Statistical Inference as Severe Testing

How to Get Beyond the Statistics Wars

Mounting failures of replication in the social and biological sciences give a new urgency to critically appraising proposed reforms. This book pulls back the cover on disagreements between experts charged with restoring integrity to science. It denies two pervasive views of the role of probability in inference: to assign degrees of belief, and to control error rates in a long run. If statistical consumers are unaware of assumptions behind rival evidence reforms, they can't scrutinize the consequences that affect them (in personalized medicine, psychology, and so on). The book sets sail with a simple tool: If little has been done to rule out flaws in inferring a claim, then it has not passed a severe test. Many methods advocated by data experts do not stand up to severe scrutiny, and are even in tension with successful strategies for blocking or accounting for cherry picking and selective reporting. Through a series of excursions, tours, and exhibits, the philosophy and history of inductive inference come alive, while philosophical tools are put to work to solve problems about science and pseudoscience, induction and falsification.

Deborah G. Mayo is Professor Emerita in the Department of Philosophy at Virginia Tech and is a visiting professor at the London School of Economics and Political Science, Centre for the Philosophy of Natural and Social Science. She is the author of *Error and the Growth of Experimental Knowledge* (1996), which won the 1998 Lakatos Prize awarded to the most outstanding contribution to the philosophy of science during the previous six years. She co-edited *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science* (2010, Cambridge University Press) with Aris Spanos, and has published widely in the philosophy of science, statistics, and experimental inference.

"In this lively, witty, and intellectually engaging book, Deborah Mayo returns to first principles to make sense of statistics. She takes us beyond statistical formalism and recipes, and asks us to think philosophically about the enterprise of statistical inference itself. Her contribution will be a welcomed addition to statistical learning. Mayo's timely book will shrink enlarged posteriors and overinflated significance, by focusing on whether our inferences have been severely tested, which is where we should be focused."

– Nathan A. Schachtman, Lecturer in Law, Columbia Law School

"Whether or not you agree with her basic stance on statistical inference, if you are interested in the subject – and all scientists ought to be – Deborah Mayo's writings are a must. Her views on inference are all the more valuable for being contrary to much current consensus. Her latest book will delight many and infuriate others but force all who are serious about these issues to think. Her capacity to jolt the complacent is second to none."

– Stephen Senn, author of *Dicing with Death*

"Deborah Mayo's insights into the philosophical dimensions of these problems are unsurpassed in their originality, their importance, and the breadth of understanding on which they are based. Here she combines perspectives from philosophy of science and the foundations of statistics to eliminate mirages produced by misunderstandings both philosophical and statistical, while putting into focus the ways in which her error-statistical approach is relevant to current problems of scientific inquiry in various disciplines."

– Kent Staley, Saint Louis University

"This book by Deborah Mayo is a timely examination of the use of statistics in science. Her severity requirement demands that the scientist provide a sharp question and related data. Absent that, the observer should withhold judgment or outright reject. It is time to get tough. Funding agencies should take note."

– S. Stanley Young, Ph.D., FASA FAAAS

Cambridge University Press
978-1-107-05413-4 — Statistical Inference as Severe Testing
Deborah G. Mayo
Frontmatter
More Information

Statistical Inference as Severe Testing

How to Get Beyond the Statistics Wars

Deborah G. Mayo

Virginia Tech

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107054134
DOI: 10.1017/9781107286184

First published 2018

Printed in the United States of America by Sheridan Books, Inc.

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Mayo, Deborah G., author.
Title: Statistical inference as severe testing : how to get beyond the statistics wars / Deborah G. Mayo (Virginia Tech).
Description: Cambridge : Cambridge University Press, 2018. | Includes bibliographical references and index.
Identifiers: LCCN 2018014718 | ISBN 9781107054134 (alk. paper)
Subjects: LCSH: Mathematical statistics. | Inference. | Error analysis (Mathematics) | Fallacies (Logic) | Deviation (Mathematics)
Classification: LCC QA276 .M3755 2018 | DDC 519.5/4–dc23
LC record available at https://lccn.loc.gov/2018014718

