STATISTICS USING STATA

An Integrative Approach

Second Edition

Building upon the success of the first edition, *Statistics Using Stata* uses the latest version of Stata to meet the needs of today's students. Engaging and accessible for students from a variety of mathematical backgrounds, this textbook integrates statistical concepts with the Stata (version 16) software package. It aligns Stata commands with examples based on real data, enabling students to understand statistics in a way that reflects statistical practice. Capitalizing on Stata's menu-driven "point and click" and program syntax interface, the chapters guide students from the comfortable "point and click" environment to the beginnings of statistical programming. Its coverage of essential topics gives instructors flexibility in curriculum planning and provides students with more advanced material to prepare for future work. Online resources – including solutions to exercises, PowerPoint slides, and Stata syntax (.do-files) for each chapter – allow students to review independently and adapt code to analyze new problems.

Sharon Lawner Weinberg is Professor of Applied Statistics and Psychology and former Vice Provost for Faculty Affairs at New York University, USA.

Sarah Knapp Abramowitz is Professor of Mathematics and Computer Science at Drew University, USA, where she is also Department Chair and Coordinator of Statistics Instruction.

Statistics Using Stata

AN INTEGRATIVE APPROACH

Second Edition

SHARON LAWNER WEINBERG

New York University

SARAH KNAPP ABRAMOWITZ

Drew University

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/9781108725835
DOI: 10.1017/9781108770163

First edition 2016
Second edition 2020

Printed in the United Kingdom by TJ International Ltd, Padstow Cornwall, 2020

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

Library of Congress Cataloging-in-Publication Data
Names: Weinberg, Sharon L., author. | Abramowitz, Sarah Knapp, 1967– author.
Title: Statistics using Stata : an integrative approach / Sharon Lawner Weinberg, New York University, Sarah Knapp Abramowitz, Drew University.
Description: Cambridge, United Kingdom ; New York, NY, USA : University Printing House, 2019. | Includes bibliographical references and index.
Identifiers: LCCN 2019019437 | ISBN 9781108725835 (alk. paper)
Subjects: LCSH: Mathematical statistics – Data processing. | Stata.
Classification: LCC QA276.4 .W455 2019 | DDC 519.50285/555–dc23
LC record available at https://lccn.loc.gov/2019019437

ISBN 978-1-108-72583-5 Paperback

Additional resources for this publication at www.cambridge.org/Stats-Stata2e

To our families

Contents

Preface

This second edition of *Statistics Using Stata: An Integrative Approach* continues to capitalize on the versatility and power of the Stata 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. Stata provides both a menu-driven and command-line approach to the analysis of data and, in so doing, facilitates the smooth transition to a more advanced course of study in statistics.

New to the Second Edition

While the first edition of this text is based on Stata's version 14, the second edition is based on version 16, the most current version of Stata. We devote an entirely new chapter (Chapter 19) on the command **putexcel** new to version 15 and highlighted in version 16, so that readers may produce reproducible reports that include professionally styled tables and figures based on the statistical analyses they have learned from throughout the text. As we know, drafting a cohesive, thoughtful, and compelling narrative is critically important for enabling work to be heard and appreciated. Through multiple examples in this new chapter we illustrate how putexcel may be used along with other Stata commands (e.g., **estpost** and **esttab**) to substantiate that narrative with professionally styled tables and figures that are concise, cleanly illustrated, and easily reproducible.

In the first edition, we gave students an important conceptual understanding of power and its importance in designing a research study. In the second edition, we extend that understanding to a practical level by providing readers with knowledge about how to use Stata to carry out a power analysis for each of the inferential methods covered in the text. In particular, we illustrate using Stata's **power** command how to link the key elements of power analysis to the calculation of a minimum sample size needed to obtain a specified power to detect a statistically significant and meaningful result.

To give greater clarity to the notion of statistical control, in the second edition we expand upon the discussion of statistical control in our first edition and include examples that illustrate in what sense multiple regression allows for the statistical control of confounders. In the second edition, we elaborate upon the use of the adjusted R-squared as a measure of model goodness-of-fit to its use as an aid in model selection. We also include discussion of cross-validation and the AIC and BIC statistics as more modern

approaches to model selection that are frequently referenced in the data science and research literature.

Because students often have a difficult time understanding and interpreting statistical interaction within the multiple regression setting, in the second edition we devote an entirely new chapter to this topic. In this new Chapter 17, we emphasize the importance of visualization via the versatile **margins** and **marginsplot** commands, which are relatively new additions to Stata. We also include in this edition the testing of simple effects following a significant interaction. We do so as a way to help uncover and elucidate the nuanced nature of the interaction obtained. Finally, although we have made a substantial number of changes in the second edition from the first, the second edition, like its predecessor, embraces and is motivated by the following important guiding principles.

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 his/her 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, the textbook's website contains several data sets used throughout the text. Two are large sets of real data that we make repeated use of in both worked-out examples and end-of-chapter exercises. One data set 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 data sets (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 data sets 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 latest version of Stata (version 16) into the presentation of statistical material using a highly integrative approach that reflects practice. Students learn Stata 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 Stata within each chapter, all chapters have an associated .do-file,

that may be accessed from the textbook's website (www.cambridge.org/Stats-Stata2e) along with other ancillary materials and resources. The .do-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 .do-files by extracting and modifying commands from them. Emphasizing data workflow management throughout the text using the Stata .do-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 robust methods of estimation based on resampling using the bootstrap, 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 non-parametric 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 19 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.

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, Daphna Harel, Jennifer Hill, Chuck Huber, 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.