Bruce Blaine

Introductory Applied Statistics

With Resampling Methods & R

Springer

Introductory Applied Statistics

Bruce Blaine

Introductory Applied Statistics

With Resampling Methods & R

Springer

Bruce Blaine
Statistics Program
University of Rochester
Rochester, NY, USA

ISBN 978-3-031-27740-5 ISBN 978-3-031-27741-2 (eBook)
https://doi.org/10.1007/978-3-031-27741-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Patti and Kate

Preface

This is a book, *I hope*, would gain the approval of John Tukey (1915–2000). I certainly have written it in ways that try to honor his legacy to statistics. John Tukey was an academic and consulting statistician of the highest stripe and a towering intellectual figure. Over his productive career, he developed a long list of innovative statistical tests and concepts, most of which are now indispensable tools for the practicing statistician. In 1962 Tukey published a paper, On the Future of Data Analysis in the *Annals of Mathematical Statistics*, which I can only imagine shocked the sensibilities of the academic establishment. In that paper, he essentially claimed that data analysis, rather than statistics, was his field of interest and practice. His paper took academic statistics to task for being too reliant on mathematical epistemologies and models, and too isolated from the sciences—including the particular data, problems, and quantitative methods of the sciences—to be of much practical use. Tukey also urged his field to develop robust statistical methods, to be as interested in exploring data as in using data to confirm hypotheses, and to harness the emerging power of computing for data analysis.

This book introduces students to the field that Tukey envisioned and inspired—a field that applies statistics through data analysis to disciplinary questions, is in dialogue with data for both confirmatory and exploratory purposes, uses statistical methods that are informed by the data and not enslaved to assumptions, and finally, integrates statistical computing skills into the learning and using of statistics.

The key features of this book include:

- *Statistics and statistical inference taught in a data analytic context*. Statistics and quantitative methods are tools for interacting with data. We have specific questions to test in data, but data also has insights to suggest to us, if we are open to them, and we need statistical methods that are suitable for both purposes. It also means that students learn some basic data wrangling tasks in the process.
- *Robust descriptive statistics*. The book strives to give students a broad portfolio of descriptive statistics for summarizing and exploring univariate and bivariate data, including robust descriptive statistics and effect size statistics.

- *Statistics taught within a study design framework*. This book doesn't organize chapters by statistical method (e.g., a t-test chapter), as is common in introductory statistics books. Rather, descriptive and inferential statistics are organized around the type of study design that produced the data. The framework includes four common bivariate models:
  - ANOVA model (categorical X/numeric Y)
  - Proportions model (categorical X/categorical Y)
  - Regression model (numeric X/numeric Y)
  - Logistic model (numeric X/categorical Y)
- To properly and fully understand statistical inference (and not merely its null hypothesis testing component), students of statistics must learn how study design elements like random sampling, manipulated versus observed predictor variables, and control of alternative explanations affect the interpretation of and conclusions about statistical procedures.
- *Statistical inference taught with resampling methods*. You can draw a straight line between Tukey's vision for statistics and modern simulation-based statistical inference. In this book, students learn and use randomization (aka permutation) and bootstrapping methods for hypothesis testing and parameter estimation. Resampling methods are intuitive and, once understood, apply logically and consistently to inferential problems across study design type. Resampling methods also free us from the limitations of theoretical probability distributions, particularly with regard to parametric assumptions. Finally, the computing skills needed to use randomization and bootstrapping are within the reach of the introductory statistics or data science student.
- *Statistical computing with R*. This book incorporates data analytic examples in R into each chapter, and assumes no prior experience with R in students. Files containing R functions and code are provided for each chapter.

Expectations for Student and Instructor

This book is written for use in undergraduate introductory statistics or applied statistics courses. Indeed, the book emerges from lessons and material I teach in my own Applied Statistics with R course. What do I mean by "applied?" It's an approach to learning statistics that is data analytic, emphasizing how statistics are used to address research questions with various kinds of data and how to do that data analysis. It assumes no mathematical background beyond algebra.

The book assumes that students will use R and RStudio on a regular basis, and have access to those programs outside of class for homework and assignments. No prior computing or programming experience is assumed in students, but developing students' statistical computing skills in R is assumed to be a learning goal of the course, so course time should be allotted for that. RStudio Cloud (aka Posit Cloud) is an excellent environment to work in: it is widely accessible outside of class and doesn't require local installs and updates. For instructors wanting students to use R Markdown for assignments, RStudio Cloud also meshes well with R Markdown.

