, *sūtra* 20) also use the term *kiñcid viśeśādhika* (“a little excess”) in describing the value of π and $\sqrt{2}$. Āryabhaṭa (5th c. CE) used the word³⁸ आसन्न (near value), which term is nicely explained by Nīlakaṇṭha in his commentary,³⁹ essentially saying that the “real value” (वास्तवी संख्या) cannot be given.

Taking cognizance of this element of non-representability fundamentally changes arithmetic. This happens, for example, in present-day computer arithmetic, where one is forced to take into account this element of non-representability, for only a finite set of numbers can be represented on a computer. Consequently, even integer arithmetic on a computer can *never* obey the rules of Peano’s arithmetic. In the case of real numbers, or floating point computer arithmetic, of course, a mechanical rule is indeed set up for rounding (for instance in the IEEE floating point standard 754 of 1986), and this means that addition in floating point arithmetic is not an associative operation,⁴⁰ so that floating point arithmetic would *never* agree with the arithmetic according to any standard formal algebraic structure such as a ring, integral domain, field etc.

In Indian tradition, this difficulty of representation connects to a much deeper philosophy of *śūnyavāda*. On the Buddhist account of the world, the world evolves according to “conditioned coorigination”. (A precise quantitative account of what this phrase means to me, and how this relates to current physics is a bit technical, and is available in the literature for those interested in it.⁴¹) The key point is that there is genuine novelty of the sort that would surprise even God, if he existed. There is no rigid linkage (no Newton’s “laws”) between present and past, the present is not implicit in the past (and cannot be calculated from knowledge of the past, even by Laplace’s demon). Accordingly, *there is genuine change; nothing stays constant*. But how does one *represent* a non-constant, continually changing entity? Note that, on Buddhist thought, this problem applies to *any* entity, for Buddhists believe nothing real can exist unchanged or constant for two instants, so there is no constant entity whatsoever which is permanent or persists unchanged.

This creates a difficulty even with the most common utterances, such as the statement “when I was a boy”, for I have changed since I was a boy, and now have a different size, gray hair etc. The linguistic representation however suggests that underlying these changes, there is something constant, the “I” to which these changes happen. Buddhists, however, denied the existence of any constant, unchanging essence or soul for it was neither empirically manifest, nor could it be inferred: the boy and I are really two different individuals with some common

³⁸ *Gaṇita* 10, trans. K. S. Shukla cited earlier.

³⁹ *Āryabhaṭīya bhāṣya*, commentary on *Gaṇita* 10, ed. Sambasiva Sastry, cited earlier, p. 56.

⁴⁰ For an example of how this happens, see C. K. Raju, “Computers, mathematics education, and the alternative epistemology of the calculus in the Yuktibhāṣa”, *Philosophy East and West*, 51(3) 2001, pp. 325–62.

⁴¹ The idea is to use functional differential equations of mixed-type to represent physical time evolution. This leads to spontaneity. See C. K. Raju, *Time: Towards a Consistent Theory*, Kluwer Academic, 1994, chp. 5b. Also, C. K. Raju, “Time Travel and the Reality of Spontaneity”, *Foundations of Physics* 36 (2006) pp. 1099–1113.

memories. However, while Buddhists accept the reality of impermanence, there is a practical problem of representation in giving a unique name to each individual at each instant. Consider, for example, Ashoka. No one, not even the Buddhists, describe him as Ashoka1, Ashoka2, and so on, with one number for each instant of his life, which cumbersome nomenclature would require some billion different names even on the gross measure of one second as an atomic instant of time. Therefore, for practical purposes, Buddhists recognize the *paucity of names*, and still use a single name to *represent* a whole procession of individuals. This “constancy” of the representation is implicitly understood in the same sense as the constancy of the partial sums of an infinite series: namely, one neglects some small differences as irrelevant to the context. That is, on the Buddhist view of constant change, the customary representation of an individual, used in everyday parlance, as in the statement “when I was a boy”, can be obtained only by *neglecting* the changes involved (my size, my gray hair, etc) as inconsequential or irrelevant in the context, and which changes are hence discarded as “non-representable” (for the practical purpose of mundane conversation, in natural language).

So, from the Buddhist perspective of impermanence, mundane linguistic usage necessarily involves such neglect of “inconsequential” things, no matter what one wants to talk about. Note the contrast from the idealistic Platonic and Neoplatonic belief. Plato and Neoplatonists believed in the existence of ideal and unchanging or eternal and constant entities (soul, mathematical truths). Within this idealistic frame, mundane linguistic usage (as in the statement “when I was a boy”) admits a simple justification in straightforward sense that change happens to some underlying constant or ideal entity. But this possibility is not available within Buddhism, which regards such underlying ideal entities as fictitious and erroneous, and can, therefore, only speak about non-constant entities, *as if* they were constant. The dot on the piece of paper is all we have, it is the idealization of a geometric point which is erroneous. (Apart from the idealist position, the formalist perspective of set theory also fails, for Buddhist logic is not two valued. But I have dealt with this matter in detail, elsewhere, and we will see this in more detail below.)

Thus, *śūnyavāda* or zeroism provides a new way to get over the “fear of small numbers”. It was, I believe, Borel who raised the question of the meaning of small numbers such as 10^{-200} . On the *śūnyavāda* perspective, we can discard such numbers as practically convenient. (We have nothing better, no “ideal” or “perfect” way of doing things.) We are not obliged to give a general or universal rule for this, though we can adopt *convenient* practices.

What this amounts to is a realist and fallibilist position. All knowledge (including mathematical knowledge) is fallible.⁴² Therefore, when given an excessively

⁴²If mathematical proof is treated as fallible, the criterion of falsifiability would need modification. When a theory fails a test, it is no longer clear what has been refuted: (a) the hypothesis or (b) the deduction connecting hypothesis to consequences. C. K. Raju, “Proofs and refutations in mathematics and physics”, in: *History and Philosophy of Science*, ed. P. K. Sen, PHISPC (to appear).

small number we may discard it, as in customary practice, or in computer arithmetic. (Unlike computer arithmetic, where one requires a rule, with human arithmetic, we can allow the “excessive smallness” of the number to be determined by the context.) It is possible, that this leads to a wrong decision. If enough evidence accumulates to the contrary, we revise our decision. It is the search for immutable and eternal truths that has to be abandoned. Such eternal truths are appropriate to religion not any kind of science.

Thus the traditional Indian understanding of mathematics, using zeroism, dispenses with the need for convergence, limits, or supertasks, and rehabilitates the frequentist interpretation of probability, in the sense that it provides a fresh answer to a long-standing philosophical difficulty in the Western tradition.

SUBJECTIVE PROBABILITIES AND THE UNDERLYING LOGIC OF SENTENCES

Probabilities of singular events

Of course, there are other problems with the frequentist interpretation: for example, it does not apply to single events, for which one might want to speak of probability. The classic example is that of a single footprint on a deserted beach (or the origin of life). There is some probability, of course, that someone came in a helicopter and left that single footprint just to mystify philosophers. But, normally, one would regard it as a natural phenomenon and seek a natural explanation for it.

In this context there is an amusing account from Indian Lokāyata tradition, which is the counterpart of the Epicurean perspective in Greek tradition. Here, a man seeking to convert his girlfriend to his philosophical perspective, goes about at night carrying a pair of wolf’s paws. He makes footprints with these paws. His aim is to demonstrate the fallibility of inference. He argues that by looking at the footprints, learned people will infer that a wolf was around, and they will be wrong. (We recall that the Lokāyata believed that the only reliable principle of proof was the empirically manifest.)

More seriously, such singular events pose a serious problem today in quantum mechanics, where the “probability interpretation of the wave function” is called into play to explain interference of probabilities exhibited by single objects. A typical illustration of such interference is the two-slit diffraction pattern that is observed even when it is practically assured that electrons are passing through the slit one at a time. Understanding the nature of quantum probabilities has become a major philosophical problem, and we describe below some attempts that have been made to understand this problem by connecting it to philosophies and logics prevalent in ancient Indian tradition.

Quantum mechanics, Boolean algebra, and the logic of propositions

In the 1950's there was a novel attempt to connect the foundations of probability theory to the Jain logic of *syādavāda*, by three influential academicians from India: P. C. Mahalanobis,⁴³ founder of the Indian Statistical Institute, J. B. S. Haldane,⁴⁴ who had moved to that institution, and D. S. Kothari,⁴⁵ Chairman of the University Grants Commission. Subsequently, the quasi truth-functional logic used in the structured-time interpretation of quantum mechanics⁴⁶ was connected to Buddhist logic.⁴⁷

To understand these attempts, first of all, let us connect them to the more common (Kolmogorov) understanding of probability as a positive measure of total mass 1 defined on a Boolean σ -algebra (usually of Borel sets of a topological space). The common definition typically requires set theory, as we saw above, to facilitate the various supertasks that are required, whether for the construction of formal real numbers as Dedekind cuts, or as equivalence classes of Cauchy sequences, or for the notion of convergence required by the law of large numbers. However, from a philosophical perspective it is more convenient to use *statements* instead of *sets* (though the two are obviously interconnected). Thus, instead of defining probabilities over measurable *sets*, it is more natural to define probabilities over a Boolean algebra of *statements*. This, incidentally, suits the subjectivist interpretation, for the probability of a statement could then be taken to indicate the degree of (general) subjective belief in that statement (or the objective propensity of that statement to be true, whatever that means).

The immediate question, however, is that of the *algebraic structure* formed by these statements. First of all, we can set aside the specifically σ -algebra aspect, for we have already dealt with the notion of convergence and supertasks above. For the purposes of this section we will focus on the Boolean algebra part. Why should probability be defined over a Boolean algebra?

The answer is obviously that if we have a 2-valued logic of sentences, then a Boolean algebra is what we naturally get from the usual notion of “and”, “or”, “not”, which are used to define the respective set-theoretic operations of intersection, union and complementation. What is not obvious is why these “usual” notions should be used, or why logic should be 2-valued.

Quantum mechanics (and especially the problem of the probabilities of singular events in it) provides a specific *empirical* reason to call the Boolean algebra into question. With probabilities defined on a Boolean algebra, joint distributions of

⁴³P. C. Mahalanobis, ‘The Foundations of Statistics (A Study in Jaina Logic)’, *Dialectica* **8**, 1954, pp. 95–111; reproduced in *Sankhya, Indian Journal of Statistics*, **18**, 1957, pp. 183–94.

⁴⁴J. B. S. Haldane, ‘The *Syādavāda* system of Predication’, *Sankhya, Indian Journal of Statistics*, **18**, 1957, pp. 195–200.

⁴⁵D. S. Kothari, ‘Modern Physics and *Syādavāda*’, Appendix IV D in *Formation of the Theoretical Fundamentals of Natural Science* vol. 2 of *History of Science and Technology in Ancient India*, by D. P. Chattopadhyaya, Firma KLM, Calcutta, 1991, pp. 441–48.

⁴⁶C. K. Raju, *Time: Towards a Consistent Theory*, cited above, chp. 6B, “Quantum Mechanical Time”.

⁴⁷C. K. Raju, *The Eleven Pictures of Time*, Sage, 2003

random variable are assured to exist. This is, however, known to *not* happen in quantum mechanics. (We will not go into details, since our primary concern here is with Indian tradition, and not quantum mechanics. However, this author has explained the detailed relation to quantum mechanics elsewhere, at both a technical⁴⁸ and a non-technical level.⁴⁹) The Hilbert space formulation of quantum mechanics starts with the premise that quantum probabilities cannot be defined on a Boolean algebra, since joint distributions do not exist. The appropriate algebraic structure is taken to be that of the lattice of subspaces (or projections) of a Hilbert space (although there are numerous other opinions about what the exact algebraic structure ought to be).

The usual definition of a random variable as a measurable function actually requires only the *inverse* function, which is a homomorphism which preserves the algebraic structure. In the Hilbert space context, this definition of a random variable as a homomorphism (on a lattice, not an algebra) naturally leads one to identify random variables with projection-valued measures (spectral measures). By the spectral theorem, such measures correspond to densely-defined, self-adjoint operators in this Hilbert space. Since the lattice of projections is non-distributive, these random variables do not admit joint distributions. This corresponds to the more common assertion (“uncertainty principle”) that dynamical (random) variables (self-adjoint operators) which do not commute cannot be simultaneously measured.

To return to the question of logic, unlike in India, where different types of logic have been in existence for over 2500 years, from pre-Buddhist times,⁵⁰ the West took cognizance of the existence of logics that are not 2-valued, only from the 1930’s onwards, starting with Łukasiewicz who proposed a 3-valued logic, where the truth values could be interpreted as “true”, “false”, and “indeterminate”. Could such a 3-valued logic account for quantum probabilities? This question was first investigated by Reichenbach, in an unsuccessful interpretation of quantum mechanics.

The 3 Indian academics mentioned above also interpreted the Jain logic of *syādvāda* (perhaps-ism) as a 3-valued logic (Haldane), and explored 3-valued logic as a philosophical basis for formulating probabilities (Mahalanobis), and interpreting quantum mechanics (Kothari). Haldane’s idea related to perception. With repeated experiments, something on the threshold of perception (such as a sound) may be perceptible sometimes, and sometimes not. In such cases, the “indeterminate” truth value should be assigned to the statement that the “something” is perceptible. Mahalanobis’ idea was that this third truth value was already a rudimentary kind of probability, for it expressed the notion of “perhaps”. Kothari’s idea was to try and explain quantum mechanics on that basis (though he overlooks Reichenbach’s earlier unsuccessful attempt).

