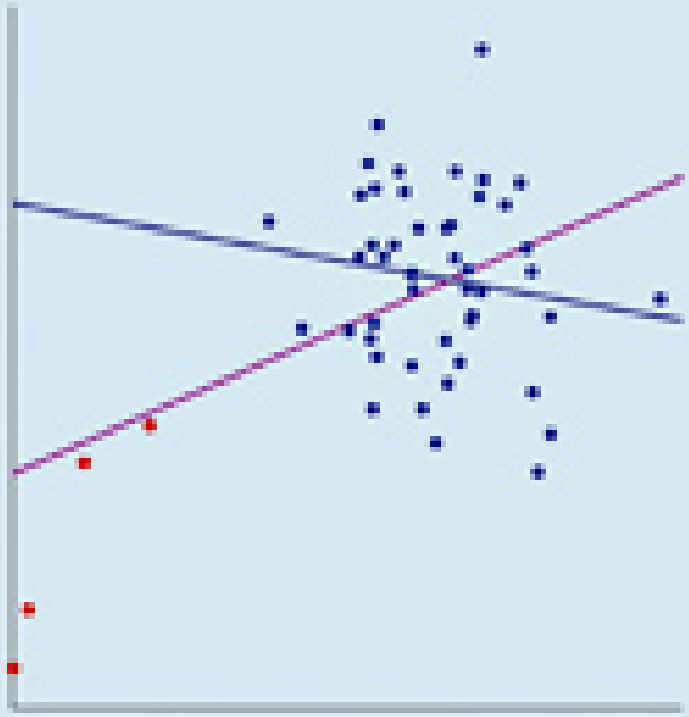
Statistical Models and Causal Inference

A Dialogue with the Social Sciences



David A. Freedman

Edited by

David Collier • Jasjeet S. Sekhon • Philip B. Stark

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521195003

This page intentionally left blank

STATISTICAL MODELS AND CAUSAL INFERENCE

A Dialogue with the Social Sciences

David A. Freedman presents here a definitive synthesis of his approach to causal inference in the social sciences. He explores the foundations and limitations of statistical modeling, illustrating basic arguments with examples from political science, public policy, law, and epidemiology. Freedman maintains that many new technical approaches to statistical modeling constitute not progress, but regress. Instead, he advocates a "shoe-leather" methodology, which exploits natural variation to mitigate confounding and relies on intimate knowledge of the subject matter to develop meticulous research designs and eliminate rival explanations. When Freedman first enunciated this position, he was met with skepticism, in part because it was hard to believe that a mathematical statistician of his stature would favor "low-tech" approaches. But the tide is turning. Many social scientists now agree that statistical technique cannot substitute for good research design and subject matter knowledge. This book offers an integrated presentation of Freedman's views.

David A. Freedman (1938–2008) was Professor of Statistics at the University of California, Berkeley. He was a distinguished mathematical statistician whose theoretical research included the analysis of martingale inequalities, Markov processes, de Finetti's theorem, consistency of Bayes estimators, sampling, the bootstrap, and procedures for testing and evaluating models and methods for causal inference. Freedman published widely on the application—and misapplication—of statistics in works within a variety of social sciences, including epidemiology, demography, public policy, and law. He emphasized exposing and checking the assumptions that underlie standard methods, as well as understanding how those methods behave when the assumptions are false—for example, how regression models behave when fitted to data from randomized experiments. He had a remarkable talent for integrating carefully honed statistical arguments with compelling empirical applications and illustrations. Freedman was a member of the American Academy of Arts and Sciences, and in 2003 he received the National Academy of Science's John J. Carty Award for his "profound contributions to the theory and practice of statistics."

David Collier is Robson Professor of Political Science at the University of California, Berkeley. He is co-author of *Rethinking Social Inquiry: Diverse Tools, Shared Standards* (2004) and co-editor of *The Oxford Handbook of Political Methodology* (2008) and *Concepts and Method in Social Science* (2009). He is a member of the American Academy of Arts and Sciences and was founding president of the Organized Section for Qualitative and Multi-Method Research of the American Political Science Association.

Jasjeet S. Sekhon is Associate Professor of Political Science at the University of California, Berkeley. His research interests include elections, applied and computational statistics, causal inference in observational and experimental studies, voting behavior, public opinion, and the philosophy and history of science. Professor Sekhon received his Ph.D. in 1999 from Cornell University and was a professor at Harvard University in the Department of Government from 1999 to 2005.

Philip B. Stark is Professor of Statistics at the University of California, Berkeley. His research centers on inference (inverse) problems, primarily in physical science. He is especially interested in confidence procedures tailored for specific goals and in quantifying the uncertainty in inferences that rely on simulations of complex physical systems. Professor Stark has done research on the Big Bang, causal inference, the U.S. Census, earthquake prediction, election auditing, the geomagnetic field, geriatric hearing loss, information retrieval, Internet content filters, nonparametrics (confidence procedures for function and probability density estimates with constraints), the seismic structure of the Sun and Earth, spectroscopy, and spectrum estimation.

Cover illustration: The data are from the Intersalt study of the relationship between salt intake and blood pressure, discussed in Chapter 9. The horizontal axis is urine salt level. The vertical axis is systolic blood pressure. Each dot represents the median value among subjects at one of 52 research centers in 32 countries. The four red dots correspond to two centers in Brazil that studied Indian tribes (Yanomamo and Xingu), a center in Papua New Guinea, and a center in Kenya. The two lines are least-squares regression lines. The purple line is fitted to all the data—the red dots and the blue dots. The blue line is fitted only to the blue dots. If all the data are included, median blood pressure is positively associated with median excreted salt. If only the blue dots are included, median blood pressure has a weak negative association with median salt. These data have been considered evidence that increasing salt intake increases blood pressure. The difference between the two regression lines suggests that any link between salt intake and blood pressure is weak. Chapter 9 discusses this and other shortcomings of the Intersalt study.

Statistical Models and Causal Inference

A Dialogue with the Social Sciences

David A. Freedman

Edited by

David Collier
University of California, Berkeley

Jasjeet S. Sekhon
University of California, Berkeley

Philip B. Stark
University of California, Berkeley

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521195003

First published in print format 2010

ISBN-13 978-0-511-68733-4 eBook (Adobe Reader)

ISBN-13 978-0-521-19500-3 Hardback

ISBN-13 978-0-521-12390-7 Paperback

Contents

Preface xi

Editors' Introduction: Inference and Shoe Leather xiii

Part I
Statistical Modeling: Foundations and Limitations

1. Issues in the Foundations of Statistics:
Probability and Statistical Models 3

Bayesians and frequentists disagree on the meaning of probability and other foundational issues, but both schools face the problem of model validation. Statistical models have been used successfully in the physical and life sciences. However, they have not advanced the study of social phenomena. How do models connect with reality? When are they likely to deepen understanding? When are they likely to be sterile or misleading?

2. Statistical Assumptions as Empirical Commitments 23

Statistical inference with convenience samples is risky. Real progress depends on a deep understanding of how the data were generated. No amount of statistical maneuvering will get very far without recognizing that statistical issues and substantive issues overlap.

3. Statistical Models and Shoe Leather 45

Regression models are used to make causal arguments in a wide variety of applications, and it is time to evaluate the results. Snow's work on cholera is a success story for causal inference based on nonexperimental data, which was collected through great expenditure of effort and shoe leather. Failures are also discussed. Statistical technique is seldom an adequate substitute for substantive knowledge of the topic, good research design, relevant data, and empirical tests in diverse settings.

Part II
Studies in Political Science, Public Policy, and Epidemiology

4. Methods for Census 2000 and Statistical Adjustments 65

The U.S. Census is a sophisticated, complex undertaking, carried out on a vast scale. It is remarkably accurate. Statistical adjustments are likely to introduce more error than they remove. This issue was litigated all the way to the Supreme Court, which in 1999 unanimously supported the Secretary of Commerce's decision not to adjust the 2000 Census.