ISBN 978-1-107-05413-4 Hardback
ISBN 978-1-107-66464-7 Paperback

Additional resources for this publication at www.cambridge.org/mayo

To George W. Chatfield
for his magnificent support

Cambridge University Press
978-1-107-05413-4 — Statistical Inference as Severe Testing
Deborah G. Mayo
Frontmatter
More Information

Itinerary

Preface

The Statistics Wars

Today's "statistics wars" are fascinating: They are at once ancient and up to the minute. They reflect disagreements on one of the deepest, oldest, philosophical questions: How do humans learn about the world despite threats of error due to incomplete and variable data? At the same time, they are the engine behind current controversies surrounding high-profile failures of replication in the social and biological sciences. How should the integrity of science be restored? Experts do not agree. This book pulls back the curtain on why.

Methods of statistical inference become relevant primarily when effects are neither totally swamped by noise, nor so clear cut that formal assessment of errors is relatively unimportant. Should probability enter to capture degrees of belief about claims? To measure variability? Or to ensure we won't reach mistaken interpretations of data too often in the long run of experience? Modern statistical methods grew out of attempts to systematize doing all of these. The field has been marked by disagreements between competing tribes of frequentists and Bayesians that have been so contentious – likened in some quarters to religious and political debates – that everyone wants to believe we are long past them. We now enjoy unifications and reconciliations between rival schools, it will be said, and practitioners are eclectic, prepared to use whatever method "works." The truth is, long-standing battles still simmer below the surface in questions about scientific trustworthiness and the relationships between Big Data-driven models and theory. The reconciliations and unifications have been revealed to have serious problems, and there's little agreement on which to use or how to interpret them. As for eclecticism, it's often not clear what is even meant by "works." The presumption that all we need is an agreement on numbers – never mind if they're measuring different things – leads to pandemonium. Let's brush the dust off the pivotal debates, walk into the museums where we can see and hear such founders as Fisher, Neyman, Pearson, Savage, and many others. This is to simultaneously zero in on the arguments between metaresearchers – those doing research on research – charged with statistical reforms.

Statistical Inference as Severe Testing

Why are some arguing in today's world of high-powered computer searches that statistical findings are mostly false? The problem is that high-powered methods can make it easy to uncover impressive-looking findings even if they are false: spurious correlations and other errors have not been severely probed. We set sail with a simple tool: If little or nothing has been done to rule out flaws in inferring a claim, then it has not passed a *severe test*. In the severe testing view, probability arises in scientific contexts to assess and control how capable methods are at uncovering and avoiding erroneous interpretations of data. That's what it means to *view statistical inference as severe testing*. A claim is severely tested to the extent it has been subjected to and passes a test that probably would have found flaws, were they present. You may be surprised to learn that many methods advocated by experts do not stand up to severe scrutiny, are even in tension with successful strategies for blocking or accounting for cherry picking and selective reporting!

The severe testing perspective substantiates, using modern statistics, the idea Karl Popper promoted, but never cashed out. The goal of *highly well-tested* claims differs sufficiently from *highly probable* ones that you can have your cake and eat it too: retaining both for different contexts. Claims may be "probable" (in whatever sense you choose) but terribly tested by these data. In saying we may view statistical inference as severe testing, I'm not saying statistical inference is always about formal statistical testing. The testing metaphor grows out of the idea that before we have evidence for a claim, it must have passed an analysis that could have found it flawed. The probability that a method commits an erroneous interpretation of data is an *error probability*. Statistical methods based on error probabilities I call *error statistics*. The value of error probabilities, I argue, is not merely to control error in the long run, but because of what they teach us about the source of the data in front of us. The concept of severe testing is sufficiently general to apply to any of the methods now in use, whether for exploration, estimation, or prediction.

