Principles of Statistical Analysis

This compact course is written for the mathematically literate reader who wants to learn to analyze data in a principled fashion. The language of mathematics enables clear exposition that can go quite deep, quite quickly, and naturally supports an axiomatic and inductive approach to data analysis. Starting with a good grounding in probability, the reader moves to statistical inference via topics of great practical importance – simulation and sampling, as well as experimental design and data collection – that are typically displaced from introductory accounts. The core of the book then covers both standard methods and such advanced topics as multiple testing, meta-analysis, and causal inference.

ERY ARIAS-CASTRO is a professor in the Department of Mathematics and in the Halıcıoğlu Data Science Institute at the University of California, San Diego, where he specializes in theoretical statistics and machine learning. His education includes a bachelor’s degree in mathematics and a master’s degree in artificial intelligence, both from École Normale Supérieure de Cachan (now École Normale Supérieure Paris-Saclay) in France, as well as a Ph.D. in statistics from Stanford University in the United States.

"With the rapid development of data-driven decision making, statistical methods have become indispensable in countless domains of science, engineering, and management science, to name a few. Ery Arias-Castro's excellent text gives a self-contained and remarkably broad exposition of the current diversity of concepts and methods developed to tackle the challenges of data science. Simply put, everyone serious about understanding the theory behind data science should be exposed to the topics covered in this book."

—Philippe Rigollet, Professor
Department of Mathematics, Massachusetts Institute of Technology

"A course on statistical modeling and inference has been a staple of many first-year graduate engineering programs. While there are many excellent textbooks on this subject, much of the material is inspired by models of physical systems, and as such these books deal extensively with parametric inference. The emerging data revolution, on the other hand, requires an engineering student to develop an understanding of statistical inference rooted in problems inspired by data-driven applications, and this book fills that need. Arias-Castro weaves together diverse concepts such as data collection, sampling, and inference in a unified manner. He lucidly presents the mathematical foundations of statistical data analysis, and covers advanced topics on data analysis. With over 700 problems and computer exercises, this book will serve the needs of beginner and advanced engineering students alike."

—Venkatesh Saligrama, Professor
Data Science Faculty Fellow, Department of Electrical and Computer Engineering,
Department of Computer Science (by courtesy), Boston University

"In this book, aimed at senior undergraduates or beginning graduate students with a reasonable mathematical background, the author proposes a self-contained and yet concise introduction to statistical analysis. By putting a strong emphasis on the randomization principle, he provides a coherent and elegant perspective on modern statistical practice. Some of the later chapters also form a good basis for a reading group. I will be recommending this excellent book to my collaborators."

—Nicolas Verzelen, Associate Professor
Mathematics, Computer Science, Physics, and Systems Department,
University of Montpellier

"This text is highly recommended for undergraduate students wanting to grasp the key ideas of modern data analysis. Arias-Castro achieves something that is rare in the art of teaching statistical science – he uses mathematical language in an intelligible and highly helpful way, without surrendering key intuitions of statistics to formalism and proof. In this way, the reader can get through an impressive amount of material without, however, ever getting into muddy waters."

—Richard Nickl, Professor
Statistical Laboratory, Cambridge University

INSTITUTE OF MATHEMATICAL STATISTICS TEXTBOOKS

IMS Textbooks give introductory accounts of topics of current concern suitable for advanced courses at master's level, for doctoral students and for individual study. They are typically shorter than a fully developed textbook, often arising from material created for a topical course. Lengths of 100–290 pages are envisaged. The books typically contain exercises.

In collaboration with the International Society for Bayesian Analysis (ISBA), selected volumes in the IMS Textbooks series carry the "with ISBA" designation at the recommendation of the ISBA editorial representative.

