Deborah G. Mayo

Error and the Growth of Experimental Knowledge

THE UNIVERSITY OF CHICAGO PRESS
Chicago and London

Preface

DESPITE THE CHALLENGES TO AND CHANGES IN traditional philosophy of science, one of its primary tasks continues to be to explain, if not also to justify, scientific methodologies for learning about the world. To logical empiricist philosophers (Carnap, Reichenbach) the task was to show that science proceeds by objective rules for appraising hypotheses. To that end many attempted to set out formal rules termed inductive logics and confirmation theories. Alongside these stood Popper's method of appraisal based on falsification: evidence was to be used to falsify claims deductively rather than to build up inductive support. Both inductivist and falsificationist approaches were plagued with numerous, often identical, philosophical problems and paradoxes. Moreover, the entire view that science follows impartial algorithms or logics was challenged by Kuhn (1962) and others. What methodological rules there are often conflict and are sufficiently vague as to "justify" rival hypotheses. Actual scientific debates often last for several decades and appear to require, for their adjudication, a variety of other factors left out of philosophers' accounts. The challenge, if one is not to abandon the view that science is characterized by rational methods of hypothesis appraisal, is either to develop more adequate models of inductive inference, or else to find some new account of scientific rationality.

This was the problem situation in philosophy of science when, as a graduate student, I grew interested in these issues. With my background in logic and mathematics, and challenged by the problems formulated by Kyburg, Salmon, and others, I was led to pursue the first option—attempt to develop a more adequate account of inductive inference.

I might never have sauntered into that first class on mathematical statistics had the Department of Statistics not been situated so closely to the Department of Philosophy at the University of Pennsylvania. There I discovered, expressed in the language of statistics, the very problems of induction and confirmation that were so much in the minds of the philosophers of science nearby. The more I learned, the more I suspected that understanding how these statistical methods worked would offer up solutions to the vexing problem of how we

learn about the world in the face of error. But the similarity of the goals of philosophy of science and statistics, any more than the physical proximity of their departments on that campus, did not diminish the large gulf that existed between philosophical work on induction and the methods and models of standard statistical practice.

While logical-empiricist systems of inductive logic, despite a few holdouts, were largely being abandoned, most philosophers of statistics viewed the role of statistics as that of furnishing a set of formal rules or "logic" relating given evidence to hypotheses. The dominant example of such an approach on the contemporary philosophical scene is based on one or another Bayesian measure of support or confirmation. With the Bayesian approach, what we have learned about a hypothesis H from evidence e is measured by the conditional probability of H given e using Bayes's theorem. The cornerstone of the Bayesian approach is the use of prior probability assignments to hypotheses, generally interpreted as an agent's subjective degrees of belief. In contrast, the methods and models of classical and Neyman-Pearson statistics (e.g., statistical significance tests, confidence interval methods) that seemed so promising to me eschewed the use of prior probabilities where these could not be based on objective frequencies. Probability enters instead as a way of characterizing the experimental or testing process itself: to express how reliably it discriminates between alternative hypotheses and how well it facilitates learning from error. These probabilistic properties of experimental procedures are *error probabilities*.

Not only was there the controversy raging between Bayesians and error statisticians, but philosophers of statistics of all stripes were also full of criticisms of Neyman-Pearson error statistics and had erected a store of counterintuitive inferences apparently licensed by those methods. Before I could hope to utilize error statistical ideas in grappling with the problems of the rationality of science, clearly these criticisms would have to be confronted. Some proved recalcitrant to an easy dismissal, and this became the task of my doctoral dissertation. This "detour," fascinating in its own right, was a main focus for the next few years.

The result of grappling with these problems was a reformulation of standard Neyman-Pearson statistics that avoided the common misinterpretations and seemed to reflect the way these methods are used in practice. By the time that attempt grew into the experimental testing account of this book, the picture had diverged sufficiently from the Neyman-Pearson model to warrant some new name. Since it retains the centerpiece of standard Neyman-Pearson methods—the funda-

mental use of error probabilities—error-probability statistics, or just *error statistics,* seems about right.

My initial attempt to reformulate Neyman-Pearson statistics, however relevant to the controversy being played out within the confines of philosophy of statistics, was not obviously so for those who had largely abandoned that way of erecting an account of science, or so I was to learn, thanks to a question or challenge put forth by Larry Laudan in 1984.