5. On "Solutions" to the Ecological Inference Problem 83

Gary King's book, *A Solution to the Ecological Inference Problem*, claims to offer "realistic estimates of the uncertainty of ecological estimates." Applying King's method and three of his main diagnostics to data sets where the truth is known shows that his diagnostics cannot distinguish between cases where estimates are accurate and those where estimates are far off the mark. King's claim to have arrived at a solution to this problem is premature.

6. Rejoinder to King 97

King's method works with some data sets but not others. As a theoretical matter, inferring the behavior of subgroups from aggregate data is generally impossible: The relevant parameters are not identifiable. King's diagnostics do not discriminate between probable successes and probable failures.

7. Black Ravens, White Shoes, and Case Selection: Inference with Categorical Variables 105

Statistical ideas can clarify issues in qualitative analysis such as case selection. In political science, an important argument about case selection evokes Hempel's Paradox of the Ravens. This paradox can be resolved by distinguishing between population and sample inferences.

8. What is the Chance of an Earthquake? 115

Making sense of earthquake forecasts is surprisingly difficult. In part, this is because the forecasts are based on a complicated mixture of geological maps, rules of thumb, expert opinion, physical models, stochastic models, and numerical simulations, as well as geodetic, seismic, and paleoseismic data. Even the concept of probability is hard to define in this

context. Other models of risk for emergency preparedness, as well as models of economic risk, face similar difficulties.

9. Salt and Blood Pressure:
Conventional Wisdom Reconsidered 131

Experimental evidence suggests that the effect of a large reduction in salt intake on blood pressure is modest and that health consequences remain to be determined. Funding agencies and medical journals have taken a stronger position favoring the salt hypothesis than is warranted, demonstrating how misleading scientific findings can influence public policy.

10. The Swine Flu Vaccine and Guillain-Barré Syndrome:
A Case Study in Relative Risk and Specific Causation 151

Epidemiologic methods were developed to prove general causation: identifying exposures that increase the risk of particular diseases. Courts of law often are more interested in specific causation: On balance of probabilities, was the plaintiff's disease caused by exposure to the agent in question? There is a considerable gap between relative risks and proof of specific causation because individual differences affect the interpretation of relative risk for a given person. This makes specific causation especially hard to establish.

11. Survival Analysis: An Epidemiological Hazard? 169

Proportional-hazards models are frequently used to analyze data from randomized controlled trials. This is a mistake. Randomization does not justify the models, which are rarely informative. Simpler methods work better. This discussion matters because survival analysis has introduced a new hazard: It can lead to serious mistakes in medical treatment. Survival analysis is, unfortunately, thriving in other disciplines as well.

Part III
New Developments: Progress or Regress?

12. On Regression Adjustments in Experiments
with Several Treatments 195

Regression adjustments are often made to experimental data to address confounders that may not be balanced by randomization. Since randomization does not justify the models, bias is likely. Neither are the usual variance calculations to be trusted. Neyman's non-parametric model

serves to evaluate regression adjustments. A bias term is isolated, and conditions are given for unbiased estimation in finite samples.

13. Randomization Does Not Justify Logistic Regression 219

The logit model is often used to analyze experimental data. Theory and simulation show that randomization does not justify the model, so the usual estimators can be inconsistent. Neyman's non-parametric setup is used as a benchmark: Each subject has two potential responses, one if treated and the other if untreated; only one of the two responses can be observed. A consistent estimator is proposed.

14. The Grand Leap 243

A number of algorithms purport to discover causal structure from empirical data with no need for specific subject-matter knowledge. Advocates have no real success stories to report. These algorithms solve problems quite removed from the challenge of causal inference from imperfect data. Nor do they resolve long-standing philosophical questions about the meaning of causation.

15. On Specifying Graphical Models for Causation, and the Identification Problem 255

Causal relationships cannot be inferred from data by fitting graphical models without prior substantive knowledge of how the data were generated. Successful applications are rare because few causal pathways can be excluded a priori.

16. Weighting Regressions by Propensity Scores 279

The use of propensity scores to reduce bias in regression analysis is increasingly common in the social sciences. Yet weighting is likely to increase random error in the estimates and to bias the estimated standard errors downward, even when selection mechanisms are well understood. If investigators have a good causal model, it seems better just to fit the model without weights. If the causal model is improperly specified, weighting is unlikely to help.

17. On the So-Called "Huber Sandwich Estimator" and "Robust Standard Errors" 295

In applications where the statistical model is nearly correct, the Huber Sandwich Estimator makes little difference. On the other hand, if the model is seriously in error, the parameters being estimated are likely to be meaningless, except perhaps as descriptive statistics.

18. Endogeneity in Probit Response Models 305

The usual Heckman two-step procedure should not be used for removing endogeneity bias in probit regression. From a theoretical perspective this procedure is unsatisfactory, and likelihood methods are superior. Unfortunately, standard software packages do a poor job of maximizing the biprobit likelihood function, even if the number of covariates is small.

19. Diagnostics Cannot Have Much Power Against General Alternatives 323

Model diagnostics cannot have much power against omnibus alternatives. For instance, the hypothesis that observations are independent cannot be tested against the general alternative that they are dependent with power that exceeds the level of the test. Thus, the basic assumptions of regression cannot be validated from data.

Part IV
Shoe Leather Revisited

20. On Types of Scientific Inquiry: The Role of Qualitative Reasoning 337

Causal inference can be strengthened in fields ranging from epidemiology to political science by linking statistical analysis to qualitative knowledge. Examples from epidemiology show that substantial progress can derive from informal reasoning, qualitative insights, and the creation of novel data sets that require deep substantive understanding and a great expenditure of effort and shoe leather. Scientific progress depends on refuting conventional ideas if they are wrong, developing new ideas that are better, and testing the new ideas as well as the old ones. Qualitative evidence can play a key role in all three tasks.

References and Further Reading 357

Index 393

Preface

David A. Freedman presents in this book the foundations of statistical models and their limitations for causal inference. Examples, drawn from political science, public policy, law, and epidemiology, are real and important.

A statistical model is a set of equations that relate observable data to underlying parameters. The parameters are supposed to characterize the real world. Formulating a statistical model requires assumptions. Rarely are those assumptions tested. Indeed, some are untestable in principle, as Freedman shows in this volume. Assumptions are involved in choosing which parameters to include, the functional relationship between the data and the parameters, and how chance enters the model. It is common to assume that the data are a simple function of one or more parameters, plus random error. Linear regression is often used to estimate those parameters. More complicated models are increasingly common, but all models are limited by the validity of the assumptions on which they ride.

Freedman's observation that statistical models are fragile pervades this volume. Modeling assumptions—rarely examined or even enunciated—fail in ways that undermine model-based causal inference. Because of their unrealistic assumptions, many new techniques constitute not progress but regress. Freedman advocates instead "shoe leather" methods, which identify and exploit natural variation to mitigate confounding and which require intimate subject-matter knowledge to develop appropriate research designs and eliminate rival explanations.

Freedman assembled much of this book in the fall of 2008, shortly before his death. His goal was to offer an integrated presentation of his views on applied statistics, with case studies from the social and health sciences, and to encourage discussion of those views. We made some changes to Freedman's initial selection of topics to reduce length and broaden coverage. The text has been lightly edited; in a few cases chapter titles have been altered. The source is cited on the first page of each chapter and in the reference list, which has been consolidated at the end. When available, references to unpublished articles have been updated with the published versions. To alert the reader, chapter numbers have been added for citations to Freedman's works that appear in this book.

Many people deserve acknowledgment for their roles in bringing these ideas and this book to life, including the original co-authors and acknowledged reviewers. Colleagues at Berkeley and elsewhere contributed valuable suggestions, and Janet Macher provided astute assistance in editing the manuscript. Donald W. DeLand converted Chapters 3 and 8 into TeX. Josephine Marks also converted files and edited the references. Ed Parsons of Cambridge University Press helped shape the project and moved it to press with amazing speed. Above all, we admire David Freedman's tenacity and lucidity during his final days, and we are deeply grateful for his friendship, collaboration, and tutelage.

David Collier, Jasjeet S. Sekhon, and Philip B. Stark
Berkeley, California
July 2009

Companion website

http://statistics.berkeley.edu/~freedman/Dialogue.htm

Supplementary material, including errata, will be posted to the companion website.

