Cambridge University Press & Assessment
978-1-107-18514-2 — Probability Theory and Statistical Inference
Empirical Modeling with Observational Data 2nd Edition
Aris Spanos
Frontmatter
More Information

Probability Theory and Statistical Inference

Doubt over the trustworthiness of published empirical results is not unwarranted and is often a result of statistical misspecification: invalid probabilistic assumptions imposed on data. Now in its second edition, this bestselling textbook offers a comprehensive course in empirical research methods, teaching the probabilistic and statistical foundations that enable the specification and validation of statistical models, providing the basis for an informed implementation of statistical procedure to secure the trustworthiness of evidence. Each chapter has been thoroughly updated, accounting for developments in the field and the author's own research. The comprehensive scope of the textbook has been expanded by the addition of a new chapter on the Linear Regression and related statistical models. This new edition is now more accessible to students of disciplines beyond economics and includes more pedagogical features, with an increased number of examples as well as review questions and exercises at the end of each chapter.

ARIS SPANOS is Wilson E. Schmidt Professor of Economics at Virginia Polytechnic Institute and State University. He is the author of *Statistical Foundations of Econometric Modelling* (Cambridge, 1986) and, with D. G. Mayo, *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science* (Cambridge, 2010).

Probability Theory and Statistical Inference

Empirical Modeling with Observational Data

Second Edition

Aris Spanos
Virginia Tech
(Virginia Polytechnic Institute & State University)

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/9781107185142

DOI: 10.1017/9781316882825

First published 1999
Third printing 2007
Second edition 2019

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

Library of Congress Cataloging-in-Publication data
Names: Spanos, Aris, 1952– author.
Title: Probability theory and statistical inference : empirical modelling with observational data / Aris Spanos (Virginia College of Technology).
Description: Cambridge ; New York, NY : Cambridge University Press, 2019. | Includes bibliographical references and index.
Identifiers: LCCN 2019008498 (print) | LCCN 2019016182 (ebook) | ISBN 9781107185142 | ISBN
Subjects: LCSH: Probabilities – Textbooks. | Mathematical statistics – Textbooks.
Classification: LCC QA273 (ebook) | LCC QA273 .S6875 2019 (print) | DDC 519.5–dc23
LC record available at https://lccn.loc.gov/2019008498

ISBN 978-1-107-18514-2 Hardback
ISBN 978-1-316-63637-4 Paperback

To my grandchildren Nicholas, Jason, and Evie,
my daughters Stella, Marina, and Alexia, and my
wife Evie for their unconditional love and support

Contents

Preface to the Second Edition

The original book, published 20 years ago, has been thoroughly revised with two objectives in mind. First, to make the discussion more compact and coherent by avoiding repetition and many digressions. Second, to improve the methodological coherence of the proposed empirical modeling framework by including material pertaining to foundational issues that has been published by the author over the last 20 years or so in journals on econometrics, statistics, and philosophy of science. In particular, this revised edition brings out more clearly several crucial distinctions that elucidate empirical modeling, including (a) the statistical vs. the substantive information/model, (b) the modeling vs. the inference facet of statistical analysis, (c) testing within and testing outside the boundary of a statistical model, and (d) pre-data vs. post-data error probabilities. These distinctions shed light on several foundational issues and suggest solutions. In addition, the comprehensiveness of the book has been improved by adding Chapter 14 on the linear regression and related models.

The current debates on the "*replication crises*" render the methodological framework articulated in this book especially relevant for today's practitioner. A closer look at the debates (Mayo, 2018) reveals that the non-replicability of empirical evidence problem is, first and foremost, a problem of *untrustworthy evidence* routinely published in prestigious journals. The current focus of that literature on the abuse of significance testing is rather misplaced, because it is only a part of a much broader problem relating to the mechanical application of statistical methods without a real understanding of their assumptions, limitations, proper implementation, and interpretation of their results. The abuse and misinterpretation of the p-value is just symptomatic of the same uninformed implementation that contributes majorly to the problem of untrustworthy evidence. Indeed, the same uninformed implementation often ensures that untrustworthy evidence is routinely replicated, when the same mistakes are repeated by equally uninformed practitioners! In contrast to the current conventional wisdom, it is argued that a major contributor to the untrustworthy evidence problem is *statistical misspecification*: invalid probabilistic assumptions imposed on one's data, another symptom of the same uninformed implementation. The primary objective of this book is to provide the necessary probabilistic foundation and the overarching modeling framework for an informed and thoughtful application of statistical methods, as well as the proper interpretation of their inferential results. The emphasis is placed less on the mechanics of the application of statistical methods, and more on understanding their assumptions, limitations, and proper interpretation.

