Preface to the Second Edition

In overall structure, the second edition has changed very little from the first. It still aims to provide an introduction to the foundations and applications of statistics to upper-level undergraduate students. The general approach remains "probability for statistics" rather than "probability, then statistics", so statistical ideas appear early, even as the probability necessary to support them is being introduced. The need for users of statistics to be fluent with computational tools has only increased since the first edition appeared, and R continues to be used throughout the text as a computational tool both for performing analyses in applications and for exploring the foundational material.

Nevertheless, the second edition includes some changes that are worth noting.

(1) The inclusion of R code has been improved through the use of `knitr`, which did not exist at the time I began the first edition but has now replaced my inferior, homespun solution. I encourage my students to take advantage of `knitr` as well, typically through R Markdown within RStudio.

(2) The R code has been changed throughout to reflect my increased sensitivity to coding style and the availability of new R packages, including `mosaic`, which has subsumed a large fraction of what was in `fastR`; the packages of Hadley Wickham's `tidyverse`, which provide a suite of tools for working with data; and `ggformula`, a new package that provides a formula interface to `ggplot2` and is used for nearly all of the plots in the book. The introduction to R in Chapter 1 and in Appendix A has been substantially rewritten as a result.

    The accompanying package for the second edition is, naturally enough, called `fastR2`; `fastR` will continue to exist, but users of the second edition should migrate to the new package.

(3) Chapters 2 and 3 have changed the least. Additions include a section on multinomial distributions, an introduction to multivariate normal distributions, and a very brief introduction to exponential families.

(4) Chapter 4 now includes an introduction to the bootstrap and a new section on the delta method and its application for propagation of uncertainty.

(5) Where it comes up, the duality between confidence intervals and p-values is explained better and pivotal quantities are mentioned.

(6) The second edition includes more matrix algebra. This material can be omitted by those who prefer the geometric, "one vector at a time" approach, but for those who want more matrix algebra, it is there now. Appendix C has been somewhat expanded to support the matrix algebra. In Chapter 6, it is not necessary to read all of the sections on how to fit a simple linear model (least squares, likelihood, projection, and matrix algebra), but each adds something to the overall picture.

(7) Chapter 5 makes use of the `maxLik` package rather than `nlm()` to find maximum likelihood estimates. The `fastR2` package includes a `maxLik2()` function that stores a bit more information and makes it possible to create plots of the log-likelihood function. Chapter 5 also includes an explanation of Zermelo's algorithm for fitting what has come to be called the Bradley-Terry model. As in the first edition, this topic can easily be omitted, but I hope it gives readers a sense for the flexibility of the maximum likelihood method.

(8) The introduction to Bayesian inference in Chapter 5 now includes an example illustrating the grid method for approximating the posterior distribution. There isn't time here to cover Markov chain Monte Carlo (MCMC) methods, but the grid method provides a taste of the computational aspects of modern Bayesian analyses.

(9) Chapter 7 introduces added variable plots and effect plots and the functions in the `effects` package that make them easy to create and concludes with an introduction to non-linear models.

(10) Appendix D has been removed from the printed text. An electronic version is available on the companion website listed below.

Along the way I have attempted to tidy things up a bit, editing explanations and notation here and there in an attempt to improve the exposition. I apologize in advance for any new errors introduced while trying to eradicate old ones.

Companion Website

Additional material related to this book is available online at

`http://www.ams.org/bookpages/amstext-28` .

Any bugs in the `fastR2` package or errors in the text can be reported at

`https://github.com/rpruim/fastR2/issues` .

Acknowledgments

It remains a pleasure to work with my colleagues at the American Mathematical Society. Sergei Gelfand, who replaced Edward Dunne as the primary editor for the second edition, and Christine Thivierge helped shepherd the second edition through

the production process. The production editor was once again Arlene O'Sean. Her keen eye has helped catch countless little things that together make a big difference. And special thanks go to David Jones, without whose technical support I would not have been able to bring my `knitr` and RStudio workflow through the AMS production process.