Getting Beyond the Statistics Wars

Thomas Kuhn's remark that only in the face of crisis "do scientists behave like philosophers" (1970), holds some truth in the current statistical crisis in science. Leaders of today's programs to restore scientific integrity have their own preconceptions about the nature of evidence and inference, and about "what we really want" in learning from data. Philosophy of science can also alleviate such conceptual discomforts. Fortunately, you needn't accept the severe testing view in order to employ it as a tool for bringing into focus the crux of all these issues. It's a tool for excavation, and for keeping us afloat in the marshes and quicksand that often mark today's controversies. Nevertheless, important consequences will follow once this tool is used. First there will be a reformulation of existing tools (tests, confidence intervals, and others) so as to avoid misinterpretations and abuses. The debates on statistical inference generally concern inference after a statistical model and data statements are in place, when in fact the most interesting work involves the local inferences needed to get to that point. A primary asset of error statistical methods is their contributions to designing, collecting, modeling, and learning from data. The severe testing view provides the much-needed link between a test's error probabilities and what's required for a warranted inference in the case at hand. Second, instead of rehearsing the same criticisms over and over again, challengers on all sides should now begin by grappling with the arguments we trace within. Kneejerk assumptions about the superiority of one or another method will not do. Although we'll be excavating the actual history, it's the methods themselves that matter; they're too important to be limited by what someone 50, 60, or 90 years ago thought, or to what today's discussants *think* they thought.

Who is the Reader of This Book?

This book is intended for a wide-ranging audience of readers. It's directed to consumers and practitioners of statistics and data science, and anyone interested in the methodology, philosophy, or history of statistical inference, or the controversies surrounding widely used statistical methods across the physical, social, and biological sciences. You might be a researcher or science writer befuddled by the competing recommendations offered by large groups ("megateams") of researchers (should *P*-values be set at 0.05 or 0.005, or not set at all?). By viewing a contentious battle in terms of a difference in goals – finding highly probable versus highly well-probed hypotheses – readers can see why leaders of rival tribes often talk right past each other. A fair-minded assessment may finally be possible. You may have a skeptical bent, keen to hold the experts accountable. Without awareness of the assumptions behind proposed reforms you can't scrutinize consequences that will affect you, be it in medical advice, economics, or psychology.

Your interest may be in improving statistical pedagogy, which requires, to begin with, recognizing that no matter how sophisticated the technology has become, the nature and meaning of basic statistical concepts are more unsettled than ever. You could be teaching a methods course in psychology wishing to intersperse philosophy of science in a way that is both serious and connected to immediate issues of practice. You might be an introspective statistician, focused on applications, but wanting your arguments to be on surer philosophical grounds.

Viewing statistical inference as severe testing will offer philosophers of science new avenues to employ statistical ideas to solve philosophical problems of induction, falsification, and demarcating science from pseudoscience. Philosophers of experiment should find insight into how statistical modeling bridges gaps between scientific theories and data. Scientists often question the relevance of philosophy of science to scientific practice. Through a series of excursions, tours, and exhibits, tools from the philosophy and history of statistics will be put directly to work to illuminate and solve problems of practice. I hope to galvanize philosophers of science and experimental philosophers to further engage with the burgeoning field of data science and reproducibility research.

Fittingly, the deepest debates over statistical foundations revolve around very simple examples, and I stick to those. This allows getting to the nitty-gritty logical issues with minimal technical complexity. If there's disagreement even there, there's little hope with more complex problems. (I try to use the notation of discussants, leading to some variation.) The book would serve as a one-semester course, or as a companion to courses on research methodology, philosophy of science, or interdisciplinary research in science and society. Each tour gives a small set of central works from statistics or philosophy, but since the field is immense, I reserve many important references for further reading on the CUP-hosted webpage for this book, www.cambridge.org/mayo.

Relation to Previous Work

While (1) philosophy of science provides important resources to tackle foundational problems of statistical practice, at the same time, (2) the statistical method offers tools for solving philosophical problems of evidence and inference. My earlier work, such as *Error and the Growth of Experimental Knowledge* (1996), falls under the umbrella of (2), using statistical science for philosophy of science: to model scientific inference, solve problems about evidence (problem of induction), and evaluate methodological rules (does more weight accrue to a hypothesis if it is prespecified?). *Error and Inference* (2010), with its joint work and exchanges with philosophers and statisticians, aimed to bridge the two-way street of (1) and (2). This work, by contrast, falls under goal (1): tackling foundational problems of statistical practice. While doing so will constantly find us entwined with philosophical problems of inference, it is the arguments and debates currently engaging practitioners that take the lead for our journey.

