Introduction to Probability and Statistics for Data Science

Introduction to Probability and Statistics for Data Science provides a solid course in the fundamental concepts, methods, and theory of statistics for students in statistics, data science, biostatistics, engineering, and physical science programs. It teaches students to understand, use, and build on modern statistical techniques for complex problems. The authors develop the methods from both an intuitive and mathematical angle, illustrating with simple examples how and why the methods work. More complicated examples, many of which incorporate data and code in R, show how the method is used in practice. Through this guidance, students get the big picture about how statistics works and can be applied. This text covers more modern topics such as regression trees, large-scale hypothesis testing, bootstrapping, MCMC, time series, and fewer theoretical topics such as the Cramer–Rao lower bound and the Rao–Blackwell theorem. It features more than 250 high-quality figures, 180 of which involve actual data. Data and R code are available on the book's website so that students can reproduce the examples and complete hands-on exercises.

Steven E. Rigdon is Professor of Biostatistics at Saint Louis University. He is a fellow of the American Statistical Association and is the author of *Statistical Methods for the Reliability of Repairable Systems*, *Calculus*, 8th and 9th editions, *Monitoring the Health of Populations by Tracking Disease Outbreaks* (2020), and *Design of Experiments for Reliability Achievement* (2022). He has received the Waldo Vizeau Award for technical contributions to quality, the Soren Bisgaard Award, and the Paul Simon Award for linking teaching and research. He is also Distinguished Research Professor Emeritus at Southern Illinois University Edwardsville.

Ronald D. Fricker, Jr. is Vice Provost for Faculty Affairs at Virginia Tech, where he has served as head of the Department of Statistics, Senior Associate Dean in the College of Science, and, subsequently, interim dean of the college. He is the author of *Introduction to Statistical Methods for Biosurveillance* (2013) and, with Steve Rigdon, *Monitoring the Health of Populations by Tracking Disease Outbreaks* (2020). He is a fellow of the American Statistical Association, a fellow of the American Association for the Advancement of Science, and an elected member of the Virginia Academy of Science, Engineering, and Medicine.

Douglas C. Montgomery is Regents' Professor and ASU Foundation Professor of Engineering at Arizona State University. He is an Honorary Member of the American Society for Quality, a fellow of the American Statistical Association, a fellow of the Institute of Industrial and Systems Engineering, and a fellow of the Royal Statistical Society. He is the author of 15 other books, including *Design and Analysis of Experiments*, 10th edition (2020) and *Design of Experiments for Reliability Achievement* (2022). He has received the Shewhart Medal, the Distinguished Service Medal, and the Brumbaugh Award from the ASQ, the Deming Lecture Award from the ASA, the Greenfield Medal from the Royal Statistical Society, and the George Box Medal from the European Network for Business and Industrial Statistics.

"This book serves as an excellent resource for students with diverse backgrounds, offering a thorough exploration of fundamental topics in statistics. The clear explanation of concepts, methods, and theory, coupled with an abundance of practical examples, provides a solid foundation to help students understand statistical principles and bridge the gap between theory and application. This book offers invaluable insights and guidance for anyone seeking to master the principles of statistics. I highly recommend adopting this book for my future statistics class."

Haijun Gong, *Saint Louis University*

"Professors Rigdon, Fricker and Montgomery have put together an impressive volume that covers not only basic probability and basic statistics, but also includes extensions in a number of directions, all of which have immediate relevance to the work of practitioners in quantitative fields. Suffused with common sense and insights about real data and problems, it is both approachable and precise. I'm excited about the inclusion of material on power and on multiple testing, both of which will help users become smarter about what their analyses can do, and I applaud their omission of too much theory. I also appreciate their use of R and of real data. This would be an excellent text for undergraduate or graduate-level data analysts."

Sam Buttrey, *Naval Postgraduate School (NPS)*

"This is a comprehensive and rich book that extends foundational concepts in statistics and probability in easily accessible form into data science as an integrated discipline. The reader applies and validates theoretical concepts in R and connects results from R back to the theory across many methods: from descriptive statistics to Bayesian models, time series, generalized linear models and more. Thoroughly enjoyable!"

Oliver Schabenberger, *Virginia Tech Academy of Data Science*

Introduction to Probability and Statistics for Data Science

with R

Steven E. Rigdon
Saint Louis University

Ronald D. Fricker, Jr.
Virginia Polytechnic Institute and State University

Douglas C. Montgomery
Arizona State 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/highereducation/isbn/9781107113046

DOI: 10.1017/9781316286166

When citing this work, please include a reference to the DOI 10.1017/9781316286166

First published 2025

Printed in Mexico by Litográfica Ingramex, S.A. de C.V.