In the new "theory change" movement that Laudan promoted, theory testing does not occur apart from appraising an entire *paradigm* (Kuhn), *research program* (Lakatos), or *tradition* (Laudan). In striking contrast to logical empiricist models and their contemporary (Bayesian) variants, the rational theory change models doubt that the technical machinery of inductive logic and statistical inference can shed much light on the problems of scientific rationality. It was perhaps Laudan's skepticism that drove me to pursue explicitly the task that had led me to philosophy of statistics in the first place—to utilize an adequate account of statistical inference in grappling with philosophical problems about evidence and inference. More than that, it was his persistent call to test our accounts against the historical record of science that led me to investigate a set of experimental episodes in science. I found, however, that little is learned from merely regarding historical episodes as instances or counterinstances of one or another philosophical thesis about science. That sort of historical approach fails to go deep enough to uncover the treasures often buried within historical cases.

What proved to be a gold mine for me was studying the nitty-gritty details of the data collection and analysis from experimental episodes. Here one can unearth a handful of standard or "canonical" strategies by which a host of noisy raw data may be turned into far more reliable modeled data. These investigations afforded me several shortcuts for how to relate statistical methods to full-bodied scientific inquiries. The rationale of statistical methods and models is found in their capacity to systematize strategies for learning from data, and thereby for furthering the growth of experimental knowledge.

In contrast to the global inductive approaches—a rule for any given data and hypothesis—so attractive to philosophers, I favor a model of experimental learning that is more of a piecemeal approach, whereby one question may be asked at a time in carrying out, modeling, and interpreting experiments—even to determine what "the data" are. The idea of viewing experimental inquiry in terms of a series of distinct models was influenced by experience with statistics as well as by early exposure to a seminal paper by Patrick Suppes (1969). By

insisting on a global measure of evidential-relationship, philosophers have overlooked the value of piecemeal error-statistical tools for filling out a series of models that link data to experiments. But how are the pieces intertwined so that the result is genuinely *ampliative?* Unlocking this puzzle occupied me for some time. One way to describe how statistical methods function, I came to see, is that they enable us, quite literally, to learn from error. A main task of this book is to develop this view.

The view that we learn from error, while commonplace, has been little explored in philosophy of science. When philosophers of science do speak of learning from error—most notably in the work of Popper—they generally mean simply that when a hypothesis is put to the test of experiment and fails, we reject it and attempt to replace it with another. Little is said about what the different types of errors are, what specifically is learned when an error is recognized, how we locate precisely what is at fault, how our ability to detect and correct errors grows, and how this growth is related to the growth of scientific knowledge. In what follows, I shall explore the possibility that addressing these questions provides a fresh perspective for understanding how we learn about the world through experiment.

Readers who wish to read the concluding overview in advance may turn to chapter 13.

Recent trends in philosophy of science lead me to think that the time is ripe for renewing the debate between Bayesian and error statistical epistemologies of experiment. Two main trends I have in mind are (1) the effort to link philosophies of science to actual scientific practice and scrutinize methodologies of science empirically or naturalistically, and (2) the growing interest in experiment by philosophers, historians, and sociologists of science.

Although methods and models from error statistics continue to dominate among experimental practitioners who use statistics, the Bayesian Way has increasingly been regarded as the model of choice among philosophers looking to statistical methodology. Given the current climate in philosophy of science, readers unfamiliar with philosophy of statistics may be surprised to find philosophers (still) declaring invalid a widely used set of experimental methods, rather than trying to explain why scientists evidently (still) find them so useful. This has much less to do with any sweeping criticisms of the standard approach than with the fact that the Bayesian view strikes a resonant chord with the logical-empiricist gene inherited from early work in confirmation and induction. In any event, it is time to remedy this situation. A genuinely adequate philosophy of experiment will only emerge if it is not at odds with statistical practice in science.