Editors' Introduction: Inference and Shoe Leather

David Collier, Jasjeet S. Sekhon, and Philip B. Stark

Drawing sound causal inferences from observational data is a central goal in social science. How to do so is controversial. Technical approaches based on statistical models—graphical models, non-parametric structural equation models, instrumental variable estimators, hierarchical Bayesian models, etc.—are proliferating. But David Freedman has long argued that these methods are not reliable. He demonstrated repeatedly that it can be better to rely on subject-matter expertise and to exploit natural variation to mitigate confounding and rule out competing explanations.

When Freedman first enunciated this position decades ago, many were skeptical. They found it hard to believe that a probabilist and mathematical statistician of his stature would favor "low-tech" approaches. But the tide is turning. An increasing number of social scientists now agree that statistical technique cannot substitute for good research design and subject-matter knowledge. This view is particularly common among those who understand the mathematics and have on-the-ground experience.

Historically, "shoe-leather epidemiology" is epitomized by intensive, door-to-door canvassing that wears out investigators' shoes. In contrast, advocates of statistical modeling sometimes claim that their methods can salvage poor research design or low-quality data. Some suggest that their algorithms are general-purpose inference engines: Put in data, turn the crank, out come quantitative causal relationships, no knowledge of the subject required.

This is tantamount to pulling a rabbit from a hat. Freedman's conservation of rabbits principle says "to pull a rabbit from a hat, a rabbit must first be placed in the hat."[1] In statistical modeling, assumptions put the rabbit in the hat.

Modeling assumptions are made primarily for mathematical convenience, not for verisimilitude. The assumptions can be true or false—usually false. When the assumptions are true, theorems about the methods hold. When the assumptions are false, the theorems do not apply. How well do the methods behave then? When the assumptions are "just a little wrong," are the results "just a little wrong"? Can the assumptions be tested empirically? Do they violate common sense?

Freedman asked and answered these questions, again and again. He showed that scientific problems cannot be solved by "one-size-fits-all" methods. Rather, they require shoe leather: careful empirical work tailored to the subject and the research question, informed both by subject-matter knowledge and statistical principles. Witness his mature perspective:

> Causal inferences can be drawn from nonexperimental data. However, no mechanical rules can be laid down for the activity. Since Hume, that is almost a truism. Instead, causal inference seems to require an enormous investment of skill, intelligence, and hard work. Many convergent lines of evidence must be developed. Natural variation needs to be identified and exploited. Data must be collected. Confounders need to be considered. Alternative explanations have to be exhaustively tested. Before anything else, the right question needs to be framed.
>
> Naturally, there is a desire to substitute intellectual capital for labor. That is why investigators try to base causal inference on statistical models. The technology is relatively easy to use, and promises to open a wide variety of questions to the research effort. However, the appearance of methodological rigor can be deceptive. The models themselves demand critical scrutiny. Mathematical equations are used to adjust for confounding and other sources of bias. These equations may appear formidably precise, but they typically derive from many somewhat arbitrary choices. Which variables to enter in the regression? What functional form to use? What assumptions to make about parameters and error terms? These choices are seldom dictated either by data or prior scientific knowledge. That is why judgment is so critical, the opportunity for error so large, and the number of successful applications so limited.[2]

Causal inference from randomized controlled experiments using the intention-to-treat principle is not controversial—provided the inference is based on the actual underlying probability model implicit in the randomization. But some scientists ignore the design and instead use regression to analyze data from randomized experiments. Chapters 12 and 13 show that the result is generally unsound.

Nonexperimental data range from "natural experiments," where Nature provides data as if from a randomized experiment, to observational studies where there is not even a comparison between groups. The epitome of a natural experiment is Snow's study of cholera, discussed in Chapters 3 and 20. Snow was able to show—by expending an enormous amount of shoe leather—that Nature had mixed subjects across "treatments" in a way that was tantamount to a randomized controlled experiment.

To assess how close an observational study is to an experiment requires hard work and subject-matter knowledge. Even without a real or natural experiment, a scientist with sufficient expertise and field experience may be able to combine case studies and other observational data to rule out possible confounders and make sound inferences.

Freedman was convinced by dozens of causal inferences from observational data—but not hundreds. Chapter 20 gives examples, primarily from epidemiology, and considers the implications for social science. In Freedman's view, the number of sound causal inferences from observational data in epidemiology and social sciences is limited by the difficulty of eliminating confounding. Only shoe leather and wisdom can tell good assumptions from bad ones or rule out confounders without deliberate randomization and intervention. These resources are scarce.

Researchers who rely on observational data need qualitative and quantitative evidence, including case studies. They also need to be mindful of statistical principles and alert to anomalies, which can suggest sharp research questions. No single tool is best: They must find a combination suited to the particulars of the problem.

Freedman taught students—and researchers—to evaluate the quality of information and the structure of empirical arguments. He emphasized critical thinking over technical wizardry. This focus shines through two influential textbooks. His widely acclaimed undergraduate text, *Statistics*,[3] transformed statistical pedagogy. *Statistical Models: Theory and Practice*,[4] written at the advanced undergraduate and graduate level, presents standard techniques in statistical modeling and explains their shortcomings. These texts illuminate the sometimes tenuous relationship between statistical theory and scientific applications by taking apart serious examples.

The present volume brings together twenty articles by David Freedman and co-authors on the foundations of statistics, statistical modeling, and causal inference in social science, public policy, law, and epidemiology. They show when, why, and by how much statistical modeling is likely to fail. They show that assumptions are not a good substitute for subject-matter knowledge and relevant data. They show when qualitative, shoe-leather approaches may well succeed where modeling will not. And they point out that in some situations, the only honest answer is, "we can't tell from the data available."

This book is the perfect companion to *Statistical Models*. It covers some of the same topics in greater depth and technical detail and provides more case studies and close analysis of newer and more sophisticated tools for causal inference. Like all of Freedman's writing, this compilation is engaging and a pleasure to read: vivid, clear, and dryly funny. He does not use mathematics when English will do. Two-thirds of the chapters are relatively non-mathematical, readily accessible to most readers. The entire book—except perhaps a few proofs—is within the reach of social science graduate students who have basic methods training.

Freedman sought to get to the bottom of statistical modeling. He showed that sanguine faith in statistical models is largely unfounded. Advocates of modeling have responded by inventing escape routes, attempts to rescue the models when the underlying assumptions fail. As Part III of this volume makes clear, there is no exit: The fixes ride on *other* assumptions that are often harder to think about, justify, and test than those they replace.

This volume will not end the modeling enterprise. As Freedman wrote, there will always be "a desire to substitute intellectual capital for labor" by using statistical models to avoid the hard work of examining problems in their full specificity and complexity. We hope, however, that readers will find themselves better informed, less credulous, and more alert to the moment the rabbit is placed in the hat.

Notes

1. See, e.g., Freedman and Humphreys (1999), p. 102.

2. Freedman (2003), p. 19. See also Freedman (1999), pp. 255–56.

3. David Freedman, Robert Pisani, and Roger Purves (2007). *Statistics*, 4th edn. New York: Norton.

4. David A. Freedman (2009). *Statistical Models: Theory and Practice*, rev. edn. New York: Cambridge.

Part I

Statistical Modeling: Foundations and Limitations

1

Issues in the Foundations of Statistics: Probability and Statistical Models

> "Son, no matter how far you travel, or how smart you get, always remember this: Someday, somewhere, a guy is going to show you a nice brand-new deck of cards on which the seal is never broken, and this guy is going to offer to bet you that the jack of spades will jump out of this deck and squirt cider in your ear. But, son, do not bet him, for as sure as you do you are going to get an ear full of cider."
>
> — Damon Runyon[1]

ABSTRACT. *After sketching the conflict between objectivists and subjectivists on the foundations of statistics, this chapter discusses an issue facing statisticians of both schools, namely, model validation. Statistical models originate in the study of games of chance and have been successfully applied in the physical and life sciences. However, there are basic problems in applying the models to social phenomena; some of the difficulties will be pointed out. Hooke's law will be contrasted with regression models for salary discrimination, the latter being a fairly typical application in the social sciences.*

Foundations of Science (1995) 1: 19–39. With kind permission from Springer Science+Business Media.