⁴⁸C. K. Raju, *Time: Towards a Consistent Theory*, cited above, chp. 6B, “Quantum Mechanical Time”.

⁴⁹C. K. Raju, *The Eleven Pictures of Time*, Sage, 2003.

⁵⁰B. M. Baruah, *A History of Pre-Buddhist Indian Philosophy*, cited above.

Buddhist and quasi truth-functional logic

While Haldane's interpretation is clear enough within itself, it is not clear that it accurately captures the logic of *syādvavāda*. Thus, the Jain tradition grew in the vicinity of the Buddhist tradition (the Buddha and Mahavira were contemporaries). However, Buddhist logic is not 3-valued. For example, in the *Dīgha Nikāya*, the Buddha asserts the existence of 4 alternatives (*catuskoṭi*): (1) The world is finite; (2) the world is not finite (= infinite); (3) the world is both finite and infinite; and (4) the world is neither finite nor infinite.⁵¹

This logic of 4-alternatives does *not* readily fit into a multi-valued truth-functional framework. Especially, the third alternative, which is of the form $A \wedge \neg A$, is a contradiction within 2-valued logic, and difficult to understand even within the frame of 3-valued logic, where it cannot ever be “true”. The reason why 3-valued logic is not appropriate for quantum probabilities is roughly this: in the case of the two-slit experiment, what is being asserted is that it is true that the electron passed through *both* slit A *and* slit B, and not that in reality it passed through only one slit, but we do not *know* which slit it passed through. What is being asserted is that we know that Schrödinger's cat is both alive and dead, as in the third alternative above, and not that it is either alive or dead, but we do not know which is the case.

However, the 3rd alternative of the Buddhist logic of 4 alternatives (*catuskoṭi*) makes perfect sense with a quasi truth-functional logic. The standard semantics here uses the Tarski-Wittgenstein notion of logical “world”, as “all that is the case”. On this “possible-world semantics” one assigns truth values (either true or false) to all atomic statements: such an assignment of truth values represents the possible facts of the world (at one instant of time), or a “possible world”. This enables the interpretation of modal notions such as possibility and necessity: a statement is “possible” if it is true in *some* possible worlds, and “necessary” if it is true in *all* possible worlds (tautology). In fact, Haldane appeals to precisely this sort of semantics, in his interpretation of Jain logic, except that his “worlds” are chronologically sequential. Thus, A is true at one instant of time, while $\neg A$ is true at *another* instant of time—there is nothing paradoxical about a cat which is alive now, and dead a while later. However, this, as we have observed, is not appropriate to model the situation depicted by quantum mechanics.

With quantum mechanics what we require are multiple logical worlds attached to a *single* instant of time. Parallel computing provides a simple and concrete desktop model of this situation, with each processor represented by a separate logical world. The meaningfulness of a quasi truth-functional logic is readily grasped in this situation where multiple (logical) worlds are chronologically simultaneous and not sequential, for this allows a statement to be simultaneously both true and false. That is, with multiple (2-valued) logical worlds attached to a *single* instant

⁵¹Brahmajāla sutta of the *Dīgha Nikāya*. (Hindi trans.) Rahul Sāṅkṛityāyana and Jagdish Kāshyapa. Delhi: Parammitra Prakashan, 2000, pp. 8–9; (English trans.) Maurice Walshe, Boston: Wisdom Publication, 1995, pp. 80–81. For a more detailed exposition, see C. K. Raju, “Logic”, in *Encyclopedia of Non-Western Science, Technology and Medicine*, Springer, 2008.

of time, it is meaningful to say that A is true in one world while $\neg A$ is *simultaneously* true in another. So, a statement may be simultaneously both true and false, without trivializing the theory, or making it inconsistent. From our immediate perspective, the important thing is this: such a quasi truth-functional logic leads on the one hand to an algebraic structure appropriate to quantum probabilities, which structure is not a Boolean algebra.⁵² On the other hand, Buddhist logic (*catuskoṭi*) naturally admits an interpretation as a quasi truth-functional logic. Thus, Buddhist logic (understood as quasi truth-functional) leads to just the sort of probabilities that seem to be required by quantum mechanics.

This quasi truth-functional logic, corresponding to simultaneous multiple worlds, is not a mere artificial and *post-facto* construct imposed on either quantum mechanics or Buddhist thought. From the viewpoint of physics, quasi truth-functional logic arises naturally by considering the nature of time. This is best understood through history. Hoping to make “rigorous” the imported calculus, and the notion of derivative with respect to time required for his “laws”, Newton made time metaphysical (“*absolute, true, and mathematical time*” which “flows equably *without relation to anything external*”⁵³). Eventually, this intrusion of metaphysics and religious belief into physics had to be eliminated, from physics, through a revised *physical* definition of the measure of time; that directly led to the special theory of relativity.⁵⁴ A correct mathematical understanding of relativity, shows that physical time evolution must be described by functional differential equations (and not ordinary differential equations or partial differential equations). The further elimination of the theological understanding of causality in physics makes these functional differential equations of mixed-type. The resulting picture of physical time evolution⁵⁵ is remarkably similar to the core Buddhist notion of “conditioned coorigination”: where the future is *conditioned* by the past, but *not* decided by it. There is genuine novelty. Thus, the relation of the quasi truth-functional logic to the revised notion of time, in physics, parallels the relation of Buddhist logic to the Buddhist notion of “conditioned coorigination” (*paticca samuppāda*). Note that this last notion differs from the common notion of “causality” used in Western thought, with which it is commonly confounded.

Of course, formal Western mathematics (and indeed much of Western philosophy) is likely to be a long-term casualty of any departure from 2-valued logic. In fact, the very idea that logic (or the basis of probability) is not culturally universal, and may not be empirically certain, unsettles a large segment of Western thought,

⁵²See the main theorem in C. K. Raju “Quantum mechanical Time”, cited above.

⁵³I. Newton, *The Mathematical Principles of Natural Philosophy*, A. Motte’s translation revised by Florian Cajori, Encyclopedia Britannica, Chicago, 1996, p. 8.

⁵⁴For an exposition of Poincaré’s philosophical analysis of the notion of time which led to the special theory of relativity, see C. K. Raju, “Einstein’s time”, *Physics Education* (India), **8**(4) (1992) pp. 293–305. A proper clock was defined by *postulating* the velocity of light to be a constant. This had nothing to do with any experiment. See, also, C. K. Raju, “The Michelson-Morley experiment”, *Physics Education* (India) **8**(3) (1991) pp. 193–200.

⁵⁵C. K. Raju, “Time travel and the reality of spontaneity”, *Found. Phys.* **36** (2006) pp. 1099–1113.

and its traditional beliefs about induction and deduction.

This page intentionally left blank

INDEX

1-, 2-, n -random, *see* randomness

A -recursion, 461

a posteriori, 901

abbreviation, 955

Abhandlunguen de Fries'schen Schule,
1159n

Abramovich, F., 602

absolutely normal, 658

acceptability, 474

acceptance rules, 475

acceptance-based inference, 474, 475,
484

accommodates, 984

accuracy, 987, 991–993, 997

accurate prediction, 991, 997

Achinstein, P., 420

acronym, 955

action, 268, 269

acyclic linear causal model, 1011n

ad hoc, 984

ad hoc hypothesis, 395

ad hoc grammar, 868

Adams' thesis, 127

Adams's principle, 465

Adams, E. W., 127, 128, 131, 465

addition rule, 80

additive mixed models, 1131, 1132,
1138, 1142

additivity, 406

adjunction, 1108

adjustable, 989

adjustable parameters, 984, 988

adjusted, 989

admissibility, 292

admissible, 1008

advantage, 299

advantage of rejecting, 299

AFL, 921, 922

AIC, 2, 3, 16, 17, 18, 24, 25, 43, 45,
110, 426, 535, 541, 545, 583,
584–586, 588–590, 592–594,
598–600, 602, 603, 889, 896,
932–934, 946, 994, 996, 1131,
1132, 1135–1138

Airy, G., 1162n

Akaike, H., 535, 541, 583, 592, 593

Albert, J. H., 481

algorithm, 645, 645n, 661–663, 665,
667, 676–677, 681, 684, 687,
690, 706, 708

algorithmic

complexity, *see* complexity

information theory, 648n

probability, 648n, 676, 683, 688

randomness, *see* randomness

algorithmic complexity, 912, 913

algorithmic information theory, 902,
903, 910, 938, 946, 947, 955

algorithmic probability, 912, 939

algorithmic randomness, 34, 38

alien intelligence, 958

Allais paradox, 734

almost sure convergence result, 112

alternative hypothesis, 87

amount of information, 270, 284, 285

analogical method for many proper-
ties, 481

analogical methods, 479

analogy, 762

analogy by proximity, 481

analogy by simlairty, 479

analytical hierarchy, 662, 696

ancient India, 43

ancillarity, 819

ANOVA, 588

- anti-realist attitude, 986
- approximation to the truth, 484
- arbitrary functions
 - method of, 1169
- argument from coincidence, 163
- arithmetic coding, 876
- arithmetical hierarchy, 662, 694, 707
- Armendt, B., 123
- Arntzenius, F., 122, 127
- Ars conjectandi*, 1151, 1152, 1157n, 1170
- artificial intelligence, 901
- Āryabhaṭa, 1179
- assertabilities, 127
- association, 816
- association rules, 1112
- associationist psychology, 1151n
- assumptions, 1128–1130
- asymptotic approximation, 282, 283
- asymptotic behaviour, 281
- asymptotic normality of maximum likelihood estimators, 1048, 1062
- attribution, 1131, 1139
- Austin, J. L., 1149, 1169, 1170
- Australian Football League (AFL), 921, 922
- auto-correlation, 1126, 1134, 1135, 1138
- autonomous, 119
- average error rate, 506
- axiom of regularity, 342
- axiom of sufficiency, 564
- axiomatic basis, 268
- axiomatic foundations, 264
- axiomatic system, 263
- axioms, 269

- Bacchus, F., 122
- background probability, 145, 146, 148
- Bacon, F., 392
- Baird, D., 4, 5, 45
- ban, 950
- Bandyopadhyay, P. S., 32, 44, 92
- Bartlett, P., 603
- Barwise, J., 461

- basic intervals, 649, 651, 667, 668, 673, 691, 693, 704
- basic Ockham efficiency theorem, 1006
- Bayes estimator, 295
- Bayes factor(s), 301, 425, 483, 493n, 584, 594, 607, 608, 612–614, 933, 988
- Bayes factor argument, 989, 1000
- Bayes factor explanation, 1007
- Bayes factor model selection, 889
- Bayes' Information Criterion, *see* BIC
- Bayes' rule, 59, 87, 583, 598, 599, 719, 721
- Bayes' theorem, 103, 139, 234, 255, 272–275, 277, 281, 288, 334–335, 345, 352–359, 363, 366, 371, 373, 375, 378, 395, 415, 416, 498, 537, 572, 599, 715, 726, 901, 929, 1050
- Bayes, T., 392, 448, 455, 752, 770, 772
- Bayesian, 14, 25, 263, 272, 500, 541, 557, 561, 572–574, 576, 577, 901, 913, 987, 990, 1027, 1134, 1135, *see also* Bayes' theorem
 - inference, 756, 766
 - logic, 755
 - statistical inference, 765, 768
 - statistics, 764, 767
- Bayesian agents, 1016
- Bayesian approach, 87, 264, 473, 475
- Bayesian bias, 945, 946
- Bayesian central limit theorem, 1028, 1051, 1052, 1063
- Bayesian confirmation, 6
- Bayesian confirmation theory, 5, 8, 333–387, 391
- Bayesian credence, 1017
- Bayesian credible regions, 293
- Bayesian epistemology, 307
 - objective, 307
- Bayesian expansion, 1020
- Bayesian experimental design, 1020