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

A Cataloging-in-Publication data record for this book is available from the Library of Congress.

ISBN 978-1-107-11304-6 Hardback
ISBN 978-1-009-56835-7 Paperback

Additional resources for this publication at www.cambridge.org/ProbStatsforDS.

Steve Rigdon

To my wife Pat, who has always supported me and been at my side.

Ron Fricker

To my spouse, Christine: *Tu ventus sub alis meis es*.

And to my first statistics professor, Randy Spoeri: You introduced me to the subject and made it fun.

Doug Montgomery

To Cheryl, who has always supported and encouraged me. And to the memory of my first statistics professor, Ray Myers, mentor, colleague, collaborator, and friend.

Contents

Preface

This book is designed for students in statistics, data science, biostatistics, engineering, and mathematics programs who need a solid course in the fundamental concepts, methods, and theory of statistics. Our goal is to give students enough background in the methods and theory of statistics that they can understand modern techniques used in statistics and be able to apply them in the practice of data science.

We had to make some difficult choices regarding topic coverage. We do cover the important concepts of statistics, including maximum likelihood, the information matrix, power, etc., because these are needed for a student to be a successful statistician. When we cover maximum likelihood estimation, we specifically cover the method of approximating the maximum of the (log) likelihood function. Nowadays, data are so plentiful that we are often faced with testing multiple null hypotheses. Holm's method and the Benjamini–Hochberg method are derived and applied to real problems. There are a number of statistical methods that were developed in the late twentieth and early twenty-first centuries, including regression trees, large-scale hypothesis testing, methods of cross-validation, the bootstrap, Markov chain Monte Carlo, and others. We address the optimal selection of levels of a predictor variable to maximize the information we obtain; this leads to an introduction to the topic of optimal design. With some exceptions, these techniques have not found their way into introductory textbooks, especially those that emphasize theory. Throughout, we have tried to include topics that a statistician would use in the practice of statistics and to cover these thoroughly. We don't develop every aspect of statistical theory; for example, we cover very little of the limit theorems in statistics (convergence in probability, convergence in distribution, almost sure convergence, Slutsky's theorem, etc.). We don't cover the Cramer–Rao lower bound or the Rao–Blackwell theorem. We cover joint continuous distributions using multiple integration, but we do not go into great depth.

The emphasis is on modern methods of statistical inference. We develop enough theory so that students will understand these methods. If a statistician or data scientist is to work effectively with practitioners, it is up to the statistician to be the one to explain how methods work, what assumptions underlie the methods, what the limitations are, and how (or whether) the assumptions can be checked. Subject matter experts (i.e., the nonstatisticians) are not trained to do this. This is why it is important for students of statistics to understand the underlying theory behind the methods.

The flip side of our approach is that we do not develop theory for theory's sake. No theory is developed for the purpose that it might be usable in a future course. We have found that students

who understand probability and the foundational concepts of statistical theory can understand and use advanced statistical methods. Without a solid grounding on the theory and concepts of statistics it is difficult to pick up new methods.

Calculus is used in a number of places in the book, so students will need at least one or two semesters of calculus. There are a few uses of multiple integrals when we discuss joint continuous distributions, and for these the third semester of calculus will be needed. An instructor can skip these topics or sidestep the use of multiple integrals. We use calculus when it is necessary, for example in getting expected values of continuous random variables. We use R throughout the book. Although we do cover an introduction to R, it would be helpful if students had some prior background in R.

We use data extensively throughout the book. Most of the data sets are real (although at times we give small data sets to introduce a method). Many of these data sets are large. In most cases, we have provided a csv (comma separated values) file for the data. We also provide the R code used in the book to analyze the data sets that we provide. This can be found at: `www.cambridge.org/ProbStatsforDS`

While the book's website contains information about getting R up and running, we offer the following advice about loading in data sets and packages. First, it is always good practice to set the working directory to the directory on your computer that contains your data files. You can do this with the `setwd()` command. For example,

```
setwd("C:/Users/Documents/Rfiles")
```

will force R to read (write) files from (to) this directory. Note two things: (1) the path must be enclosed in quotes, and (2) subdirectories are indicated by forward slashes, not backslashes. Second, many of the methods we apply in this book require special R packages to run. These packages are collections of functions, dataframes, etc. Before you can use a package you must (1) install it, and (2) load it in during each R session. To install a package, such as `dplyr`, `type`

```
install.packages("dplyr")
```

Then, every time you start a new R session, you will have to load this package using

```
library(dplyr)
```

You need only install a package once on your computer, but you must call `library()` each time you begin an R session.

If you type `library` for a package you haven't installed, you will get an error. For example, if you haven't installed the `testassay` package and if you type `library(testassay)`, then you will get an error like this:

```
Error in library(testassay) : there is no package called 'testassay'
```

The remedy is to first install the package by typing `install.packages("testassay")` and then typing `library(testassay)`. If you ever get an error like the following

```
Error in arrange(df, y) : could not find function "arrange"
```

there is a good chance you forgot to load the package that contains the function `arrange()`, which is in the `dplyr` package. The remedy is to first type `library(dplyr)`.

Most two-semester courses will include a fairly standard first semester, which would likely cover the following chapters:

Semester 1

Chapter	Topic
1	Introduction
2	Data Visualization
3	Basic Probability
4	Random Variables
5	Discrete Distributions
6	Continuous Distributions
7	About Data and Data Collection
8	Sampling Distributions
9	Point Estimation
10	Confidence Intervals
11	Hypothesis Testing
12	Hypothesis Tests for Two or More Populations

The choice of topics for a second course would depend on the nature of the course. For example, our book could be used in a mathematical statistics course that emphasizes applications of statistics without sacrificing any of the underlying theory. Such a course could use the following material in the second term:

Semester 2

Chapter	Topic
13	Hypothesis Tests for Categorical Data
14	Regression
15	Bayesian Methods
17	The Jackknife and Bootstrap
18	Generalized Linear Models and Regression Trees
20	Large-Scale Hypothesis Testing

For a course that leans toward data science, the second semester coverage might include:

Semester 2

Chapter	Topic
13	Hypothesis Tests for Categorical Data
14	Regression
16	Time Series Methods
17	The Jackknife and Bootstrap
18	Generalized Linear Models and Regression Trees
19	Cross-Validation and Estimates of Prediction Error
20	Large-Scale Hypothesis Testing

A course for scientists or engineers could include selected topics in the above chapters, with additional methods from Chapter 15. For example, a course in biostatistics might emphasize the sections on logistic regression, discrimination, and classification since these are frequently used in medical and public health research. Such a course could minimize or skip material on regression trees. Instructors

could also use this as a textbook for a one-semester course by selecting (and omitting) material in the early part of the book. For example, the following chapters could be covered in a one-semester course:

One-semester course emphasizing statistics

Chapter	Topic
1	Introduction
2	Data Visualization (omitting data visualization for survey data, geospatial data, and network data)
3	Basic Probability
4	Random Variables
5	Discrete Distributions (possibly omitting the hypergeometric and multinomial distributions)
6	Continuous Distributions (possibly skipping the Weibull, Beta distributions, and the sections on transformations, moment-generating functions, and QQ plots)
7	About Data and Data Collection (hitting just the main ideas)
8	Sampling Distributions (skipping the proof of the Central Limit Theorem)
9	Point Estimation
10	Confidence Intervals
11	Hypothesis Testing
12	Hypothesis Tests for Two or More Populations
13 or 14	Hypothesis Tests for Categorical Data/Regression

For situations where students have had a prior course on statistics (possibly one that did not use calculus), a course could be designed to emphasize data science:

One-semester course emphasizing data science

Chapter	Topics
4–6	Select topics in these chapters to bring students up to speed
7	About Data and Data Collection (hitting just the main ideas)
8	Sampling Distributions (skipping the proof of the Central Limit Theorem)
9	Point Estimation
10	Confidence Intervals
11	Hypothesis Testing
12	Hypothesis Tests for Two or More Populations
13	Hypothesis Tests for Categorical Data
14	Regression
17.	The Jackknife and Bootstrap
18.	Generalized Linear Models and Regression Trees
20.	Large-Scale Hypothesis Testing

This book was typeset in LaTeX using a modified version of The Legrand Orange Book template originally created by Mathias Legrand and modified by Vel and the authors.

We would like to thank Emily Rigdon for LaTeXing much of the material in the book and Gary Smith for his careful reading and editing of the manuscript. We would also like to thank the staff at Cambridge, especially Lauren Cowles, Maggie Jeffers, and Lucy Edwards for their help in molding this book into what it has become, and for their patience through the process.

Steven E. Rigdon

Ronald D. Fricker, Jr.

Douglas C. Montgomery