1.1 What is probability?

For a contemporary mathematician, probability is easy to define, as a countably additive set function on a σ-field, with a total mass of one. This definition, perhaps cryptic for non-mathematicians, was introduced by A. N. Kolmogorov around 1930, and has been extremely convenient for mathematical work; theorems can be stated with clarity, and proved with rigor.[2]

For applied workers, the definition is less useful; countable additivity and σ-fields are not observed in nature. The issue is of a familiar type—what objects in the world correspond to probabilities? This question divides statisticians into two camps:

(i) the "objectivist" school, also called the "frequentists,"
(ii) the "subjectivist" school, also called the "Bayesians," after the Reverend Thomas Bayes (England, c. 1701–61) (Bayes, 1764).

Other positions have now largely fallen into disfavor; for example, there were "fiducial" probabilities introduced by R. A. Fisher (England, 1890–1962). Fisher was one of the two great statisticians of the century; the other, Jerzy Neyman (b. Russia, 1894; d. U.S.A., 1981), turned to objectivism after a Bayesian start. Indeed, the objectivist position now seems to be the dominant one in the field, although the subjectivists are still a strong presence. Of course, the names are imperfect descriptors. Furthermore, statisticians agree amongst themselves about as well as philosophers; many shades of opinion will be represented in each school.

1.2 The objectivist position

Objectivists hold that probabilities are inherent properties of the systems being studied. For a simple example, like the toss of a coin, the idea seems quite clear at first. You toss the coin, it will land heads or tails, and the probability of heads is around 50%. A more exact value can be determined experimentally, by tossing the coin repeatedly and taking the long run relative frequency of heads. In one such experiment, John Kerrich (a South African mathematician interned by the Germans during World War II) tossed a coin 10,000 times and got 5067 heads: The relative frequency was $5067/10{,}000 = 50.67\%$. For an objectivist such as myself, the probability of Kerrich's coin landing heads has its own existence, separate from the data; the latter enable us to estimate the probability, or test hypotheses concerning it.

The objectivist position exposes one to certain difficulties. As Keynes said, "In the long run, we are all dead." Heraclitus (also out of context)

is even more severe: "You can't step into the same river twice." Still, the tosses of a coin, like the throws of a die and the results of other such chance processes, do exhibit remarkable statistical regularities. These regularities can be described, predicted, and analyzed by technical probability theory. Using Kolmogorov's axioms (or more primitive definitions), we can construct statistical models that correspond to empirical phenomena; although verification of the correspondence is not the easiest of tasks.

1.3 The subjectivist position

For the subjectivist, probabilities describe "degrees of belief." There are two camps within the subjectivist school, the "classical" and the "radical." For a "classical" subjectivist, like Bayes himself or Laplace—although such historical readings are quite tricky—there are objective "parameters" which are unknown and to be estimated from the data. (A parameter is a numerical characteristic of a statistical model for data—for instance, the probability of a coin landing heads; other examples will be given below.) Even before data collection, the classical subjectivist has information about the parameters, expressed in the form of a "prior probability distribution."

The crucial distinction between a classical subjectivist and an objectivist: The former will make probability statements about parameters—for example, in a certain coin-tossing experiment, there is a 25% chance that the probability of heads exceeds .67. However, objectivists usually do not find that such statements are meaningful; they view the probability of heads as an unknown constant, which either is—or is not—bigger than .67. In replications of the experiment, the probability of heads will always exceed .67, or never; 25% cannot be relevant. As a technical matter, if the parameter has a probability distribution given the data, it must have a "marginal" distribution—that is, a prior. On this point, objectivists and subjectivists agree; the hold-out was R. A. Fisher, whose fiducial probabilities come into existence only after data collection.

"Radical" subjectivists, like Bruno de Finetti or Jimmie Savage, differ from classical subjectivists and objectivists; radical subjectivists deny the very existence of unknown parameters. For such statisticians, probabilities express degrees of belief about observables. You pull a coin out of your pocket, and—Damon Runyon notwithstanding—they can assign a probability to the event that it will land heads when you toss it. The braver ones can even assign a probability to the event that you really will toss the coin. (These are "prior" probabilities, or "opinions.") Subjectivists can also "update" opinions in the light of the data; for example, if the coin is tossed ten times, landing heads six times and tails four times, what is the

chance that it will land heads on the eleventh toss? This involves computing a "conditional" probability using Kolmogorov's calculus, which applies whether the probabilities are subjective or objective.

Here is an example with a different flavor: What is the chance that a Republican will be president of the U.S. in the year 2025? For many subjectivists, this is a meaningful question, which can in principle be answered by introspection. For many objectivists, this question is beyond the scope of statistical theory. As best I can judge, however, complications will be found on both sides of the divide. Some subjectivists will not have quantifiable opinions about remote political events; likewise, there are objectivists who might develop statistical models for presidential elections, and compute probabilities on that basis.[3]

The difference between the radical and classical subjectivists rides on the distinction between parameters and observables; this distinction is made by objectivists too and is often quite helpful. (In some cases, of course, the issue may be rather subtle.) The radical subjectivist denial of parameters exposes members of this school to some rhetorical awkwardness; for example, they are required not to understand the idea of tossing a coin with an unknown probability of heads. Indeed, if they admit the coin, they will soon be stuck with all the unknown parameters that were previously banished.[4]

1.3.1 Probability and relative frequency

In ordinary language, "probabilities" are not distinguished at all sharply from empirical percentages—"relative frequencies." In statistics, the distinction may be more critical. With Kerrich's coin, the relative frequency of heads in 10,000 tosses, 50.67%, is unlikely to be the exact probability of heads; but it is unlikely to be very far off. For an example with a different texture, suppose you see the following sequence of ten heads and ten tails:

T H T H T H T H T H T H T H T H T H T H.

What is the probability that the next observation will be a head? In this case, relative frequency and probability are quite different.[5]

One more illustration along that line: United Airlines Flight 140 operates daily from San Francisco to Philadelphia. In 192 out of the last 365 days, Flight 140 landed on time. You are going to take this flight tomorrow. Is your probability of landing on time given by 192/365? For a radical subjectivist, the question is clear; not so for an objectivist or a classical subjectivist. Whatever the question really means, 192/365 is the wrong answer—if you are flying on the Friday before Christmas. This is Fisher's "relevant subset" issue; and he seems to have been anticipated

by von Mises. Of course, if you pick a day at random from the data set, the chance of getting one with an on-time landing is indeed 192/365; that would not be controversial. The difficulties come with (i) extrapolation and (ii) judging the exchangeability of the data, in a useful Bayesian phrase. Probability is a subtler idea than relative frequency.[6]

1.3.2 Labels do not settle the issue

Objectivists sometimes argue that they have the advantage, because science is objective. This is not serious; "objectivist" statistical analysis must often rely on judgment and experience: Subjective elements come in. Likewise, subjectivists may tell you that objectivists (i) use "prior information," and (ii) are therefore closet Bayesians. Point (i) may be granted. The issue for (ii) is how prior information enters the analysis, and whether this information can be quantified or updated the way subjectivists insist it must be. The real questions are not to be settled on the basis of labels.

1.4 A critique of the subjectivist position

The subjectivist position seems to be internally consistent, and fairly immune to logical attack from the outside. Perhaps as a result, scholars of that school have been quite energetic in pointing out the flaws in the objectivist position. From an applied perspective, however, the subjectivist position is not free of difficulties either. What are subjective degrees of belief, where do they come from, and why can they be quantified? No convincing answers have been produced. At a more practical level, a Bayesian's opinion may be of great interest to himself, and he is surely free to develop it in any way that pleases him; but why should the results carry any weight for others?

To answer the last question, Bayesians often cite theorems showing "inter-subjective agreement." Under certain circumstances, as more and more data become available, two Bayesians will come to agree: The data swamp the prior. Of course, other theorems show that the prior swamps the data, even when the size of the data set grows without bounds—particularly in complex, high-dimensional situations. (For a review, see Diaconis and Freedman 1986.) Theorems do not settle the issue, especially for those who are not Bayesians to start with.