- Bayesian explanation, 989, 991
- Bayesian explanation of Ockham's razor, 988
- Bayesian inductive logic, 473, 475
- Bayesian inference, 493n, 895
- Bayesian MML, 930
- Bayesian model average, 603
- Bayesian model selection, 25
- Bayesian net, 327, 909, 927, 928, 964
- Bayesian network(s), 109, 909
- Bayesian paradigm, 264, 265, 272, 273, 283, 302
- Bayesian posterior, 932
- Bayesian posterior probabilities, 517
- Bayesian prior(s), 920, 921, 930, 942, 944
- Bayesian statistics, 3, 8, 9, 256, 473, 475, 485, 988
- Bayesian universal distribution, 879
- Bayesian updating, 217, 220
- Bayesianism, 2, 5, 8, 11, 45, 103, 217, 233, 234, 307, 558, 715, 716, 913, 948, 949
- Bayesians, 510, 535, 635
- Bayesian, *see* degree(s) of belief
- beg the question, 994
- beg the question in favor of simplicity, 990
- behavioristic rationale in testing, 157
- belief, 266, 514
 - degree of, 334, 363, 366, 368, 370, *see also* degree of belief
 - formation, 248
 - function, 344–345, 365–368, 370–371
 - strength, 335, 344–345, 363, 365–368, 370–371
- belief-control, 252
- bell-shaped curve, 76
- Bennett, D., 34, 38, 45
- Bennett, J. G., 92
- Berger, J. O., 602
- Berkeley, G., 46
- Berkson, J., 826, 827
- Berksonian bias, 827
- Bernardo's objective Bayesianism, 10
- Bernardo, J. M., 9, 45, 583
- Bernoulli
 - distribution, 1060
 - model, 879–882
 - random variable, 1030, 1035, 1037
 - Bernoulli trial(s), 80, 271, 286, 1037
- Bernoulli, D., 43, 266, 275, 1149, 1152
- Bernoulli, J., 441, 447, 1150–1152, 1154, 1157, 1170
- Berry paradox, 676, 681, 682
- Berry-Esseen theorem, 1030, 1035, 1038, 1052
- Bertrand's box paradox, 255, 713, 714, 716, 718, 719, 722
- Bertrand's paradox, 449, 953
- Bertrand, J., 43, 449–451, 1164, 1165, 1168
- beta distribution, 275, 1044, 1046, 1051
- beta function, 1052
- betting argument(s), 1010, 1013
- betting quotient, 245
- betting strategy
 - capital, *see* martingales
 - gambling, 646, 654–656, 669–671, 671n, 672, 673, 698, 699, 701, 702, 706
 - non-montotonic, 701–702
- Bhaskara II, 1176, 1177
- bias, 546, 825, 990, 992, 993, 1103
- bias-variance trade-off, 992
- biased design, 503
- biasing path, 824, 825
- biasing paths, 826
- BIC, 24, 25, 110, 425, 536, 583, 584, 586–589, 597–601, 603, 617, 889, 939, 940
- BIL, 484
- Billingsley, P., 115
- binary lambda calculus, 681
- binomial distribution, 71, 79, 80, 82, 200, 202, 203, 212, 216, 275,

- 282, 286, 297, 1030, 1034–1037, 1051
- binomial parameter, 1051
- binomial random variable, 1035
- bionic eye, 965
- Birkhoff Ergodic Theorem, 661, 703–704
- Blackwell, D., 117, 483
- bleen, 955
- Blume, J., 11, 12, 14, 19, 45
- Bode's law, 1155n
- Boik, R., 38
- Bolzano, B., 444–449, 451, 453, 454, 456, 458, 459
- Bonferroni adjustment, 506
- Bonferroni method, 1107
- Boole, G., 1160, 1171
- Boolean algebra, 459, 460
- boosting, 950–952
- boosting priors, 951
- bootstrap, 195
- Borel normality, 656–660, 672, 690–692, 704, 706
- Borel's strong law, 641, 648, 655–657, 659, 661, 669, 672, 673, 706
 - of large numbers, 35
- Borel, É, 35, 111, 645n, 656, 658, 1167, 1171
- Borel-Cantelli lemma, 674
- borrow strength, 1075, 1087
- Boulton, D. M., 909, 936
- boundary rule, 940
- breaking strength, 268
- Breiman, L., 603
- Brier scores, 248
- BS, 478, 481, 485
- Buffon, G.-L. L., Comte de, 1151n
- Burckhardt, J., 1168n, 1169
- bus number problem, 932, 933, 947
- butter, 1102
- $c^\dagger$, 454, 455
- C. S. Peirce on relative frequency, 155
- cable guy paradox, 731
- Calas case, 1153
- Calcul de probabilités*, 1164
- calculus
 - Descartes' idiosyncratic reaction, 1184
- calibration, 247, 249–251, 304, 322–325
- Calude, C. S., 655
- cancer, 1077, 1080
- Cantelli, F., 656
- Cantor space, 648–649, 651–653, 667, 691, 703, 705
- Cardano, G., 634, 1150
- cardiac modelling, 966
- Carnap Q -predicates, 454
- Carnap, R., 8, 99, 118, 307, 315, 335, 364, 370, 396, 441, 453, 455–459, 463, 473, 476–481, 481n, 751, 752, 756, 761–764, 772, 1150
- Carnapian logic, 766
- Cartwright, N., 108, 223
- Casella, G., 583
- catch-up phenomenon, 897
- catuskoti*, 1193, 1194
- causal, 958, 959
 - conclusions, 998
 - decision theory, 142, 240
 - diagrams, 822, 828
 - discovery, 997, 998, 1000
 - flips, 1012
 - identification, 817
 - inference, 25, 26, 28, 813, 816, 820
 - laws, 816
 - Markov condition, 109
 - models, 813
 - networks, 1012
 - pathways, 824
 - systems, 822
 - theories, 997, 1012
- causality, 813, 819, 959, 960
- causation, 218, 813, 814
- cause and effect, 815

- cautious monotony, 310
- center of mass, 1123
- central limit theorem (CLT), 38, 64, 79, 80, 82, 86, 1027–1029, 1039, 1045, 1061, 1132
- chain rule, 881
- chaining inferences, 317
- Chaitin's omega (Ω), 691–692, 694–696
- Chaitin, G. J., 36, 635, 645n, 647, 675, 683, 896, 909, 912
- Chakrabarti, A., 24, 25, 45, 592–594
- Chalmers, D., 127
- Champernowne sequence; Champernowne number, 657–660
- chance, 139–142, 146
 - as species of subjective probability, 1169
- characteristic function, 1060, 1062, 1065
- Charnigo, R., 38, 41, 46
- Cherry, S., 32, 44
- Chervonenkis, A., 946
- Chevalier de Méré, 441
- χ^2 test, 202, 203, 205, 207
- χ^2 distribution, 1054
- Chinese room, 956
- Chomsky, N., 866
- choosing, 989
- Christensen, D., 123
- Church's thesis, *see* Church-Turing thesis
- Church, A., 35, 36, 645n, 663, 672
- Church-Kleene, 462n
- Church-Turing thesis, 663, 690, 706–708
- circular, 1021
 - arguments, 1000
- Clasekens, G., 603
- classical interpretation, 106
 - of probability, 713
- classical statisticians, 991, 997
- classical statistics, 3, 256
- classical/error statistics, 45
- classification graphs, 964
- classification tree, 946, 950, 964
- climate change, 1121, 1122
- climate science, 41
- clustering, 965
- code, 865
 - book, 937, 938
 - length function, 875
 - words, 874
- codes and probability distributions, 869
- coding prior, 934, 938, 939
- coding schemes, 947, 949, 950
- cognitive decision theory, 475
- cognitive utility, 475, 484
- coherence, 274, 303, 464
- coherent distributions, 1016
- Coletti, G., 443, 466
- collider, 824
 - bias, 827
- common cause principle, 108
- communal intelligence, 958
- comparisons, 1009
- compatible, 989
- competing hypotheses, 553
- complement of any event, 55
- complement rule, 56
- complete code, 875
- complete coding, 949
- complete ignorance, 408
- completing the code, 949
- complex hypotheses, 560
- complex theory, 986, 989, 990, 1008
- complexity, 585, 635, 637, 866, 994, 988, 996
 - algorithmic, 645, 679, 680n, 681–683, 706
 - equivalent programs, 679
 - finite programs, 677–678
 - infinite, 894
 - Kolmogorov, 647, 648, 673n, 675–688
 - oscillation, 683
 - plain algorithmic (C), 677, 679–681, 683–685, 690, 696

- invariance theorem for, 679
 - optimal program, 679, 682, 685
 - universal, 679
- prefix-free (K), 679, 683–687
 - invariance theorem for, 685
 - optimal function, 685–686, 691, 694
 - universal, 685
- component risk, 1091
- composite hypotheses, 523
- composite likelihood, 519
- compressibility, compressible, 646, 647, 677–680, 686–689
 - b -compressible, 1-compressible, 678, 680, 686
 - compression factor, 677, 678
 - compression programs, 687
 - decompression, 676
- compression, 865, 901, 902, 957
- computability, computable, 645, 648, 665–667, 672, 690
 - for strings, 667
 - real numbers, 653, 692, 698
 - relations, 694
- computable hypotheses, 873
- computably enumerable (c.e.), 665, 667–668, 682, 690, 694, 700
 - martingale, *see* martingales
 - relative, 697
- computably random, *see* randomness, computable
- computational techniques, 985
- computer simulation, 997
- Comte, A., 1153, 1154
- conclusion conjunction, 310
- conditional dependence, 1003
- conditional distribution, 1027, 1043, 1058
 - sampling, 1044
- conditional empirical complexities, 1004
- conditional excluded middle, 128
- conditional expectation, 105
- conditional frequentist, 564
- conditional independence, 105
- conditional likelihood, 519
- conditional measures of uncertainty, 266
- conditional probability, 28, 29, 57–59, 99, 143, 240, 967
 - as a random variable, 115
 - as primitive, 118
 - axioms, 338
- conditionality, 819
- conditionalization, 104, 121, 130, 236, 243, 252, 255, 987
- conditioning, 563, 826
- conditions with probability zero, 111
- Condorcet, M. J. A. N. de, 1149, 1151n, 1153, 1154, 1164, 1171
- confidence interval(s), 1027, 1035, 1037–1039, 1041, 1043, 1055, 1155
 - estimation, 82
 - procedure, 86
 - vs. severity reasoning, 181
- confidence regions, 304
- confidence set, 1092–1094
- confirmation, 11, 109, 474, 483, 484, 517, 984
- confirmation theory, Bayesian, 333–387
- confirmational inference, 474, 475, 484
- conflict, 313
- confounders, 825
- confounding, 817, 824
 - path, 825, 826
- conglomerability, 125, 143
- conglomerative property, 468
- conjecture, 945
- Connor, R. J., 478
- consequence, 269
- conservation law, 996
- conservation theory, 1003
- consistency, 292
 - maintenance procedure, 319
- constrained sample space, 57
- constrained theory, 993
- constraint graph, 326
- context-free grammar, 866

- contingency tables, 481
- continuity correction, 1035, 1038
- continuous random variable, 72, 75
- continuum of inductive methods, 457n
- controlled experiments, 998
- conventional hypothesis testing, 299
- convergence
 - and supertasks, 1185
 - Bayesian, 336, 357, 362, 372–373
 - in distribution, 1029, 1058
 - in probability, 1013, 1029, 1058
 - in the limit, 985, 1021
 - to the truth, 1000, 1003n, 1004, 1005
 - in the limit, 986
- convex, 1082
- Copeland-Erdős number, 658
- Copernicus, 983, 984
- core second-order EP, 318
- correlation coefficients, 206
- corroboration, 483
- Costantini, D., 480, 481n
- countable additivity, 340–341, 412
- countable admissible set, 461
- countable infinity of outcomes, 412
- counterfactual(s), 814, 828, 986
 - prediction, 999
- Cournot, A. A., 43, 1154–1156, 1158, 1159, 1163, 1165, 1171
- Cours de philosophie positive*, 1153
- Cousin, V., 1154
- cover, 313
- coverage, 1037, 1039, 1043, 1055, 1056
- Cox, R. T., 397
- creativity, 967
- credal net(s), 325, 328, 329
- credal network(s), 319
- credal probability, 249
- credence, 143, 144, 146
- credibility, 1150
- credible intervals, 1052, 1056
- credible region(s), 292, 297, 304
- criterion of evidence, 715
- critical reasoning, 64
- critical region, 257
- critical value, 300
- Critique of Pure Reason*, 1158, 1166
- croix* vs. *pile*, 1152
- cross classified populations, 486, 487
- cross-validation, 996, 1108
- cubic regression splines, 1133
- cumulant, 1032, 1053–1055, 1061
 - generating function, 1032, 1060, 1061, 1065
- cumulative distribution function, 1028, 1031, 1035, 1036, 1041, 1043, 1044, 1046, 1052–1055, 1057, 1058, 1066
- cumulative hierarchy, 461
- curve fitting, 17, 19, 21, 110, 486
- curve of best fit, 423
- cut-points, 969
- cutoff subtraction, 1015
- Czech book theorem, 398n
- d'Alembert, J. le Rond, 43, 1152, 1166
- d'Amador, R., 1154
- Darwin, C., 930, 931
- Dasgupta, A., 34, 35, 38, 45
- Daston, L. J., 1151n
- data, 493, 901
 - compression, 885
 - distribution, 64
 - dredging, 527
 - mining, 3, 41, 1099
- database normalisation, 964
- Dawid, A. P., 25, 449
- de Finetti's lottery, 412
- de Finetti, B., 270, 335, 344, 1150, 1169–1171
- De Morgan, A., 1160, 1161, 1164, 1166, 1171
- decidability, 1001
 - decision, 233, 234, 517
 - graphs, 946, 964
 - problem, 268, 295, 298
 - rule, 997
 - theory, 110, 142, 268, 558