I have also been fortunate again to have colleagues who have volunteered to point out ways that the book could be improved. I especially want to thank the following colleagues who read one or more chapters of the second edition and provided valuable feedback:

Douglas Armstead
SUNY Cortland

Alan Arnholt
Appalachian State University

Bernadine Cochran
University of Pikeville

David Gurney
Southeastern Louisiana University

Johanna Hardin
Pomona College

Marcos Minoru Hasegawa
Universidade Federal do Paraná, Brazil

Stacy Hoehn
Franklin College

Nicholas Horton
Amherst College

Kimberly Muller
Lake Superior State University

Julia Seaman
Babson Survey Research Group

Bradley Warner
US Air Force Academy

Jeffrey Witmer
Oberlin College

Andrew Zieffler
University of Minnesota

Nicholas Zoller
Southern Nazarene University

Preface to the First Edition

Intended Audience

As the title suggests, this book is intended as an introduction to both the foundations and applications of statistics. It is an introduction in the sense that it does not assume a prior statistics course. But it is not introductory in the sense of being suitable for students who have had nothing more than the usual high school mathematics preparation. The target audience is undergraduate students at the equivalent of the junior or senior year at a college or university in the United States.

Students should have had courses in differential and integral calculus, but not much more is required in terms of mathematical background. In fact, most of my students have had at least another course or two by the time they take this course, but the only courses that they have all had is the calculus sequence. The majority of my students are not mathematics majors. I have had students from biology, chemistry, computer science, economics, engineering, and psychology, and I have tried to write a book that is interesting, understandable, and useful to students with a wide range of backgrounds and career goals.

This book is suitable for what is often a two-semester sequence in "mathematical statistics", but it is different in some important ways from many of the books written for such a course. I was trained as a mathematician first, and the book is clearly mathematical at some points, but the emphasis is on the statistics. Mathematics and computation are brought in where they are useful tools. The result is a book that stretches my students in different directions at different times – sometimes statistically, sometimes mathematically, sometimes computationally.

The Approach Used in This Book

Features of this book that help distinguish it from other books available for such a course include the following:

- The use of R, a free software environment for statistical computing and graphics, throughout the text.

  Many books claim to integrate technology, but often technology appears to be more of an afterthought. In this book, topics are selected, ordered, and discussed in light of the current practice in statistics, where computers are an indispensable tool, not an occasional add-on.

  R was chosen because it is both powerful and available. Its "market share" is increasing rapidly, so experience with R is likely to serve students well in their future careers in industry or academics. A large collection of add-on packages is available, and new statistical methods are often available in R before they are available anywhere else.

  R is open source and is available at the Comprehensive R Archive Network (CRAN, `http://cran.r-project.org`) for a wide variety of computing platforms at no cost. This allows students to obtain the software for their personal computers – an essential ingredient if computation is to be used throughout the course.

  The R code in this book was executed on a 2.66 GHz Intel Core 2 Duo MacBook Pro running OS X (version 10.5.8) and the current version of R (version 2.12). Results using a different computing platform or different version of R should be similar.

- An emphasis on practical statistical reasoning.

  The idea of a statistical study is introduced early on using Fisher's famous example of the lady tasting tea. Numerical and graphical summaries of data are introduced early to give students experience with R and to allow them to begin formulating statistical questions about data sets even before formal inference is available to help answer those questions.

- Probability *for* statistics.

  One model for the undergraduate mathematical statistics sequence presents a semester of probability followed by a semester of statistics. In this book, I take a different approach and get to statistics early, developing the necessary probability as we go along, motivated by questions that are primarily statistical. Hypothesis testing is introduced almost immediately, and p-value computation becomes a motivation for several probability distributions. The binomial test and Fisher's exact test are introduced formally early on, for example. Where possible, distributions are presented as statistical models first, and their properties (including the probability mass function or probability density function) derived, rather than the other way around. Joint distributions are motivated by the desire to learn about the sampling distribution of a sample mean.