For instructors, some prior experience with the R language and environment is necessary, but your R skills need not be deep or extensive, especially if you have colleagues to whom you can turn with questions and support. If you are new to R or are just learning, teaching with R is the best way to learn the language. Students need support too, and that can be provided in a variety of ways, both in and outside of class.

Chapter Teaching and Learning Goals

Chapter 1 serves two purposes: 1. Get students at the statistics "workbench" immediately, so they can interact with data and R from the first sessions in the course, and 2. Introduce foundational statistics and statistical concepts and vocabulary that can be built upon in subsequent chapters. This is done largely within the context of summarizing a numeric variable, although late in the chapter I do address how we summarize a categorical variable.

Chapter 2 introduces the statistical modeling framework and study design elements that inform data analysis and interpretations throughout the chapters. I introduce the general linear model, how different combinations of X and Y result in different models, how a statistical model guides analysis, what statistical lenses we can/should use to examine a particular question, how we explore data and allow data to teach us, and how sampling and study design elements figure into data analysis and interpretation.

Chapters 3, 4, 5, and 6 starts with the ANOVA model (Chap. 3) because it is a common and intuitive model for beginning students, and the mean difference is a friendly starting point for learning about effect size statistics. Also, some concepts are foundational and re-purposed in subsequent models (e.g., the mean difference -> risk difference). In Chapter 3, I want to teach students that a statistic is a lens through which one can summarize the effect of a treatment variable, with lots of interesting aspects of a group difference to explore.

Chapter 4 is organized much like Chap. 3 and introduces the proportions model, as well as foundational concepts that show up again in the logistic model: probabilities, proportions, and odds—and the statistics built with them.

Chapter 5 moves from group differences (Chaps. 3 and 4) as the quantity of interest—and the statistics with which we summarize those differences—to linear relationships (Chaps. 5 and 6). Here I teach the basics of least squares simple linear regression but maintain the connection of those particular methods to the analytic framework (relationship summary and exploration, robust/nonrobust statistics, study design elements, and interpretation).

Chapter 6 extends some of the statistical concepts from Chapt. 4, and also builds on least squares regression ideas to teach logistic simple regression. The structure and pedagogy of the chapter reflects the previous 3.

Chapters 3, 4, 5, and 6 form a coherent unit that might be called “statistics and data analysis in bivariate models.” At this point, it would be reasonable to do some formal assessment of knowledge, via testing and/or a larger assignment or mini-project that requires students to use these methods and write up a data analytic report.

Chapter 7 introduces statistical inference and focuses on null hypothesis testing methods. It is important for students to learn statistical tools and data analytic skills for descriptive (both confirmatory/summary and exploratory) analysis, without the clutter and complications of establishing “statistical significance.” The idea is to help students realize that much of data analysis is descriptive, and we learn a lot about an X-Y relationship through descriptive tools. In this chapter, students learn the logic and principles of hypothesis testing, randomization testing as a resampling method, and how to interpret a p-value.

Chapter 8 focuses on parameter estimation and closes the loop on the effect size statistics covered in Chapters 3, 4, 5, and 6, in that statistics are estimates of some “true” relationship size (e.g., mean difference in the population) and our ultimate interest is, in fact, that parameter. The logic and methods of bootstrapping as a second resampling method are taught, along with two bootstrapped confidence intervals (t interval w/bootstrapped standard error, percentile interval). This chapter also addresses factors that affect interval estimation.

Chapters 7 and 8 introduce resampling methods for inference, but the examples are simple. Chapter 9 applies those methods to each of the models covered in Chaps. 3, 4, 5, and 6, with full data analytic examples. Students therefore learn covering how randomization and bootstrapping are applied to the particularities of each statistical model.

Chapters 7, 8 and 9 form a unit that might be called “statistical inference in bivariate models.” At this point, it would be reasonable to again do some formal assessment of knowledge, via testing and/or a larger assignment. Additionally, students are now prepared to do a data analytic project, the parameters of which can be freed or constrained to suit the instructor and circumstances.

Chapter 10 covers descriptive and inferential analysis in repeated-measures data, and focuses on the simplest form of a repeated-measures design—the pre-post study. A pre-post design is very commonly used by researchers, and often is the type of study a student will come up with when given the opportunity to design a study of their own to test some hypothesis. Pre-post designs and data, however, have statistical and inferential issues that set them apart from the designs covered in Chaps. 3, 4, 5, and 6 and 9, and thus this material is somewhat set aside in Chap. 10 and can be thought of as additional material.

Rochester, NY, USA Bruce Blaine

Contents