- tree(s), 946, 947, 950, 964
- deduction(s), 957
 - regarded as inferior to induction by Lokāyata, 1186
- deductive
 - argument, 93
 - consequence relation, 91
 - inference, 53, 473, 957
 - logic, 1001
 - validity, 53
- defeated, 1012
- defective code, 875
- deferent, 983
- definability, definable, 648, 662, 671n
- definition of empirical simplicity, 1002
- degree(s) of belief, 143, 144, 146, 243, 245, 250, 475, 987, 989, 1016
- degree of freedom, 208
- degree of order, 485
- degrees of falsifiability, 1002n
- degrees of testability, 856, 859
- Delampady, M., 597
- delta method, 1027, 1032, 1033, 1062
- Dempster's rule of combination, 432
- density, 937
- Descartes, R., 43, 392, 1157, 1170
 - idiosyncratic reaction to Indian calculus, 1184
 - cogito*, 989
- description
 - algorithmic or effective, 646, 671n, 676–678, 681, 687–688
 - complexity, 676
 - prefix-free, 683–684, 686
- descriptive data analysis, 67
- descriptive set theory, 667
- descriptive statistics, 70
- desiderata, 516
- desirability, 110
- Destutt de Tracy, A. L. C., 1154
- detection, 1131, 1139
- deterministic, 996
 - questions, 1007
- Diaconis, P., 122
- diagnostic test, 497
- diagnosticity, 106
- DIC, 594
- dice
 - how was the game played in early India, 1180
 - hymn in *Rgveda*, 1180
 - in early India and martingale bets, 1180
 - in *Mahabharata* epic, 1181
 - science of, related to sampling in *Mahabharata*, 1182
 - story of Nala and Rtuparna*, 1182
- Dickson, M., 4, 5, 45
- Die Zeit Constantins des Grossen*, 1168n
- difference between probability and statistics, 62
- difference set, 313
- digamma function, 297
- dilution, 623
- direct inference, 309
 - likelihood, 339, 348–349, 359
 - principle, 988
- direct instruction(s), 957
- direct standardization, 821
- directed acyclic graph, 822
- directed paths, 824
- directional angular data, 964
- Dirichlet distribution, 477–479, 481, 485
- discrete random variable, 72
- discriminative learning, 963
- disjunctions of theories, 1022
- disjunctive beliefs, 1014
- disposition, 139, 140
- distance from knowledge, 1009
- distance from the truth, 484
- divergence, 516, 519n
- divination, 633
- DNA string alignment, 955
- doctrine of chances, 1150, 1151
- domination, 1008
- Donkin, W. F., 1162, 1171
- doomsday paradox, 726

- double sense
 - probability has according to Cournot, 1156
- Dowe, D. L., 22, 46, 110
- dual additive measure, 410
- Dubbins, L., 117, 308
- Duhem, P., 4, 223, 346
- Duhemian problems, 161
- Dupin, C., 1153
- Dutch Book, 5
 - argument(s), 102, 246, 398, 990
 - diachronic, 121, 123
- E. Coli*, 1096
- Earman, J., 335, 373–374
- Easwaran, K., 29, 45, 120
- Eddington, A. S., 583, 586
- Edgeworth expansion, 1053–1055, 1065, 1066
- Edwards, A. W. F., 361
- Eells, E., 108, 366
- effective
 - computability, *see* computability
 - description, *see* description
 - ergodic map, 703–705
 - full-measure, 673–674, 706
 - measure-zero, 673–674, 697–698
 - relative version, 693–694
 - open, 667–668, 673n, 692–693
 - relations, 694–695
 - relative version, 693
 - uniformly, 667–668, 673–675, 693, 697, 698
 - pursuit of the truth, 1000, 1021
 - specifiability, 646–647, 671n, 672
 - topological notions, 667–668
- effects, 813
- efficiency, 935, 1005
- efficient convergence, 1005
- efficient markets, 967
- efficient pursuit curve, 1017
- Einstein's theory of gravitation, 586
- Einstein, A., 583, 586, 587, 930, 931
- elapsed time, 1010
- elimination of nuisance parameters, 277
- Ellis, R. L., 43, 1159, 1160, 1162, 1163, 1171
- Ellsberg's urns, 404
- elimination of dominated alternatives, 994
- Elo system, 964
- elusive model paradox, 960, 961, 963
- empirical
 - adequacy, 1003n
 - Bayes, 593, 1080, 1086, 1087, 1092
 - complexity, 1003, 1004
 - complexity set, 1005
 - effects, 1002
 - estimate, 991
 - laws, 984
 - method, 1004
 - presupposition, 1002
 - risk minimization (ERM), 851, 858
 - strategy, 1004
 - support, 483
 - theory, 1002
 - world, 1004
- endpoints, 504
- ensemble, 1095
 - risk, 1091
- ENSO, 1127–1129, 1136, 1137, 1139
- entailment, 753, 755, 771, 1018
 - non-ampliative, 760
- entropy, 285, 875, 909, 921, 923, 966, 969
- Entscheidungsproblem, 939, 960, 963
- enumerative induction, 391n
- EP-calibrated objective Bayesianism, 327
- EP-OBE, 328
- epicycle, 983
- epistemic
 - consideration, 1002
 - goal, 250
 - justification, 249, 251, 254
 - motivation, 1009

- probability, 475, 476, 1150
 - calibrated objective Bayesian epistemology (EP-OBE), 328
 - reasons, 246–248
 - taste, 256
 - utility, 248
- Epstein, L. D., 481
- equidistribution, 660–661, 704–705
 - Weyl's theorem, 645n, 660, 704
- equipossible cases, 1150, 1151
- equivocation, 322, 324
- equivocator, 324, 325
- ergodic, 660–661, 703–705
- ERM, 853
- error(s), 560, 757, 759, 770, 987, 1021
 - probabilities, 4, 155, 518, 519, 521
 - pre-data vs. post-data, 167
 - type I and II, 167
 - vs. posterior probabilities, 179
- rate, 556, 1125
 - of hypothesis testing, 498
- statistical philosophy, 154, 160
- statistical test, 165
- statistics, 2–4, 11, 153
 - and background knowledge, 159
 - and objectivity, 158
 - and the error statistical philosophy, 153
 - is self-correcting, 189
 - mistakes in inference, 160
 - philosophy for, 165
- ESP, 625
- Essai sur les fondements de nos connaissances*, 1158
- estimated correlation, 997
- estimated distance from the truth, 484
- estimated verisimilitude, 484
- estimation, 64, 82, 91, 265, 294, 536, 752, 756, 759, 761, 763, 764, 766, 886
 - of a proportion, 266
 - problem, 82
- estimator(s), 536
 - (statistic), 83
- Euclidean distance, 1016, 1017
- evaluation, 241
- event(s), 55, 237, 266
- eventual informativeness, 1005
- evidence, 6, 11, 12, 14, 16, 18, 25, 252, 515, 516, 987, 1138, 1140
- evidence function, 14, 15, 516, 524, 571
- evidential decision theory, 142
- evidential framework, 495
- evidential probability, 307, 329
 - second-order, 307
- evidential quantity, 494
- evidentialism, 14, 16, 527
- exchangeability, 270, 271, 457, 458, 762, 763n, 768, 1040, 1059
- exchangeable, 480
 - events, 456
 - multinomial method, 477, 478
- expansions of belief, 1010
- expectation, 110
- expected distance from the truth, 485
- expected loss, 269
- expected retraction times, 1014
- expected squared error, 536
- expected utility, 233, 240, 244, 269
- expected value, 73, 74, 105
- expected verisimilitude, 485, 486
- experiment, 55, 79
- experimental data, 991
- experimental design, 964
- experimental science, 814
- expert probability, 107
- explanation, 903, 912, 930, 931, 956, 966, 984, 1021
- Exposition de la theorie des chances et des probabilités*, 1155, 1158
- expression, 1019
- extra principle, 984
- extraneous variation, 817
- Fabius, J., 478
- factor analysis, 965

- Fagin, R., 455
 fair coin model, 646, 652–656
 fairness, 633, 638
 faithfulness, 109
 fallacy of acceptance, 176
 fallacy of rejection, 168
 arising from a large sample size, 174
 False Discovery Rate, 1107
 false oracle, 946
 falsifiability, 855
 falsification, 849
 Fisher, 158
 Popper, 158
 falsificationism, 221, 858, 859
 falsificationist, 517
 approach, 515
 family of normal distribution, 77
 family-wise probability, 505
 Fan, J., 602
 Feinberg, S. E., 481
 Ferguson, T., 483
 Fermat, P. de, 634
 Fermi-Dirace, 452n
 Festa, R., 8, 9, 45, 473, 475, 478, 479n, 480, 481, 483, 485–487
 fictionalism, 968
 fiducial probability, 760n
 Fieller-Creasy problem, 32, 737, 738, 745, 746
 fine tuning, 958
 Fine, K., 457n
 Finetti, B. de, 1, 6n, 8, 8n, 10, 99, 102, 112, 120, 126, 129, 398, 441–444, 456, 457, 459, 460n, 464–469, 475, 477, 752, 763n, 768
 finite additivity, 412
 finite Gaussian mixture models, 945
 finite mixture model, 935
 finite parameter space, 272
 finite population central limit theorem, 1041
 finite state automata, 955
 first part, 939
 first problem of the priors, 408
 first-order approximations, 281
 first-order effect, 1002
 Fisher information matrix, 282, 289, 880, 942, 943, 953
 Fisher, R. A., 1, 4, 90, 206–210, 212, 214, 221, 347, 361, 475, 634, 752, 755, 759, 760, 763, 766, 772, 818, 819, 1170, 1171
 Fitelson, B., 335, 357, 483
 five types of questions in statistics, 63
 fixed bias, 1000
 fixed-level testing, 89
 flat priors, 291
 Forbes, J. D., 1161, 1162
Formal Logic, 1160, 1161n
 Forster, M. R., 17, 18, 22, 45, 110, 535n, 545, 584, 589, 598, 599
 foundations, 265
 of decision theory, 295
 Fourier coefficients, 591, 592
 Fourier expansion, 591
 Frömmichen, K. H., 1158n
 free logic, 469
 free parameter(s), 991, 994, 999
 Frege, G., 46, 459
 frequency, 1150
 interpretation, 634
 of probability, 73
 limiting, 648, 656, 657, 660, 670n, 669–673, 698, 699, 701–704, 706
 relative, 655, 656, 669, 672
 view, 635–637
 frequentism, 106, 541, 556
 frequentist(s), 11, 263, 510, 536, 554, 556, 557, 563, 564, 577, 635, 642, 646, 668, 669, 670n, 697, 699, 703, 704, 706, 707, 1134, 1135
 calibration, 267

- confidence intervals, 294
- coverage probability, 292
- inference, 493n
- interpretation of probability, 57
- properties of Bayesian reference procedures, 292
- statistics, 154, 475, 895
 - Neyman-Pearson vs. Fisher, 156
 - Bayesian discrepancy, 738
- Freudian psychology, 855
- Fries, J. F., 1159, 1171
- FSMML, 940
- fully consistent disjunctions, 1018
- funnel, 994
- fuzzy disjunction, 1018
- fuzzy propositions, 1019

- Gödel, K., 46, 642, 663
- Gödel number(ing), 459n, 461, 665–666, 679
- Gödel's incompleteness theorem, 648, 682–683
- Gaifman's condition, 322, 341
- Gaifman, H., 107, 341, 373, 460
- Galileo, 595, 596, 634, 1150
- GAM, 1131
- gambling, 238
 - strategy, *see* betting strategy
 - system, 654–656, 670, 692, 699
- games, 633–635
- gaming, 634
- Gamma density, 280
- Gaṇita Sāra Saṃgraha*, 1178
- gap or no gap, 932, 941
- gappy, 932, 941, 947, 957
- Gardner, M., 587
- Garibaldi, U., 481n
- Gaussian, 935
 - competition(s), 920, 922
 - distribution, 70
 - mixture modelling, 965
- Gauvreau, K., 13
- GC-method, 478–481, 485

- gene, 1095, 1096
- generality, 304
- generalization
 - Carnapian method, *see* GC-method
 - cross-validation, 1133
 - of conditionalisation, 253
 - exponential family, 275
 - probability conjunction rule, 60
- generative learning, 963
- geometric, 1088
 - distribution, 1037
 - heuristic, 1081, 1082
 - random variable, 1037
- geometry, 1018
- Gettier, E., 1014, 1015, 1021, 1022
- Ghosh, J. K., 24, 25, 45, 46n, 583, 588, 592–594, 597, 599
- Gibbs sampler, 305
- Gillies, D., 449
- global reliability, 521
- globalistic approach, 477
- Glymour's "bootstrap", 394
- Glymour, C., 366
- God, 991
- Godambe, V. P., 44
- Gödel's incompleteness theorem, 960
- Goldstein, M., 107
- Good, I. J., 478, 481, 483, 483n, 1166
- Goodman's new riddle of induction, 859n
- Goodman, N., 364, 857n, 859n
- goodness of fit, 17, 866, 901
- Gossett, W., 205–207
- Gouraud, C., 1154
- Grünwald, P., 21–23, 46, 110
- grand mean, 1077, 1080, 1084
- graphical analyses, 65
- graphical model(s), 909, 928, 947, 963
- graphical summaries, 65
- gravitational field, 267, 278
- great human intellects, 930
- Greaves, H., 121
- Greenland, S., 25, 26, 28, 46
- Greenwood, M., 43