My own experience suggests that neither decision-makers nor their statisticians do in fact have prior probabilities. A large part of Bayesian statistics is about what you would do *if* you had a prior.[7] For the rest, statisticians make up priors that are mathematically convenient or attractive. Once used, priors become familiar; therefore, they come to be accepted

as "natural" and are liable to be used again. Such priors may eventually generate their own technical literature.

1.4.1 Other arguments for the Bayesian position

Coherence. Well-known theorems, including one by Freedman and Purves (1969), show that stubborn non-Bayesian behavior has costs. Your opponents can make a "dutch book," and extract your last penny—if you are generous enough to cover all the bets needed to prove the results.[7] However, most of us don't bet at all; even the professionals bet on relatively few events. Thus, coherence has little practical relevance. (Its rhetorical power is undeniable—who wants to be incoherent?)

Rationality. It is often urged that to be rational is to be Bayesian. Indeed, there are elaborate axiom systems about preference orderings, acts, consequences, and states of nature, whose conclusion is—that you are a Bayesian. The empirical evidence shows, fairly clearly, that those axioms do not describe human behavior at all well. The theory is not descriptive; people do not have stable, coherent prior probabilities.

Now the argument shifts to the "normative": If you were rational, you would obey the axioms and be a Bayesian. This, however, assumes what must be proved. Why would a rational person obey those axioms? The axioms represent decision problems in schematic and highly stylized ways. Therefore, as I see it, the theory addresses only limited aspects of rationality. Some Bayesians have tried to win this argument on the cheap: To be rational is, by definition, to obey their axioms. (Objectivists do not always stay on the rhetorical high road either.)

Detailed examination of the flaws in the normative argument is a complicated task, beyond the scope of the present article. In brief, my position is this. Many of the axioms, on their own, have considerable normative force. For example, if I am found to be in violation of the "sure thing principle," I would probably reconsider.[9] On the other hand, taken as a whole, decision theory seems to have about the same connection to real decisions as war games do to real wars.

What are the main complications? For some events, I may have a rough idea of likelihood: One event is very likely, another is unlikely, a third is uncertain. However, I may not be able to quantify these likelihoods, even to one or two decimal places; and there will be many events whose probabilities are simply unknown—even if definable.[10] Likewise, there are some benefits that can be assessed with reasonable accuracy; others can be estimated only to rough orders of magnitude; in some cases, quantification may not be possible at all. Thus, utilities may be just as problematic as priors.

The theorems that derive probabilities and utilities from axioms push the difficulties back one step.[11] In real examples, the existence of many states of nature must remain unsuspected. Only some acts can be contemplated; others are not imaginable until the moment of truth arrives. Of the acts that can be imagined, the decision-maker will have preferences between some pairs but not others. Too, common knowledge suggests that consequences are often quite different in the foreseeing and in the experiencing.

Intransitivity would be an argument for revision, although not a decisive one; for example, a person choosing among several job offers might well have intransitive preferences, which it would be a mistake to ignore. By way of contrast, an arbitrageur who trades bonds intransitively is likely to lose a lot of money. (There is an active market in bonds, while the market in job offers—largely nontransferable—must be rather thin; the practical details make a difference.) The axioms do not capture the texture of real decision making. Therefore, the theory has little normative force.

The fallback defense. Some Bayesians will concede much of what I have said: The axioms are not binding; rational decision-makers may have neither priors nor utilities. Still, the following sorts of arguments can be heard. The decision-maker must have some ideas about relative likelihoods for a few events; a prior probability can be made up to capture such intuitions, at least in gross outline. The details (for instance, that distributions are normal) can be chosen on the basis of convenience. A utility function can be put together using similar logic: The decision-maker must perceive some consequences as very good, and big utility numbers can be assigned to these; he must perceive some other consequences as trivial, and small utilities can be assigned to those; and in between is in between. The Bayesian engine can now be put to work, using such approximate priors and utilities. Even with these fairly crude approximations, Bayesian analysis is held to dominate other forms of inference: That is the fallback defense.

Here is my reaction to such arguments. Approximate Bayesian analysis may in principle be useful. That this mode of analysis dominates other forms of inference, however, seems quite debatable. In a statistical decision problem, where the model and loss function are given, Bayes procedures are often hard to beat, as are objectivist likelihood procedures; with many of the familiar textbook models, objectivist and subjectivist procedures should give similar results if the data set is large. There are sharp mathematical theorems to back up such statements.[12] On the other hand, in real problems—where models and loss functions are mere

approximations—the optimality of Bayes procedures cannot be a mathematical proposition. And empirical proof is conspicuously absent.

If we could quantify breakdowns in model assumptions, or degrees of error in approximate priors and loss functions, the balance of argument might shift considerably. The rhetoric of "robustness" may suggest that such error analyses are routine. This is hardly the case even for the models. For priors and utilities, the position is even worse, since the entities being approximated do not have any independent existence—outside the Bayesian framework that has been imposed on the problem.

De Finetti's theorem. Suppose you are a radical subjectivist, watching a sequence of 0's and 1's. In your prior opinion, this sequence is exchangeable: Permuting the order of the variables will not change your opinion about them. A beautiful theorem of de Finetti's asserts that your opinion can be represented as coin tossing, the probability of heads being selected at random from a suitable prior distribution. This theorem is often said to "explain" subjective or objective probabilities, or justify one system in terms of the other.[13]

Such claims cannot be right. What the theorem does is this: It enables the subjectivist to discover features of his prior by mathematical proof, rather than introspection. For example, suppose you have an exchangeable prior about those 0's and 1's. Before data collection starts, de Finetti will prove to you by pure mathematics that in your own opinion the relative frequency of 1's among the first n observations will almost surely converge to a limit as $n \to \infty$. (Of course, the theorem has other consequences too, but all have the same logical texture.)

This notion of "almost surely," and the limiting relative frequency, are features of your opinion not of any external reality. ("Almost surely" means with probability 1, and the probability in question is your prior.) Indeed, if you had not noticed these consequences of your prior by introspection, and now do not like them, you are free to revise your opinion—which will have no impact outside your head. What the theorem does is to show how various aspects of your prior opinion are related to each other. That is all the theorem can do, because the conditions of the theorem are conditions on the prior alone.

To illustrate the difficulty, I cite an old friend rather than making a new enemy. According to Jeffrey (1983, p. 199), de Finetti's result proves "your subjective probability measure [is] a certain mixture or weighted average of the various possible objective probability measures"—an unusually clear statement of the interpretation that I deny. Each of Jeffrey's "objective" probability measures governs the tosses of a p-coin, where p is your limiting relative frequency of 1's. (Of course, p has a probability

distribution of its own, in your opinion.) Thus, p is a feature of your opinion, not of the real world: The mixands in de Finetti's theorem are "objective" only by terminological courtesy. In short, the "p-coins" that come out of de Finetti's theorem are just as subjective as the prior that went in.

1.4.2 To sum up

The theory—as developed by Ramsey, von Neumann and Morgenstern, de Finetti, and Savage, among others—is great work. They solved an important historical problem of interest to economists, mathematicians, statisticians, and philosophers alike. On a more practical level, the language of subjective probability is evocative. Some investigators find the consistency of Bayesian statistics to be a useful discipline; for some (including me), the Bayesian approach can suggest statistical procedures whose behavior is worth investigating. But the theory is not a complete account of rationality, or even close. Nor is it the prescribed solution for any large number of problems in applied statistics, at least as I see matters.

1.5 Statistical models

Of course, statistical models are applied not only to coin tossing but also to more complex systems. For example, "regression models" are widely used in the social sciences, as indicated below; such applications raise serious epistemological questions. (This idea will be developed from an objectivist perspective, but similar issues are felt in the other camp.)

