Data Analysis and Graphics Using R, Second Edition

Join the revolution ignited by the ground-breaking R system! Starting with an introduction to R, covering standard regression methods, then presenting more advanced topics, this book guides users through the practical and powerful tools that the R system provides. The emphasis is on hands-on analysis, graphical display and interpretation of data. The many worked examples, taken from real-world research, are accompanied by commentary on what is done and why. A website provides computer code and data sets, allowing readers to reproduce all analyses. Updates and solutions to selected exercises are also available. Assuming basic statistical knowledge and some experience of data analysis, the book is ideal for research scientists, final-year undergraduate or graduate level students of applied statistics, and practicing statisticians. It is both for learning and for reference.

This second edition reflects changes in R since 2003. There is new material on survival analysis, random coefficient models and the handling of high-dimensional data. The treatment of regression methods has been extended, including a brief discussion of errors in predictor variables. Both text and code have been revised throughout, and where possible simplified. New graphs have been added.

JOHN MAINDONALD is Visiting Fellow at the Centre for Mathematics and its Applications, Australian National University. He has collaborated extensively with scientists in a wide range of application areas, from medicine and public health to population genetics, machine learning, economic history and forensic linguistics.

JOHN BRAUN is Associate Professor of Statistical and Actuarial Sciences, University of Western Ontario. He has collaborated with biostatisticians, biologists, psychologists and most recently has become involved with a network of forestry researchers.

Data Analysis and Graphics Using R – an Example-Based Approach

Second Edition

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

Editorial Board

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

Already published

1. *Bootstrap Methods and Their Application*, by A. C. Davison and D. V. Hinkley
2. *Markov Chains*, by J. Norris
3. *Asymptotic Statistics*, by A. W. van der Vaart
4. *Wavelet Methods for Time Series Analysis*, by Donald B. Percival and Andrew T. Walden
5. *Bayesian Methods*, by Thomas Leonard and John S. J. Hsu
6. *Empirical Processes in M-Estimation*, by Sara van de Geer
7. *Numerical Methods of Statistics*, by John F. Monahan
8. *A User's Guide to Measure Theoretic Probability*, by David Pollard
9. *The Estimation and Tracking of Frequency*, by B. G. Quinn and E. J. Hannan
10. *Data Analysis and Graphics using R*, by John Maindonald and W. John Braun
11. *Statistical Models*, by A. C. Davison
12. *Semiparametric Regression*, by D. Ruppert, M. P. Wand, R. J. Carroll
13. *Exercises in Probability*, by Loic Chaumont and Marc Yor
14. *Statistical Analysis of Stochastic Processes in Time*, by J. K. Lindsey
15. *Measure Theory and Filtering*, by Lakhdar Aggoun and Robert Elliott
16. *Essentials of Statistical Inference*, by G. A. Young and R. L. Smith
17. *Elements of Distribution Theory*, by Thomas A. Severini
18. *Statistical Mechanics of Disordered Systems*, by Anton Bovier
19. *The Coordinate-Free Approach to Linear Models*, by Michael J. Wichura
20. *Random Graph Dynamics*, by Rick Durrett

Data Analysis and Graphics Using R – an Example-Based Approach

Second Edition

John Maindonald
Centre for Mathematics and its Applications, Australian National University

and

W. John Braun
Department of Statistical and Actuarial Science, University of Western Ontario

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521861168

First published 2003
Second edition 2007

Printed in the United Kingdom at the University Press, Cambridge

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

ISBN-13 978-0-521-86116-8 hardback
ISBN-10 0-521-86116-0 hardback

It is easy to lie with statistics. It is hard to tell the truth without statistics.

[Andrejs Dunkels]

... technology tends to overwhelm common sense.

[D. A. Freedman]

For Amelia and Luke
also Shireen, Peter, Lorraine, Evan and Winifred

For Susan, Matthew and Phillip

Contents

Preface

This book is an exposition of statistical methodology that focuses on ideas and concepts, and makes extensive use of graphical presentation. It avoids, as much as possible, the use of mathematical symbolism. It is particularly aimed at scientists who wish to do statistical analyses on their own data, preferably with reference as necessary to professional statistical advice. It is intended to complement more mathematically oriented accounts of statistical methodology. It may be used to give students with a more specialist statistical interest exposure to practical data analysis.

While no prior knowledge of specific statistical methods or theory is assumed, there is a demand that readers bring with them, or quickly acquire, some modest level of statistical sophistication. Readers should have some prior exposure to statistical methodology, some prior experience of working with real data, and be comfortable with the typing of analysis commands into the computer console. Some prior familiarity with regression and with analysis of variance will be helpful.

We cover a range of topics that are important for many different areas of statistical application. As is inevitable in a book that has this broad focus, there will be investigators working in specific areas – perhaps epidemiology, or psychology, or sociology, or ecology – who will regret the omission of some methodologies that they find important.

We comment extensively on analysis results, noting inferences that seem well-founded, and noting limitations on inferences that can be drawn. We emphasize the use of graphs for gaining insight into data – in advance of any formal analysis, for understanding the analysis, and for presenting analysis results.