- Grossman, J., 18, 19, 45
 group invariance, 1166n
 grue, 2, 954, 955, 958
 guarantee
 worst-case, 870
 Gupta, A. K., 481
- Hájek, A., 28, 29, 45, 111n, 112–114, 118, 130, 131, 339, 343, 344
 Hacking, I., 44, 107, 361, 537, 1170, 1178
 Halayudha, 1175
 Haldane, J. B. S., 1191
 Hall, N., 130, 131
 Halpern, J. Y., 459n
 halting
 probability, 691–692
 relative, 694
 problem, HALT, 666–667, 939, 960, 961, 963, 967
 happiness-conducive, 1002
 hard choice, 1012
 Harman, G., 20, 46
 Harper, J., 43, 46
 Hartley, D., 1151n
 Hausman, D., 109
 Hawthorne, J., 5, 6, 8n, 11, 14, 45, 366, 368, 370, 376, 378, 381–383
 HC, 461
 heavy-tailed distribution, 79
 Hegel, 1158
 Hegelianism, 576
 Hellinger distance, 593
 Hempel's "satisfaction" criterion, 394, 392
 Hempel, C., 92, 109, 392
 Herbrand, J., 663
 hereditarily countable sets, *see* HC
 hereditarily finite sets, *see* HF
 Herschel, J. F. W., 1160–1162, 1171
 Hesse, M., 481
 HF, 461
 hierarchical classification, 947
- Higgs boson, 966, 969
 high variance, 79
 Hilbert's program, 663
 Hild, M., 121
 Hildebrand, D. K., 487
 histograms, 66
 history of science, 983
 history of the normal distribution, 76
 Hjort, N. L., 603
 Hoeting, J., 603
 Holm-Bonferroni method, 1107
 Howson, C., 8, 10, 45, 335, 368, 450, 467n, 536
 HPD regions, 297
 Huffman code, 903, 913, 938
 construction, 903, 905
 Hume, D., 2, 43, 392, 814, 1151n, 1154, 1158
 humour, 967
 Humphreys' paradox, 139, 140, 142
 Humphreys, P., 138–141
 Huygens, C., 441, 634
 hybrid Bayesian network, 947, 963
 hyperarithmetical (Δ_1^1), 463n
 hyperarithmetical hierarchy, 696
 hypothesis, 199, 201n, 203, 204, 206, 207, 210, 214, 217, 220, 223, 517, 751, 755, 757–761, 763–772, 865, 901
 testing, 64, 87, 90, 265, 294, 298, 499, 505, 738, 919, 964, 1106
 hypothetical induction, 394
 hypothetico-deductive framework, 514
 hypothetico-deductive reasoning, 225
- I_{1D} , 940
 ideal agents, 254
 ideal codes, 876
 Ideal Group (IG), 940, 941
 ideal MDL, 872
 idealizing condition, 486
 IG, 940, 941
 ignorability, 820
 strong, 820

- weak, 821
- ignorance prior, 411
- IIP, 963
- image recognition, 947
- importance sampling, 304
- imprecise hypothesis testing problem, 742
- improper, 117
 - distribution, 450
 - prior, 607, 611, 614, 618
 - function, 274
- inaccessible truths, 1003n
- inclusion bias, 827
- incompressibility, incompressible, 38, 646–647, 681, 686, 687, 689, 705–707
 - b*-incompressible, 680, 681, 686–689
- incremental measure of confirmation, 483
- indefeasibility accounts of knowledge, 1001
- independence, 104, 143
- independent and identically distributed (i.i.d.), 1109
- independently confirmable principles, 984
- indexical beliefs, 121
- indifference, 989
- indirect inference, 309
- induction, 1, 903, 912, 930, 931, 956
- inductive acceptance, 484, 485
- inductive argument(s), 55, 200
- inductive disjunctive fallacy, 410
- inductive generalisation, 394
- inductive inference, 53, 91, 93, 473, 475, 896, 901, 930, 931, 956, 985, 1099
- inductive inquiry, 1008
- inductive logic, 8, 9, 473, 474, 484, 487, 751–754, 756, 769, 772, 1001
 - ampliative, 755
 - Bayesian, 769, 771
 - Carnapian, 761, 762, 764, 767–769
 - non-ampliative, 763
 - programming, 963
- inductive probabilities, 473
- inductive rules, 480
- inductive-statistical explanation, 109
- inference, 62, 70, 751, 752, 754, 755, 759, 771, 813, 903, 912, 930, 931, 956, 962, 966
 - ampliative, 758n
 - Bayesian, 764, 769, 770
 - inductive, 761, 763
 - statistical, 772
- inference about uncertainty, 92
- inference on normal parameters, 293, 298
- inference procedure, 557, 558, 564
- inference summary, 265, 294
- inferential analysis, 67
- inferential rational in testing., 158
- infinitary language, 460
- infinitesimal probability, 111n
- information, 263, 572
 - complexity, 676, 682, 683, 689, 706
 - content, 675, 680, 680n, 681, 692
 - criteria, 523, 524
 - theory, 284, 648n, 680n, 682, 869, 902, 903, 936
 - Shannon's (entropy), 680n
 - theoretic, 264, 296, 902, 913, 936
- informationless priors, 8
- instantaneous code, 907
- integral representations, 271
- intelligence, 901, 930, 956–958
 - testing, 201, 218
- interpretations of probability, 55, 106, 264
- intersection, 56
- interval(s), 64
 - estimation, 87, 292, 297, 485
 - scale, 65
 - of values, 404

- intervening vs. conditioning, 140
- intervention(s), 987, 999
- intrinsic
 - classification, 965
 - credible regions, 298
 - discrepancy, 296, 300
 - estimation, 296
 - estimator, 287, 296
 - hypothesis testing, 299, 301
 - limit, 285
- invariance, 286, 296, 297, 409, 932, 953
- invariant minEKL, 954
- invariant prior(s), 618, 953, 954
- invariantised density, 609, 612, 615
- inverse-probability weighted, 821

- James-Stein estimation, 947, 1077, 1079, 1080, 1086, 1087, 1091, 1093
- Jaynes, E. T., 126, 325, 335, 397, 446, 450, 451
- Jeffrey, R. C., 110, 335, 344, 367
- Jeffrey conditionalization, 122, 123, 130
- Jeffrey's rule, 432
- Jeffrey's measure, 607, 609, 612, 613, 615, 618, 619, 621–624, 626
 - conditional, 618
 - marginal, 618
- Jeffreys, H., 99, 118, 335, 475, 1150, 1166, 1170, 1171
- Jeffreys' prior, 286 949, 954
- Johnson, W. E., 99
- joint probability mass function, 75
- Joyce, J., 335, 344, 593
- Judy Benjamin problem, 125
- justification, 985, 987, 1009
 - of Ockham's razor, 1021
- justifying Bayesian updating, 990

- K -triviality, 697
- Kadane, J. B., 117, 126
- Kahneman, D., 433
- Kant, I., 43, 1158, 1166, 1167
- Kelly, K., 23–25, 46, 858–860

- kernel, 274
- Keynes, J. M., 99, 118, 308n, 335, 364, 403, 451, 452, 454, 459, 1150, 1159, 1167, 1171
- king of Siam, 1165
- KL distance, *see* Kullback Leiber (KL) distance
- KL divergence, *see* Kullback Leiber (KL) divergence
- Kleene, S., 663, 673n
- knowledge discovery, 1099
- Knuth, D., 645n, 661
- Kolmogorov complexity, 21, 36, *see* complexity, 866, 873, 896, 902, 910, 912, 938, 939, 946–948, 955
- Kolmogorov's axiomatization, 100, 118
- Kolmogorov, A. N., 28, 29, 36, 99, 111, 115, 117, 120, 132, 457, 463, 635, 645n, 647, 656, 668, 669, 673n, 675, 680n, 896, 909, 912
- Koltchinski, V., 603
- Kothari, D. S., 1191
- Kraft's inequality, 691, 875, 907
- Krauss, P., 341, 460, 461, 463
- Kriyākramkarī*, 1176
- Kuhn, T., 447, 814
- Kuipers, T. A. F., 478, 480
- Kulkarni, S., 20, 46
- Kullback-Leibler, 265, 909, 914, 922–924, 927, 931, 933–935, 941, 952–954
- Kullback-Leibler discrepancy, 1132
- Kullback-Leibler distance, 17, 18, 611, 924
- Kullback Leiber (KL) divergence, 22, 518n, 519, 524, 592, 593, 599, 924, 925, 1017, 1018 599
- kurtosis, 1046, 1061
- Kyburg, H. E., 8, 92, 307, 309, 315, 317, 322, 349, 442, 443, 469

- $L(\omega_1^{CK})$, 462, 463n

- La Harpe, J.-F., 1154
 Laing, J. D., 487
 Lambalgen, M. van, 648, 669, 692, 694, 696, 698
 Lange, M., 123, 370
 language, 955, 958
 Laplace central limit theorem, 1028–1030, 1032, 1035, 1038, 1041
 Laplace method, 1064
 Laplace, P. S., 43, 446, 455, 634, 644, 654, 675, 676, 681, 687, 688, 1149, 1151n, 1153–1155, 1160, 1161, 1163, 1168, 1171
 large p, small n, 1095
 Laurent, H. H., 1164
 law of
 errors, 200
 frequencies, 703–704
 iterated logarithms, 659–660, 673, 691
 large number(s), , 373, 739. 1029, 1063, *see* Borel strong law and fear of small numbers, 1185
 law of likelihood, 15, 18, 357–361, 496, 501, 517–519, 537, 540, 541, 574
 of probability, 240, 248
 of nature, 955
 randomness, 656, 661, 673
 symmetric oscillations, 659, 672, 703
 law of total probability, 59, 103
 learning, 901
 process, 273
 theoretic paradigm, 399
 least square(s), 1073, 1088
 difference, 486
 Lebesgue measure, 648–653, 656, 692, 703, 705n
 Leblanc, H., 118
 left classical equivalence, 310
 Lehmann, E. L., 583
 Leibniz, G., 43, 1150, 1152, 1158
 Lele, S., 14–16, 45
 Letters on Probability, 1162
 level of confidence, 86, 87
 Levi, I., 113, 121, 123, 308n, 335, 349, 475
 Levin, L. A., 648n, 683
 Lewis, D., 121, 127–130, 335, 348, 400, 465, 1169
 Li, K. C., 584, 602
 likelihood, 103, 333–336, 339, 345, 347–387, 901, 1046, 1050, 1051, 1066, 1131
 based statistics, 3
 framework, 11, 14
 function, 11, 274, 497, 517, 561, 569, 571–578, sub 1029, 1048, 1051, 1056, 1064, 1067
 interval, 540
 paradigm, 496, 505
 principle, 3, 18, 183, 185, 274, 360–361, 497, 518n, 519, 553, 554, 557, 561–563, *see* LP, 564, 566–569, 571, 573–578, 771, 966
 weak vs. strong, 185
 ratio, 355–358, 361–363, 365, 369, 371–375, 377–387, 496, 497, 505, 517–520, 522, 523, 525, 528, 574
 Convergence Theorem, 357, 362–363, 369, 371–384, 387
 theory of evidence, 421
 likelihoodism/ists, 2, 11, 45, 357–361, 371, 421
 Līlāvātī, 1176
 limited information, 291
 limited translation estimator, 1091
 limiting frequency, *see* frequency
 Lindeberg-Feller theorem, 1031, 1067
 Lindenbaum algebra, 460
 Lindley's paradox, 32, 34, 302, 737, 738, 741–744
 Lindley, D. V., 481
 linear, 1131
 causal network, 1003

- combination, 79
 - model, 1043
 - regression, 593
 - transformation, 77
 - trend, 1125, 1140
- linguistic propositions, 515
- linguistic theory, 516
- LNPP-space, 936
- LNPPP, 960
- local reliability, 521, 522, 527, 528
- location model, 291
- Locke, J., 1151n
- log-likelihood, 584, 901, 931
- log-loss, 916, 952
 - compression-based competition, 922
 - scoring, 914, 928
- logarithmic divergence, 265, 281, 284
- logarithmic loss, 889, 916
- logic
 - 3-valued, 1192
 - and *syādavāda*, 1192
 - Buddhist, 1191
 - Buddhist logic as quasi truth-functional, 1193
 - Buddhist logic leads to quantum probabilities, 1194
 - catuṣkoṭi*, 1193
 - Jain, 1191
 - Jain logic as 3-valued, 1192
 - not culturally universal, 1194
 - of 4-alternatives and Schrödinger's cat, 1193
 - of 4-alternatives as quasi truth-functional, 1193
 - pre-Buddhist, 1192
 - quasi truth-functional, 1191
 - quasi truth-functional and quantum mechanics, 1191
- Logic of Chance*, 1159, 1160, 1162, 1163
- logical consistency, 1010
- logical interpretation, 106
- logical probability, 144–146, 1150
- logical probability function, 456
- logical retractions, 1020
- logically independent, 61
- logistic regression, 1049
- logit function, 1049
- long run, 990
 - convergence, 990, 997n
- loss, 1076, 1079, 1080, 1082, 1083, 1086, 1092, 1093, 1095
 - entropy, 1095
 - function, 10, 295, 398n, 560
- lot, 633
- lottery, 317, 495, 633, 636
 - paradox, 204, 722, 1108
- lowest posterior loss, 298
- LP, 567, 569, 572
- luckiness, 869, 887
 - function, 897
 - principle, 871, 872, 877
- Lugosi, G., 603
- M_G , 521–523, 526
- M_L , 521, 523, 525
- Martin-Löf randomness, *see* randomness, Martin-Löf-
- Martin-Löf, P., 645n, 646, 653, 673, 673n, 674, 675, 689, 696
- Martin-Löf-Chaitin thesis, 649, 690, 705–708
- $m^\dagger$, 454
- Méré, C. de, 634
- machine learning, 901, 1021, 1099
- MacQueen, J. B., 483
- magnetic field, 934
- Mahabharata*, 1181
 - notion of fair gambling in, 1181
 - relation of dice to sampling theory, 1182
- Mahalanobis, P. C., 1191
- Maher, P., 123, 335, 481, 485
- Makinson, D., 443n
- Mallow's statistic, 996
- manipulation(s), 813, 814
- MAP, 939, 941