The problem is not purely academic. The census suffers an undercount, more severe in some places than others; if certain statistical models are to be believed, the undercount can be corrected—moving seats in Congress and millions of dollars a year in entitlement funds (*Survey Methodology* (1992) 18(1); *Jurimetrics* (1993) 34(1); *Statistical Science* (1994) 9(4). If yet other statistical models are to be believed, the veil of secrecy can be lifted from the ballot box, enabling the experts to determine how racial or ethnic groups have voted—a crucial step in litigation to enforce minority voting rights (*Evaluation Review*, (1991) 1(6); Klein and Freedman, 1993).

1.5.1 Examples

Here, I begin with a noncontroversial example from physics, namely, Hooke's law: Strain is proportional to stress. We will have some number n of observations. For the ith observation, indicated by the subscript i, we hang weight_i on a spring. The length of the spring is measured as length_i. The regression model says that[14]

$$\text{length}_i = a + b \times \text{weight}_i + \epsilon_i. \tag{1}$$

The "error" term ϵ_i is needed because length_i will not be exactly equal to $a + b \times \text{weight}_i$. If nothing else, measurement error must be reckoned with. We model ϵ_i as a sequence of draws, made at random with replacement from a box of tickets; each ticket shows a potential error—the ϵ_i that will be realized if that ticket is the ith one drawn. The average of all the potential errors in the box is assumed to be 0. In more standard terminology, the ϵ_i are assumed to be "independent and identically distributed, with mean 0." Such assumptions can present difficult scientific issues, because error terms are not observable.

In equation (1), a and b are parameters, unknown constants of nature that characterize the spring: a is the length of the spring under no load, and b is stretchiness—the increase in length per unit increase in weight. These parameters are not observable, but they can be estimated by "the method of least squares," developed by Adrien-Marie Legendre (France, 1752–1833) and Carl Friedrich Gauss (Germany, 1777–1855) to fit astronomical orbits. Basically, you choose the values of $\hat{a}$ and $\hat{b}$ to minimize the sum of the squared "prediction errors," $\sum_i {e_i}^2$, where e_i is the prediction error for the ith observation:[15]

$$e_i = \text{length}_i - \hat{a} - \hat{b} \times \text{weight}_i. \tag{2}$$

These prediction errors are often called "residuals": They measure the difference between the actual length and the predicted length, the latter being $\hat{a} - \hat{b} \times$ weight.

No one really imagines there to be a box of tickets hidden in the spring. However, the variability of physical measurements (under many but by no means all circumstances) does seem to be remarkably like the variability in draws from a box. This is Gauss' model for measurement error. In short, statistical models can be constructed that correspond rather closely to empirical phenomena.

I turn now to social-science applications. A case study would take us too far afield, but a stylized example—regression analysis used to demonstrate sex discrimination in salaries (adapted from Kaye and Freedman, 2000)—may give the idea. We use a regression model to predict salaries (dollars per year) of employees in a firm from:

- education (years of schooling completed),
- experience (years with the firm),
- the dummy variable "man," which takes the value 1 for men and 0 for women.

Employees are indexed by the subscript i; for example, salary_i is the salary of the ith employee. The equation is[16]

$$\text{salary}_i = a + b \times \text{education}_i + c \times \text{experience}_i + d \times \text{man}_i + \epsilon_i. \tag{3}$$

Equation (3) is a statistical model for the data, with unknown parameters a, b, c, d; here, a is the "intercept" and the others are "regression coefficients"; ϵ_i is an unobservable error term. This is a formal analog of Hooke's law (1); the same assumptions are made about the errors. In other words, an employee's salary is determined as if by computing

$$(4) \qquad a + b \times \text{education} + c \times \text{experience} + d \times \text{man},$$

then adding an error drawn at random from a box of tickets. The display (4) is the expected value for salary given the explanatory variables (education, experience, man); the error term in (3) represents deviations from the expected.

The parameters in (3) are estimated from the data using least squares. If the estimated coefficient d for the dummy variable turns out to be positive and "statistically significant" (by a "t-test"), that would be taken as evidence of disparate impact: Men earn more than women, even after adjusting for differences in background factors that might affect productivity. Education and experience are entered into equation (3) as "statistical controls," precisely in order to claim that adjustment has been made for differences in backgrounds.

Suppose the estimated equation turns out as follows:

$$(5) \quad \begin{aligned} \text{predicted salary} = \$7100 &+ \$1300 \times \text{education} \\ &+ \$2200 \times \text{experience} + \$700 \times \text{man}. \end{aligned}$$

That is, $\hat{a} = \$7100$, $\hat{b} = \$1300$, and so forth. According to equation (5), every extra year of education is worth on average \$1300; similarly, every extra year of experience is worth on average \$2200; and, most important, men get a premium of \$700 over women with the same education and experience, on average.

An example will illustrate (5). A male employee with twelve years of education (high school) and ten years of experience would have a predicted salary of

$$(6) \quad \begin{aligned} &\$7100 + \$1300 \times 12 + \$2200 \times 10 + \$700 \times 1 \\ &\qquad = \$7100 + \$15{,}600 + \$22{,}000 + \$700 \\ &\qquad = \$45{,}400. \end{aligned}$$

A similarly situated female employee has a predicted salary of only

$$(7) \quad \begin{aligned} &\$7100 + \$1300 \times 12 + \$2200 \times 10 + \$700 \times 0 \\ &\qquad = \$7100 + \$15{,}600 + \$22{,}000 + \$0 \\ &\qquad = \$44{,}700. \end{aligned}$$

Notice the impact of the dummy variable: \$700 is added to (6), but not to (7).

A major step in the argument is establishing that the estimated coefficient of the dummy variable in (3) is "statistically significant." This step turns out to depend on the statistical assumptions built into the model. For instance, each extra year of education is assumed to be worth the same (on average) across all levels of experience, both for men and women. Similarly, each extra year of experience is worth the same across all levels of education, both for men and women. Furthermore, the premium paid to men does not depend systematically on education or experience. Ability, quality of education, or quality of experience are assumed not to make any systematic difference to the predictions of the model.

The story about the error term—that the ϵ's are independent and identically distributed from person to person in the data set—turns out to be critical for computing statistical significance. Discrimination cannot be proved by regression modeling unless statistical significance can be established, and statistical significance cannot be established unless conventional presuppositions are made about unobservable error terms.

Lurking behind the typical regression model will be found a host of such assumptions; without them, legitimate inferences cannot be drawn from the model. There are statistical procedures for testing some of these assumptions. However, the tests often lack the power to detect substantial failures. Furthermore, model testing may become circular; breakdowns in assumptions are detected, and the model is redefined to accommodate. In short, hiding the problems can become a major goal of model building.

Using models to make predictions of the future, or the results of interventions, would be a valuable corrective. Testing the model on a variety of data sets—rather than fitting refinements over and over again to the same data set—might be a good second-best (Ehrenberg and Bound 1993). With Hooke's law (1), the model makes predictions that are relatively easy to test experimentally. For the salary discrimination model (3), validation seems much more difficult. Thus, built into the equation is a model for nondiscriminatory behavior: The coefficient d vanishes. If the company discriminates, that part of the model cannot be validated at all.

Regression models like (3) are widely used by social scientists to make causal inferences; such models are now almost a routine way of demonstrating counterfactuals. However, the "demonstrations" generally turn out to depend on a series of untested, even unarticulated, technical assumptions. Under the circumstances, reliance on model outputs may be quite unjustified. Making the ideas of validation somewhat more precise is a serious problem in the philosophy of science. That models should

correspond to reality is, after all, a useful but not totally straightforward idea—with some history to it. Developing appropriate models is a serious problem in statistics; testing the connection to the phenomena is even more serious.[17]

1.5.2 Standard errors, t-statistics, and statistical significance

The "standard error" of $\hat{d}$ measures the likely difference between $\hat{d}$ and d, due to the action of the error terms in equation (3). The "t-statistic" is $\hat{d}$ divided by its standard error. Under the "null hypothesis" that $d = 0$, there is only about a 5% chance that $|t| > 2$. Such a large value of t would demonstrate "statistical significance." Of course, the parameter d is only a construct in a model. If the model is wrong, the standard error, t-statistic, and significance level are rather difficult to interpret.

