Statistical Hypothesis Testing in Context

Fay and Brittain present statistical hypothesis testing and compatible confidence intervals, focusing on application and proper interpretation. The emphasis is on equipping applied statisticians with enough tools – and advice on choosing among them – to find reasonable methods for almost any problem and enough theory to tackle new problems by modifying existing methods. After covering the basic mathematical theory and scientific principles, tests and confidence intervals are developed for specific types of data. Essential methods for applications are covered, such as general procedures for creating tests (e.g., likelihood ratio, bootstrap, permutation, testing from models), adjustments for multiple testing, clustering, stratification, causality, censoring, missing data, group sequential tests, and noninferiority tests. New methods developed by the authors are included throughout, such as melded confidence intervals for comparing two samples and confidence intervals associated with Wilcoxon–Mann–Whitney tests and Kaplan–Meier estimates. Examples, exercises, and the R package asht support practical use.

MICHAEL P. FAY is a Mathematical Statistician at the National Institute of Allergy and Infectious Diseases, and previously worked at the National Cancer Institute. He has served as Associate Editor for *Biometrics* and is currently an Associate Editor for *Clinical Trials* and a Fellow of the American Statistical Association. He is a coauthor on over 100 papers in statistical and medical journals and has written and maintains over a dozen R packages on CRAN.

ERICA H. BRITTAIN is Deputy Branch Chief of Biostatistics Research at the National Institute of Allergy and Infectious Diseases and has well over three decades of experience as a statistician, with previous positions at the FDA, the National Heart, Lung, and Blood Institute, and a statistical consulting company. Her applied work at the NIH and her methodological publications in statistical journals focus on innovation in clinical trial design. She frequently serves on advisory panels for the FDA and the NIH and has served as Statistical Consultant for *Nature* journals and Associate Editor for *Controlled Clinical Trials*.

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

A complete list of books in the series can be found at www.cambridge.org/statistics.
Recent titles include the following:

25. *Design of Comparative Experiments*, by R. A. Bailey
26. *Symmetry Studies*, by Marlos A. G. Viana
27. *Model Selection and Model Averaging*, by Gerda Claeskens and Nils Lid Hjort
28. *Bayesian Nonparametrics*, edited by Nils Lid Hjort *et al*.
29. *From Finite Sample to Asymptotic Methods in Statistics*, by Pranab K. Sen, Julio M. Singer and Antonio C. Pedrosa de Lima
30. *Brownian Motion*, by Peter Mrters and Yuval Peres
31. *Probability: Theory and Examples* (*Fourth Edition*), by Rick Durrett
33. *Stochastic Processes*, by Richard F. Bass
34. *Regression for Categorical Data*, by Gerhard Tutz
35. *Exercises in Probability* (*Second Edition*), by Loïc Chaumont and Marc Yor
36. *Statistical Principles for the Design of Experiments*, by R. Mead, S. G. Gilmour and A. Mead
37. *Quantum Stochastics*, by Mou-Hsiung Chang
38. *Nonparametric Estimation under Shape Constraints*, by Piet Groeneboom and Geurt Jongbloed
39. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, by Jianfeng Yao, Shurong Zheng and Zhidong Bai
40. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, by Evarist Giné and Richard Nickl
41. *Confidence, Likelihood, Probability*, by Tore Schweder and Nils Lid Hjort
42. *Probability on Trees and Networks*, by Russell Lyons and Yuval Peres
43. *Random Graphs and Complex Networks* (*Volume 1*), by Remco van der Hofstad
44. *Fundamentals of Nonparametric Bayesian Inference*, by Subhashis Ghosal and Aad van der Vaart
45. *Long-Range Dependence and Self-Similarity*, by Vladas Pipiras and Murad S. Taqqu
46. *Predictive Statistics*, by Bertrand S. Clarke and Jennifer L. Clarke
47. *High-Dimensional Probability*, by Roman Vershynin
48. *High-Dimensional Statistics*, by Martin J. Wainwright
49. *Probability: Theory and Examples* (*Fifth Edition*), by Rick Durrett
50. *Statistical Model-based Clustering and Classification*, by Charles Bouveyron *et al*.
51. *Spectral Analysis for Univariate Time Series*, by Donald B. Percival and Andrew T. Walden
52. *Statistical Hypothesis Testing in Context*, by Michael P. Fay and Erica Brittain

Statistical Hypothesis Testing in Context

Reproducibility, Inference, and Science

Michael P. Fay
National Institute of Allergy and Infectious Diseases

Erica H. Brittain
National Institute of Allergy and Infectious Diseases

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/9781108423564
DOI: 10.1017/9781108528825

