A PRACTICAL GUIDE TO DATA ANALYSIS USING R

Using diverse real-world examples, this text examines what models used for data analysis mean in a specific research context. What assumptions underlie analyses, and how can you check them?

Building on the successful *Data Analysis and Graphics Using R*, third edition (Cambridge, 2010), it expands upon topics including cluster analysis, exponential time series, matching, seasonality, and resampling approaches. An extended look at *p*-values leads to an exploration of replicability issues and of contexts where numerous *p*-values exist, including gene expression.

Developing practical intuition, this book assists scientists in the analysis of their own data, and familiarizes students in statistical theory with practical data analysis. The worked examples and accompanying commentary teach readers to recognize when a method works and, more importantly, when it doesn't. Each chapter contains copious exercises. Selected solutions, notes, slides, and R code are available online, with extensive references pointing to detailed guides to R.

JOHN H. MAINDONALD is Contract Associate at Statistics Research Associates and was previously Visiting Fellow at the Australian National University. He has had wide experience both as a university lecturer and as a quantitative problem solver, working with researchers in diverse areas. He is the author of *Statistical Computation* (1984), and the senior author of *Data Analysis and Graphics Using R* (third edition, 2010).

W. JOHN BRAUN is Professor at the University of British Columbia, where he is Director of the UBCO campus of the Banff International Research Station for Mathematical Innovation and Discovery. In 2020, he received the Statistical Society of Canada Award for Impact of Applied and Collaborative Work.

JEFFREY L. ANDREWS is Associate Professor at the University of British Columbia. He currently serves as Principal Co-director of the Master of Data Science program and President-elect of The Classification Society (TCS). He is the 2013 Distinguished Dissertation Award winner from TCS and a recipient of the 2017 Chikio Hayashi Award for Young Researchers from the International Federation of Classification Societies.

A PRACTICAL GUIDE TO DATA ANALYSIS USING R

An Example-Based Approach

JOHN H. MAINDONALD
Statistics Research Associates, Wellington, New Zealand

W. JOHN BRAUN
University of British Columbia, Okanagan

JEFFREY L. ANDREWS
University of British Columbia, Okanagan

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, 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

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009282277

DOI: 10.1017/9781009282284

First published 2024

Printed in the United Kingdom by CPI Group Ltd, Croydon CR0 4YY

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

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-009-28227-7 Hardback

Cambridge University Press & Assessment has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

For my grandchildren Luke, Amelia, and Ted

For my children, Matthew, Phillip, and Reese

For my family (Irene, Charlie, and Mia) and my parents (Dave and Marleen)

Contents

Figures

Preface

This text is designed as an aid, for learning and for reference, in the navigation of a world in which unprecedented new data sources, and tools for data analysis, are pervasive. It aims to teach, using real-world examples, a style of analysis and critique that, given meaningful data, can generate defensible analysis results. Its focus is on ideas and concepts, with extensive use of graphical presentation. It may be used to give students who have taken courses in statistical theory exposure to practical data analysis. It is designed, also, as a resource for scientists who wish to do statistical analyses on their own data, preferably with reference as necessary to professional statistical advice. It emphasizes the role of statistical design and analysis as part of the wider scientific process.

As far as possible, our account of statistical methodology comes from the coalface, where the quirks of real data must be faced and addressed. Experience in consulting with researchers in many different areas of application, in supervising research students, and in lectures to researchers, have been strong influences in the text's style and content. 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 project has been a tremendous learning experience for all three of us. As is usual, the more we learn, the more we appreciate how much more we have to learn.

The text is suitable for a style of learning where readers work through the text with a computer at their side, running the R code as and when this seems helpful. It complements more mathematically oriented accounts of statistical methodology. The appendix provides a brief account of R, primarily as a starting point for learning. We encourage readers with limited R experience to avail themselves of the wealth of instructional material on the web as well as the hardcopy resources listed in Section 1.11.

While no prior knowledge of specific statistical methods or theory is assumed, readers will need to bring with them, or quickly acquire, a modest level of statistical sophistication. Prior experience with real data, prior exposure to statistical methodology, and some prior familiarity with regression methods, will all be helpful.

Important technical terms will include *random sample*, *independence*, *dependence*, *standard deviation*, and *normal distribution*, with limited attention to formal definition. Our primary concern is with the role and meaning of this language in practical data analysis. While there will be references to theoretical results, it is not our purpose to provide a systematic account of statistical theory.[1] We make only limited use of mathematical symbolism.