Even if the model is granted, there is a further issue: The 5% is a probability for the data given the model, namely, $P\{|t| > 2 \,||\, d = 0\}$. However, the 5% is often misinterpreted as $P\{d = 0|\text{data}\}$. Indeed, this misinterpretation is a commonplace in the social-science literature, and seems to have been picked up by the courts from expert testimony.[18] For an objectivist, $P\{d = 0|\text{data}\}$ makes no sense: Parameters do not exhibit chance variation. For a subjectivist, $P\{d = 0|\text{data}\}$ makes good sense, but its computation via the t-test is grossly wrong, because the prior probability that $d = 0$ has not been taken into account: The calculation exemplifies the "base rate fallacy."

The single vertical bar "|" is standard notation for conditional probability. The double vertical bar "||" is not standard; Bayesians might want to read this as a conditional probability; for an objectivist, "||" is intended to mean "computed on the assumption that . . . ."

1.5.3 Statistical models and the problem of induction

How do we learn from experience? What makes us think that the future will be like the past? With contemporary modeling techniques, such questions are easily answered—in form if not in substance.

- The objectivist invents a regression model for the data, and assumes the error terms to be independent and identically distributed; "IID" is the conventional abbreviation. It is this assumption of IID-ness that enables us to predict data we have not seen from a training sample—without doing the hard work of validating the model.

- The classical subjectivist invents a regression model for the data, assumes IID errors, and then makes up a prior for unknown parameters.

- The radical subjectivist adopts a prior that is exchangeable or partially exchangeable, and calls you irrational or incoherent (or both) for not following suit.

In our days, serious arguments have been made from data. Beautiful, delicate theorems have been proved, although the connection with data analysis often remains to be established. And an enormous amount of fiction has been produced, masquerading as rigorous science.

1.6 Conclusions

I have sketched two main positions in contemporary statistics, objectivist and subjectivist, and tried to indicate the difficulties. Some questions confront statisticians from both camps. How do statistical models connect with reality? What areas lend themselves to investigation by statistical modeling? When are such investigations likely to be sterile?

These questions have philosophical components as well as technical ones. I believe model validation to be a central issue. Of course, many of my colleagues will be found to disagree. For them, fitting models to data, computing standard errors, and performing significance tests is "informative," even though the basic statistical assumptions (linearity, independence of errors, etc.) cannot be validated. This position seems indefensible, nor are the consequences trivial. Perhaps it is time to reconsider.

Notes

1. From "The Idyll of Miss Sarah Brown," *Collier's Magazine*, 1933. Reprinted in *Guys and Dolls: The Stories of Damon Runyon*. Penguin Books, New York, 199 pp. 14–26. The quote is edited slightly, for continuity.

2. This note will give a compact statement of Kolmogorov's axioms. Let Ω be a set. By definition, a σ-field $\mathcal{F}$ is a collection of subsets of Ω, which has Ω itself as a member. Furthermore,

(i) $\mathcal{F}$ is closed under complementation (if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$), and

(ii) $\mathcal{F}$ is closed under the formation of countable unions (if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$, then $\bigcup_i A_i \in \mathcal{F}$).

A probability P is a non-negative, real-valued function on $\mathcal{F}$ such that $P(\Omega) = 1$ and P is "countably additive": If $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$, and the sets are pairwise disjoint, in the sense that $A_i \cap A_j = 0$ for $i \neq j$, then $P(\bigcup_i A_i) = \sum_i P(A_i)$. A random variable X is an $\mathcal{F}$-measurable function on Ω. Informally, probabilists might say that Nature chooses

$\omega \in \Omega$ according to P, and shows you $X(\omega)$; the latter would be the "observed value" of X.

3. Models will be discussed in Section 1.5. Those for presidential elections may not be compelling. For genetics, however, chance models are well established; and many statistical calculations are therefore on a secure footing. Much controversy remains, for example, in the area of DNA identification (*Jurimetrics* (1993) 34(1)).

4. The distinction between classical and radical subjectivists made here is not often discussed in the statistical literature; the terminology is not standard. See, for instance, Diaconis and Freedman (1980a), Efron (1986), and Jeffrey (1983, section 12.6).

5. Some readers may say to themselves that here, probability is just the relative frequency of transitions. However, a similar but slightly more complicated example can be rigged up for transition counts. An infinite regress lies just ahead. My point is only this: Relative frequencies are not probabilities. Of course, if circumstances are favorable, the two are strongly connected—that is one reason why chance models are useful for applied work.

6. To illustrate the objectivist way of handling probabilities and relative frequencies, I consider repeated tosses of a fair coin: The probability of heads is 50%. In a sequence of 10,000 tosses, the chance of getting between 49% and 51% heads is about 95%. In replications of this (large) experiment, about 95% of the time, there will be between 49% and 51% heads. On each replication, however, the probability of heads stays the same—namely, 50%.

The strong law of large numbers provides another illustration. Consider n repeated tosses of a fair coin. With probability 1, as $n \to \infty$, the relative frequency of heads in the first n tosses eventually gets trapped inside the interval from 49% to 51%; ditto, for the interval from 49.9% to 50.1%; ditto, for the interval from 49.99% to 50.01%; and so forth. No matter what the relative frequency of heads happens to be at any given moment, the probability of heads stays the same—namely, 50%. Probability is not relative frequency.

7. Similarly, a large part of objectivist statistics is about what you would do *if* you had a model; and all of us spend enormous amounts of energy finding out what would happen if the data kept pouring in. I wish we could learn to look at the data more directly, without the fictional models and priors. On the same wish-list: We should stop pretending to fix bad designs and inadequate measurements by modeling.

8. A "dutch book" is a collection of bets on various events such that the bettor makes money, no matter what the outcome.

9. According to the "sure thing principle," if I prefer A to B given that C occurs, and I also prefer A to B given that C does not occur, I must prefer A to B when I am in doubt as to the occurrence of C.

10. Although one-sentence concessions in a book are not binding, Savage (1972 [1954], p. 59) does say that his theory "is a code of consistency for the person applying it, not a system of predictions about the world"; and personal probabilities can be known "only roughly." Another comment on this book may be in order. According to Savage (1972 [1954], pp. 61–62), "on no ordinary objectivistic view would it be meaningful, let alone true, to say that on the basis of the available evidence it is very improbable, though not impossible, that France will become a monarchy within the next decade." As anthropology of science, this seems wrong. I make qualitative statements about likelihoods and possibilities, and expect to be understood; I find such statements meaningful when others make them. Only the quantification seems problematic. What would it mean to say that $P(\text{France will become a monarchy}) = .0032$? Many objectivists of my acquaintance share such views, although caution is in order when extrapolating from such a sample of convenience.

11. The argument in the text is addressed to readers who have some familiarity with the axioms. This note gives a very brief review; Kreps (1988) has a chatty and sympathetic discussion (although some of the details are not quite in focus); Le Cam (1977) is more technical and critical; the arguments are crisp. In the axiomatic setup, there is a space of "states of nature," like the possible orders in which horses finish a race. There is another space of "consequences"; these can be pecuniary or non-pecuniary (win \$1000, lose \$5000, win a weekend in Philadelphia, etc.). Mathematically, an "act" is a function whose domain is the space of states of nature and whose values are consequences. You have to choose an act: That is the decision problem. Informally, if you choose the act f, and the state of nature happens to be s, you enjoy (or suffer) the consequence $f(s)$. For example, if you bet on those horses, the payoff depends on the order in which they finish: The bet is an act, and the consequence depends on the state of nature. The set of possible states of nature, the set of possible consequences, and the set of possible acts are all viewed as fixed and known. You are supposed to have a transitive preference ordering on the acts, not just the consequences. The sure thing principle is an axiom in Savage's setup.

12. Wald's idea of a statistical decision problem can be sketched as follows. There is an unobservable parameter θ. Corresponding to each

θ, there is a known probability distribution P_θ for an observable random quantity X. (This family of probability distributions is a "statistical model" for X, with parameter θ.) There is a set of possible "decisions"; there is a "loss function" $L(d, \theta)$ which tells you how much is lost by making the decision d when the parameter is really θ. (For example, d might be an estimate of θ, and loss might be squared error.) You have to choose a "decision rule," which is a mapping from observed values of X to decisions. Your objective is to minimize "risk," that is, expected loss.