More than any other philosophical field of which I am aware, the probability and statistics debates tend to have the vehemence usually restricted to political or religious debates. I do not hope to bring hard-core Bayesians around to my view, but I do hope to convince the large pool of tempered, disgruntled, and fallen Bayesians that a viable non-Bayesian alternative exists. Most important, I aim to promote a general change of focus in the debates so that statistical accounts are scrutinized according to how well they serve specific ends. I focus on three chief tasks to which statistical accounts can and have been put in philosophy of science: (1) modeling scientific inference, (2) solving problems about evidence and inference, and (3) performing a critique of methodological rules.

The new emphasis on experiment is of special relevance in making progress on this debate. Experiments, as Ian Hacking taught us, live lives of their own, apart from high level theorizing. Actual experimental inquiries, the new experimentalists show, focus on manifold local tasks: checking instruments, ruling out extraneous factors, getting accuracy estimates, distinguishing real effect from artifact, and estimating the effects of background factors. The error-statistical account offers a tool kit of methods that are most apt for performing the local tasks of designing, modeling, and learning from experimental data. At the same time, the already well worked out methods and models from error statistics supply something still absent: a systematic framework for making progress with the goals of the new experimentalist work.

This book is intended for a wide-ranging audience of readers interested in the philosophy and methodology of science: for practitioners and philosophers of experiment and science, and for those interested in interdisciplinary work in science and technology studies. My hope is to strengthen existing bridges and create some new bridges between these fields. I regard the book as nontechnical and open to readers without backgrounds in statistics or probability. This does not mean that it contains no formal statistical ideas, but rather that the same ideas will also be presented in semiformal and informal ways. Readers can therefore feel free to put off (temporarily or permanently) statistical discussions without worrying that they will miss the main arguments or the working of this approach. So, for example, the reader hungry for a full-blown illustration of the approach can jump to chapter 7 after wading through the first two sections of chapter 5.

I have attempted to design the book so that an idea not caught in one place will likely be caught in another. This attempt, I will confess, required me to diverge from a more usual linear approach wherein one defines the formal concepts needed, articulates an approach and then contrasts it with others, and so on. My critique of the Bayesian

Way—one key aim of the book—is woven through the book, which simultaneously addresses two interrelated aims: developing the alternative error statistical approach and tackling a number of philosophical problems about evidence and inference. In addition to this cyclical or "braided" approach, some may fault me for overlooking certain technical qualifications or for failing to mention so-and-so's recent result. Again, I admit my guilt, but this seemed a necessary trade-off to bring these ideas into the mainstream where they belong.

My intention is to take readers on a journey that lets them get a feel for the error probability way of thinking, links it to our day-to-day strategies for finding things out, and points to the direction in which a new philosophy of experiment might move. I want to identify for the reader the style of inference and argument involved—it is of a sort we perform every day in learning from errors. Once this is grasped, I believe, the appropriate way to use and interpret the formal statistical tools will then follow naturally. If anything like a full-blown philosophy of experiment is to be developed, it will depend as much on having an intuitive understanding of standard or "canonical" arguments from error as on being able to relate them to statistical models.

I began writing this book while I was a visitor at the Center for Philosophy of Science at the University of Pittsburgh in the fall of 1989. I am grateful for the stimulating environment and for conversations on aspects of this work with John Earman, Clark Glymour, Adolf Grünbaum, Nicholas Rescher, and Wesley Salmon.

My ideas were importantly shaped by Ronald Giere's defenses of Neyman-Pearson statistics in the 1970s, and he has been a wonderful resource over many years. Without his encouragement and help, especially early on, I might never have found the path or had the nerve to pursue this approach. I have benefited enormously from his scrupulous reading of earlier drafts of this manuscript and from his uncanny ability to distill, in a few elegant phrases, just what I am trying to say.

I am deeply grateful to Henry Kyburg, in whose (1984) NEH summer seminar on induction and probability many of my ideas on statistical inference crystallized. He has given me generous help with the revisions to early drafts and, more than anyone else, is to be credited with having provoked me to a bolder, clearer, and more direct exposition of my account; although he would have preferred that I perform even greater liposuction on the manuscript.

I owe the largest debt of gratitude to Wesley Salmon. My debt is both to his work, which has had a strong influence, as well as to his massive help throughout the course of this project. In countless conversations and commentaries, though I proffered only the roughest of

drafts, he gave me the benefit of his unparalleled mastery of the problems with which I was grappling. I thank him for letting me try out my views on him, for his steadfast confidence and support, and for rescuing me time and again from being stalled by one or another obstacle.