  Confidence intervals and inference for means based on t-distributions must wait until a bit more machinery has been developed, but my intention is that a student who only takes the first semester of a two-semester sequence will have a solid understanding of inference for one variable – either quantitative or categorical.

- The linear algebra middle road.

  Linear models (regression and ANOVA) are treated using a geometric, vector-based approach. A more common approach at this level is to introduce these topics without referring to the underlying linear algebra. Such an approach avoids the problem of students with minimal background in linear algebra but leads to mysterious and unmotivated identities and notions.

  Here I rely on a small amount of linear algebra that can be quickly reviewed or learned and is based on geometric intuition and motivation (see Appendix C). This works well in conjunction with R since R is in many ways vector-based and facilitates vector and matrix operations. On the other hand, I avoid using an approach that is too abstract or requires too much background for the typical student in my course.

Brief Outline

The first four chapters of this book introduce important ideas in statistics (distributions, variability, hypothesis testing, confidence intervals) while developing a mathematical and computational toolkit. I cover this material in a one-semester course. Also, since some of my students only take the first semester, I wanted to be sure that they leave with a sense for statistical practice and have some useful statistical skills even if they do not continue. Interestingly, as a result of designing my course so that stopping halfway makes some sense, I am finding that more of my students are continuing on to the second semester. My sample size is still small, but I hope that the trend continues and would like to think it is due in part because the students are enjoying the course and can see "where it is going".

The last three chapters deal primarily with two important methods for handling more complex statistical models: maximum likelihood and linear models (including regression, ANOVA, and an introduction to generalized linear models). This is not a comprehensive treatment of these topics, of course, but I hope it both provides flexible, usable statistical skills and prepares students for further learning.

Chi-squared tests for goodness of fit and for two-way tables using both the Pearson and likelihood ratio test statistics are covered after first generating empirical p-values based on simulations. The use of simulations here reinforces the notion of a sampling distribution and allows for a discussion about what makes a good test statistic when multiple test statistics are available. I have also included a brief introduction to Bayesian inference, some examples that use simulations to investigate robustness, a few examples of permutation tests, and a discussion of Bradley-Terry models. The latter topic is one that I cover between Selection Sunday and the beginning of the NCAA Division I Basketball Tournament each year. An application of the method to the 2009–2010 season is included.

Various R functions and methods are described as we go along, and Appendix A provides an introduction to R focusing on the way R is used in the rest of the book. I recommend working through Appendix A simultaneously with the first chapter – especially if you are unfamiliar with programming or with R.

Some of my students enter the course unfamiliar with the notation for things like sets, functions, and summation, so Appendix B contains a brief tour of the basic

mathematical results and notation that are needed. The linear algebra required for parts of Chapter 4 and again in Chapters 6 and 7 is covered in Appendix C. These can be covered as needed or used as a quick reference. Appendix D is a review of the first four chapters in outline form. It is intended to prepare students for the remainder of the book after a semester break, but it could also be used as an end of term review.

Access to R Code and Data Sets

All of the data sets and code fragments used in this book are available for use in R on your own computer. Data sets and other utilities that are not provided by R packages in CRAN are available in the `fastR` package. This package can be obtained from CRAN, from the companion website for this book, or from the author's website.

Among the utility functions in `fastR` is the function `snippet()`, which provides easy access to the code fragments that appear in this book. The names of the code fragments in this book appear in boxes at the right margin where code output is displayed. Once `fastR` has been installed and loaded,

```
snippet("snippet")
```
snippet

will both display and execute the code named "snippet", and

```
snippet("snippet", exec = FALSE)
```
snippet2

will display but not execute the code.

`fastR` also includes a number of additional utility functions. Several of these begin with the letter `x`. Examples include `xplot()`, `xhistogram()`, `xpnorm()`, etc. These functions add extra features to the standard functions they are based on. In most cases they are identical to their `x`-less counterparts unless new arguments are used.

Companion Website

Additional material related to this book is available online at

`http://www.ams.org/bookpages/amstext-13`.

