Springer Texts in Statistics

Series Editors:
Richard DeVeaux
Stephen E. Fienberg
Ingram Olkin

More information about this series at http://www.springer.com/series/417

Also by Richard M. Heiberger

R through Excel:
A Spreadsheet Interface for Statistics,
Data Analysis, and Graphics,
with Erich Neuwirth, Springer 2009

Computation for the Analysis of Designed Experiments, Wiley 1989

Richard M. Heiberger • Burt Holland

Statistical Analysis and Data Display

An Intermediate Course with Examples in R

Second Edition

Springer

Richard M. Heiberger
Department of Statistics
Temple University
Philadelphia, PA, USA

Burt Holland
Department of Statistics
Temple University
Philadelphia, PA, USA

ISSN 1431-875X
ISSN 2197-4136 (electronic)
Springer Texts in Statistics
ISBN 978-1-4939-2121-8
ISBN 978-1-4939-2122-5 (eBook)
DOI 10.1007/978-1-4939-2122-5

Library of Congress Control Number: 2015945945

Springer New York Heidelberg Dordrecht London

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

In loving memory of Mary Morris Heiberger

To my family: Margaret, Irene, Andrew, and Ben

Preface

1 Audience

Students seeking master's degrees in applied statistics in the late 1960s and 1970s typically took a year-long sequence in statistical methods. Popular choices of the course textbook in that period prior to the availability of high-speed computing and graphics capability were those authored by Snedecor and Cochran (1980) and Steel and Torrie (1960).

By 1980, the topical coverage in these classics failed to include a great many new and important elementary techniques in the data analyst's toolkit. In order to teach the statistical methods sequence with adequate coverage of topics, it became necessary to draw material from each of four or five text sources. Obviously, such a situation makes life difficult for both students and instructors. In addition, statistics students need to become proficient with at least one high-quality statistical software package.

This book *Statistical Analysis and Data Display* can serve as a standalone text for a contemporary year-long course in statistical methods at a level appropriate for statistics majors at the master's level and for other quantitatively oriented disciplines at the doctoral level. The topics include concepts and techniques developed many years ago and also a variety of newer tools.

This text requires some previous studies of mathematics and statistics. We suggest some basic understanding of calculus including maximization or minimization of functions of one or two variables, and the ability to undertake definite integrations of elementary functions. We recommend acquired knowledge from an earlier statistics course, including a basic understanding of statistical measures, probability distributions, interval estimation, hypothesis testing, and simple linear regression.

2 Motivation

The Second Edition in 2015 has four major changes since the First Edition in 2004 Heiberger and Holland (2004). The changes are summarized here and described in detail in Section 5.

- The computation for the Second Edition is entirely in R (R Core Team, 2015). R is a free open-source publicly licensed software environment for statistical computing and graphics. The computation for the First Edition is mostly in S-Plus, with some R and some SAS. R uses a dialect of the S language developed at Bell Labs. The R dialect is closely related to the dialect of S used by S-Plus. R is much more powerful now than it was when the First Edition was written.
- All graphs from the First Edition have been redrawn in color. There are many additional graphs new to the Second Edition. The graphs are easier to specify because they are built with the much more powerful graphical primitives that exist now and didn't exist 12 years ago. Most graphs are constructed with **lattice**, the R implementation of **trellis** graphics pioneered by S-Plus. Some, particularly in Chapter 15, are drawn using `mosaic` and related functions in the **vcd** package. Functions for the graphic displays designed for this book are included in the **HH** package available at CRAN (Heiberger, 2015).
- Most chapters in the Second Edition are similar in content to the chapters in the First Edition. There are several revised and expanded chapters and several additional appendices.
- The new appendices respond to shifts in the software landscape and/or in the assumed knowledge of computing by the intended audience since 2004.

3 Structure

The book is organized around statistical topics. Each chapter introduces concepts and terminology, develops the rationale for its methods, presents the mathematics and calculations for its methods, and gives examples supported by graphics and computer output, culminating in a writeup of conclusions. Some chapters have greater detail of presentation than others, based on our personal interests and expertise.

Our emphasis on graphical display of data is a distinguishing characteristic of this book. Many of our graphical displays appeared here for the first time. We show graphs, how to construct and interpret them, and how they relate to the tabular outputs that appear automatically when a statistical program "analyzes" a data set. The graphs are not automatic and so must be requested. Gaining an understanding of a data set is always more easily accomplished by looking at appropriately drawn

graphs than by examining tabular summaries. In our opinion, graphs are the heart of most statistical analyses; the corresponding tabular results are formal confirmations of our visual impressions.

We believe that a firm control of the language gives the analyst the tools to think about the ideal way to detect and display the information in the data. We focus our presentation on the written command languages, the most flexible descriptors of the statistical techniques. The written languages provide the opportunity for growth and understanding of the underlying techniques. The point-and-click technology of icons and menus is sometimes convenient for routine tasks. However, many interesting data analyses are not routine and therefore cannot be accomplished by pointing and clicking the icons provided by the program developers.

