STATISTICS USING R

An Integrative Approach

Using numerous examples with real data, this textbook closely integrates the learning of statistics with the learning of R. It is suitable for introductory-level learners, allows for curriculum flexibility, and includes, as an online resource, R-code script files for all examples and figures included in each chapter, for students to learn from and adapt and use in their future data analytic work. Other unique features created specifically for this textbook include an online R tutorial that introduces readers to data frames and other basic elements of the R architecture, and a CRAN library of datasets and functions that is used throughout the book. Essential topics often overlooked in other introductory texts, such as data management, are covered. The textbook includes online solutions to all end-of-chapter exercises and PowerPoint slides for all chapters as additional resources, and is suitable for those who do not have a strong background in mathematics.

Sharon Lawner Weinberg is Professor of Applied Statistics and Psychology, and the former Vice Provost for Faculty Affairs, at New York University (NYU). She is the recipient of the NYU Steinhardt Outstanding Teaching Award, and has taught statistics at both undergraduate and graduate levels. Her research has been supported by federal agencies and private foundations.

Daphna Harel is Associate Professor of Applied Statistics at New York University. She is known for her innovative approach to teaching both introductory and advanced statistics. Her research has been supported by federal agencies and foundations, such as the National Institutes for Health and the Canadian Institutes for Health Research.

Sarah Knapp Abramowitz is Professor of Mathematics and Computer Science, Department Chair, and Co-ordinator of Statistics Instruction at Drew University. She is Associate Editor of the *Journal of Statistics Education* and has presented at national conferences on topics related to the teaching of statistics.

Statistics Using R

AN INTEGRATIVE APPROACH

SHARON LAWNER WEINBERG

New York University

DAPHNA HAREL

New York University

SARAH KNAPP ABRAMOWITZ

Drew University, New Jersey

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

79 Anson Road, #06–04/06, Singapore 079906

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/9781108719148
DOI: 10.1017/9781108755351

First published 2021

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

Library of Congress Cataloging-in-Publication Data
Names: Weinberg, Sharon L., author. | Harel, Daphna, 1988– author. |
Abramowitz, Sarah Knapp, 1967– author.
Title: Statistics using R : an integrative approach / Sharon Lawner
Weinberg, Daphna Harel, Sarah Knapp Abramowitz.
Description: Cambridge ; New York, NY : Cambridge University Press, 2021. |
Includes bibliographical references and index.
Identifiers: LCCN 2019059928 (print) | LCCN 2019059929 (ebook) |
ISBN 9781108719148 (paperback) | ISBN 9781108755351 (ebook)
Subjects: LCSH: R (Computer program language) |
Mathematical statistics – Data processing. | Statistics – Data processing.
Classification: LCC QA276.45.R3 W45 2021 (print) | LCC QA276.45.R3 (ebook) |
DDC 519.50285/5133–dc23
LC record available at https://lccn.loc.gov/2019059928
LC ebook record available at https://lccn.loc.gov/2019059929

ISBN 978-1-108-71914-8 Paperback

Additional resources for this publication at www.cambridge.org/stats-r

To our families

Contents

Preface

This first edition of *Statistics Using R: An Integrative Approach* capitalizes on the versatility and power of the R software package to create a course of study that links good statistical and data science practice to the analysis of real data. It also benefits from the many years of the authors' experience teaching statistics to undergraduate students at a liberal arts university and to undergraduate and graduate students at a large research university from a variety of disciplines including education, psychology, health, and policy analysis. R is a free, open-source software, which provides a command line approach to the analysis of data. This textbook teaches readers the skills necessary to program statistical analyses using this command line approach.

Guiding Principles Underlying Our Approach

First, and perhaps most importantly, we believe that a good data analytic plan must serve to uncover the story behind the numbers; what the data tell us about the phenomenon under study. To begin, a good data analyst must know their data well and have confidence that it satisfies the underlying assumptions of the statistical methods used. Accordingly, we emphasize the usefulness of diagnostics in both graphical and statistical form to expose anomalous cases, which might unduly influence results, and to help in the selection of appropriate assumption-satisfying transformations so that ultimately we may have confidence in our findings. We also emphasize the importance of using more than one method of analysis to answer fully the question posed and understanding potential bias in the estimation of population parameters.

Second, because we believe that data are central to the study of good statistical practice, we have made available the datasets used in this textbook. The textbook's website contains the datasets used throughout the text. Also, we have published a software package, hosted on R's server, that can be downloaded as another means to access the datasets. This package, named *sur,* also contains several R functions that we have written to facilitate the application of several statistical methods.

The datasets used in this textbook provide examples of real-world data. Two are large sets of real data that we make repeated use of in both worked-out examples and end-of-chapter exercises. One dataset contains 48 variables and 500 cases from the education discipline that includes psychosocial measures such as self-concept; the other contains 49 variables and nearly 4,500 cases from the health discipline. By posing interesting questions about variables in these large, real datasets (e.g., Is there a gender difference in eighth graders' expected income at age 30? Is self-concept in eighth grade related to eighth

graders' expected income at age 30?), we are able to employ a more meaningful and contextual approach to the introduction of statistical methods and to engage students more actively in the learning process. The repeated use of these datasets also contributes to creating a more cohesive presentation of statistics; one that links different methods of analysis to each other and avoids the perception that statistics is an often confusing array of so many separate and distinct methods of analysis, with no bearing or relationship to one another.

