Statistics and Computing

Series Editors:
John Chambers
David J. Hand
Wolfgang K. Härdle

Statistics and Computing

Brusco/Stahl: Branch and Bound Applications in Combinatorial Data Analysis
Chambers: Software for Data Analysis: Programming with R
Dalgaard: Introductory Statistics with R, 2nd ed.
Gentle: Elements of Computational Statistics
Gentle: Numerical Linear Algebra for Applications in Statistics
Gentle: Random Number Generation and Monte Carlo Methods, 2nd ed.
Härdle/Klinke/Turlach: XploRe: An Interactive Statistical Computing Environment
Hörmann/Leydold/Derflinger: Automatic Nonuniform Random Variate Generation
Krause/Olson: The Basics of S-PLUS, 4th ed.
Lange: Numerical Analysis for Statisticians
Lemmon/Schafer: Developing Statistical Software in Fortran 95
Loader: Local Regression and Likelihood
Marasinghe/Kennedy: SAS for Data Analysis: Intermediate Statistical Methods
Muenchen: R for SAS and SPSS Users
Ó Ruanaidh/Fitzgerald: Numerical Bayesian Methods Applied to Signal Processing
Pannatier: VARIOWIN: Software for Spatial Data Analysis in 2D
Pinheiro/Bates: Mixed-Effects Models in S and S-PLUS
Unwin/Theus/Hofmann: Graphics of Large Datasets: Visualizing a Million
Venables/Ripley: Modern Applied Statistics with S, 4th ed.
Venables/Ripley: S Programming
Wilkinson: The Grammar of Graphics, 2nd ed.

Robert A. Muenchen

R for SAS and SPSS Users

Springer

Robert A. Muenchen
University of Tennessee
Knoxville, TN, USA
muenchen.bob@gmail.com

ISBN: 978-0-387-09417-5 e-ISBN: 978-0-387-09418-2
DOI 10.1007/978-0-387-09418-2

Library of Congress Control Number: 2008931588

Printed on acid-free paper

springer.com

Preface

While SAS and SPSS have many things in common, R is *very* different. My goal in writing this book is to help you translate what you know about SAS or SPSS into a working knowledge of R as quickly and easily as possible. I point out how they differ using terminology with which you are familiar, and show you which add-on packages will provide results most like those from SAS or SPSS. I provide many example programs done in SAS, SPSS, and R so that you can see how they compare topic by topic.

When finished, you should be able to use R to:

- Read data from various types of text files and SAS/SPSS datasets.
- Manage your data through transformations or recodes, as well as splitting, merging and restructuring data sets.
- Create publication quality graphs including bar, histogram, pie, line, scatter, regression, box, error bar, and interaction plots.
- Perform the basic types of analyses to measure strength of association and group differences, and be able to know where to turn to cover much more complex methods.

Who Is This Book For?

This book is, perhaps obviously, for people who already know a statistics package. While aimed at SAS and SPSS users, many statistics packages share their main attributes, especially the use of a rectangular dataset as their only data structure. I expect users of most other statistics packages could benefit from this book. An audience I did not expect to serve is R users wanting to learn SAS or SPSS. I have heard from quite a few of them who have said that by explaining the differences, it helped them learn in the reverse order I had anticipated. However, I explain none of the SAS or SPSS programs, only the R ones, and how the packages differ, so it is not ideal for that purpose.

Who Is This Book Not For?

I make no effort to teach statistics or graphics. Although I briefly state the goal and assumptions of each analysis, I do not cover their formulas or derivations. We have more than enough to discuss without tackling those topics too. This is also not a book about writing R functions; it is about using the thousands that already exist. We will write only a few very short functions. If you want to learn more about writing functions, I recommend John Chamber's *Software for Data Analysis: Programming with R* [1]. However, if you know SAS or SPSS, reading this book should be a nice leisurely step to take before diving into a book like that.

Practice Datasets and Programs

All the programs, datasets and files that we use in this book are available for download at http://RforSASandSPSSusers.com. Also check that site for updates and/or corrections to this book.

Acknowledgments

I am very grateful for the many people who have helped make this book possible, including the developers of the S language upon which R is based, Rick Becker, John Chambers, and Allan Wilks; the people who started R itself, Ross Ihaka and Robert Gentleman; the many other R developers for providing such wonderful tools for free and all the r-help participants who have kindly answered so many questions. Virtually all the examples I present here are modestly tweaked versions of countless posts to the r-help discussion list, as well as a few SAS-L and SPSSX-L posts. All I add is the selection, organization, explanation, and comparison to similar SAS and SPSS programs.

I am especially grateful to the people who provided advice, caught typos, and suggested improvements including: Patrick Burns, Peter Flom, Martin Gregory, Ralph O'Brien, Charilaos Skiadas, Phil Spector, Michael Wexler, and seven anonymous reviewers who provided pages of invaluable advice to a neophyte.

A special thanks goes to Hadley Wickham, who provided much guidance on his `ggplot2` graphics package. Thanks to Gabor Grothendieck, Lauri Nikkinen, and Marc Schwarz and for the r-help discussion that led to Sect. 14.15. Thanks to Gabor Grothendieck also for a detailed discussion that lead to Sect. 14.4. Thanks to Patrick Burns for his assistance with the glossary of terms in Appendix A.

My thanks also go to these people who helped compile *Appendix B: A Comparison of SAS and SPSS Products with R Packages and Functions* including: Thomas E. Adams, Jonathan Baron, Roger Bivand, Jason Burke, Patrick Burns, David L. Cassell, Chao Gai, Dennis Fisher, Bob Green, Frank E. Harrell Jr., Max Kuhn, Paul Murrell, Charilaos Skiadas, Antony Unwin, and Tobias Verbeke. Thanks to Henrique Dallazuanna for the code to count packages presented in that appendix.

I also thank SPSS Inc., especially Jon Peck for his helpful review of this book and his SPSS expertise which appears in the programs for extracting the first/last observation per group and generating data.

Finally, I am grateful to my wife, Carla Foust and sons Alexander and Conor, who put up with many lost weekends as I wrote this book.

About the Author

Robert A. Muenchen is a consulting statistician with 28 years of experience. He is currently the manager of the Statistical Consulting Center at the University of Tennessee. He holds a B.A. in Psychology and an M.S. in Statistics. Bob has conducted research for a variety of public and private organizations and has assisted on more than 1,000 graduate theses and dissertations. He has coauthored over 40 articles published in scientific journals and conference proceedings.

Bob has served on the advisory boards of SPSS Inc., the Statistical Graphics Corporation and PC Week Magazine. His suggested improvements have been incorporated into SAS, SPSS, JMP, STATGRAPHICS and several R packages.

His research interests include statistical computing, data graphics and visualization, text analysis, data mining, psychometrics and resampling.

Contents