- margin of error, 1093, 1094
- marginal
 - effects, 820
 - likelihood, 525, 607–609, 618, 621
 - mass functions, 75
 - probability, 57, 58, 901, 937, 953
- marginalization, 288, 290
 - marginalization paradox, 622
- Markov chain(s), 874
 - Monte Carlo (MCMC), 304, 940
- Markov condition, 319
- Markov decomposition, 823
- Markov model, 962
- marksman, 991, 992
- Martin-Löf, P., 36
- Martin-Löf-Chaitin thesis, 38
- martingales, 38, 673, 697–703, 706
 - computable, 700–701
 - computably enumerable, 700, 707
 - fairness condition, 698, 699
 - partial computable, 700–701
- Marxism, 855
- material theory of induction, 400
- material vs. formal view of logic, 1163
- Maximum A Posteriori (MAP), 934, 939, 941
- maximum entropy, 285, 411
- maximum likelihood, 574, 575, 584, 752, 755, 759–761, 764, 766, 932–934, 941, 947, 952, 953, 964, 1073–1077, 1079, 1080, 1082–1084, 1087, 1088, 1090–1092, 1096, 1134, 1135
 - estimation, 9, 281, 423, 557, 772, 991, 993, 1048, 1051
- maximum posterior expected entropy, 1166n
- Mayo's test severity, 521
- Mayo, D., 4, 11, 16, 19, 45
- McCarthy, D., 127
- McGee, V., 119, 132
- McMichael, A. F., 483
- MDL, 895, 944, 948–950, 964
- mean, 65, 66, 1123
 - squared error (MSE), 992
- measurable, 650–652, 657, 703–704
 - bijection, 653
- measure(s)
 - of critical tendency, 65
 - of dispersion, 68
 - of evidence, 493
 - zero, 650, 651, 651n, 654
 - full-, 653, 654, 696
 - Lebesgue, *see* Lebesgue measure
 - preserving, 703–704
- measurement error, 515, 828
- measurement of a physical constant, 267
- median, 65, 66, 1123
- medieval scholastic terminology, 1169
- Mellor, H., 1150
- memory, 957
- Mendel, G., 208–210, 212
- message length, 958
- meta Jeffreys measure, 613, 615, 621, 626
- metaphysical argument, 989
- method, 985
- methodological principles, 1004
- Metropolis algorithm, 305
- Michell, J., 1161, 1164
- microarray, 1095
- Mill's ratio, 1064
- Mill, J. S., 43, 392, 1159, 1160, 1163, 1164, 1171
- mimimal sufficient statistic, 1043
- minEKL, 932, 933, 940, 941, 953, 957
- minimal cover under difference, 313
- minimal sufficient statistic, 1060
- minimally sufficient test statistic, 259
- minimax, 996, 1008
- minimax optimal rate, 887
- minimax regret, 996
- minimizing MSE, 996
- minimum description length (MDL), 21, 865, 902, 928, 940, 944, 948, 949, 996
 - inference, 110

- minimum message length (MML), 21, 896, 901, 928, 953, 956, 962
 - inference, 110
- minimum sufficient statistic, 939
- miracle, 967
- Mis-Specification (M-S) testing, 190
 - a parametric test of independence, 192
 - and the use of the residuals, 194
 - the runs test, 191
 - vs. Neyman-Pearson (N-P) testing, 190
- Mises, R. von, 35, 36, 634, 635, 642, 645n, 646, 648, 654, 668–670, 670n, 671, 671n, 672, 699, 703–707, 1150
 - and randomness, *see* randomness, von Mises-
- misleading evidence, 13, 494, 503, 504, 518, 520, 522, 525, 526
- missing data, 820
- missing information, 285
- misspecification, 935
- mixed strategies, 1013
- mixture modelling, 928, 965
- ML plug-in code, 883
- MLE, 993
- MML, 929, 930, 933–936, 940, 941, 944–948, 950, 951, 953–956, 958, 962–964, 966–968
 - subitem estimator, 932
 - subitem mixture modelling, 947
 - subitem -like, 949
- MMLD, 940
- mnemonic, 968
- mode, 65
- model(s), 17, 514, 560, 869, 874, 991
 - averaging, 962
 - building, 813
 - complexity, 594
 - misspecification, 934, 944, 945, 949
 - order selection, 949
 - robustness, 518
 - selection, 3, 24, 25, 524, 874, 876, 887, 997
 - consistency, 888
 - vs. statistical model specification, 195
 - selection bias, 527
 - specification, 813
 - switching, 897
 - validation, 189
- modus ponens, 957
- Moivre, A. de, 634
- moment generating function, 1090
- monkey, 1101
- monotonicity, 53, 54, 1001
- monotonic reasoning, 55, 91, 93
- Montague, R., 459n
- Monte Carlo resampling, 1107
- Montmort, P. R. de, 441
- Monty Hall problem, 30, 32, 123, 720, 721
- Moore's paradox, 107
- Moore, J., 43, 46
- Morgan, A. de, 43
- Mosimann, J. E., 478
- motif de croire, 1154
- MSE, 993, 994, 996, 997
- muddy Venn diagram, 416
- Mukhopandhyay, N. D., 593
- multicategorical experiment, 480, 481
- multinomial inference, 476, 477
- multinomial method, 476, 479–481, 485
- multinomial process, 476, 479–481
- multiple comparisons, 493, 495, 504, 526
- multiple decision makers, 561
- multiple hypothesis testing, 1106
- multiple looks, 493, 495
- multiple testing, 1125
- multiplication, 80, 415
 - rule, 59, 60
 - rule of probabilities, 466
- multiplicity, 524
- multiply robust, 822

- multivariate central limit theorem, 1030, 1031, 1063
- Mura, A., 465
- mutual independence, 60, 61
- mutually exclusive events, 56, 61, 72
- mutually independent events, 80
- n*-randomness, *see* randomness
- narrowness, 415
- nat, 907, 950
- natural ban, 907
- negative binomial, 271, 1036, 1038
 - distribution, 1034, 1037, 1038
 - random variable, 1038
- nested model, 584, 588
- Neumann, J. von, 643
- neural nets, 964
- neutrality, 478
- Newcomb's data, 84
- Newcomb's measurement, 82, 83
- Newcomb's problem, 32, 729, 730
- Newton, I., 930, 1002, 1008, 1194
- Neyman vs. Carnap, 154
- Neyman, J., 1, 90, 214, 221, 475, 634, 1171
- Neyman-Pearson, 523
- Neyman-Pearson (N-P) test, 9, 167, 752, 755–759, 763, 769, 772
- Neyman-Pearson test sizes, 521
- Neyman-Scott, 932, 933, 941, 947, 957, 964
 - panel data, 947
- Nicod's principle, 254, 255
- Niiniluoto, I., 473, 475, 480, 481, 484n, 485, 486
- nit, 907, 950
- NML, 948, 949, 954
- no short path(s) assumption, 1011, 1015n
- noise, 902, 1103
- noiseless coding theorem, 875
- nominal, 64
 - scale, 65
- non-additive measures, 407
- non-Bayesian shifts, 417
- non-circular epistemic argument, 1008
- non-conglomerability, 125, 126
- non-experimental data, 1000
- non-human intelligence, 958
- non-measurable sets, 114
- non-metric scale, 65
- non-monotonic, 1001
- non-universality probability, 913
- non-identification, 814, 815, 828
- non-ignorability, 817
- non-informative prior, 283
- non-informative priors, 275, 411
- non-linear, 1131, 1138
 - trend, 1140
- non-monotonic logic(s), 311, 326, 329
- non-monotonic reasoning, 54, 91, 93, 1111
- non-parametric regression, 591, 592
- non-randomized studies, 820
- non-rejection, 257, 258
- normal, 289
 - approximations, 38
 - distribution, 70, 71, 76, 84, 200–202, 203n, 206, sub 1073–1075, 1081, 1087, 1088, 1090, 1092, 1095
 - Ockham method, 1005
 - Ockham strategies, 1008
 - parameters, 277
 - probability, 1091
- normalised maximum likelihood, 880, 949
- normative axioms, 239
- norms of inference, 557
- Norton, J. D., 7, 8, 11, 25, 45
- nose, 993
- notational brevity, 984
- novel policies, 997
- nuisance parameter(s), 277, 283, 518n, 525
- null hypotheses, 87, 89, 201, 203, 207, 210–214, 216, 218, 220, 221, 223, 553, 816

- in testing, 155
- multinomial distribution, 1096
- $O(1)$, 913
- objections to the over-fitting argument, 996
- objective/objectivity, 263, 265, 284, 303, 988
 - chance, 991, 1169
 - constraints, 990
 - invariant prior, 954
 - priors, 953, 954
 - probability, 236, 988
- objective Bayesian/ism, 9, 314, 557, 950, 953
 - epistemology, 322, 329
 - methods, 281
- observational science, 813, 826, 828
- Ockham efficiency theorem, 1007, 1018, 1014, 1016, 1021
- Ockham violator, 1009
- Ockham's razor, 24, 585, 858–860, 866, 891, 903, 928–930, 956, 983, 985–989, 991, 994, 996, 999, 1000, 1004, 1009, 1105
- old evidence, 253
- one standard deviation of the mean, 70
- one-part message, 936
- one-tailed significance test, 211
- open sets, 649–651
 - effective, *see* effective
- operational meaning, 304
- optimal efficiency, 987
- optional stopping, 186
- options, 236, 237
- Oracle, 584, 592, 600, 601
- Order 1, 913
- order of magnitude, 1058, 1059
- order statistic, 1043–1046, 1060
- ordinal, 64
 - scale, 65
- originality, 967
- orthodox statistics, 475
- outcome, 237, 245
- over-aim, 993
- over-fitting, 41, 868, 869, 993, 997, 1007, 1102
 - argument, 996, 1000
- P -value, 90
- p -value(s), 166, 213, 216, 222, 493, 495, 499, 500, 506, 517, 520–522, 739, 741, 742
 - vs. posterior probability, 179
- P-atom, 129
- Pagano, M., 13
- pairwise independence, 61
- panel data, 933, 941, 944
- paradox of indifference, 989
- paradox of the raven, 392
- parameter, 264, 270, 271, 989
 - estimation, 759
- parametric complexity, 880
- parametric regression, 591
- Pareto, 1009
- partial
 - assignment, 966
 - computable, 665–666, 668, 685, 686
 - computably random, *see* randomness, partial computable
 - correlation, 1003
 - order, 402
- particle physics, 966
- Pascal, B., 441, 634
- path-independent, 1020
- paticca samuppāda*, 1194
- Pāṭigaṇita*, 1176, 1178, 1180
- Pauler, D. K., 602
- PCCP, 128, 129
 - conditional, 129
 - function, 129
- PDO, 1127–1129, 1136, 1137, 1139
- Pearl, J., 124
- Pearson, E. S., 90, 206, 207, 214, 215, 221, 475, 634
- Pearson, K., 202, 203, 205, 207, 208

- Peirce, C. S., 1163, 1171
 penalized likelihood rule, 584
 penalized log-likelihood, 585
 pendula, 1002
per datum predictive accuracy, 545
 perfect illusion, 1003n
 permutation central limit theorem, 1045
 permutation sampling distribution, 1047
 permutation test, 818, 1043–1045, 1060
 permutation-invariant, 446
 permutations and combinations
 and Indian music, 1175
 and large numbers, 1176
 continuous tradition from *Vedas*
 to Bhaskara II, 1176
 Piñgala's rule and binomial ex-
 pansion, 1175
 typical formula for combinations,
 1177
 Peter principle, 969
 Phillips Information Criterion, 949
 philosophical can of worms, 895
 philosophy, 1021
 of statistics, 1, 2, 46, 47
 phrase structured grammars, 866
 physical probability, 1150
 Π_1^1 , 463n
 PIC, 949
 Piñgala, 1175
 pivotal inference, 557
 pivotal quantity, 1027, 1055
 place selection (rule), 669–672, 698–
 699, 701–706
 admissible, 671–672, 706
 computable, 672
 partial computable, 672
 plain algorithmic complexity, *see* com-
 plexity
 planatory power, 984
 planetary astronomy, 983
 Plato, 392
 plausibilistic inference, 474, 475, 484
 plausibility, 269, 474
 Poincaré, H., 1168, 1169, 1171
 Poinsoot, L., 1153
 point estimate, 949
 point estimation, 82–84, 292, 295, 485
 Poisson distribution, 965
 Poisson, S.-D., 43, 271, 280, 441, 1081,
 1149, 1153, 1155–1157, 1159,
 1171
 Poisson-Gamma mixture, 280
 policy predictions, 987
 political survey, 267
 polynomial
 curve, 984
 degree, 986, 1012
 law, 1010
 regression, 947, 964
 theories, 1012
 Popper dimension, 856, 857, 859
 Popper function(s), 29, 119, 122
 Popper, K., 20, 29, 99, 118, 221, 225,
 432, 473, 483, 484, 849, 854–
 856, 858–860, 1150
 fallacy in his solution of problem
 of induction, 1185
 theories of verisimilitude, 484
 population
 distribution, 80
 mean, 70, 80
 standard deviation, 84
 variance, 69, 80
 Port Royal logic, 442
 possible observations, 561
 possible world, 968
 post data, 522, 527
 Post machine, 662–666, 681–683
 instruction set, 663
 programs, 663–666, 668, 676–677,
 681–683
 universal, 666
 Post, E., 663
 post-data reliability, 519
 posterior
 cumulative distribution function,
 1056
 density, 1050, 1052, 1053, 1056,

