“How and why is computational statistics taking over the world? In this serious work of synthesis that is also fun to read, Efron and Hastie, two pioneers in the integration of parametric and nonparametric statistical ideas, give their take on the unreasonable effectiveness of statistics and machine learning in the context of a series of clear, historically informed examples.”

— Andrew Gelman, *Columbia University*

“This unusual book describes the nature of statistics by displaying multiple examples of the way the field has evolved over the past 60 years, as it has adapted to the rapid increase in available computing power. The authors’ perspective is summarized nicely when they say, ‘Very roughly speaking, algorithms are what statisticians do, while inference says why they do them.’ The book explains this ‘why’; that is, it explains the purpose and progress of statistical research, through a close look at many major methods, methods the authors themselves have advanced and studied at great length. Both enjoyable and enlightening, *Computer Age Statistical Inference* is written especially for those who want to hear the big ideas, and see them instantiated through the essential mathematics that defines statistical analysis. It makes a great supplement to the traditional curricula for beginning graduate students.”

— Rob Kass, *Carnegie Mellon University*

“This is a terrific book. It gives a clear, accessible, and entertaining account of the interplay between theory and methodological development that has driven statistics in the computer age. The authors succeed brilliantly in locating contemporary algorithmic methodologies for analysis of ‘big data’ within the framework of established statistical theory.”

— Alastair Young, *Imperial College London*

“This is a guided tour of modern statistics that emphasizes the conceptual and computational advances of the last century. Authored by two masters of the field, it offers just the right mix of mathematical analysis and insightful commentary.”

— Hal Varian, *Google*

“Efron and Hastie guide us through the maze of breakthrough statistical methodologies following the computing evolution: why they were developed, their properties, and how they are used. Highlighting their origins, the book helps us understand each method’s roles in inference and/or prediction. The inference–prediction distinction maintained throughout the book is a welcome and important novelty in the landscape of statistics books.”

— Galit Shmueli, *National Tsing Hua University*

“A masterful guide to how the inferential bases of classical statistics can provide a principled disciplinary frame for the data science of the twenty-first century.”

— Stephen Stigler, *University of Chicago, author of*
Seven Pillars of Statistical Wisdom

"*Computer Age Statistical Inference* offers a refreshing view of modern statistics. Algorithmics are put on equal footing with intuition, properties, and the abstract arguments behind them. The methods covered are indispensable to practicing statistical analysts in today's big data and big computing landscape."

— Robert Gramacy, *The University of Chicago Booth School of Business*

"Every aspiring data scientist should carefully study this book, use it as a reference, and carry it with them everywhere. The presentation through the two-and-a-half-century history of statistical inference provides insight into the development of the discipline, putting data science in its historical place."

— Mark Girolami, *Imperial College London*

"Efron and Hastie are two immensely talented and accomplished scholars who have managed to brilliantly weave the fiber of 250 years of statistical inference into the more recent historical mechanization of computing. This book provides the reader with a mid-level overview of the last 60-some years by detailing the nuances of a statistical community that, historically, has been self-segregated into camps of Bayes, frequentist, and Fisher yet in more recent years has been unified by advances in computing. What is left to be explored is the emergence of, and role that, big data theory will have in bridging the gap between data science and statistical methodology. Whatever the outcome, the authors provide a vision of high-speed computing having tremendous potential to enable the contributions of statistical inference toward methodologies that address both global and societal issues."

— Rebecca Doerge, *Carnegie Mellon University*

"In this book, two masters of modern statistics give an insightful tour of the intertwined worlds of statistics and computation. Through a series of important topics, Efron and Hastie illuminate how modern methods for predicting and understanding data are rooted in both statistical and computational thinking. They show how the rise of computational power has transformed traditional methods and questions, and how it has pointed us to new ways of thinking about statistics."

— David Blei, *Columbia University*

Absolutely brilliant. This beautifully written compendium reviews many big statistical ideas, including the authors' own. A must for anyone engaged creatively in statistics and the data sciences, for repeated use. Efron and Hastie demonstrate the ever-growing power of statistical reasoning, past, present, and future.