First published 2022

Printed in the United Kingdom by TJ Books Ltd, Padstow Cornwall

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

Library of Congress Cataloging-in-Publication Data
Names: P. Fay, Michael, author. | Brittain, Erica H., author.
Title: Statistical hypothesis testing in context : reproducibility, inference, and science /
Michael P. Fay and Erica H. Brittain.
Description: Cambridge, United Kingdom ; New York, NY, USA :
Cambridge University Press, 2022. | Series: Cambridge series in statistical and
probabilistic mathematics | Includes bibliographical references and index.
Identifiers: LCCN 2021044789 | ISBN 9781108423564 (hardback) |
ISBN 9781108437677 (paperback) | ISBN 9781108528825 (epub)
Subjects: LCSH: Statistical hypothesis testing. | BISAC: MATHEMATICS /
Probability & Statistics / General
Classification: LCC QA277 .P43 2022 | DDC 519.5/6–dc23/eng/20211117
LC record available at https://lccn.loc.gov/2021044789

ISBN 978-1-108-42356-4 Hardback

Contents

Preface

This book is about statistical hypothesis testing and related statistics such as p-values and confidence intervals (CIs), which together are known as the *frequentist* (or *classical*) school of statistics. The focus is both scientific and practical, so there is substantial discussion of study design and causality with a broad scope to attack many types of applications. The frequentist school of statistics is founded on bounding the probability of errors from false claims and its theory is based on imagining replicating studies. Currently, there is a crisis in reproducibility in that scientists and others are becoming more aware that many claims made from scientific studies do not turn out to be reproducible.

A problem is that statistical hypothesis tests and p-values, the basic tools of frequentist statistics, are often misunderstood and misused, and this misuse has contributed to the reproducibility crisis. The American Statistical Association has joined the conversation in a strong way by publishing a statement on p-values with accompanying commentaries (Wasserstein and Lazar, 2016), as well as publishing a special issue of *The American Statistician* on "Moving to a World Beyond '$p < 0.05$'" (Wasserstein et al., 2019). While the statistical community can agree that p-values are often misunderstood and can unify in correcting some common mistakes, there is less unity on the proper response. Many voices in those commentaries and the special issue propose that the answer to this reproducibility crisis and the misuse of p-values is to use different methods of statistics to make scientific inferences. Others state that the issue is not inherent in the p-value statistic itself, and most of its misuse problems would accompany many if not all of the alternative statistics proposed (see, e.g., Benjamini, 2016). We are in the latter camp. We believe if the problem is reproducibility[1] then the answer is not to replace p-values, but to better educate users about their correct use and limitations. After all, the frequentist school is founded on the idea of keeping the rate of false claims low, and hence will, *if properly applied*, lead to better reproducibility in science. This book is largely about the proper use of p-values and confidence intervals in scientific research.

The level of mathematical difficulty in the book is mixed. There are some very heavy mathematics that are appropriate for a graduate-level course in statistics (so we assume some familiarity with basic probability and statistics at the level of, for example, Casella and Berger (2002)), yet there are many practical and simple examples that are used to convey the

[1] The term reproducibility has multiple meanings in the science literature (Plesser, 2018), our use of the term is nontechnical. In the book, we will use precise technical language about certain reproducibility properties. For example, we emphasize the validity of the confidence interval, where a valid 95% confidence interval procedure means that if we repeat the procedure multiple times, we expect that at least 95% of the time we will cover the true parameter.

ideas without using mathematical notation. Our professional experience is not from teaching statistics in a classroom, but from extensive experience in applying statistics in the medical environment. We have over 60 years of combined experience, primarily working for the US National Institutes of Health (NIH). Many of the applied problems we have encountered have motivated us to develop valid and appropriate statistical methods to address those problems, and thankfully the environment at the NIH encouraged that work and allowed us the time to write this book. We envision the use of this book as a reference book for applied statisticians or for graduate students of statistics.

The book is structured with the applied researcher in mind. The chapters are defined so that researchers handling a specific type of data may skip to the appropriate chapter, as long as they are already aware of the material in Chapters 2 and 3. Each chapter begins with an overview, so that readers may get a flavor of what is in the chapter and decide what sections will be helpful to them at that time. Specifically, before mathematically theoretical sections, we discuss why we thought it was useful to go into that level of abstraction. This gives readers insight into whether it makes more sense for them to delve into the slower reading of mathematical notation, or perhaps skip that section. Finally, each chapter ends with a summary section that outlines the methods introduced therein and provides context and recommendations for use, and this is written at a relatively nontechnical level. Because the applied statistician must be able to handle all kinds of data inference problems, the focus of the book is for a generalist who needs to know some about many areas of statistics. A solid theoretical foundation is useful for the applied statistician as well, because by knowing the theory well, new tests and confidence intervals may be derived using slight adjustments to meet new situational needs. There will be few mathematical theorems, and when provided we will only give the conditions of their applicability, and no proofs. For the applied researcher, many chapters will serve as introductions or reviews of statistical areas. For example, Chapter 10 has only a few pages on the bootstrap, Chapter 14 covers testing from models, and Chapter 15 covers causal inference. These chapters are there because the applied researcher needs to be aware of them, but it is expected that to learn these topics in a deep way, the reader will need to supplement them with further study with the references given.

