Foundations of Agnostic Statistics

Reflecting a sea change in how empirical research has been conducted over the past three decades, *Foundations of Agnostic Statistics* presents an innovative treatment of modern statistical theory for the social and health sciences. This book develops the fundamentals of what the authors call agnostic statistics, which considers what can be learned about the world without assuming that there exists a simple generative model that can be known to be true. Aronow and Miller provide the foundations for statistical inference for researchers unwilling to make assumptions beyond what they or their audience would find credible. Building from first principles, the book covers topics including estimation theory, regression, maximum likelihood, missing data, and causal inference. Using these principles, readers will be able to formally articulate their targets of inquiry, distinguish substantive assumptions from statistical assumptions, and ultimately engage in cutting-edge quantitative empirical research that contributes to human knowledge.

Peter M. Aronow is an associate professor of Political Science, Public Health (Biostatistics), and Statistics and Data Science at Yale University and is affiliated with the university's Institution for Social and Policy Studies, Center for the Study of American Politics, Institute for Network Science, and Operations Research Doctoral Program.

Benjamin T. Miller is a doctoral candidate in Political Science at Yale University. Mr. Miller holds a BA in Economics and Mathematics from Amherst College (2012).

Foundations of Agnostic Statistics

PETER M. ARONOW
Yale University

BENJAMIN T. MILLER
Yale University

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/9781107178915

DOI: 10.1017/9781316831762

First published 2019 (version 4, February 2023)

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

Library of Congress Cataloging-in-Publication data
Names: Aronow, Peter M., author. | Miller, Benjamin T., author.
Title: Foundations of agnostic statistics / Peter M. Aronow, Benjamin T. Miller.
Description: New York : Cambridge University Press, 2019.
Identifiers: LCCN 2018039877 | ISBN 9781107178915 (hardback) |
ISBN 9781316631140 (paperback)
Subjects: LCSH: Quantitative research. | Statistics. |
BISAC: POLITICAL SCIENCE / General.
Classification: LCC QA76.9.Q36 A76 2019 |
DDC 001.4/2–dc23 LC record available at https://lccn.loc.gov/2018039877

ISBN 978-1-107-17891-5 Hardback
ISBN 978-1-316-63114-0 Paperback

For Our Parents

I do not pretend to know where many ignorant men are sure—that is all that agnosticism means.

— Clarence Darrow

Contents

Tables and Figures

TABLES

FIGURES

Introduction

> *Humans are allergic to change. They love to say, "We've always done it this way." I try to fight that. That's why I have a clock on my wall that runs counter-clockwise.*
>
> —Grace Hopper

The last three decades have seen a marked change in the manner in which quantitative empirical inquiry in the social and health sciences is conducted. Sometimes dubbed the "credibility revolution," this change has been characterized by a growing acknowledgment that the evidence that researchers adduce for their claims is often predicated on unsustainable assumptions. Our understanding of statistical and econometric tools has needed to change accordingly. We have found that conventional textbooks, which often begin with incredible modeling assumptions, are not well suited as a starting point for credible research.

We designed this book as a first course in statistical inquiry to accommodate the needs of this evolving approach to quantitative empirical research. Our book develops the fundamentals of what we call *agnostic statistics*. With agnostic statistics, we attempt to consider what can be learned about the world without assuming that there exists a simple generative model that can be known to be true. We provide the foundations for statistical inference for researchers unwilling to make assumptions beyond what they or their audience would find credible.

Under the agnostic paradigm, there is little magic required for statistical inquiry. Armed with the tools developed in this book, readers will be able to critically evaluate the credibility of both applied work and statistical methods under the assumptions warranted by their substantive context. Additionally, building from the principles established in the book, readers will be able to formally articulate their targets of inquiry, distinguish substantive assumptions from statistical assumptions, and ultimately engage in cutting-edge quantitative empirical research that contributes to human knowledge.

WHAT IS IN THIS BOOK?

In Part I (Probability, Chapters 1 and 2), we begin by presenting a canonical mathematical formulation of randomness: the notion of probability. We can neatly describe random events using a set-theoretic formalization that largely agrees with our intuitions. Furthermore, when events can be quantified, we can represent random generative processes with *random variables*. We show that probability theory provides us with a clear language for describing features of random generative processes, even when the structure of those processes is not fully known.

In Part II (Statistics, Chapters 3, 4, and 5), we engage with data. If the researcher can collect data produced by repeated, independent draws from some random generative process, we can learn about some of the characteristics of the process that generated them without any further assumptions. We can estimate features of this random generative process (for example, "our guess of the average height in this population is 5.6 feet"), and we can even make probabilistic statements describing the uncertainty of our estimates (for example, "we can state with 95% confidence that our guess lies within 0.2 feet of the true average height"). Simple statistical methods for estimation and inference (based on the *plug-in principle*), including standard tools such as ordinary least squares regression and maximum likelihood estimation, allow us to approximate these features without assuming the validity of a restrictive model.

In Part III (Identification, Chapters 6 and 7), we show how the statistical foundations of an agnostic approach to statistics naturally allow us to draw ties between features of a probability distribution and substantive processes. This task, of course, necessitates detailed knowledge of the process at hand. We discuss assumptions with clear substantive interpretations that allow us to generalize from the statistical model to broader phenomena, including missing data and causal inference. These *identification* assumptions can be viewed as separable from those embedded in the agnostic approach to statistical inference, thus laying bare the sources of our inferential leverage.

WHAT DO I NEED TO KNOW TO READ THIS BOOK?