Key Features of the Book

- It offers a seamless integration of probability theory and statistical inference with a view to elucidating the interplay between deduction and induction in "learning from data" about observable phenomena of interest using statistical procedures.
- It develops frequentist modeling and inference from first principles by emphasizing the notion of a statistical model and its adequacy (the validity of its probabilistic assumptions vis-à-vis the particular data) as the cornerstone for reliable inductive inference and trustworthy evidence.
- It presents frequentist inference as well-grounded procedures whose optimality is assessed by their capacity to achieve genuine "learning from data."
- It focuses primarily on the skills and the technical knowledge one needs to be able to begin with substantive questions of interest, select the relevant data carefully, and proceed to establish trustworthy evidence for or against hypotheses or claims relating to the questions of interest. These skills include understanding the statistical information conveyed by data plots, selecting appropriate statistical models, as well as validating them using misspecification testing before any inferences are drawn.
- It articulates reasoned responses to several charges leveled against several aspects of frequentist inference by addressing the underlying foundational issues, including the use and abuse of p-values and confidence intervals, Neyman–Pearson vs. Fisher testing, and inference results vs. evidence that have bedeviled frequentist inference since the 1930s. The book discusses several such foundational issues/problems and proposes ways to address them using an error statistical perspective grounded in the concept of severity. Methodological issues discussed in this book include rebuttals to widely used, ill-thought-out arguments for ignoring statistical misspecification, as well as principled responses to certain Bayesian criticisms of the frequentist aproach.
- Its methodological perspective differs from the traditional textbook perspective by bringing out the perils of curve-fitting and focusing on the key question: How can empirical modeling lead to "learning from data" about phenomena of interest by giving rise to trustworthy evidence?

NOTE: All sections marked with an asterisk ($*$) can be skipped at first reading without any serious interruption in the flow of the discussion.

Acknowledgments

More than any other person, Deborah G. Mayo, my colleague and collaborator on many foundational issues in statistical inference, has helped to shape my views on several methodological issues addressed in this book; for that and the constant encouragement, I'm most grateful to her. I'm also thankful to Clark Glymour, the other philosopher of science with whom I had numerous elucidating and creative discussions on many philosophical issues discoursed in the book. Thanks are also due to Sir David Cox for many discussions that helped me appreciate the different perspectives on frequentist inference. Special thanks are also due to my longtime collaborator, Anya McGuirk, who contributed majorly in puzzling out several thorny issues discussed in this book. I owe a special thanks to Julio Lopez for his insightful comments, as well as his unwavering faith in the coherence and value of the proposed approach to empirical modeling. I'm also thankful to Jesse Bledsoe for helpful comments on chapter 13 and Mariusz Kamienski for invaluable help on the front cover design.

I owe special thanks to several of my former and current students over the last 20 years, who helped to improve the discussion in this book by commenting on earlier drafts and finding mistakes and typos. They include Elena Andreou, Andros Kourtellos, Carlos Elias, Maria Heracleous, Jason Bergtold, Ebere Akobundu, Andreas Koutris, Alfredo Romero, Niraj Pouydal, Michael Michaelides, Karo Solat, and Mohammad Banasaz.

Symbols

$\mathbb{N}$	–	set of natural numbers $\mathbb{N}:=\{1, 2, ..., n, ...\}$
$\mathbb{R}$	–	the set of real numbers; the real line $(-\infty, \infty)$
$\mathbb{R}^n$		$:=\overbrace{\mathbb{R} \times \mathbb{R}\times \cdots \times \mathbb{R}}^{n \text{ times}}$
$\mathbb{R}_+$	–	the set of positive real numbers; the half real line $(0, \infty)$
$f(x; \boldsymbol{\theta})$	–	density function of X with parameters $\boldsymbol{\theta}$
$F(x; \boldsymbol{\theta})$	–	cumulative distribution function of X with parameters $\boldsymbol{\theta}$
$\mathsf{N}(\mu, \sigma^2)$	–	Normal distribution with mean μ and variance σ^2
$\mathcal{E}$	–	Random Experiment (RE)
S	–	outcomes set (sample space)
$\Im$	–	event space (a $\sigma-$field)
$\mathbb{P}(.)$	–	probability set function
$\sigma(X)$	–	minimal sigma-field generated by X

Acronyms

AR(p)	–	Autoregressive model with p lags
CAN	–	Consistent, Asymptotically Normal
cdf	–	cumulative distribution function
CLT	–	Central Limit Theorem
ecdf	–	empirical cumulative distributrion function
GM	–	Generating Mechanism
IID	–	Indepedent and Identically Distributed
LS	–	Least-Squares
ML	–	Maximum Likelihood
M-S	–	Mis-Specification
N-P	–	Neyman-Pearson
PMM	–	Parametric Method of Moments
SLLN	–	Strong Law Large Numbers
WLLN	–	Weak Law Large Numbers
UMP	–	Uniformly Most Powerful