The data sets that we use as a vehicle for demonstrating statistical methodology have been generated by researchers in many different fields, and have in many cases featured in published papers. As far as possible, our account of statistical methodology comes from the coalface, where the quirks of real data must be faced and addressed. Features that may challenge the novice data analyst have been retained. The diversity of examples has benefits, even for those whose interest is in a specific application area. Ideas and applications that are useful in one area often find use elsewhere, even to the extent of stimulating new lines of investigation. We hope that our book will stimulate such cross-fertilization.

To summarize: the strengths of this book include the directness of its encounter with research data, its advice on practical data analysis issues, the inclusion of code that reproduces analyses, careful critiques of analysis results, attention to graphical and other

presentation issues, and the use of examples drawn from across the range of statistical applications.

John Braun wrote the initial drafts of Subsections 4.7.3, 4.7.4, 5.5.3, 6.8.5, 8.4.1 and Section 9.3. Initial drafts of remaining material were, mostly, from John Maindonald's hand. A substantial part was derived, intially, from the lecture notes of courses for researchers, at the University of Newcastle (Australia) over 1996–1997 and at The Australian National University over 1998–2001. Both of us have worked extensively over the material in these chapters. John Braun has taken primary responsibility for maintenance of the *DAAG* package.

The R system

We use the R system for the computations. The R system implements a dialect of the influential S language, developed at AT&T Bell Laboratories by Rick Becker, John Chambers and Allan Wilks, which is the basis for the commercial S-PLUS system. It follows S in its close linkage between data analysis and graphics. Versions of R are available, at no charge, for 32-bit versions of Microsoft Windows, for Linux and other Unix systems, and for the Macintosh. It is available through the Comprehensive R Archive Network (CRAN). Go to `http://cran.r-project.org/`, and find the nearest mirror site.

The development model used for R has proved highly effective in marshalling high levels of computing expertise for continuing improvement, for identifying and fixing bugs, and for responding quickly to the evolving needs and interests of the statistical community. Oversight of "base R" is handled by the R Core Team, whose members are widely drawn internationally. Use is made of code, bug fixes and documentation from the wider R user community. Especially important are the large number of packages that supplement base R, and that anyone is free to contribute. Once installed, these attach seamlessly into the base system.

Many of the analyses offered by R's packages were not, 10 years ago, available in any of the standard statistical packages. What did data analysts do before we had such packages? Basically, they adapted more simplistic (but not necessarily simpler) analyses as best they could. Those whose skills were unequal to the task did unsatisfactory analyses. Those with more adequate skills carried out analyses that, even if not elegant and insightful by current standards, were often adequate. Tools such as are available in R have reduced the need for the adaptations that were formerly necessary. We can often do analyses that better reflect the underlying science. There have been challenging and exciting changes from the methodology that was typically encountered in statistics courses 10 or 15 years ago.

In the ongoing development of R, priorities have been: the provision of good data manipulation abilities; flexible and high-quality graphics; the provision of data analysis methods that are both insightful and adequate for the whole range of application area demands; seamless integration of the different components of R; and the provision of interfaces to other systems (editors, databases, the web, etc.) that R users may require.

Ease of use is important, but not at the expense of power, flexibility and checks against answers that are potentially misleading.

Depending on the user's level of skill with R, there will be some relatively routine tasks where another system may seem simpler to use. Note however the availability of interfaces, notably John Fox's *Rcmdr*, that give a graphical user interface (GUI) to a limited part of R. Such interfaces will develop and improve as time progresses. They may in due course, for many users, be the preferred means of access to R. Be aware that the demand for simple tools will commonly place limitations on the tasks that can, without professional assistance, be satisfactorily undertaken.

Primarily, R is designed for scientific computing and for graphics. Among the packages that have been added are many that are not obviously statistical – for drawing and coloring maps, for map projections, for plotting data collected by balloon-born weather instruments, for creating color palettes, for working with bitmap images, for solving sudoko puzzles, for creating magic squares, for reading and handling shapefiles, for solving ordinary differential equations, for processing various types of genomic data, and so on. Check through the list of R packages that can be found on any of the CRAN sites, and you may be surprised at what you find!

The citation for John Chambers' 1998 Association for Computing Machinery Software award stated that S has "forever altered how people analyze, visualize and manipulate data." The R project enlarges on the ideas and insights that generated the S language. We are grateful to the R Core Team, and to the creators of the various R packages, for bringing into being the R system – this marvellous tool for scientific and statistical computing, and for graphical presentation. We list at the end of the reference section the authors and compilers of packages that have been used in this book.

Influences on the modern practice of statistics

The development of statistics has been motivated by the demands of scientists for a methodology that will extract patterns from their data. The methodology has developed in a synergy with the relevant supporting mathematical theory and, more recently, with computing. This has led to methodologies and supporting theory that are a radical departure from the methodologies of the pre-computer era.