Statistical analysis relies heavily on mathematical models. An understanding of the mathematics underlying a model is important only to the extent that it helps in understanding, and where possible in checking, what the model means in the context from which the data came. Is it reasonable to assume that observations are independent? What are the influences, perhaps the time sequence in which the data were collected, that might place this assumption in question? This is just one of the issues, but a very important one, that data analysts need to consider. Comments made by John W. Tukey emphasize the importance, in statistical training and practice, of wrestling with what the models used mean in the context of data that has been presented for analysis:

> ... Statistics is a science ... and it is no more a branch of mathematics than are physics, chemistry and economics; for if its methods fail the test of experience – not the test of logic – they are discarded.
> [Tukey (1953), quoted by Brillinger (2002)]

The methods that we cover have wide application. The datasets, many of which have featured in published papers, are drawn from many different fields. They reflect a journey in learning and understanding, alike for the authors and for those with whom they have worked, that has ranged widely over many different research areas. We hope that our text will stimulate the cross-fertilization that occurs when ideas and applications that have proved effective in one area find use elsewhere, perhaps even leading to new lines of investigation.

To summarize: The strengths of this book include the directness of its encounter with research data, its advice on practical data analysis issues, careful critiques of analysis results, the use of modern data analysis tools and approaches, the use of simulation and other computer-intensive methods where these provide insight or give results that are not otherwise available, attention to graphical and other presentation issues, the use of examples drawn from across the range of statistical applications, the links that it makes into the debate over reproducibility in science, and the inclusion of code that reproduces analyses.

A substantial part of the first edition of *Data Analysis and Graphics Using R* (Maindonald and Braun, 2003) was derived, initially, from the lecture notes of courses for researchers that the first author presented, at the University of Newcastle (Australia) over 1996–1997 and at Australian National University from 1998, through until formal retirement and beyond. It was a privilege to have contacts, arising from consulting work and lectures, across the University. Those contacts were extended as a result of short courses on R-based analysis that were offered,

[1] For an overview of the theory of statistical inference, see, for example, Cox (2006).

across a wide variety of Australian government and academic institutions, between 2003 and 2014.

Influences on the Modern Practice of Statistics

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. As documented in Gigerenzer et al. (1989, *The Empire of Chance*), differences in historical development have led to some differences in practice between research areas.

Statistical methods have found wide use, but they have also been widely misused. There has been a widespread reliance on "black box" approaches, used without due consideration of the reasonableness of assumptions made, or attention to diagnostic checks, or attention to the processes that generated the data. In experimental work, the use of p-values and other statistics has too often become a substitute for the checks that independent replication provides on the total experimental process. There has been a renewed attention, both in the wider scientific community and in the statistical community, to the interplay between scientific methodology and statistical design and analysis. Critical reexamination of current scientific processes, and of the role of statistical analysis within those processes, can help ensure that the demands of scientific rationality do in due course win out over accidents of historical development and all-too-human failures to maintain critical standards.

New Data Analysis Tools

The methodology has developed in a synergy with the relevant supporting mathematical theory and, more recently, with computing. This has led to major advances on the methodologies of the precomputer era. "Data Science," or perhaps "Statistical Science," is a good name for the mix of tools and skills required for effective data analysis. Data analysts now have at their disposal vastly new powerful tools than were available even 20 years ago, for exploratory analysis of regression data, for choosing between alternative models, for diagnostic checks, for handling nonlinearity, for assessing the predictive power of models, and for graphical presentation. New computing tools 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. Machine learning and related methodologies emphasize new types of data, new data analysis demands, new data analysis tools, and datasets that may be of unprecedented size. Textual data and image data offer interesting new challenges.

The traditional concerns of professional data analysts remain as important as ever. Irrespective of the size of dataset, questions of data quality, of relevance to the issues that are under investigation, and of the way that the data have been sampled, remain as important as ever. Implicit or explicit claims that results generalize to a relevant wider target population must be justified.

Students in first or second year university courses, in such areas as geography or biology or politics or psychology or business studies, are increasingly likely to encounter R. It is finding its way into the upper levels of secondary schools. While this is to be encouraged, students do need to understand that such courses are at the start of an adventure in statistical understanding. 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 datasets. No one should be embarrassed that they have difficulty with analyses that involve ideas that professional statisticians may take seven or eight years of training and experience to master.

