Statistics and Computing

Series Editors:

J. Chambers

D. Hand

W. 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
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.

Peter Dalgaard

Introductory Statistics with R

Second Edition

Springer

Peter Dalgaard
Department of Biostatistics
University of Copenhagen
Denmark
p.dalgaard@biostat.ku.dk

ISSN: 1431-8784
ISBN: 978-0-387-79053-4 e-ISBN: 978-0-387-79054-1
DOI: 10.1007/978-0-387-79054-1

Library of Congress Control Number: 2008932040

Printed on acid-free paper

springer.com

To Grete, for putting up with me for so long

Preface

R is a statistical computer program made available through the Internet under the General Public License (GPL). That is, it is supplied with a license that allows you to use it freely, distribute it, or even sell it, as long as the receiver has the same rights and the source code is freely available. It exists for Microsoft Windows XP or later, for a variety of Unix and Linux platforms, and for Apple Macintosh OS X.

R provides an environment in which you can perform statistical analysis and produce graphics. It is actually a complete programming language, although that is only marginally described in this book. Here we content ourselves with learning the elementary concepts and seeing a number of cookbook examples.

R is designed in such a way that it is always possible to do further computations on the results of a statistical procedure. Furthermore, the design for graphical presentation of data allows both no-nonsense methods, for example `plot(x,y)`, and the possibility of fine-grained control of the output's appearance. The fact that R is based on a formal computer language gives it tremendous flexibility. Other systems present simpler interfaces in terms of menus and forms, but often the apparent user-friendliness turns into a hindrance in the longer run. Although elementary statistics is often presented as a collection of fixed procedures, analysis of moderately complex data requires ad hoc statistical model building, which makes the added flexibility of R highly desirable.

R owes its name to typical Internet humour. You may be familiar with the programming language C (whose name is a story in itself). Inspired by this, Becker and Chambers chose in the early 1980s to call their newly developed statistical programming language S. This language was further developed into the commercial product S-PLUS, which by the end of the decade was in widespread use among statisticians of all kinds. Ross Ihaka and Robert Gentleman from the University of Auckland, New Zealand, chose to write a reduced version of S for teaching purposes, and what was more natural than choosing the immediately preceding letter? Ross' and Robert's initials may also have played a role.

In 1995, Martin Maechler persuaded Ross and Robert to release the source code for R under the GPL. This coincided with the upsurge in Open Source software spurred by the Linux system. R soon turned out to fill a gap for people like me who intended to use Linux for statistical computing but had no statistical package available at the time. A mailing list was set up for the communication of bug reports and discussions of the development of R.

In August 1997, I was invited to join an extended international core team whose members collaborate via the Internet and that has controlled the development of R since then. The core team was subsequently expanded several times and currently includes 19 members. On February 29, 2000, version 1.0.0 was released. As of this writing, the current version is 2.6.2.

This book was originally based upon a set of notes developed for the course in Basic Statistics for Health Researchers at the Faculty of Health Sciences of the University of Copenhagen. The course had a primary target of students for the Ph.D. degree in medicine. However, the material has been substantially revised, and I hope that it will be useful for a larger audience, although some biostatistical bias remains, particularly in the choice of examples.

In later years, the course in Statistical Practice in Epidemiology, which has been held yearly in Tartu, Estonia, has been a major source of inspiration and experience in introducing young statisticians and epidemiologists to R.

This book is not a manual for R. The idea is to introduce a number of basic concepts and techniques that should allow the reader to get started with practical statistics.

In terms of the practical methods, the book covers a reasonable curriculum for first-year students of theoretical statistics as well as for engineering students. These groups will eventually need to go further and study more complex models as well as general techniques involving actual programming in the R language.

For fields where elementary statistics is taught mainly as a tool, the book goes somewhat further than what is commonly taught at the undergraduate level. Multiple regression methods or analysis of multifactorial experiments are rarely taught at that level but may quickly become essential for practical research. I have collected the simpler methods near the beginning to make the book readable also at the elementary level. However, in order to keep technical material together, Chapters 1 and 2 do include material that some readers will want to skip.

The book is thus intended to be useful for several groups, but I will not pretend that it can stand alone for any of them. I have included brief theoretical sections in connection with the various methods, but more than as teaching material, these should serve as reminders or perhaps as appetizers for readers who are new to the world of statistics.

Notes on the 2nd edition

The original first chapter was expanded and broken into two chapters, and a chapter on more advanced data handling tasks was inserted after the coverage of simpler statistical methods. There are also two new chapters on statistical methodology, covering Poisson regression and nonlinear curve fitting, and a few items have been added to the section on descriptive statistics. The original methodological chapters have been quite minimally revised, mainly to ensure that the text matches the actual output of the current version of R. The exercises have been revised, and solution sketches now appear in Appendix D.

Acknowledgements

Obviously, this book would not have been possible without the efforts of my friends and colleagues on the R Core Team, the authors of contributed packages, and many of the correspondents of the e-mail discussion lists.

I am deeply grateful for the support of my colleagues and co-teachers Lene Theil Skovgaard, Bendix Carstensen, Birthe Lykke Thomsen, Helle Rootzen, Claus Ekstrøm, Thomas Scheike, and from the Tartu course Krista Fischer, Esa Läära, Martyn Plummer, Mark Myatt, and Michael Hills, as well as the feedback from several students. In addition, several people, including Bill Venables, Brian Ripley, and David James, gave valuable advice on early drafts of the book.

Finally, profound thanks are due to the free software community at large. The R project would not have been possible without their effort. For the

typesetting of this book, TEX, LATEX, and the consolidating efforts of the LATEX2e project have been indispensable.

Peter Dalgaard
Copenhagen
April 2008

Contents