4 Computation

In the First Edition, and again in the Second Edition, the code and data for all examples and figures in the book is available for download.

For the Second Edition, the datasets and R code will be distributed as the R package **HH** through CRAN (Heiberger, 2015).

For the First Edition, the download containing S-Plus, R, and SAS code was initially (in 2004) available from my web site. In 2007, the R code was placed on CRAN (the Comprehensive R Archive Network) as the R package **HH**. In 2009, the S-Plus code was placed on CSAN (the Comprehensive S Archive Network) as the S-Plus package **HH** (Heiberger, 2009).

All datasets in the **HH** package are documented in the book.

4.1 R

R (R Core Team, 2015) is free, publicly licensed, extensible, open-source software. The R language is a dialect of the S language (Becker et al., 1988), similar to that used by S-Plus (Insightful Corp., 2002; TIBCO Software Inc., 2010). Much code (both functions and examples) written for one will also work in the other. R has been increasing its reach—within academia, industry, government, and internationally. Please see Appendix A for information on downloading and using R.

The S language was originally developed at Bell Labs in the 1970s. The Association for Computing Machinery (ACM) awarded John M. Chambers of Bell Labs the 1998 Software System Award for developing the S system.

The R language is an exceptionally well-developed tool for statistical research and analysis, that is for exploring and designing new techniques of analysis, as well as for analysis. The trellis graphics implementation in R's **lattice** package is especially strong for statistical graphics, the output of data analysis through which both the raw data and the results are displayed for the analyst and the client.

R is available by download. The developers are The R Development Core Team, an international group that includes John Chambers and other former Bell Labs researchers.

4.2 The HH Package in R

An important feature of this book is its graphical displays of statistical analyses. For the Second Edition, the **HH** functions for graphing have been rewritten using the more powerful graphing infrastructure that is now available in the **lattice** package in R. The package version number has been changed from the **HH_2.3.x** series to the **HH_3.1-x** series to reflect the redesign. The First Edition had black-and-white figures in print, even though the software at that time produced color figures. In the Second Edition all figures, both in print and in the eBook edition, are in color.

Please see Appendix B for information on working with the **HH** package.

R graphics have much improved since the time of the First Edition. The **lattice** graphics package for plotting coordinated sets of displays was in its infancy when we wrote the First Edition, not yet as capable as the equivalent **trellis** graphics system in S-Plus, and specifically not capable of all the figures in the book. Now **lattice** is much more powerful than **trellis**, and can be even further extended with the capabilities since encoded in the **latticeExtra** package (Sarkar and Andrews, 2013).

The R package system was also not as extensive at that time, and the S-Plus package system did not yet exist. The code and examples for the First Edition of the book were distributed as a zip file on my website and accessible through the Springer website. The code and examples were revised and distributed as an R package **HH** beginning in 2007, and as an S-Plus package in 2009, when S-Plus created their package system. I have continually maintained and extended the software.

4.3 S-Plus, now called S+

S+ is still available, but less commonly used. TIBCO, the owner of S+ is now distributing a Developer's Edition of R called TERR (TIBCO Enterprise Runtime for R) based on their new enterprise-grade, high-performance statistical engine (TIBCO

Software Inc., 2014). The design goal of TERR is to be able to install all R packages. As of July 2014, TERR had not yet implemented their graphics system. Once their graphics system is implemented, **HH_3.1-x** will work with TERR.

The older version of **HH** (Heiberger, 2009), designed for the First Edition of this book, continues to work with S+.

4.4 SAS

SAS is an important statistical computing system in industry. All the code from our First Edition still works. My own personal work has become more highly R-focused. I have chosen to drop most of the SAS discussion and examples from the body of the Second Edition.

Some SAS material is still in the body of the Second Edition. Now-standard terminology introduced by SAS, primarily the notation for "Types" of Sums of Squares described in Section 13.6, is referenced and described. The notation of the SAS `MODEL` statement is similar to the notation of the R model formula. Comparisons of the two notations are in Sections 9.4.1, 12.13.1, 12.15, 12.A, 13.4, and 13.5.

All datasets in the Second Edition can be used with SAS. See Appendix H for details.

5 Chapters in the Second Edition

5.1 Revised Chapters

All graphs from the First Edition have been redrawn in color and with the use of much more powerful graphical primitives that didn't exist 12 years ago.

There are many additional graphs new to the Second Edition.

Chapters 3 and 5 have many new figures, most built with the `NTplot` function. The graphs, showing significance and power of hypothesis tests for the normal and *t* distributions, produced by this single function cover most of the standard first semester introductory Statistics course.