Other Books in the Series (*with ISBA)

1. *Probability on Graphs*, by Geoffrey Grimmett
2. *Stochastic Networks*, by Frank Kelly and Elena Yudovina
3. *Bayesian Filtering and Smoothing*, by Simo Särkkä
4. *The Surprising Mathematics of Longest Increasing Subsequences*, by Dan Romik
5. *Noise Sensitivity of Boolean Functions and Percolation*, by Christophe Garban and Jeffrey E. Steif
6. *Core Statistics*, by Simon N. Wood
7. *Lectures on the Poisson Process*, by Günter Last and Mathew Penrose
8. *Probability on Graphs (Second Edition)*, by Geoffrey Grimmett
9. *Introduction to Malliavin Calculus*, by David Nualart and Eulàlia Nualart
10. *Applied Stochastic Differential Equations*, by Simo Särkkä and Arno Solin
11. **Computational Bayesian Statistics*, by M. Antónia Amaral Turkman, Carlos Daniel Paulino, and Peter Müller
12. *Statistical Modelling by Exponential Families*, by Rolf Sundberg
13. *Two-Dimensional Random Walk: From Path Counting to Random Interlacements*, by Serguei Popov
14. *Scheduling and Control of Queueing Networks*, by Gideon Weiss

Principles of Statistical Analysis

Learning from Randomized Experiments

ERY ARIAS-CASTRO
University of California, San Diego

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

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

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/9781108489676
DOI: 10.1017/9781108779197

First published 2022

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

ISBN 978-1-108-48967-6 Hardback
ISBN 978-1-108-74744-8 Paperback

Cambridge University Press 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.

I would like to dedicate this book to some professors that have, along the way, inspired, supported, and mentored me in my studies and academic career, and to whom I am eternally grateful:

David L. Donoho
my doctoral thesis advisor

Persi Diaconis
my first co-author on a research article

Yves Meyer
my master's thesis advisor

Robert Azencott
my undergraduate thesis advisor

In controlled experimentation it has been found not difficult to introduce explicit and objective randomization in such a way that the tests of significance are demonstrably correct. In other cases we must still act in the faith that Nature has done the randomization for us. [...] We now recognize randomization as a postulate necessary to the validity of our conclusions, and the modern experimenter is careful to make sure that this postulate is justified.

Ronald A. Fisher
International Statistical Conferences, 1947

Contents

Part I Elements of Probability Theory

Part II Practical Considerations

Part III Elements of Statistical Inference

Preface

This book is intended for the mathematically literate reader who wants to understand how to analyze data in a principled fashion. The language of mathematics allows for a more concise, and arguably clearer exposition that can go quite deep, quite quickly, and naturally accommodates an axiomatic and inductive approach to data analysis, which is the raison d'être of the book. To elaborate, the book starts with a preliminary foundation in probability theory, continues with an intermezzo of sampling and data collection, and finally moves to statistical inference – the core of the book which includes, in addition to standard topics, more advanced ones such as multiple testing, meta-analysis, and causal inference. The book thus provides a self-contained exposition of fundamental principles and methods of statistical analysis, covering topics which are typically displaced from introductory, general accounts. Emphasis is on inference, and more exploratory approaches to data analysis such as clustering and dimensionality reduction are not covered.

The compact treatment is grounded in mathematical theory and concepts, and is fairly rigorous, even though measure theoretic matters are kept in the background, and most proofs are left as problems. In fact, much of the learning is accomplished through embedded problems – around 700 of them! Some problems call for mathematical derivations, and assume a certain comfort with calculus, or even real analysis. Other problems require basic programming on a computer.

Structure

The book is divided into three parts. The introduction to probability, in Part I, stands as the mathematical foundation for statistical inference. Indeed, without a solid foundation in probability and, in particular, a good understanding of how experiments are modeled, there is no clear distinction between descriptive and inferential analyses. The exposition there is quite standard. It starts by introducing Kolmogorov's axioms, which are instantiated in the context of discrete sample spaces. The narrative then transitions to a comprehensive discussion of distributions on the real line, both discrete and continuous, and also multivariate. This is followed by an introduction of the basic concentration inequalities and limit theorems. (A construction of the Lebesgue integral is not included, and measure-theoretic matters are mostly avoided.) Part I ends with a brief discussion of Markov chains and related stochastic processes.