— Carl Morris, *Harvard University*

Computer Age Statistical Inference

The twenty-first century has seen a breathtaking expansion of statistical methodology, both in scope and in influence. “Big data,” “data science,” and “machine learning” have become familiar terms in the news, as statistical methods are brought to bear upon the enormous data sets of modern science and commerce. How did we get here? And where are we going?

This book takes us on an exhilarating journey through the revolution in data analysis following the introduction of electronic computation in the 1950s. Beginning with classical inferential theories – Bayesian, frequentist, Fisherian – individual chapters take up a series of influential topics: survival analysis, logistic regression, empirical Bayes, the jackknife and bootstrap, random forests, neural networks, Markov chain Monte Carlo, inference after model selection, and dozens more. The distinctly modern approach integrates methodology and algorithms with statistical inference. The book ends with speculation on the future direction of statistics and data science.

BRADLEY EFRON is Max H. Stein Professor, Professor of Statistics, and Professor of Biomedical Data Science at Stanford University. He has held visiting faculty appointments at Harvard, UC Berkeley, and Imperial College London. Efron has worked extensively on theories of statistical inference, and is the inventor of the bootstrap sampling technique. He received the National Medal of Science in 2005 and the Guy Medal in Gold of the Royal Statistical Society in 2014.

TREVOR HASTIE is John A. Overdeck Professor, Professor of Statistics, and Professor of Biomedical Data Science at Stanford University. He is coauthor of *Elements of Statistical Learning*, a key text in the field of modern data analysis. He is also known for his work on generalized additive models and principal curves, and for his contributions to the R computing environment. Hastie was awarded the Emmanuel and Carol Parzen prize for Statistical Innovation in 2014.

INSTITUTE OF MATHEMATICAL STATISTICS MONOGRAPHS

IMS Monographs are concise research monographs of high quality on any branch of statistics or probability of sufficient interest to warrant publication as books. Some concern relatively traditional topics in need of up-to-date assessment. Others are on emerging themes. In all cases the objective is to provide a balanced view of the field.

Other Books in the Series

1. *Large-Scale Inference,* by Bradley Efron
2. *Nonparametric Inference on Manifolds,* by Abhishek Bhattacharya and Rabi Battacharya
3. *The Skew-Normal and Related Families*, by Adelchi Azzalini
4. *Case-Control Studies,* by Ruth H. Keogh and D. R. Cox
5. *Computer Age Statistical Inference*, by Bradley Efron and Trevor Hastie

Computer Age Statistical Inference

Algorithms, Evidence, and Data Science

BRADLEY EFRON
Stanford University, California

TREVOR HASTIE
Stanford University, California

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, 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

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107149892

First published 2016
8th printing 2018

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

ISBN 978-1-107-14989-2 Hardback

Cambridge University Press & Assessment has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

To Donna and Lynda

Contents

Preface

Statistical inference is an unusually wide-ranging discipline, located as it is at the triple-point of mathematics, empirical science, and philosophy. The discipline can be said to date from 1763, with the publication of Bayes' rule (representing the philosophical side of the subject; the rule's early advocates considered it an argument for the existence of God). The most recent quarter of this 250-year history—from the 1950s to the present—is the "computer age" of our book's title, the time when computation, the traditional bottleneck of statistical applications, became faster and easier by a factor of a million.

The book is an examination of how statistics has evolved over the past sixty years—an aerial view of a vast subject, but seen from the height of a small plane, not a jetliner or satellite. The individual chapters take up a series of influential topics—generalized linear models, survival analysis, the jackknife and bootstrap, false-discovery rates, empirical Bayes, MCMC, neural nets, and a dozen more—describing for each the key methodological developments and their inferential justification.

Needless to say, the role of electronic computation is central to our story. This doesn't mean that every advance was computer-related. A land bridge had opened to a new continent but not all were eager to cross. Topics such as empirical Bayes and James–Stein estimation could have emerged just as well under the constraints of mechanical computation. Others, like the bootstrap and proportional hazards, were pureborn children of the computer age. Almost all topics in twenty-first-century statistics are now computer-dependent, but it will take our small plane a while to reach the new millennium.

Dictionary definitions of statistical inference tend to equate it with the entire discipline. This has become less satisfactory in the "big data" era of immense computer-based processing algorithms. Here we will attempt, not always consistently, to separate the two aspects of the statistical enterprise: algorithmic developments aimed at specific problem areas, for instance random forests for prediction, as distinct from the inferential arguments offered in their support.