A comparison with the setup in note 11 may be useful. The "state of nature" seems to consist of the observable value of X, together with the unobservable value θ of the parameter. The "consequences" are the decisions, and "acts" are decision rules. (The conflict in terminology is regrettable, but there is no going back.) The utility function is replaced by L, which is given but depends on θ as well as d.

The risk of a Bayes' procedure cannot be reduced for all values of θ; any "admissible" procedure is a limit of Bayes' procedures ("the complete class theorem"). The maximum-likelihood estimator is "efficient"; and its sampling distribution is close to the posterior distribution of θ by the "Bernstein–von Mises theorem," which is actually due to Laplace. More or less stringent regularity conditions must be imposed to prove any of these results, and some of the theorems must be read rather literally; Stein's paradox and Bahadur's example should at least be mentioned.

Standard monographs and texts include Berger (1985), Berger and Wolpert (1988), Bickel and Doksum (1977), Casella and Berger (1990), Ferguson (1967), Le Cam (1986), Lehmann and Casella (2003), Lehmann and Romano (2005), and Rao (1973). The Bernstein–von Mises theorem is discussed in Le Cam and Yang (1990) and Prakasa Rao (1987). Of course, in many contexts, Bayes procedures and frequentist procedures will go in opposite directions; for a review, see Diaconis and Freedman (1986). These references are all fairly technical.

13. Diaconis and Freedman (1980a,b; 1981) review the issues and the mathematics. The first-cited paper is relatively informal; the second gives a version of de Finetti's theorem applicable to a finite number of observations, with bounds; the last gives a fairly general mathematical treatment of partial exchangeability, with numerous examples, and is more technical. More recent work is described in Diaconis and Freedman (1988, 1990).

The usual hyperbole can be sampled in Kreps (1988, p. 145): de Finetti's theorem is "the fundamental theorem of statistical inference—the theorem that from a subjectivist point of view makes sense out of most statistical procedures." This interpretation of the theorem fails to distin-

guish between what is assumed and what is proved. It is the assumption of exchangeability that enables you to predict the future from the past, at least to your own satisfaction—not the conclusions of the theorem or the elegance of the proof. Also see Section 1.5. If you pretend to have an exchangeable prior, the statistical world is your oyster, de Finetti or no de Finetti.

14. The equation holds for quite a large range of weights. With large enough weights, a quadratic term will be needed in equation (1). Moreover, beyond some point, the spring passes its "elastic limit" and snaps. The law is named after Robert Hooke, England, 1653–1703.

15. The residual e_i is observable, but is only an approximation to the disturbance term ϵ_i in (1); that is because the estimates $\hat{a}$ and $\hat{b}$ are only approximations to the parameters a and b.

16. Such equations are suggested, somewhat loosely, by "human capital theory." However, there remains considerable uncertainty about which variables to put into the equation, what functional form to assume, and how error terms are supposed to behave. Adding more variables is no panacea: Freedman (1983) and Clogg and Haritou (1997).

17. For more discussion in the context of real examples, with citations to the literature of model validation, see Freedman (1985, 1987, 1991 [Chapter 3], 1997). Many recent issues of *Sociological Methodology* have essays on this topic. Also see Oakes (1990), who discusses modeling issues, significance tests, and the objectivist-subjectivist divide.

18. Some legal citations may be of interest (Kaye and Freedman 2000): *Waisome v. Port Authority*, 948 F.2d 1370, 1376 (2d Cir. 1991) ("Social scientists consider a finding of two standard deviations significant, meaning there is about 1 chance in 20 that the explanation for a deviation could be random"); *Rivera v. City of Wichita Falls*, 665 F.2d 531, 545 n.22 (5th Cir. 1982) ("A variation of two standard deviations would indicate that the probability of the observed outcome occurring purely by chance would be approximately five out of 100; that is, it could be said with a 95% certainty that the outcome was not merely a fluke."); *Vuyanich v. Republic Nat'l Bank*, 505 F. Supp. 22 271 (N.D. Tex. 1980), vacated and remanded, 723 F.2d 1195 (5th Cir. 1984) ("if a 5% level of significance is used, a sufficiently large t-statistic for the coefficient indicates that the chances are less than one in 20 that the true coefficient is actually zero.").

An example from the underlying technical literature may also be of interest. According to (Fisher 1980, p. 717), "in large samples, a t-statistic of approximately two means that the chances are less than one in twenty that the true coefficient is actually zero and that we are

observing a larger coefficient just by chance . . . . A t-statistic of approximately two and one half means the chances are only one in one hundred that the true coefficient is zero . . . ." No. If the true coefficient is zero, there is only one chance in one hundred that $|t| > 2.5$. (Frank Fisher is a well-known econometrician who often testifies as an expert witness, although I do not believe he figures in any of the cases cited above.)

Acknowledgments

I would like to thank Dick Berk, Cliff Clogg, Persi Diaconis, Joe Eaton, Neil Henry, Paul Humphreys, Lucien Le Cam, Diana Petitti, Brian Skyrms, Terry Speed, Steve Turner, Amos Tversky, Ken Wachter, and Don Ylvisaker for many helpful suggestions—some of which I could implement.

2

Statistical Assumptions as Empirical Commitments

With Richard A. Berk

ABSTRACT. *Statistical inference with convenience samples is a risky business. Technical issues and substantive issues overlap. No amount of statistical maneuvering can get very far without deep understanding of how the data were generated. Empirical generalizations from a single data set should be viewed with suspicion. Rather than ask what would happen in principle if the study were repeated, it is better to repeat the study—as is standard in physical science. Indeed, it is generally impossible to predict variability across replications of an experiment without replicating the experiment, just as it is generally impossible to predict the effect of intervention without actually intervening.*

2.1 Introduction

Researchers who study punishment and social control, like those who study other social phenomena, typically seek to generalize their findings from the data they have to some larger context: In statistical jargon, they

Law, Punishment, and Social Control: Essays in Honor of Sheldon Messinger (2005) 2nd edn. T. G. Blomberg and S. Cohen, eds. Aldine de Gruyter, pp. 235–54.

generalize from a sample to a population. Generalizations are one important product of empirical inquiry. Of course, the process by which the data are selected introduces uncertainty. Indeed, any given data set is but one of many that could have been studied. If the data set had been different, the statistical summaries would have been different, and so would the conclusions, at least by a little.

How do we calibrate the uncertainty introduced by data collection? Nowadays, this question has become quite salient, and it is routinely answered using well-known methods of statistical inference, with standard errors, t-tests, and P-values, culminating in the "tabular asterisks" of Meehl (1978). These conventional answers, however, turn out to depend critically on certain rather restrictive assumptions, for instance, random sampling.[1]

When the data are generated by random sampling from a clearly defined population, and when the goal is to estimate population parameters from sample statistics, statistical inference can be relatively straightforward. The usual textbook formulas apply; tests of statistical significance and confidence intervals follow.

If the random-sampling assumptions do not apply, or the parameters are not clearly defined, or the inferences are to a population that is only vaguely defined, the calibration of uncertainty offered by contemporary statistical technique is in turn rather questionable.[2]

Thus, investigators who use conventional statistical technique turn out to be making, explicitly or implicitly, quite restrictive behavioral assumptions about their data collection process. By using apparently familiar arithmetic, they have made substantial empirical commitments; the research enterprise may be distorted by statistical technique, not helped. At least, that is our thesis, which we will develop in the pages that follow.

Random sampling is hardly universal in contemporary studies of punishment and social control. More typically, perhaps, the data in hand are simply the data most readily available (e.g., Gross and Mauro 1989; MacKenzie 1991; Nagin and Paternoster 1993; Berk and Campbell 1993; Phillips and Grattet 2000; White 2000). For instance, information on the use of prison "good time" may come from one prison in a certain state. Records on police use of force may be available only for encounters in which a suspect requires medical attention. Prosecutors' charging decisions may be documented only after the resolution of a lawsuit.

"Convenience samples" of this sort are not random samples. Still, researchers may quite properly be worried about replicability. The gen-