Third, to facilitate the analysis of these data, and to provide students with a platform for actively engaging in the learning process associated with what it means to be a good researcher and data analyst, we incorporate the functionality of R version 3.5.3 into the presentation of statistical material using a highly integrative approach that reflects practice. Students learn R along with each new statistical method covered, thereby allowing them to apply their newly learned knowledge to the real world of applications. In addition to demonstrating the use of R within each chapter, all chapters have an associated .R file, that may be accessed from the textbook's website (www.cambridge.org/stats-r) along with other ancillary materials and resources. The .R files are designed to allow students not only to replicate all worked-out examples within a chapter, but also to reproduce the figures embedded in a chapter, and to create their own .R files by extracting and modifying commands from them. Emphasizing data workflow management throughout the text using the .R file allows students to begin to appreciate one of the key ingredients in being a good researcher. Of course, another key ingredient to being a good researcher is content knowledge, and toward that end, we have included in the text a more comprehensive coverage of essential topics in statistics not covered by other textbooks at the introductory level, including regression to the mean, the weighted mean, and potential sources of bias in the estimation of population parameters based on the analysis of data from quasi-experimental designs. We also include a discussion of Simpson's paradox, counterfactuals, and other issues related to research design in an entire chapter (Chapter 12) devoted to this topic.

Fourth, in accordance with our belief that the result of a null hypothesis test (to determine whether an effect is real or merely apparent) is only a means to an end (to determine whether the effect is important or useful) rather than an end in itself, we stress the need to evaluate the magnitude of an effect if it is deemed to be real, and of drawing clear distinctions between statistically significant and substantively significant results. Toward this end, we introduce the computation of standardized measures of effect size as common practice following a statistically significant result. While we provide guidelines for evaluating, in general, the magnitude of an effect, we encourage readers to think more subjectively about the magnitude of an effect, bringing into the evaluation their own knowledge and expertise in a particular area.

Finally, we believe that a key ingredient of an introductory statistics text is a lively, clear, conceptual, yet rigorous approach. We emphasize conceptual understanding through an exploration of both the mathematical principles underlying statistical methods and real-world applications. We use an easy-going, informal style of writing that we have found gives readers the impression that they are involved in a personal conversation with the authors. And, we sequence concepts with concern for student readiness, reintroducing topics in a spiraling manner to provide reinforcement and promote the transfer of learning.

Overview of Content Coverage and Intended Audience

Along with the earlier topics mentioned, the inclusion of linear and nonlinear transformations, diagnostic tools for the analysis of model fit, tests of inference, an in-depth discussion of interaction and its interpretation in both two-way analysis of variance and multiple regression, and nonparametric statistics, the text provides a highly comprehensive coverage of essential topics in introductory statistics, and in so doing gives instructors flexibility in curriculum planning and provides students with more advanced material for future work in statistics. In addition, the text includes a large bibliography of references to relevant books and journal articles, and many end-of-chapter exercises with detailed answers on the textbook's website.

The book, consisting of 18 chapters, is intended for use in a one- or two-semester introductory applied statistics course for the behavioral, social, or health sciences at either the graduate or undergraduate level, or as a reference text as well. It is not intended for readers who wish to acquire a more theoretical understanding of mathematical statistics. To offer another perspective, the book may be described as one that begins with modern approaches to exploratory data analysis (EDA) and descriptive statistics, and then covers material similar to what is found in an introductory mathematical statistics text, designed, for example, for undergraduates in math and the physical sciences, but stripped of calculus and linear algebra, and grounded instead in data examples. Thus, theoretical probability distributions, Bayes' theorem, the law of large numbers, sampling distributions and the central limit theorem are all covered, but in the context of solving practical and interesting problems.

Finally, to facilitate learning R, an online tutorial has been written to accompany the textbook. The tutorial covers the basic use and manipulation of datasets, also referred to as data frames in R. The activities covered in the tutorial are designed to help students understand the examples in each chapter and to complete end-of-chapter exercises in the textbook. It also is designed to teach some basic coding skills that will be helpful when working with data frames in R, complete more complex analyses, and, ultimately, to learn the larger statistical concepts covered throughout the textbook.

Acknowledgments

This book has benefited from the many helpful comments of our New York University and Drew University students, too numerous to mention by name, and from the insights and suggestions of several colleagues. For their help, we would like to thank (in alphabetical order) Chris Apelian, Ellie Buteau, Chris Casement, Miao Chi, Sarah Friedman, Andrea Hassler, Jennifer Hill, Michael Karchmer, Steve Kass, Jon Kettenring, Linda Lesniak, Yi Lu, Kathleen Madden, Isaac Maddow-Zimet, Meghan McCormick, Joel Middleton, and Marc Scott. Of course, any errors or shortcomings in the text remain the responsibility of the authors.