In terms of the use of this book for a graduate course in statistical hypothesis testing, we contrast it to one of the best books on the topic, Lehmann and Romano (2005). When using statistical hypothesis tests in a scientific problem, there are both scientific issues and mathematical issues. Loosely speaking, the mathematical issues relate to determining the appropriate statistical test and its properties given the assumed probability models represented by the null and alternative hypotheses. The scientific issues relate to the appropriateness of those probability models for answering the scientific question, and the presentation of the results such that the scientific community will properly interpret those results. Lehmann and Romano (2005) is an excellent book that focuses on the mathematical issues of statistical hypothesis testing. Their book discusses important properties of tests once one has already defined the null and alternative hypotheses (for example, validity, consistency, uniformly most powerful test, asymptotic relative efficiency). Lehmann and Romano (2005) has extensive discussions on how one determines the optimal test given the hypotheses assumed, including different ways that optimality may be defined (e.g., uniformly most powerful unbiased [UMPU] tests). In this book, we focus on some of those mathematical issues (validity), but much less on others (UMPU tests). We will spend more time on expanding the scope of applicability. For example, our book covers censoring and causality which are not covered in Lehmann and Romano (2005). In other words, we discuss more scientific issues – how to design a study to answer a scientific question, and how to find a set of probability models (encompassing the null and alternative hypotheses) that can plausibly be applied to the study.

We have tried to include all of the essential ideas needed for applied frequentist statistical hypothesis testing. But, of course, the selection of what is essential is subjective, so we will try to be clear about our biases upfront. We are biased toward ensuring valid inferences, so some may think we overly emphasize exact methods and de-emphasize asymptotics; however, we have included asymptotics where we think it is helpful. We are biased toward experiments over observational studies, so many of the causal discussions will be heavily focused on experiments. Part of this is our professional experience that is shaped heavily by randomized clinical trials. Our bias is toward calculating two one-sided hypotheses instead of one two-sided hypothesis, because, for example, if we find two treatments differ, we usually want to know which one is better. Our bias in applications is toward biomedical ones, rather than other scientific, economic, or business applications. Finally, we hope the reader will forgive our bias in presenting methods that we have worked on ourselves, but we believe that many of those methods have wide applicability. Despite these biases, we believe a quick perusal of the Table of Contents would convince most statisticians that we have included most of the important ideas needed for the applied statistician wanting to make frequentist inferences, and we even have included Chapter 21 at the end that compares frequentist hypothesis testing to some Bayesian approaches.

Although every effort has been made to be accurate, please inform us if you find a mistake (mfay@niaid.nih.gov). We will keep a list of corrections at https://github.com/michaelpfay/StatHypTest.

We have both had the good fortune to work in the Biostatistics Research Branch of the National Institute of Allergy and Infectious Diseases since the early 2000s, where the environment is intellectually vibrant, as well as warm and cooperative. In addition to engaging collaboration with doctors and other scientists at the Institute, we have time and space to work on statistical methodology; allowing us to write this book. We particularly want to thank our Branch Chief Dean Follmann who supported this endeavor, and who has always been a great collaborator and source of wisdom. Michael Proschan also stands out; he has a deep knowledge of mathematical statistics, probability, and clinical trials and has been wonderful to work with. We are grateful for those two and all our Biostatistics Research Branch colleagues from whom we have learned so much and have shared many enjoyable lunches.

We thank several anonymous reviewers who provided comments on early drafts of the book. We thank Dianna Gillooly for encouragement throughout the several years it took to finish this book, and a great title.

Erica Brittain would also like to thank all her colleagues from throughout her career for making her a better statistician, but most of all, for enhancing her love of statistics. She feels especially blessed by the many amazing mentors she had during her education and in the early years of her career, particularly Tom Fleming, Ed Davis, and Janet Wittes. And, finally, she is beyond grateful to her family: her son Tim, her daughter Rebecca, and her husband John Hyde (who doubles as her favorite and wisest statistical consultant) for making everything worthwhile.

Michael Fay wants to thank Rocky Feuer, Barry Graubard, Janet Wittes, and Larry Freedman for great mentorship early in his career, and Joanna Shih for being a wonderful collaborator. He thanks Sally Hunsberger for her dual role as a practical advisor on all issues related to clinical trials, and as a life partner. On the latter, the list of thanks needed is too long; the short version is that his life is much fuller and more fun with her around. Finally, he wishes that his and Sally's children, Paul and Sara, will find their work life as interesting and fulfilling as he has through biostatistics.