Join me, then, on a series of six excursions and 16 tours, during which we will visit three leading museums of statistical science and philosophy of science, and engage with a host of tribes marked by family quarrels, peace treaties, and shifting alliances.[1]

[1] A bit of travel trivia for those who not only read to the end of prefaces, but check its footnotes: two museums will be visited twice, one excursion will have no museums. With one exception, we engage current work through interaction with tribes, not museums. There's no extra cost for the 26 souvenirs: A–Z.

Acknowledgments

I am deeply grateful to my colleague and frequent co-author, Aris Spanos. More than anyone else, he is to be credited for encouraging the connection between the error statistical philosophy of Mayo (1996) and statistical practice, and for developing an error statistical account of misspecification testing. I thank him for rescuing me time and again from being stalled by one or another obstacle. He has given me massive help with the technical aspects of this book, and with revisions to countless drafts of the entire manuscript.

My ideas were importantly influenced by Sir David Cox. My debt is to his considerable work on statistical principles, his involvement in conferences in 2006 (at Virginia Tech) and 2010 (at the London School of Economics), and through writing joint papers (2006,[2] 2010[3]). I thank him for his steadfast confidence in this project, and for discussions leading to my identifying the unsoundness in arguments for the (strong) Likelihood Principle (Mayo 2014b) – an important backdrop to the evidential interpretation of error probabilities that figures importantly within.

I have several people to thank for valuable ideas on many of the topics in this book through extensive blog comments (errorstatistics.com), and/or discussions on portions of this work: John Byrd, Nancy Cartwright, Robert Cousins, Andrew Gelman, Gerd Gigerenzer, Richard Gill, Prakash Gorroochurn, Sander Greenland, Brian Haig, David Hand, Christian Hennig, Thomas Kepler, Daniël Lakens, Michael Lew, Oliver Maclaren, Steven McKinney, Richard Morey, Keith O'Rourke, Caitlin Parker, Christian Roberts, Nathan Schachtman, Stephen Senn, Cosma Shalizi, Kent Staley, Niels Waller, Larry Wasserman, Corey Yanofsky, and Stanley Young.

Older debts recalled are to discussions and correspondence with George Barnard, Ronald Giere, Erich Lehmann, Paul Meehl, and Karl Popper.

Key ideas in this work grew out of exchanges with Peter Achinstein, Alan Chalmers, Clark Glymour, Larry Laudan, Alan Musgrave, and John Worrall, published in Mayo and Spanos (2010). I am grateful for the stimulating conversations on aspects of this research during seminars and conferences at the London School of Economics in 2008, 2010, 2012, and 2016, Centre for Philosophy of Natural and Social Sciences. I thank Virginia Tech and Doug Lind, chair of the philosophy department, for support and professional accommodations which were essential to this project. I obtained valuable feedback

[2] *First Symposium on Philosophy, History, and Methodology of Error (ERROR 06)*, Virginia Tech.
[3] *Statistical Science & Philosophy of Science: Where Do/Should They Meet in 2010 (and Beyond)?* The London School of Economics and Political Science, Centre for the Philosophy of Natural and Social Science.

from graduate students of a 2014 seminar (with A. Spanos) on Statistical Inference and Modeling at Virginia Tech.

I owe special thanks to Diana Gillooly and Cambridge University Press for supporting this project even when it existed only as a ten-page summary, and for her immense help throughout. I thank Esther Migueliz, Margaret Patterson, and Adam Kratoska for assistance in the production and preparation of this manuscript. For the figures in this work, I'm very appreciative for all Marcos Jiménez' work. I am grateful to Mickey Mayo for graphics for the online component. I thank Madeleine Avirov, Mary Cato, Michael Fay, Nicole Jinn, Caitlin Parker, and Ellen Woodall for help with the copy-editing. For insightful comments and a scrupulous review of this manuscript, copy-editing, library, and indexing work, I owe mammoth thanks to Jean Anne Miller. For other essential support, I am indebted to Melodie Givens and William Hendricks.

I am grateful to my son, Isaac Chatfield, for technical assistance, proofing, and being the main person to cook real food. My deepest debt is to my husband, George W. Chatfield, for his magnificent support of me and the study of E.R.R.O.R.S.[4] I dedicate this book to him.

[4] Experimental Reasoning, Reliability, and the Objectivity, and Rationality of Science.