- 1057
- distribution, 274, 275, 929, 1027, 1050–1053, 1056, 1057, 1064, 1067, 1086, 1087, 1169 *see also* posterior probability
- mean, 295, 953
- mode, 295
- model probabilities, 607, 609, 615
- odds, 584, 612, 614
- probability, 234, 252, 347, 351–359, 361–364, 369, 371–373, 375, 379, 387, 477, 493n, 500, 572, 584, 765–768, 770, 771, 901, 933 *see also* posterior distribution
- quantile, 295
- ratio, 988
- potential outcome, 814, 815, 828
- potential probability statement, 312
- power, 526, 757, 759, 770
 - of a test, 167
- practical certainties, 314
- pragmatic
 - considerations, 1002
 - instrumentalist, 997
 - motivation, 1009
 - reasons, 246
 - virtues, 986
- pre-data, 522
- precise fractions
 - comparison of Indian and Roman numeration, 1179
 - in late Greek tradition and Ptolemy, 1179
- precision, 401
- prediction, 11, 17, 20, 28, 267, 280, 813, 814, 869, 903, 930, 931, 962, 966, 983, 993, 1000
 - logic, 487
- predictive
 - accuracy, 17, 18, 20, 535, 541, 913, 952, 986, 1021
 - distribution, 280, 615
 - distributions, 279
 - inaccuracy, 996
 - inference, 279
 - MDL, 881, 889
 - probability, 478–481
- predictivistic approach, 477
- preference(s), 236, 238, 245, 269
 - formation, 242
- prefix code(s), 870, 874, 907
- prefix-free
 - coding, 683–684
 - complexity, *see* complexity set, 683–685, 691, 692
- premise disjunction, 310
- prequential forecasting system, 882, 889, 890
- prequential model validation, 890
- preserving content, 1020
- prevalence, 273, 498
- principal components, 1049, 1050
- principal principle, 107, 138, 140, 144, 245, 348–349, 400
- principle of direct probability, 107
- principle of indifference, 112, 408, 458, 713–718, 720, 767, 1166n
- principle of parsimony, 22, 585, 866
- principle of positive instantial relevance, 457n
- principle of total evidence, 772
- prior bias, 987
 - toward simplicity, 986
- prior degree of belief, 987
- prior density, 1067
- prior distribution, 275, 478, 479, 481, 577, 1051, 1086, 1087
- prior odds, 613
- prior probability, 234, 252, 335, 336, 347, 350–365, 370–373, 387, 477, 498, 558, 572, 576, 584, 760n, 764–769, 771, 901, 988, 990, 991, 994
 - distribution, 272
- prior ratio, 988
- prior-free decision rule, 994
- priors constructed by some formal rule,

- 411
- prisoner's dilemma, 734
- probabilistic
 - causation, 108
 - competition, 921
 - conclusion, 91–93
 - diagnosis, 266, 273, 275
 - independence, 60, 61, 204
 - indifferentism, 1019
 - inference, 91–93
 - logic, 319, 320
 - prediction, 915
 - prediction competition, 922
 - predictions, 921, 927
 - process, 199, 212
 - reasoning, 93
 - reasoning in ancient India, 44
 - retractions, 1020
 - scoring, 913
- probabilistic/statistical reasoning, 54
- probabilities of conditionals, 127, 967
- probability, 43, 199, 201–204, 206–217, 220, 263, 265, 266, 322
 - as long-run frequency, 739
 - assertions, 460
 - assessment, 291
 - axiom, 246
 - density function, 75, 1046, 1048, 1050, 1054, 1055, 1058, 1059, 1065, 1066
 - distribution, 71–73, 79, 269
 - failure of frequentist interpretation, 1183
 - frequentist interpretation and supertasks, 1185, 1186
 - function, 100, 322, 323
 - has a double sense according to Cournot, 1155
 - its role in inductive inference, 153
 - mass function, 73, 1035–1038, 1041, 1046, 1048, 1050, 1058, 1059
 - matching priors, 293
 - model, 57, 64, 71, 80, 263, 272
 - not ampliative, but estimates may be, 1185
 - of single events and quantum mechanics, 1191
 - of single events and subjectivist interpretation, 1190
 - of observing misleading evidence, 501, 504
 - on lattice of subspaces of a Hilbert space, 1192
 - over a Boolean algebra of statements, 1191
 - paradoxes, 32
 - quantum interference, 1190
 - probability space, 100
 - /statistical reasoning, 53
 - theory, 264, 271
- probability₁, 441
- probability₂, 441
- probable verisimilitude, 486
- problem
 - of induction, 2, 19, 46, 1004
 - of observational errors, 486
 - of old evidence, 6, 402
 - of subjectivity, 14
 - of the reference class, 312
 - solving, 1004
- process variation, 515
- product of normal means, 293
- profile likelihood, 518, 519, 525, 525n
- promiscuous grammar, 868
- propensity, 139–142, 146, 1154
 - interpretation, 106
 - scores, 821
- proper likelihood, 525n
- property holds always, 1006
- proportion of infected people, 287
- propositions, 237
- prosecutor's fallacy, 179
- protein folding, 965
- pseudo-randomness, *see* randomness
- psychological probability, 1150, 1166
- psychologism, 1159
- Ptolemaic theory, 983
- Ptolemy, 984

- public decision making, 283, 303
- pun, 967
- pure (deterministic) strategy, 1013
- pursuit of the truth, 1001, 1002, 1016
- puzzle, 986, 1000
- Q -predicate, 455
- quadratic loss, 292
- qualitative evidence, 323
- qualitative theory, 486
- qualitative variables, 65
- quality assurance, 268, 280
- quantitative evidence, 323, 324
- quantity of interest, 283
- quantum mechanics
 - non-existence of joint probability distributions, 1191
 - probability interpretation of wave-function, 1190
 - structured-time interpretation, 1191
- Quetelet, A., 1149, 1162
- Quine, W. V., 4, 223, 346
- quiz shows, 920
- Rényi, A., 466
- Radon-Nikodym theorem, 116
- Raftery, A., 603
- Raju, C. K., 43, 44, 46
- Ramsey test, 127
- Ramsey, F. P., 127, 335, 344, 1442, 459, 150, 1171
- random, 962
 - methods, 1014
 - number generators, 636
 - sample, 270, 271
 - sampling, 1109
 - strategies, 1021
 - variable(s), 71, 105
- randomization, 817, 823, 826, 1039
- randomized experiments, 818
- randomness, 34–36
 - n -randomness, 662, 694–696
 - 1-random, 662, 695, 696
 - 2-random, 694–696
 - absolute, 662
 - algorithmic, 645–648, 648n, 662, 705, 706
 - arithmetical, 695
 - Church, 672, 701
 - computable, 700–702
 - existence of random strings, 680, 687
 - hyperarithmetical, 696
 - Kolmogorov-, 696
 - Kolmogorov-Chaitin, 688–690, 696, 706, 707
 - Kolmogorov-Loveland, 701–703, 707
 - Martin-Löf, 646–648, 653, 673–675, 689–690, 695, 697, 698, 700–703, 705, 705n, 706, 707
 - Mises-Wald-Church, 671–673, 690, 702, 704
 - of sequences and strings, 641–648, 653–662, 668–676, 680–681, 683, 686–708
 - partial computable, 700–702
 - pseudo-randomness, 642, 643, 661
 - random real number, 691
 - random walk, 658–659, 672
 - relative, 692–694
 - Schnorr, 697–698, 701, 702, 705n
 - Solovay, 673–675, 689
 - stochastic, 653
 - laws for, 655–657, 659, 661, 672, 689, 706
 - von Mises, 646, 648, 668–673, 698, 703–707
- Ratio*, 1159n
- ratio, 64
 - analysis, 111, 132
 - formula, 100
- rational reconstruction, 473, 483, 485, 486
- rationality, 990
- rationality update, 987
- real values, 402, 405
- realism, 20
- Recherches sur la probabilité des juge-*

- ments*, 1153, 1154
- recursively enumerable, 461
- red herring, 962
- reference
 - analysis, 10, 264, 283, 284
 - class, 1104
 - class problem, 41
 - credible intervals, 294
 - distribution, 284
 - formula, 309
 - marginal density, 616, 624
 - marginal likelihood, 617, 621
 - measure, 615
 - model probabilities, 607, 616, 621
 - posterior, 283
 - posterior odds, 624, 626
 - prior(s), 283, 285, 286, 288, 289, 293, 303, 411, 615
 - and the LP, 186
- referendum, 275, 276
- refined MDL, 897
- reflection principle, 107
- refutability, 1001
- refute and rescale, 415
- refuted, 1012
- regression, 816, 1088
 - analysis, 486
 - coefficients, 206
 - model, 623
- regret, 869, 870, 996
 - worst-case, 870
- regular, 111
- regularity condition, 1027, 1048, 1050, 1054, 1056, 1064, 1066
- Reichenbach, H., 108, 1150
- rejection, 257, 258, 299
- relative frequency, 55, 247, 266
- relative frequency interpretation of probability, 397
- relative frequentists, 217
- relative randomness, 638
- relative satisfiability, 445
- relativity to evidence, 256
- relativization, 693
- relativized conditional probability, 144
- relevant events, 268
- relevant statistics, 314
- reliabilism, 520
- reliability, 499, 516, 521, 523
- reliable indication, 985, 1000, 1021
- Rényi, A., 99, 118
- repeated examination of accumulating data, 502
- representation theorem(s), 239, 270, 271, 303
- representative sampling, 1110
- response bias, 827
- restricted posterior, 277
- restricted problem, 1004
- restricted reference priors, 291
- retraction(s), 277, 298, 986, 987, 1001, 1021
 - complexity, 1001
 - degrees, 1014
 - efficient, 1007
 - times, 1009
 - minimization, 1002
- retrograde motion, 983, 984
- Rgveda*
 - hymn on dice, 1180
- richness, 313, 314
- rifle welded, 997
- right weakening, 310
- right/wrong errors, 952
- right/wrong scoring, 919, 950, 951
- rigidity, 467
- risk, 1076, 1082–1084, 1086, 1087, 1090, 1091, 1093, 1095
- Rissanen, J. J., 602, 896, 948, 950
- Robbins, 41
- robust regression, 1129
- Roeper, P., 118
- Romeijn, J. W., 8, 9, 46, 481
- Rooij, S. de, 21–23, 46
- Rosenkrantz, R., 335
- Rosenthal, H., 487
- Royall's questions, 513
- Royall, R., 3, 4, 10–12, 16, 19, 359,

- 361, 540
 Royer-Collard, P. P., 1154
 rule of succession, 455
 Russell, B., 46, 1150
- safety belt, 268, 280
 Sakamoto, Y., 535n
 Salmon, W., 108
 Samanta, T., 583, 588, 597
 sample, 199–202, 205, 206, 209, 210,
 216, 218, 222–224, 757
 sample mean, 70, 80
 proportion, 82
 size, 80, 87, 501, 503
 space, 55, 57, 72, 274, 510, 753–
 756, 758–761, sub 763, 764,
 769–771
 standard deviation, 84, 86
 statistic, 199
 variance, 68, 69
 sampling, 199
 distribution, 64, 80, 83, 84, 86,
 294, 1027, 1041, 1053
 sanity check, 870
 satisfaction set, 310
 statistical inconsistency, 947
 Savage's representation theorem, 259
 Savage, L. J., 335, 344, 361, 373, 398,
 442, 475, 1150, 1170
 scale invariance, 516n
 Schervish, M. J., 117
 Schnorr's theorem, 647, 648, 688, 689,
 696, 707
 Schnorr, C. P., 689, 697, 698
 Schulte, O., 859n
 Schwarz, G., 110, 536, 541, 597, 940
Scientific inference, 1166
 scientific
 communication, 283
 data analysis, 299
 evidence, 493
 hypotheses, 556
 method, 983
 realism, 858
 reporting, 303
 theory, 983
 scoring probabilities, 914
 scoring rule, 398n
 Scott, D., 341, 460, 461, 463
 Scozzafava, R., 443, 466
 Scriven, M., 22
 Searle, J. R., 956
 second part, 939
 second problem of the priors, 428
 second-order effect, 1002
 second-order evidential probability, 315
 Seemann, T., 922
 segmentation, 969
 Seidenfeld, T., 117, 126, 308n, 449
 selection bias, 827
 self-ratification, 240
 self-supporting, 241
 semantic tableaux, 459
Sense and Sensibilia, 1169
 sensitivity, 106, 497, 498
 analysis, 275
 to the prior, 610, 611
 sequences, 642–646, 648–651, 653, 656,
 658, 671, 683, 690, 692
 sequential
 analysis, 525
 clustering, 965
 prediction strategy, 882
 sampling, 525
 severe test, definition, 164
 severely tested, 984
 severity, 4
 and fallacies of rejection, 168
 and testing model assumptions,
 193
 in the case of accepting the null,
 177
 in the case of rejecting the null,
 169
 vs. power, 172
 severity principle
 weak vs. full, 162
 severity rationale, 162

