# STATISTICS WITH R

## A BEGINNER'S GUIDE

online resources

ROBERT STINEROCK

'The book provides an excellent guide to statistics and the R programming language for beginners. Its wide range is exceptional to meet readers' needs in the era of big data and data science being the future. Brilliant!'

**Shaomin Wu, University of Kent**

'Stinerock provides a much-needed, easy-to-follow introduction to statistics and the R programming language. Any reader wishing to master and implement the statistical methods needed to derive meaning from data in today's challenging information-rich environment will benefit from this insightful, exciting, and profoundly useful text.'

**Morris B. Holbrook, Columbia University**

'The ability to code, analyze, and derive insights from vast amounts of data are critical skills in today's world of big data. This well-written book provides an excellent introduction to statistics and R programming language.'

**Sunil Gupta, Harvard Business School**

With carefully cultivated, jargon-free pedagogy featuring a mix of text, visuals, and off the page learning, this book offers students a step-by-step guide to statistical language. Featuring resources for reflection, revision, and practice, it is ideal for anyone hoping to:

- complete an introductory course in statistics
- prepare for more advanced statistical courses
- gain transferable analytical skills needed to interpret research from across the social sciences
- learn technical skills needed to present data visually
- acquire a basic competence in using R.

Supported by exercises, data sets, formulae lists, accessible definitions, software screenshots, and author-created screencasts, the book gives readers the conceptual foundation to use applied statistical methods in everyday research.

**Robert Stinerock** is Professor of Marketing and Quantitative Methods at the *Universidade Nova de Lisboa* in Lisbon.


**https://study.sagepub.com/stinerock**

SAGE www.sagepublishing.com
Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

Cover Design • Shaun Mercier

ISBN-13: 978-1-4739-2489-5

9 781473 924895

# STATISTICS WITH R

STINEROCK

SAGE

# STATISTICS WITH R

## A BEGINNER'S GUIDE

ROBERT STINEROCK

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483



First published 2018



Editor: Jai Seaman
Assistant editor: Alysha Owen
Production editor: Ian Antcliff
Copyeditor: Richard Leigh
Indexer: Martin Hargreaves
Marketing manager: Susheel Gokarakonda
Cover design: Shaun Mercier
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in the UK

**Library of Congress Control Number: 2016951887**

**British Library Cataloguing in Publication data**

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

ISBN 978-1-4739-2489-5
ISBN 978-1-4739-2490-1 (pbk)

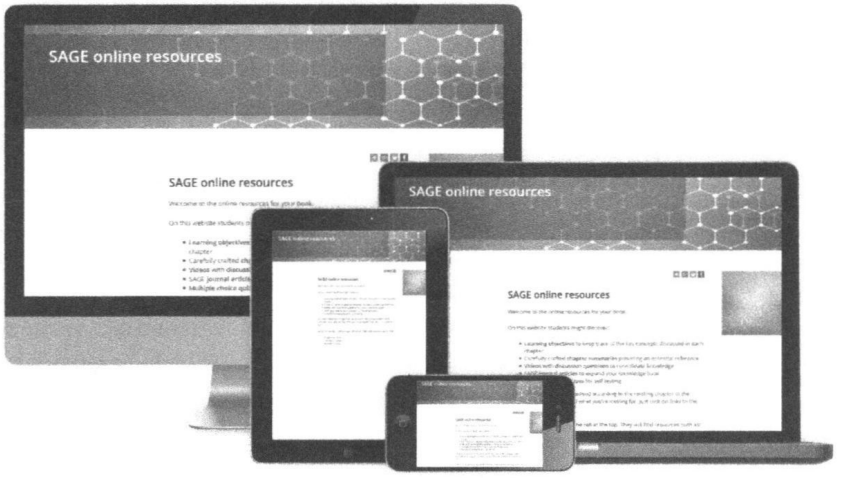
# STATISTICS WITH R

Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 800 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company's continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

To Jyoti. For being there for me through thick and thin.

# ONLINE RESOURCES




*Statistics with R: A Beginner's Guide* is supported by a wealth of online resources for both students and lecturers to aid in study and support teaching, which are available at https://study.sagepub.com/stinerock

## FOR STUDENTS

**Author-made screencasts** give you deeper insight into the key statistical ideas and R functions discussed in each chapter and show you first-hand how to work through some of the examples in the book. The **RStudio projects** the author used in the screencasts are also available so you can follow along on your own computer.

**Datasets and all R Scripts from the book** ready to upload into R and RStudio generate meaningful information to help you master your statistics and data analysis skills.

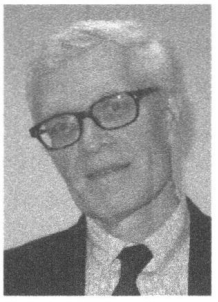
**Exercises** and **multiple choice questions** test your knowledge of key concepts and provide a helpful revision tool for assignments and exams while the **answers to in-text exercises** allow you to check your work and make sure you're on track.

## FOR INSTRUCTORS

**PowerPoint slides** featuring figures, tables, and key topics from the book can be downloaded and customized for use in your own presentations.

**Exercise testbanks** containing questions related to the key concepts in each chapter can be downloaded and used in class or for homework and exams.

# ABOUT THE AUTHOR



Robert Stinerock has more than 30 years of experience teaching statistics and probability to students at both the undergraduate as well as graduate level. He currently teaches statistics at three different universities: the Executive MBA program at Baruch College of the City University of New York; the Quantitative Finance program at the Stevens Institute of Technology in Hoboken, New Jersey; and the Faculdade de Economia, Universidade Nova de Lisboa, in Lisbon, Portugal.

He has received several awards for excellence in the classroom: the *Stevens Howe School Outstanding Undergraduate Teacher Award* (2006); the *Stevens Alumni Association Outstanding Teacher Award* (2005); and the *Fairleigh Dickinson Distinguished Faculty Award for Teaching* (1995).

He has published numerous research articles in academic journals, most recently in the *Journal of Macromarketing*, the *Journal of Business Research*, and *Geoforum*. This is his first book.

He earned his Bachelor's, Master's, and Ph.D. degrees, all from Columbia University.

He and his wife, Jyoti (a native of Mumbai, India), live in New York City.

# ACKNOWLEDGMENTS

I am profoundly grateful to all of those who have contributed to the conception and completion of this project; without those contributions, I can honestly say that this book would never have been started, much less completed.

I would like to thank the many special people at SAGE. My first editor, Katie Metzler, shepherded the proposal from inception through the review process to acceptance by the editorial board. Jai Seaman has been the editor with whom I have worked during the last two years, and from the beginning she has been encouraging and patient, even when I did not meet every deadline. Alysha Owen has been that hands-on, go-to person who has responded to every question and problem with a cheerful, can-do attitude that kept me on track even when progress was slow. On the editing side, Ian Antcliff has proven to be a creative and helpful problem-solver, and Richard Leigh has shown that he edits with skill and attention-to-detail. I am fortunate indeed to have had the privilege of working with such dedicated people.

I am also grateful to those in the software community who have created and distributed all their amazing software, free of charge. Special thanks go to both the people at the R Core Team as well as at RStudio. It is no exaggeration to say that you are changing the world.

My friend, Eric Novik, deserves special recognition as someone who has contributed to many aspects of this project. Eric first introduced me to R nearly 10 years ago, and since then has been not only a technical advisor but also a problem solver. Eric also encouraged me to introduce scripting within the context of the RStudio Project, something that is now a feature of the book's screencasts. Eric's influence has been indispensable because he has made the project better than I could have on my own.

I am also indebted to Gary Bronson (of Fairleigh Dickinson University) for providing timely, helpful advice when I sought his help early in this project. Gary himself has published 15 books on programming languages. When he speaks, I listen.

A special word of thanks goes to Donald G. Morrison (of Columbia and UCLA) who, back in 1982, first introduced me to the power and beauty of statistics. I know I speak for many of my Ph.D. classmates, as well as for myself, when I say that Don has been a role model *par excellence*, mentor, and friend. For that, Don, I thank you.

My debt of gratitude to my family is more than I can express here. My grandmother and mother-in-law have been the ones who made my life possible. There is never a day when I do not think of you both; your influence remains with me after all these years. Finally, how I tricked my wife, Jyoti, into marrying me is something I will never understand. During our decades together, her strength of character, integrity, loyalty, fierce courage, honesty, and generosity-of-heart have been her examples for me to live by. Jyoti, you have been the best thing that ever happened to me.

# PREFACE

The practice of statistics is primarily concerned with the development and application of methods for collecting, presenting, and analyzing data for the purpose of converting it into useful information. The ability to use these methods effectively is especially important today when the role of data shapes such an important part of our world. Acquiring an understanding of both the strengths and limitations of the most widely-used statistical methods, and then using these methods competently, now form an important part of the foundation of such diverse fields as economics, political science, psychology, sociology, finance, marketing, data mining, production management, quality control, and many other commercial, scientific, and engineering areas. Although one need not be a rigorously trained statistician to understand and apply the statistical methods presented in this book, a grasp of the conceptual foundation underpinning *statistics as a way of thinking* is required.

## THE NEED FOR AN R-BASED FIRST-YEAR STATISTICS BOOK

The principal feature of this book is that it provides step-by-step instruction in the use of the R statistical language that is both accessible and comprehensive for those individuals wishing to (1) complete an introductory-level course in statistics, (2) prepare for more advanced statistical courses, (3) gain the necessary analytic skills now required for other disciplines such as the management, natural and social sciences, and (4) acquire a basic competence in the use of R.

R is an extremely powerful, highly versatile, widely used statistical programming system that is free and downloadable. Its particular strengths are in statistical modeling, graphical presentations, and its ability to program user-defined functions. In the last decade, R has gained wide acceptance among statisticians and data scientists in industry, academia, research organizations, and public sectors.

To say that R is only a statistical package, however, is to underclaim all that R can do. Because it offers capabilities beyond those featured by other statistical packages, R should not be characterized as a statistical package at all but rather as a *statistical environment.* While it is true that R is a powerful statistical programming and modeling language, it also has the following features and capabilities:

1. R includes a wide-variety of techniques for generating all sorts of publication-quality graphical presentations and data visualization results.
2. R is a versatile and powerful data-manipulation tool.

3. R includes access to thousands of free, downloadable, add-on packages contributed by practitioners and researchers from different fields that greatly expand the scope and types of problems R can work on. Currently, more than 11,000 user-contributed R packages are available at http://www.r-project.org/, and more are being added every day.
4. R provides a connection to a worldwide community of practitioners and researchers-programmers, data scientists, and statisticians, among others-from almost every conceivable field. These naturally occurring communities are connected and sustained not only by way of social media, blogs, and websites but also through specialized R groups and conferences.
5. R provides a mailing list to which one may subscribe free-of-charge that deals with many commonly encountered problems as well as frequently asked questions.
6. R has the ability to read from, and write to, most of the other commercially available software packages including but not limited to Excel, JMP, Minitab, SAS, SPSS, Stata, and Systat. It has the ability to move data and graphics easily between itself and these other environments.
7. R is platform-independent in that it runs on all the major operating systems including Windows, Mac OS X, and Unix.
8. The source code can be downloaded so that anyone can examine the internal workings of its routines and packages and make modifications, if desired.

To elaborate further as to why the use of R features so prominently in this book, there are two main reasons, one pedagogical, the other professional.

First, to answer why I provide such extensive coverage of a statistical language in an introductory-level book, I concede that in former times this aspect might not have been included at all. A few years ago, many faculty felt that since statistics and probability are essentially conceptual in nature, the real value-added to students of being able to think statistically and probabilistically overrode any consideration of students learning to operate software, especially since such instruction would have been viewed as an unnecessary encroachment on time that might be spent more productively on learning the core ideas of statistics. Although some authors and faculty still adhere to this position, I do not take this view. On the contrary, I believe that the ability to think programmatically, developed in part by learning how to translate the requirements of a given statistical problem into the syntax of a statistical programming language, can actually help students and readers learn, absorb, and apply the statistical methods they will encounter throughout the book.

Second, as to why I choose R over the more ubiquitous Excel, I would add that since today's world of work is more challenging than ever, the people entering that world must be armed with the broadest and most current skill-set possible. Today, the ability to use a programming language like R provides students with a clear advantage in the competitive arenas of both university admissions and subsequent employment. At the very least, just as it allows them to add another line to their vita (under computer skills), it also gives them something to talk about in admissions or job interviews. In my experience at the Stevens Institute of Technology and the Universidade Nova de Lisboa, some of our graduating students have reported that their knowing a bit of R gave them a differentiating advantage over other students with whom they were competing for positions in the financial or pharmaceutical industries. As one corporate recruiter recently said to one of our students after making an offer of employment upon graduation, "that you actually have some experience working with R means that we would not have to spend so much time training