We expect that the readers of this book will have had some exposure to the ideas of probability and statistics at the undergraduate level; while not required, it will significantly ease the readers' experience with the book. Some mild calculus will be used—nothing much beyond partial derivatives and integrals, and even then, numerical methods will typically suffice for anything more complicated than a polynomial. Some elementary set theory will also be required for our exposition of probability theory. Notably, we avoid the gratuitous use of linear algebra, and readers need not have prior training in order to engage with the text. Concepts from more advanced areas of mathematics

(such as measure theory) appear in some technical footnotes, but these can be safely ignored by readers not yet fully comfortable with these subjects. We try to include the proofs of as many of the theorems and principles in this book as possible, though we omit those that would be tedious or require advanced mathematics not covered here.

HOW SHOULD I READ THIS BOOK?

This book is intended to serve as an introductory graduate-level course in statistics as well as a reference work for more experienced researchers. We have therefore attempted to write this book in a manner that is both accessible to readers with minimal background in statistics and useful to those with more advanced training. One way in which we attempt to strike this balance is through extensive use of footnotes to provide clarification and commentary on technical points. These are provided mainly to answer questions that some sophisticated readers might raise and to note how certain concepts and theorems can be extended. Some readers may find these notes useful, but in general, they can be safely ignored by readers new to the subject. Similarly, we provide references at the end of each chapter to other texts that discuss the subjects we cover in greater detail.

Though this is not primarily a book on probability theory, it does make extensive and rigorous use of the concepts and theorems of this foundational area of mathematics. Thus, our treatment of probability in Part I is abbreviated but mathematically dense. Mathematically sophisticated readers (that is, those comfortable with the concepts and notation of calculus, basic set theory, and proofs) should have little difficulty learning the essentials of probability theory from these chapters. For readers who have already been exposed to mathematical probability theory, these chapters should serve as a review of the concepts that will be important for the rest of this book. Once we have laid this technical groundwork, the mathematics ease in Parts II and III.

Many of the ideas in Part I are essential for understanding the fundamentals of this book, and our treatment of them may be somewhat unfamiliar for readers whose prior training is in applied econometrics or data analysis. We therefore recommend that even relatively statistically sophisticated readers (and also readers otherwise uninterested in probability theory) read the contents of these chapters, as their presentation will inform our discussion of more advanced and applied topics in subsequent chapters. For readers with neither previous exposure to probability theory nor fluency in college-level mathematics, we strongly recommend consulting an undergraduate textbook on probability theory.[1]

[1] We recommend Chapter 1 of Wasserman (2004) for a concise treatment, though we are also fond of Part IV of Freedman, Pisani, and Purves (1998) as a very friendly introduction to the basics of probability theory. For a more thorough treatment of mathematical probability theory, we recommend Blitzstein and Hwang (2014) and Wackerly, Mendenhall, and Scheaffer (2008).

Finally, some common mathematical notation used in this book will not be defined in the main text. However, definitions of many of these notations are included in the Glossary of Mathematical Notation. Readers with little background in advanced mathematics or statistics may want to begin by reviewing and familiarizing themselves with the concepts and notation in this appendix. In addition, we provide a Glossary of Common Abbreviations, which gives the meanings of all common abbreviations used in this book, along with references to where they are defined in the text.

WHO HELPED TO WRITE THIS BOOK?

We thank the following readers and research assistants for their valuable contributions to this book: Ellen Alpert, Laura Balzer, Tommaso Bardelli, Jonathon Baron, Paul C. Bauer, Kassandra Birchler, Xiaoxuan Cai, Alex Coppock, Forrest Crawford, Naoki Egami, Germán Feierherd, Robin Gomila, Don Green, Anand Gupta, Josh Kalla, Sarah Hamerling, Erin Hartman, Jennifer Hill, Will Hunt, Jun Hwang, Donald Lee, Daniel Masterson, Mary McGrath, Adelaide McNamara, Joel Middleton, Avi Nuri, Lilla Orr, Betsy Levy Paluck, Raja Panjwani, Kyle Peyton, Thomas Richardson, Jamie Robins, Cyrus Samii, Fredrik Sävje, Collin Schumock, Matt Shafer, Vivien Shotwell, Pavita Singh, Brandon Stewart, Eric Tchetgen-Tchetgen, Dustin Tingley, Teppei Yamamoto, the participants of the *Theory of Agnostic Statistics: A Discussion* conference at Yale University, and the reviewers and editorial staff at *Cambridge University Press*. We thank Jens Hainmueller and Dominik Hangartner for generous data sharing. We owe a special debt of gratitude to Winston Lin, whose comments and insights guided us at every stage of the book—from conception to development to completion. In fact, our book's title is an homage to Lin (2013), itself referencing the "agnostic regression" of Angrist and Imbens (2002). We would also be remiss not to specially acknowledge the many contributions of Molly Offer-Westort, which included coding, technical edits, figures, and important guidance on framing. We thank our editor, Robert Dreesen, whose patience, encouragement, and wisdom helped us turn an idea into a manuscript and a manuscript into a book.

We also thank our classes, the Yale students in the first year graduate quantitative methods sequence for political science (PLSC 500 and PLSC 503), who were invaluable in shaping the contents of this book. In fact, this book originated from the lecture notes for these courses, in which one of us (Aronow) was the instructor and the other (Miller) was a student and later a teaching fellow. We began work on the book in earnest during the summer of 2014, and it quickly became a fully collaborative effort. We view our contributions to this book as equal and inseparable.