The questions that data analysis is designed to answer can often be stated simply. This may encourage the layperson, or even scientists doing their own analyses, to believe that the answers are similarly simple. Commonly, they are not. Be prepared for unexpected subtleties. Comments made by Stephen Senn are apt:

> I've been studying statistics for over 40 years and still don't understand it. The ease with which non-statisticians master it is staggering.

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 using available analysis tools. The best any analysis can do is to highlight the information in the data.

The R System

Work on R started in the early 1990s, as a project of Ross Ihaka and Robert Gentleman, when both were at the University of Auckland (New Zealand). The R system implements a dialect of the S language, developed at AT&T by John Chambers and colleagues. Section 1.4 in Chambers (2008) describes the history. Versions of R are available, at no charge, for Microsoft Windows, for Linux and other Unix systems, and for Macintosh systems. It is available through the Comprehensive R Archive Network (CRAN). Go to http://cran.r-project.org/, and find the nearest mirror site. A huge range of packages, contributed by specialists in many different areas, supplement base R. The development model has proved effective in marshaling 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. The R Task Views web page[2] lists packages that handle some of the more common R applications. It has become an increasing challenge to keep pace with the new and/or improved abilities that R packages, new and old, continue to develop. Those who rely heavily on R for their day-to-day work will do well to keep attuned to major changes and developments.

The R system has brought into a common framework a huge range of abilities that extend beyond the data analysis and associated data manipulation and graphics abilities that are the focus of this text. Examples include drawing and coloring maps, reading and handling shapefiles, map projections, plotting data collected by balloon-borne weather instruments, creating color palettes, manipulating bitmap images, solving sudoku puzzles, creating magic squares, solving ordinary differential equations, and processing various types of genomic data. Help files and

[2] https://cran.r-project.org/web/views/.

vignettes that are included with packages are a large reservoir of information on the methodologies that they implement.

There are several graphical user interfaces (GUIs) that can be highly helpful in accessing a restricted range of R abilities – examples are *BlueSky*, *Rcmdr*, *R-Instat*, *jamovi*, and *rattle*. Access to the fill range of abilities that R and R packages make available will require use of the command line.

RStudio is a widely used R interactive development environment (IDE) for tasks that include viewing history, debugging, managing the workspace, package management, and data input and output. It has features that greatly assist project management and package development.

Among systems that have the potential to challenge R's dominance for data analysis, Julia (`julialang.org/`) seems particularly interesting. Relative to R, it has high computational efficiency. It has the potential to develop or adapt a range of packages that together match what R packages offer.

Changes and Additions from Data Analysis and Graphics Using R

Chapters 1–5 of *Data Analysis and Graphics Using R*, third edition (Maindonald and Braun, 2010) have been amalgamated and condensed somewhat into Chapters 1–3 of the present book. Here, the focus has moved, from including extensive R tutorial content in the text, to pointing users to the extensive R help resources now available both on the web and in printed form. Supplementary content available online includes R Markdown scripts, one for each chapter, that can be processed to reproduce all computer output, including tables and graphs. This content is available at `https://jhmaindonald.github.io/PGRcode`.

Concerns about reproducibility (or, in the terminology we prefer, "replicability"), especially in wet laboratory biology and in psychology, have attracted extensive attention in the pages of *Nature*, *Science*, *The Economist*, psychology journals, and elsewhere. The uses and limitations of p-values have been an important part of the discussion. Chapter 1 now has a much extended discussion of their use and role, leading on to the wider discussion of replicability issues. Information statistics (AIC, AICc, and BIC) get more detailed attention.

The treatment of p-values extends to noting the new possibilities that arise when there are, potentially, hundreds, or thousands, or more, p-values. The false discovery rate estimates that are then available are more informative, and relate more directly to the questions that are commonly of experimental interest, than p-values. The new Section 9.5 takes up these ideas as they apply to the analysis of RNA-Seq gene expression data.

Other topics that get new or increased attention include: the modeling of extra-binomial or extra-Poisson variation; exponential time series, including their use in forecasting; seasonality; spline smooths with time series error terms; fitting monotonic increasing or decreasing response curves; and quantile regression automatic choice of smoothing parameter.