I would like to acknowledge a number of individuals to whom I owe special thanks for substantive help with particular pieces of this book. I am indebted to Larry Laudan for lengthy discussions about my reworking of Kuhnian normal science in chapter 2, and I benefited greatly from Teddy Seidenfeld's recommendations and criticisms of the statistical ideas in chapters 3, 5, and 10. I thank Clark Glymour for help in clarifying his position in chapter 9. For insights on an early draft of chapter 11 on Peirce, I am grateful to Isaac Levi. An older debt recalled in developing the key concept of severe tests is to Alan Musgrave. My understanding of the views on novel predictions and my coming to relate them to severe tests (in chapters 6 and 8) grew directly out of numerous conversations with him while he was a visitor at Virginia Tech in 1986. I am grateful to communication with Karl Popper around that time, which freed me, in chapter 6, to clearly distinguish my severity concept from his.

Larry Laudan's influence, far more than the citations indicate, can be traced throughout most of the chapters; I benefited much from having him as a colleague from 1983 to 1987.

The epistemology of experiment developed in this book is broadly Peircean and I would like to acknowledge my debt to the scholarship of C. S. Peirce. Through the quandaries of virtually every chapter, his work served much like a brilliant colleague and a kindred spirit.

A portion of the rewriting took place while I was a visiting professor at the Center for Logic and Philosophy of Science at the University of Leuven, Belgium, in 1994, and I learned much from the different perspectives of colleagues there. I want to thank especially Herman Roelants, chair of the Department of Philosophy, for valuable comments on this work.

Many others gave helpful comments and criticisms on portions of this work: Richard Burian, George Barnard, George Chatfield, Norman Gilinsky, I. J. Good, Marjorie Grene, Ian Hacking, Gary Hardcastle, Valerie Hardcastle, Paul Hoyningen-Huene, William Hendricks, Joseph Pitt, J. D. Trout, and Ronald Workman.

At different times, I was greatly facilitated in the research and writing for this book by the support I received from an NEH fellowship for college teachers, an NEH summer stipend, and an NSF grant (Studies in Science, Technology, and Society Scholars Award). I am grateful to Virginia Tech and to Joseph Pitt, chair of the philosophy department,

for helping to accommodate my leaves and for endorsing this project.

Portions of chapters 3, 8, 11, and 12 have appeared in previously published articles; I thank the publishers for permission to use some of this material: "The New Experimentalism, Topical Hypotheses, and Learning From Error," in *PSA 1994*, vol. 1, edited by D. Hull, M. Forbes, and R. Burian (East Lansing, Mich.: Philosophy of Science Association, 1994), 270–79; "Novel Evidence and Severe Tests," *Philosophy of Science* 58 (1991): 523–52; "Did Pearson Reject the Neyman-Pearson Philosophy of Statistics?" *Synthese* 90 (1992): 233–62; and "The Test of Experiment: C. S. Peirce and E. S. Pearson," in *Charles S. Peirce and the Philosophy of Science: Papers from the 1989 Harvard Conference*, edited by E. Moore (Tuscaloosa, Ala.: University of Alabama Press, 1983), 161–74.

I owe special thanks to Susan Abrams and the University of Chicago Press for supporting this project even when it existed only as a ten-page summary, and for considerable help throughout. I thank David Hull for his careful and instructive review of this manuscript, Madeleine Avirov for superb copyediting, and Stacia Kozlowski for a good deal of assistance with the manuscript preparation.

I obtained valuable feedback on this manuscript from the graduate students of my seminar Foundations of Statistical Inference in 1995, especially Mary Cato, Val Larson, Jean Miller, and Randy Ward. For extremely valuable editorial and library assistance over several years, I thank Mary Cato. For organizational help and superb childcare, I am indebted to Cristin Brew and Wendy Turner.

I am grateful to my father, Louis J. Mayo, for first sparking my interest in philosophy as a child; I regret that he passed away before completion of this book. I am thankful to my mother, Elizabeth Mayo, for understanding my devotion to this project, and to my son, Isaac, for not minding using the backs of discarded drafts as drawing paper for five years. My deepest debt is to my husband, George W. Chatfield, to whom this book is dedicated.