Some utilitarian, but absolutely critical, aspects of probability and statistics are discussed in Part II. These include probability sampling and pseudo-random number generation – the practical side of randomness; as well as survey sampling and experimental design – the practical side of data collection.

Part III is the core of the book. It attempts to build a theory of statistical inference from first principles. The foundation is randomization, either controlled by design or assumed to be natural. In either case, randomization provides the essential randomness needed to justify probabilistic modeling. It naturally leads to conditional inference, and allows for causal inference. In this framework, permutation tests play a special, almost canonical role. Monte Carlo sampling, performed on a computer, is presented as an alternative to complex mathematical derivations, and the bootstrap is then introduced as an accommodation when the sampling distribution is not directly available and has to be estimated.

What is not here

I do not find normal models to be particularly compelling: unless there is a central limit theorem at play, there is no real reason to believe numerical data are normally distributed. Normal models are thus mentioned only in passing. More generally, parametric models are not emphasized – except for those that arise naturally in some experiments.

The usual emphasis on parametric inference is, I find, misplaced and misleading, as it can be (and often is) introduced independently of how the data were gathered, thus creating a chasm that separates the design of experiments and the analysis of the resulting data. Bayesian modeling is, consequently, not covered beyond basic definitions in the context of average risk optimality. Linear models and time series are not discussed in any detail. As is typically the case for an introductory book, especially of this length and at this level, there is only a hint of abstract decision theory, and multivariate analysis is omitted entirely.

Cambridge University Press & Assessment
978-1-108-48967-6 — Principles of Statistical Analysis
Ery Arias-Castro
Frontmatter
More Information

How to use this Book

The idea for this book arose from a dissatisfaction with how statistical analysis is typically taught at the undergraduate and master's levels, coupled with an inspiration for weaving a narrative, which I find more compelling.

This narrative was formed over years of teaching statistics at the University of California, San Diego, in particular an undergraduate-level course on *computational statistics* focusing on resampling methods of inference. As it stands, however, the book is perhaps best used for independent study.

The reader is invited to progress through the book in the order in which the material is presented, working on the problems as they come, and saving those that seem harder for later. If an experienced instructor or tutor is available as an occasional guide, it is worthwhile to tackle even the harder problems when they are encountered.

Although the text emphasizes a conceptual understanding of data analysis, it is also grounded in practice. A large number of articles in the applied sciences are cited with the intention of providing the reader with a sense of how statistics is used in real life. In addition, a companion R notebook is provided to facilitate the transition from theory to practice. It is available from the author's webpage [`https://math.ucsd.edu/~eariasca`].

Intention

The book introduces, what I believe, are essential concepts that I would want a student graduating with a bachelor's or master's degree in statistics to have been exposed to, even if only in passing.

My main hope in writing this book is that it seduces mathematically minded people into learning more about statistical analysis, at least for their personal enrichment, particularly in this age of artificial intelligence, machine learning, and data science more broadly.

Acknowledgements

Ian Abramson, Emmanuel Candès, Persi Diaconis, David Donoho, Larry Goldstein, Gábor Lugosi, Dimitris Politis, Philippe Rigollet, Joseph Romano, Jason Schweinsberg, and Nicolas Verzelen, provided feedback or pointers on particular topics. Clément Berenfeld, Zexin Pan, and Yunhao Sun, as well as two anonymous referees and a copyeditor, Benjamin Johnson, retained by Cambridge, helped proofread the manuscript (all remaining errors are, of course, mine). Diana Gillooly, and later Becca Grainger, were my main contacts at Cambridge, offering expert guidance through the publishing process. I am grateful for all this generous support.

I looked at a number of textbooks in Probability and Statistics, and also lecture notes. The most important ones are cited in the text.

I used a number of freely available resources contributed by countless people all over the world. The manuscript was typeset in LaTeX (using TeXShop, TeXstudio, TeXworks, BibDesk, and JabRef) and the figures were produced using R via RStudio. The companion R notebook was written using Bookdown. I made extensive use of Google Scholar, Wikipedia, and some online discussion lists such as StackExchange.