Included there are

- an errata list,
- additional instructions, with links, for installing R and the R packages used in this book,
- additional examples and problems,
- additional student solutions,
- additional material – including a complete list of solutions – available only to instructors.

Acknowledgments

Every author sets out to write the perfect book. I was no different. Fortunate authors find others who are willing to point out the ways they have fallen short of their goal and suggest improvements. I have been fortunate.

Most importantly, I want to thank the students who have taken advanced undergraduate statistics courses with me over the past several years. Your questions and comments have shaped the exposition of this book in innumerable ways. Your enthusiasm for detecting my errors and your suggestions for improvements have saved me countless embarrassments. I hope that your moments of confusion have added to the clarity of the exposition.

If you look, some of you will be able to see your influence in very specific ways here and there (happy hunting). But so that you all get the credit you deserve, I want to list you all (in random order, of course): Erin Campbell, John Luidens, Kyle DenHartigh, Jessica Haveman, Nancy Campos, Matthew DeVries, Karl Stough, Heidi Benson, Kendrick Wiersma, Dale Yi, Jennifer Colosky, Tony Ditta, James Hays, Joshua Kroon, Timothy Ferdinands, Hanna Benson, Landon Kavlie, Aaron Dull, Daniel Kmetz, Caleb King, Reuben Swinkels, Michelle Medema, Sean Kidd, Leah Hoogstra, Ted Worst, David Lyzenga, Eric Barton, Paul Rupke, Alexandra Cok, Tanya Byker Phair, Nathan Wybenga, Matthew Milan, Ashley Luse, Josh Vesthouse, Jonathan Jerdan, Jamie Vande Ree, Philip Boonstra, Joe Salowitz, Elijah Jentzen, Charlie Reitsma, Andrew Warren, Lucas Van Drunen, Che-Yuan Tang, David Kaemingk, Amy Ball, Ed Smilde, Drew Griffioen, Tim Harris, Charles Blum, Robert Flikkema, Dirk Olson, Dustin Veldkamp, Josh Keilman, Eric Sloterbeek, Bradley Greco, Matt Disselkoen, Kevin VanHarn, Justin Boldt, Anthony Boorsma, Nathan Dykhuis, Brandon Van Dyk, Steve Pastoor, Micheal Petlicke, Michael Molling, Justin Slocum, Jeremy Schut, Noel Hayden, Christian Swenson, Aaron Keen, Samuel Zigterman, Kobby Appiah-Berko, Jackson Tong, William Vanden Bos, Alissa Jones, Geoffry VanLeeuwen, Tim Slager, Daniel Stahl, Kristen Vriesema, Rebecca Sheler, and Andrew Meneely.

I also want to thank various colleagues who read or class-tested some or all of this book while it was in progress. They are

Ming-Wen An
Vassar College

Alan Arnholdt
Appalacian State University

Stacey Hancock
Clark University

Jo Hardin
Pomona College

Nicholas Horton
Smith College

Laura Kapitula
Calvin College

Daniel Kaplan
Macalester College

John Kern
Duquesne University

Kimberly Muller
Lake Superior State University

Ken Russell
University of Wollongong, Australia

Greg Snow
Intermountain Healthcare

Nathan Tintle
Hope College

Interesting data make for interesting statistics, so I want to thank colleagues and students who helped me locate data for use in the examples and exercises in this book, especially those of you who made original data available. In the latter cases, specific attributions are in the documentation for the data sets in the `fastR` package.

Thanks also go to those at the American Mathematical Society who were involved in the production of this book: Edward Dunne, the acquisitions editor with whom I developed the book from concept to manuscript; Arlene O'Sean, production editor; Cristin Zannella, editorial assistant; and Barbara Beeton, who provided TEXnical support. Without their assistance and support the final product would not have been as satisfying.

Alas, despite the efforts of so many, this book is still not perfect. No books are perfect, but some books are useful. My hope is that this book is both useful and enjoyable. A list of those (I hope few) errors that have escaped detection until after the printing of this book will be maintained at

`http://www.ams.org/bookpages/amstext-13`.

My thanks in advance to those who bring these to my attention.