Very broadly speaking, algorithms are what statisticians do while inference says why they do them. A particularly energetic brand of the statistical enterprise has flourished in the new century, *data science*, emphasizing algorithmic thinking rather than its inferential justification. The later chapters of our book, where large-scale prediction algorithms such as boosting and deep learning are examined, illustrate the data-science point of view. (See the epilogue for a little more on the sometimes fraught statistics/data science marriage.)

There are no such subjects as Biological Inference or Astronomical Inference or Geological Inference. Why do we need “Statistical Inference”? The answer is simple: the natural sciences have nature to judge the accuracy of their ideas. Statistics operates one step back from Nature, most often interpreting the observations of natural scientists. Without Nature to serve as a disinterested referee, we need a system of mathematical logic for guidance and correction. Statistical inference is that system, distilled from two and a half centuries of data-analytic experience.

The book proceeds historically, in three parts. The great themes of classical inference, Bayesian, frequentist, and Fisherian, reviewed in Part I, were set in place before the age of electronic computation. Modern practice has vastly extended their reach without changing the basic outlines. (An analogy with classical and modern literature might be made.) Part II concerns early computer-age developments, from the 1950s through the 1990s. As a transitional period, this is the time when it is easiest to see the effects, or noneffects, of fast computation on the progress of statistical methodology, both in its theory and practice. Part III, “Twenty-First-Century topics,” brings the story up to the present. Ours is a time of enormously ambitious algorithms (“machine learning” being the somewhat disquieting catchphrase). Their justification is the ongoing task of modern statistical inference.

Neither a catalog nor an encyclopedia, the book’s topics were chosen as apt illustrations of the interplay between computational methodology and inferential theory. Some missing topics that might have served just as well include time series, general estimating equations, causal inference, graphical models, and experimental design. In any case, there is no implication that the topics presented here are the only ones worthy of discussion.

Also underrepresented are asymptotics and decision theory, the “math stat” side of the field. Our intention was to maintain a technical level of discussion appropriate to Masters’-level statisticians or first-year PhD students. Inevitably, some of the presentation drifts into more difficult waters, more from the nature of the statistical ideas than the mathematics. Readers who find our aerial view circling too long over some topic shouldn't hesitate to move ahead in the book. For the most part, the chapters can be read independently of each other (though there is a connecting overall theme). This comment applies especially to nonstatisticians who have picked up the book because of interest in some particular topic, say survival analysis or boosting.

Useful disciplines that serve a wide variety of demanding clients run the risk of losing their center. Statistics has managed, for the most part, to maintain its philosophical cohesion despite a rising curve of outside demand. The center of the field has in fact moved in the past sixty years, from its traditional home in mathematics and logic toward a more computational focus. Our book traces that movement on a topic-by-topic basis. An answer to the intriguing question "What happens next?" won't be attempted here, except for a few words in the epilogue, where the rise of data science is discussed.

Acknowledgments

We are indebted to Cindy Kirby for her skillful work in the preparation of this book, and Galit Shmueli for her helpful comments on an earlier draft. At Cambridge University Press, a huge thank you to Steven Holt for his excellent copy editing, Clare Dennison for guiding us through the production phase, and to Diana Gillooly, our editor, for her unfailing support.

Bradley Efron
Trevor Hastie
Department of Statistics
Stanford University
May 2016

Notation

Throughout the book the numbered † sign indicates a technical note or reference element which is elaborated on at the end of the chapter. There, next to the number, the page number of the referenced location is given in parenthesis. For example, `lowess` in the notes on page 11 was referenced via a $\dagger_1$ on page 6. Matrices such as $\mathbf{\Sigma}$ are represented in bold font, as are certain vectors such as $\boldsymbol{y}$, a data vector with n elements. Most other vectors, such as coefficient vectors, are typically not bold. We use a dark green `typewriter` font to indicate data set names such as `prostate`, variable names such as `prog` from data sets, and `R` commands such as `glmnet` or `locfdr`. No bibliographic references are given in the body of the text; important references are given in the endnotes of each chapter.