Changes in the lme4 package for fitting mixed-effects models, and the implementation of the Kenward–Roger approach that is now available in the *afex* package, have required substantial rewrites. In Chapter 7, there is a new section on "A Mixed Model with a Betabinomial Error." The treatment of Principal Component Analysis and of multi-dimensional scaling is now followed by a new section on hierarchical and other forms of clustering.

The treatment of causal inference from observational data has been greatly extended to discuss the role of matching. There is some limited attention to the use of multiple imputation to fill in missing values in data where some observations are incomplete.

Source Files That Combine Text and R Code

Drafts of this text were created from *Sweave* source files that combine marked up code and text into one document, in a form that could then be processed using Yihui Xie's *knitr* package to give the LaTeX files and associated R output and figures from which this text was generated. Rerunning and checking of code is a built-in part of the process, making the revising and updating of text and code easier and less error prone. The R Markdown plain text format, designed to be easier for novices to learn and master, can can be processed using *knitr* abilities in a very similar way. R Markdown is widely used for creating online content, for papers and books, and for the vignettes that many R packages use to supplement help pages. See `https://rmarkdown.rstudio.com/`.

Acknowledgements

The prefaces to the three editions of *Data Analysis and Graphics Using R* give names of those who provided helpful comment. For this new text, James Cone has provided useful comments. Trish Scott has helped with copyediting. Discussions on the R-help and R-devel email lists have contributed greatly to insight and understanding. The failings that remain are, naturally, our responsibility.

This text has drawn on data from many different sources. Following the references is a list of data sources (individuals and/or organizations) that we wish to thank and acknowledge. Thanks are due also to the many researchers whose discussions with us have helped stimulate thinking and understanding, and who in many instances have given us access to their data. We apologize to anyone that we may have inadvertently failed to acknowledge.

Too often, data that have become the basis for a published paper are not made available in any form of public record. The data may not find their way into any permanent record, and cease to be available for checking the analysis, for work that builds on what can be learned when data from multiple sources are brought together, to try a new form of analysis, or for use in teaching. In areas where data are as a matter of course kept available for future researchers to use, this has been a major contributor to advances in scientific understanding. Those benefits can and should extend more widely. Thanks are due to Beverley Lawrence for her efforts as copy-editor, and to Cambridge University Press staff who assisted us through the copy-editing and publication process – Roger Astley, Natalie Tomlinson, Anna Scriven, and Clare Dennison.

Notes for Readers

For many readers, a largely “learn as one goes” approach to mastering what they need to know of R will work well. For this, they can look for the mix of sources of tutorial content that works best for them – online tutorial content such as is noted in Section 1.11, books and other printed material, results from web searches, and such guidance as is provided in Appendix A. We encourage readers who are new to R to skim over the content of Appendix A before or as they work through the first chapter.

A complete set of R code, together with other supplementary material, is available from `https://jhmaindonald.github.io/PGRcode`.

Graphs and Graphics Packages

In Chapter 1, simplified code is given for figures that do not involve relatively complicated code. In later chapters, code is given only for those figures that are specifically targeted at the methodology under discussion.

The main graphics packages that will be used are the base *graphics* package, *lattice* and *latticeExtra*, and *ggplot2*. The `plot()` and related functions in base graphics directly generate a plot. With *lattice* and *ggplot2* functions, an alternative to directly creating a plot is to save the output as a graphics object that can be further updated and/or modified before use to create a plot.

Accessing Data and Functions from Packages

A number of packages are automatically loaded, with their functions and datasets then available, at the start of a new R session. For functions and datasets in packages that are not already available, there is a choice between using `library()` or an equivalent to make all datasets and functions from the package available, or using code such as `lattice::xyplot()` (execute the lattice function from the *lattice* package) or `DAAG::cuckoos` (the `cuckoos` dataset from the *DAAG* package) whenever such a function or dataset is required.

Conventions

Starred headings identify more technical discussions that can be skipped at a first reading. Item numbers for more technical and/or challenging exercises are likewise starred.

Comments, prefaced by `#` or for extra emphasis by `##`, will often be included in code chunks. Where code is included in comments, it will be surrounded by back quotes, as in `` `species ~ length` `` in the final line of code that now follows:

```
## Code for a stripped down version of Figure 1.1A
library(latticeExtra)  # The 'lattice' package will be loaded & attached also
cuckoos <- DAAG::cuckoos
## Panel A: Dotplot without species means added
dotplot(species ~ length, data=cuckoos)  ## `species ~ length` is a 'formula'
```