Chapter 11 "Multiple Regression—Regression Diagnostics" has a new section 11.3.7 "Residuals vs Leverage" to discuss one of the panels produced by R's `plot.lm` function that was not in the similar S-Plus function.

Chapter 15 "Bivariate Statistics—Discrete Data" has undergone major revision. The examples are now centered on `mosaic` graphics, using the **vcd** package that was not available when the First Edition was written.

Section 15.8 "Example—Adverse Experiences" is new. The discussion focuses on the Adverse Effects dotplot, and shows how multi-panel plots graphical displays can replace pages of tabular data. The discussion is based on the work in which I participated while at research leave at GSK (Amit et al., 2008).

Section 15.9 "Likert Scale Data" is new. This section is based on my recent work with Naomi Robbins (Heiberger and Robbins, 2014). Rating scales, such as Likert scales and semantic differential scales, are very common in marketing research, customer satisfaction studies, psychometrics, opinion surveys, population studies, and numerous other fields. We recommend diverging stacked bar charts as the primary graphical display technique for Likert and related scales. We discuss the perceptual issues in constructing the graphs. Many examples of plots of Likert scales are given.

5.2 Revised Appendices

We have made major changes to the Appendices. There are more appendices now and the previous appendices have been restructured and expanded. The description of the Second Edition appendices is in Section 1.3.5.

6 Exercises

Learning requires that the student work a fair selection of the exercises provided, using, where appropriate, one of the statistical software packages we discuss. Beginning with the exercises in Chapter 5, even when not specifically asked to do so, the student should routinely plot the data in a way that illuminates its structure, and state all assumptions made and discuss their reasonableness.

Acknowledgments: First Edition

We are indebted to many people for providing us advice, comments, and assistance with this project. Among them are our editor John Kimmel and the production staff at Springer, our colleagues Francis Hsuan and Byron Jones, our current and former students (particularly Paolo Teles who coauthored the paper on which Chapter 18 is based, Kenneth Swartz, and Yuo Guo), and Sara R. Heiberger. Each of us gratefully

acknowledges the support of a study leave from Temple University. We are also grateful to Insightful Corp. for providing us with current copies of S-Plus software for ourselves and our student, and to the many professionals who reviewed portions of early drafts of this manuscript.

Philadelphia, PA, USA Richard M. Heiberger
Philadelphia, PA, USA Burt Holland
July 2004

Acknowledgments

We are indebted to many additional people for support in preparing the Second Edition. Our editors at Springer Jon Gurstelle (now at Wiley), Hannah Bracken, and Michael Penn encouraged the preparation of this Second Edition. Alicia Strandberg at Villanova University used a preliminary version of this edition with two of her classes. She and her students provided excellent feedback and suggestions for the preparation of this material. I also used drafts of this edition in my own courses at Temple University and incorporated the classes' feedback into the revision.

We are grateful to the R Core and the many R users and contributors who have provided the software we use so heavily in our graphical and tabular analyses.

The material in the new section on Adverse Effects is based on the work with the GSK team investigating graphics for safety data in clinical trials, particularly coauthors Ohad Amit and Peter W. Lane.

The material in the new section on Likert scale plots is based on the work with Naomi Robbins.

The First Edition was coauthored by Burt Holland. Even though Burt died in 2010, I am writing this second preface mostly in the plural. Burt's voice is present in much of the text of the Second Edition. Most of the numbered chapters have essentially the same content as in the First Edition.

The new sections and the Appendices in the Second Edition are entirely by me. All graphs in this edition are newly drawn by me using the more powerful graphics infrastructure that is now available in R.

I had several discussions with Kenneth Swartz when I was initially considering writing this edition and at various points along the way.

Barbara Bloomfield provided me overall support in everything. She also responded to my many queries on stylistic and appearance issues in the revised manuscript and graphs.

Philadelphia, PA, USA Richard M. Heiberger
October 2015

Contents

Author Bios

Richard M. Heiberger is Professor Emeritus in the Department of Statistics of Temple University, an elected Fellow of the American Statistical Association, and a former Chair of the Section on Statistical Computing of the American Statistical Association. He was Graduate Chair for the Department of Statistics and Acting Associate Vice Provost for the University. He participated in the design of the linear model and analysis of variance functions while on research leave at Bell Labs. He has taught short courses at the Joint Statistics Meetings, the American Statistical Association Conference on Statistical Practice, the R Users Conference, and the Deming Conference on Applied Statistics. He has consulted with several pharmaceutical companies.

Burt Holland was Professor in the Department of Statistics of Temple University, an elected Fellow of the American Statistical Association, Chair of the Department of Statistics of Temple University, and Chair of Collegial Assembly of the Fox School. He has taught short courses at the Joint Statistics Meetings and the Deming Conference on Applied Statistics. He has made many contributions to linear modeling and simultaneous statistical inference. He frequently served as consultant to medical investigators. He developed a very popular General Education course on Statistics and the News.