Statistics is a young discipline. Only in the 1920s and 1930s did the modern framework of statistical theory, including ideas of hypothesis testing and estimation, begin to take shape. Different areas of statistical application have taken these ideas up in different ways, some of them starting their own separate streams of statistical tradition. Gigerenzer *et al*. (1989, "The Empire of Statistics") examine the history, commenting on the different streams of development that have influenced practice in different research areas.

Separation from the statistical mainstream, and an emphasis on "black box" approaches, have contributed to a widespread exaggerated emphasis on tests of hypotheses, to a neglect of pattern, to the policy of some journal editors of publishing only those studies that show a statistically significant effect, and to an undue focus on the individual study. Anyone

who joins the R community can expect to witness, and/or engage in, lively debate that addresses these and related issues. Such debate can help ensure that the demands of scientific rationality do in due course win out over influences from accidents of historical development.

New tools for effective data analysis

We have drawn attention to advances in statistical computing methodology. These have led to new powerful tools for exploratory analysis of regression data, for choosing between alternative models, for diagnostic checks, for handling non-linearity, for assessing the predictive power of models, and for graphical presentation. In addition, we have new computing tools that make it straightforward to move data between different systems, to keep a record of calculations, to retrace or adapt earlier calculations, and to edit output and graphics into a form that can be incorporated into published documents.

The best any analysis can do is to highlight the information in the data. No amount of statistical or computing technology can be a substitute for good design of data collection, for understanding the context in which data are to be interpreted, or for skill in the use of statistical analysis methodology. Statistical software systems are one of several components of effective data analysis.

The questions that statistical analysis is designed to answer can often be stated simply. This may encourage the layperson to believe that the answers are similarly simple. Often, they are not. Be prepared for unexpected subtleties. Effective statistical analysis requires appropriate skills, beyond those gained from taking one or two undergraduate courses in statistics. There is no good substitute for professional training in modern tools for data analysis, and experience in using those tools with a wide range of data sets. No-one should be embarrassed that they have difficulty with analyses that involve ideas that professional statisticians may take 7 or 8 years of professional training and experience to master.

Changes in this second edition

This new edition takes account of changes in R since 2003. There is new material on survival analysis, random coefficient models and the handling of high-dimensional data. The treatment of regression methods has been extended, including in particular a brief discussion of errors in predictor variables. Both the text and R code have been extensively revised. Code has, wherever possible, been simplified. Some examples have been reworked. There are changes to some graphs, and new graphs have been added.

Acknowledgments

Many different people have helped us with this project. Winfried Theis (University of Dortmund, Germany) and Detlef Steuer (University of the Federal Armed Forces, Hamburg, Germany) helped with technical aspects of working with LaTeX, with setting up a cvs server to manage the LaTeX files, and with helpful comments. Lynne Billard

(University of Georgia, USA), Murray Jorgensen (University of Waikato, NZ) and Berwin Turlach (University of Western Australia) gave valuable help in the identification of errors and text that required clarification. Susan Wilson (Australian National University) gave welcome encouragement. Duncan Murdoch (University of Western Ontario) helped set up the *DAAG* package, and has supplied valuable technical advice. Thanks also to Cath Lawrence (Australian National University) for her Python program that allowed us to extract the R code, as and when required, from our LaTeX files; this has now at length become an R function. Many of the tables in this book were generated, in first draft form, using the `xtable()` function from the *xtable* package for R.

For this second edition, Brian Ripley (University of Oxford) has gone through the manuscript and made extensive comments, leading to important corrections and improvements. We are most grateful to him, and to others who have commented on the manuscript. Alan Welsh (Australian National University) has been helpful in working through points where it has seemed difficult to get the emphasis right. Once again, Duncan Murdoch has given much useful technical advice. Others who have made helpful comments and/or pointed out errors include Jeff Wood (Australian National University), Nader Tajvidi (University of Lund), Paul Murrell (University of Auckland, on Section 14.11), Graham Williams (`http://www.togaware.com`, on Chapter 1) and Yang Yang (University of Western Ontario, on Chapter 10). The failings that remain are, naturally, our responsibility.

A strength of this book is the extent to which it has drawn on data from many different sources. We give a list, following the list of references for the data near the end of the book, of individuals and/or organizations to whom we are grateful for allowing use of data. We are grateful to those who have allowed us to use their data. At least these data will not, as often happens once data have become the basis for a published paper, gather dust in a long-forgotten folder! We are grateful, also, to the many researchers who, in their discussions with us, have helped stimulate our thinking and understanding. We apologize if there is anyone that we have inadvertently failed to acknowledge.

Diana Gillooly of Cambridge University Press, taking over from David Tranah for this new edition, has been a marvellous source of advice and encouragement throughout the revision process.

Conventions

Text that is R code, or output from R, is printed in a verbatim text style. For example, in Chapter 1 we will enter data into an R object that we call `austpop`. We will use the `plot()` function to plot these data. The names of R packages, including our own *DAAG* package, are printed in italics.

Starred exercises and sections identify more technical items that can be skipped at a first reading.

Solutions to exercises

Solutions to selected exercises, R scripts that have all the code from the book and other supplementary materials are available via the link given at `http://www.maths.anu.edu.au/~johnm/r-book`