- vs. behavioristic rationale, 163
- Shafer–Dempster calculus, 407
- Shannon entropy, 927
- Shannon, C. E., 284, 680n, 875
- Shao, J., 584
- sharp, 1018
 - hypothesis testing, 300–302
- shattering, 20, 851–853
- Shibata, R., 584, 592
- Shimony, A., 108
- shooting contest, 993
- shooting room paradox, 725, 726
- short path problem, 1012
- short run, 990
- short simplicity paths, 1015
- short-run truth-indication, 986
- shrinkage, 933, 1082, 1084, 1093
- Shtarkov, Y. M., 896
- SIC, 601
- Sierpinski, W., 658
- significance level, 258
- Σ_1 , 461
- σ -algebra, 100
- σ -field, 100
- σ_1^0 , 461
- σ_2^0 , 461
- significance, 752, 757, 758n, 759, 769, 770
 - significance test(s), 4, 90, 155, 199, 200, 204, 205, 208, 210–215, 217, 219, 220–225, 257, 499, 500, 553, 554
- similarity, 480
- simple, 1018
 - hypotheses, 560, 562
 - law, 985
 - models, 522
 - random sample, 71
 - theory, 989
- simpler theory, 988
- simplest UTM, 953
- simplex, 1016
- simplicity, 2, 17, 19, 20, 22–24, 858–860, 901, 983, 987
 - biases, 1020
 - degrees, 1012
- Simpson's paradox, 29, 31, 32, 93, 124, 310, 724, 726
- Simpson, E. H., 124
- skeptical attitude, 986
- skeptical path(s), 1003, 1004
- skewed, 66
- skewness, 1034, 1061
- Skyrms, B., 1n, 111n, 121, 123, 335, 344, 370, 480, 481, 481n, 483, 1169
- Sleeping Beauty problem, 726, 728
- Slutsky's theorem, 1027, 1032–1034, 1045, 1062, 1063
- Smith, A. F. M., 583
- SMML, 934, 937
- SMML code-book, 938
- Smullyan, R., 459, 463
- Snir, M., 373
- Snob program, 965
- Sober, E., 17, 18, 22, 38, 45, 110, 535n, 584, 589, 598, 599
- Solomonoff, R. J., 36, 645n, 647, 648n, 675, 683, 866, 896, 909, 912, 930, 954
- Solomonoff-Kolmogorov-Chaitin invariance theorem, 678
- Solovay, R. M., 675
- space shuttle Challenger, 60
- Spanos, A., 4, 11, 16, 19, 45
- sparsity, 602
- spatial clustering, 965
- spatial correlation, 1134, 1138
- spatial-temporal correlation, 1126, 1138
- spatial-temporal error, 1133
- specific conditioning logic, 418
- specificity, 313, 314, 497, 498
- speed of light, 82
- Speigelhalter, D. J., 594
- spending of test size, 525
- Spirtes, P., 25, 26, 46, 109
- square-root inverted gamma, 278
- squared error, 991

- Śridhar, 1176
 Srinivasan, C., 38, 41, 46
 SRM, 947
 St. Petersburg paradox, 731
 stability of knowledge, 1001
 Stalnaker's hypothesis, 128
 Stalnaker, R. C., 128, 129
 stalwartness, 1005
 violators, 1010
 standard deviation, 64, 68
 standard error, 207, 224
 standard normal distribution, 86
 standardization, 820
 state description(s), 396, 453–456
 states, 237
 statistic, 80
Statistical Methods and Scientific Inference, 1170
 statistical
 consistency, 934, 935, 940, 945, 952, 963
 evidence, 493
 hypotheses, 517
 inconsistency, 944, 954, 957
 inference, 2, 3, 62, 64, 80, 82, 270, 474, 493, 517, 1022
 invariance, 933–935, 944, 945
 learning theory, 19, 849, 850, 854, 858, 860, 946, 947
 likelihood function, 939
 link, 997
 model, 154, 517
 paradigm, 493
 paradox(es), 29, 30, 737
 techniques, 985
 test, 234, 257
 verisimilitude, 487
 statistically
 consistent, 936, 945
 inconsistent, 941
 invariant, 932, 934, 936, 940, 942, 945
 statistics, 493, 1021
 steel post, 992
 Steel, D., 20, 46
 Stegmüller, W., 477, 478
 Stein estimator, 41
 Stein's paradox, 3, 38, 39, 41, 290
 Stein, C., 39
 stem cells, 966
 Sterzinger, O., 1149
 stochasticity; stochastic, 672, 698, 702, 703
 Church, 672, 701, 702
 for finite strings, 681, 687
 Kolmogorov-Loveland-, 701–703, 705
 laws, *see* randomness
 Mises-Wald-Church, 672, 701–705
 randomness, *see* randomness
 Stone, M., 588, 594
 stopping rule(s), 434, 502, 519, 626
 stopping rule principle, 187
 stream of experience, 1004
 strength, 314
 of evidence, 82, 250, 254, 494, 520
 strict coherence, 108
 Strict Minimum Message Length (SMML), 936, 937–939
 Strict Strict MML (SSMML), 938
 strings, 643–649, 656–658, 664, 667–668, 675–689, 691–692
Strong Inference, 514
 strong argument for Ockham's razor, 1011
 strong normality, 659–660
 strongly analogical method, 480, 481
 strongly beats, 1006, 1007
 structural axioms, 239
 structural constraints, 324
 structural equations, 816
 structural risk minimisation (SRM), 894, 996, 997n
 structure-description, 455, 456
 struture of the decision problem, 269
 Student, 206, 277, 294
 Student's *t*-test, 206

- subjective and objective
 - Bayes/Bayesian, 895, 988
 - Bayes compromise, 614
 - interpretation, 106
 - model weights, 614
 - prior, 517, 558
 - probability, 143, 144, 146, 235, 236
 - reversed senses, 1169, 1170
 - view, 637
- subjectivists, 217, 635
- subjectivity, 16
- submodel, 584
- sufficiency, 565, 566n, 567, 571, 575
- sufficient state, 568
- sufficient statistic(s), 19, 258, 275, 939, 1043, 1060
- sufficient-component cause model, 828
- summary statistics, 66
- śūnyavāda*
 - and zeroism, 1188
- supertask
 - and convergence, 1185
 - defined, 1183
 - involved in any notion of limits, 1183
 - involved in representing a formal real number, 1187
 - made possible by set theory, 1183
 - not involved in Indian way to sum infinite series, 1187
 - related to neglect of small numbers, 1185
- Suppes, P., 108
- support curve, 523
- support interval, 523
- support vector machines (SVMs), 946, 947, 964,
- sure-thing principle, 269
- Suśruta, 1176
- swarm intelligence, 958
- sweet spot, 993
- switch code, 889
- syādvāda*, 1192
- symmetry, 67, 996
- System of Logic*, 1159, 1163
- systematic errors, 813
- T*-test, 205
- TAIG, 940
- tail area probability, 499
- Taper, M., 14–16, 45
- target, 991, 993
 - formula, 309
- Tarski, A., 459, 461, 662
- tautology, 56
- Teng, C. M., 41, 46
- test, 757, 758, 769–772, 984
 - of significance, 82, 87
 - statistic, 89, 257, 258, 300
 - statistic in testing, 156
- testability, 849, 854, 860
- testing, 82
 - hypothesis, 91
 - the value of a normal mean, 301
- tests for randomness, 637, 638
- The Civilization of the Renaissance in Italy*, 1169
- the 68-95-99.7 rule, 76
- theoretical question, 1002
- theoretical terms, 245
- theoretical truth, 1000
- Théorie analytique*, 1160
- theory of inductive probability, 476
- thermometer, 985
- three basic rules, 56
- three key assumptions, 86
- three measures of central tendency, 64
- three standard deviations of the mean, 70
- time's arrow, 966, 969
- TIP, 478, 480, 481, 485
- topological structure, 1002
- tracking, 1000
- Traite du calcul des probabilités*, 1164
- transformation invariance, 516n
- transitivity, 405

- of preference, 269
- Treatise on Probability*, 1171
- trend detection, 1128
- trends, 993
- trial, 199
- triviality results, 129
- triviality theorem, 465
- true and causal theory, 999
- true theory(ies), 987, 991, 1021
- truncated sequential design, 503
- truth, 514, 519, 522–524, 985
 - finding, 869, 885
 - conducive, 987, 1002
 - conduciveness, 985, 986, 990, 991, 1000
 - indicator, 985
 - related virtues, 12
- truthlikeness, 515
- Tsao, C. A., 32, 34, 46
- Turgot, A.-R.-J., 1153
- Turing
 - degrees, 697
 - equivalence, 697, 698
 - reducibility, 697
 - machine(s), 461, 873, 902, 903, 909–913, 939, 946, 955, 960, 967, 990
- Turing, A., 658, 663
- Tversky, A., 433
- two envelope (or exchange) problem, 126, 732–733, 734
- two standard deviations of the mean, 70
- two-part, 939
 - codes, 878
 - compression, 902, 957
 - construction, 912
 - file compression, 913
 - inference, 903
 - Kolmogorov complexity, 939
 - message, 902, 936, 948, 949
 - MML, 912
- two-sided test, 90
- two-tailed significance test, 211
- Type I error, 90, 499, 506, 520, 526
- Type II error, 90, 506, 520
- typical performance, 994
- typicality, 646–647, 673, 705–707
- Über die Lehre des Wahrscheinlichen*, 1158n
- Ulam, S., 645
- unbiased estimate, 292,
- unbiased estimator, 17, 535, 536, 541
- unbiased thermometer, 538
- uncertain evidence, 431
- uncertain inference, 91
- uncertainty, 263–265, 269, 271
- unconditional probability, 58, 99
- undecidability, 960, 967
- under- and overfitting, 868, 894, 993, 1102
- under-aim, 993
- under-determination of theory, 161, 515, 814
- unified explanation, 984
- uniform, 985
 - code, 867, 870
 - distribution, 1051
 - prior, 285, 932
- uniformity of nature, 2, 996
- uniformly effective open, *see* effective open
- uninformative priors, 1166n
- union, 55
- uniquely decodable code, 874
- uniqueness, 951
 - property, 916
 - property of log-loss scoring, 914
- unit cube, 1018
- unity, 984
- universal bound, 501, 522
- universal code, 869, 876, 877
 - for the integers, 872
- universal comparability, 401, 403
- universal generalisations, 430
- universal Turing machine (UTM), 911, 912, 939, 948, 953, 961

- unobserved common cause, 999
- unpredictability, 646–647, 654, 657, 670, 700, 705–707
 - maximal, 647
- unreliability of inference
 - and statistical misspecifications, 190
- unstable, 1009
- Urbach, P., 335, 368, 450, 536
- useful models, 987
- utility, 238
 - function, 269, 560
 - maximisation, 240
 - of smallpox vaccination, 1152
 - or loss function, 558
- vague probability, 113
- valuation, 1018
- van de Geer, S. A., 603
- van Fraassen, B., 107, 113, 125, 131
- Vapnik, V. N., 602, 858, 946
- Vapnik-Chenovenkis (VC) dimension, 20, 849, 851–854, 856–860, 946, 947
- variance, 64, 546, 992, 993, 1103
 - of a random variable, 74
- vector of interest, 276
- Venn, J., 43, 634, 1149, 1150, 1159, 1160, 1162, 1163, 1165, 1171
- verifiability, 1001
- verisimilitude, 473, 484–486, 515, 516
- verisimilitudinarian BIL, 484, 485
- Ville, J. A., 36, 672
- Vineberg, S., 30–32, 46
- violations of eventual informativeness, 1010
- virtue, 968
- virus, 266, 273
- Voltaire (Arouet, F. M.), 43, 1153
- von Mises circular distribution, 964, 965
- von Mises, R., *see* Mises, R. von
- von Mises-Fisher spherical distribution, 964
- von Neumann trick, 653, 670
- von Neumann, J., *see* Neumann, J. von
- Waerden, B. L. van der, 655
- Wald, A., 671
- Wallace non-universality problem, 913
- Wallace, C. S., 110, 121, 909, 930, 936
- Walley, P., 113
- wave theory of light, 1008
- WCP, 567
- weak conditionality principle, 563–565, *see* WCP, 569, 571, 572n, 575
- weak evidence, 520, 526
- weak prequential principle, 890
- weak sufficiency principle, 564–566, *see* WSP, 571, 572n, 575
- weakly analogical exchangeable methods, 481
- weakly analogical method, 480
- weakly beaten, 1006
- Webb, G., 893
- Wegener, A., 931
- weighted average squared distance, 74
- weighted averages
 - and arbitrage in early India, 1178
 - and mathematical expectation, 1178
- Weirich, P., 5, 11, 34, 45
- welded rifle, 993
- Weyl, H., 645n, 660
- Wheeler, G., 10, 45
- William of Ockham, 1170
- Williamson, J., 10, 45, 335
- Williamson, T., 128
- wishful thinking, 1002
- Woodward, J. F., 109
- worst-case retraction cost, 1005
- WSP, 567
- yoga, 769
- z-score, 201
- z-test, 200, 201
- Zabell, S. L., 43, 46, 122

zero-one law, 652, 654

zeroism, 1185

and *śūnyavāda*, 1188

and conditioned coorigination, 1188

and fear of small numbers, 1189

and impermanence, 1189

as a realist and fallibilist position, 1189

how does one represent a continually changing entity?, 1188

impossibility of representing a continually changing entity, 1188

non-representability changes arithmetic, 1188

non-representability of a formal real number, 1187

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank