

# **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

© 2008 Springer Science+Business Media, LLC

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

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,  $\text{T}_\text{E}\text{X}$ ,  $\text{L}^\text{A}\text{T}_\text{E}\text{X}$ , and the consolidating efforts of the  $\text{L}^\text{A}\text{T}_\text{E}\text{X}2\text{e}$  project have been indispensable.

Peter Dalgaard  
Copenhagen  
April 2008

# Contents

<b>Preface</b>	<b>vii</b>
<b>1 Basics</b>	<b>1</b>
1.1 First steps . . . . .	1
1.1.1 An overgrown calculator . . . . .	3
1.1.2 Assignments . . . . .	3
1.1.3 Vectorized arithmetic . . . . .	4
1.1.4 Standard procedures . . . . .	6
1.1.5 Graphics . . . . .	7
1.2 R language essentials . . . . .	9
1.2.1 Expressions and objects . . . . .	9
1.2.2 Functions and arguments . . . . .	11
1.2.3 Vectors . . . . .	12
1.2.4 Quoting and escape sequences . . . . .	13
1.2.5 Missing values . . . . .	14
1.2.6 Functions that create vectors . . . . .	14
1.2.7 Matrices and arrays . . . . .	16
1.2.8 Factors . . . . .	18
1.2.9 Lists . . . . .	19
1.2.10 Data frames . . . . .	20
1.2.11 Indexing . . . . .	21
1.2.12 Conditional selection . . . . .	22
1.2.13 Indexing of data frames . . . . .	23
1.2.14 Grouped data and data frames . . . . .	25

1.2.15	Implicit loops . . . . .	26
1.2.16	Sorting . . . . .	27
1.3	Exercises . . . . .	28
<b>2</b>	<b>The R environment</b>	<b>31</b>
2.1	Session management . . . . .	31
2.1.1	The workspace . . . . .	31
2.1.2	Textual output . . . . .	32
2.1.3	Scripting . . . . .	33
2.1.4	Getting help . . . . .	34
2.1.5	Packages . . . . .	35
2.1.6	Built-in data . . . . .	35
2.1.7	<code>attach</code> and <code>detach</code> . . . . .	36
2.1.8	<code>subset</code> , <code>transform</code> , and <code>within</code> . . . . .	37
2.2	The graphics subsystem . . . . .	39
2.2.1	Plot layout . . . . .	39
2.2.2	Building a plot from pieces . . . . .	40
2.2.3	Using <code>par</code> . . . . .	42
2.2.4	Combining plots . . . . .	42
2.3	R programming . . . . .	44
2.3.1	Flow control . . . . .	44
2.3.2	Classes and generic functions . . . . .	46
2.4	Data entry . . . . .	46
2.4.1	Reading from a text file . . . . .	47
2.4.2	Further details on <code>read.table</code> . . . . .	50
2.4.3	The data editor . . . . .	51
2.4.4	Interfacing to other programs . . . . .	52
2.5	Exercises . . . . .	53
<b>3</b>	<b>Probability and distributions</b>	<b>55</b>
3.1	Random sampling . . . . .	55
3.2	Probability calculations and combinatorics . . . . .	56
3.3	Discrete distributions . . . . .	57
3.4	Continuous distributions . . . . .	58
3.5	The built-in distributions in R . . . . .	59
3.5.1	Densities . . . . .	59
3.5.2	Cumulative distribution functions . . . . .	62
3.5.3	Quantiles . . . . .	63
3.5.4	Random numbers . . . . .	64
3.6	Exercises . . . . .	65
<b>4</b>	<b>Descriptive statistics and graphics</b>	<b>67</b>
4.1	Summary statistics for a single group . . . . .	67
4.2	Graphical display of distributions . . . . .	71
4.2.1	Histograms . . . . .	71

4.2.2	Empirical cumulative distribution . . . . .	73
4.2.3	Q-Q plots . . . . .	74
4.2.4	Boxplots . . . . .	75
4.3	Summary statistics by groups . . . . .	75
4.4	Graphics for grouped data . . . . .	79
4.4.1	Histograms . . . . .	79
4.4.2	Parallel boxplots . . . . .	80
4.4.3	Stripcharts . . . . .	81
4.5	Tables . . . . .	83
4.5.1	Generating tables . . . . .	83
4.5.2	Marginal tables and relative frequency . . . . .	87
4.6	Graphical display of tables . . . . .	89
4.6.1	Barplots . . . . .	89
4.6.2	Dotcharts . . . . .	91
4.6.3	Piecharts . . . . .	92
4.7	Exercises . . . . .	93
<b>5</b>	<b>One- and two-sample tests</b>	<b>95</b>
5.1	One-sample $t$ test . . . . .	95
5.2	Wilcoxon signed-rank test . . . . .	99
5.3	Two-sample $t$ test . . . . .	100
5.4	Comparison of variances . . . . .	103
5.5	Two-sample Wilcoxon test . . . . .	103
5.6	The paired $t$ test . . . . .	104
5.7	The matched-pairs Wilcoxon test . . . . .	106
5.8	Exercises . . . . .	107
<b>6</b>	<b>Regression and correlation</b>	<b>109</b>
6.1	Simple linear regression . . . . .	109
6.2	Residuals and fitted values . . . . .	113
6.3	Prediction and confidence bands . . . . .	117
6.4	Correlation . . . . .	120
6.4.1	Pearson correlation . . . . .	121
6.4.2	Spearman's $\rho$ . . . . .	123
6.4.3	Kendall's $\tau$ . . . . .	124
6.5	Exercises . . . . .	124
<b>7</b>	<b>Analysis of variance and the Kruskal-Wallis test</b>	<b>127</b>
7.1	One-way analysis of variance . . . . .	127
7.1.1	Pairwise comparisons and multiple testing . . . . .	131
7.1.2	Relaxing the variance assumption . . . . .	133
7.1.3	Graphical presentation . . . . .	134
7.1.4	Bartlett's test . . . . .	136
7.2	Kruskal-Wallis test . . . . .	136
7.3	Two-way analysis of variance . . . . .	137

7.3.1	Graphics for repeated measurements . . . . .	140
7.4	The Friedman test . . . . .	141
7.5	The ANOVA table in regression analysis . . . . .	141
7.6	Exercises . . . . .	143
<b>8</b>	<b>Tabular data</b>	<b>145</b>
8.1	Single proportions . . . . .	145
8.2	Two independent proportions . . . . .	147
8.3	$k$ proportions, test for trend . . . . .	149
8.4	$r \times c$ tables . . . . .	151
8.5	Exercises . . . . .	153
<b>9</b>	<b>Power and the computation of sample size</b>	<b>155</b>
9.1	The principles of power calculations . . . . .	155
9.1.1	Power of one-sample and paired $t$ tests . . . . .	156
9.1.2	Power of two-sample $t$ test . . . . .	158
9.1.3	Approximate methods . . . . .	158
9.1.4	Power of comparisons of proportions . . . . .	159
9.2	Two-sample problems . . . . .	159
9.3	One-sample problems and paired tests . . . . .	161
9.4	Comparison of proportions . . . . .	161
9.5	Exercises . . . . .	162
<b>10</b>	<b>Advanced data handling</b>	<b>163</b>
10.1	Recoding variables . . . . .	163
10.1.1	The <code>cut</code> function . . . . .	163
10.1.2	Manipulating factor levels . . . . .	165
10.1.3	Working with dates . . . . .	166
10.1.4	Recoding multiple variables . . . . .	169
10.2	Conditional calculations . . . . .	170
10.3	Combining and restructuring data frames . . . . .	171
10.3.1	Appending frames . . . . .	172
10.3.2	Merging data frames . . . . .	173
10.3.3	Reshaping data frames . . . . .	175
10.4	Per-group and per-case procedures . . . . .	178
10.5	Time splitting . . . . .	179
10.6	Exercises . . . . .	183
<b>11</b>	<b>Multiple regression</b>	<b>185</b>
11.1	Plotting multivariate data . . . . .	185
11.2	Model specification and output . . . . .	187
11.3	Model search . . . . .	190
11.4	Exercises . . . . .	193

<b>12 Linear models</b>	<b>195</b>
12.1 Polynomial regression . . . . .	196
12.2 Regression through the origin . . . . .	198
12.3 Design matrices and dummy variables . . . . .	200
12.4 Linearity over groups . . . . .	202
12.5 Interactions . . . . .	206
12.6 Two-way ANOVA with replication . . . . .	207
12.7 Analysis of covariance . . . . .	208
12.7.1 Graphical description . . . . .	209
12.7.2 Comparison of regression lines . . . . .	212
12.8 Diagnostics . . . . .	218
12.9 Exercises . . . . .	224
<b>13 Logistic regression</b>	<b>227</b>
13.1 Generalized linear models . . . . .	228
13.2 Logistic regression on tabular data . . . . .	229
13.2.1 The analysis of deviance table . . . . .	234
13.2.2 Connection to test for trend . . . . .	235
13.3 Likelihood profiling . . . . .	237
13.4 Presentation as odds-ratio estimates . . . . .	239
13.5 Logistic regression using raw data . . . . .	239
13.6 Prediction . . . . .	241
13.7 Model checking . . . . .	242
13.8 Exercises . . . . .	247
<b>14 Survival analysis</b>	<b>249</b>
14.1 Essential concepts . . . . .	249
14.2 Survival objects . . . . .	250
14.3 Kaplan–Meier estimates . . . . .	251
14.4 The log-rank test . . . . .	254
14.5 The Cox proportional hazards model . . . . .	256
14.6 Exercises . . . . .	258
<b>15 Rates and Poisson regression</b>	<b>259</b>
15.1 Basic ideas . . . . .	259
15.1.1 The Poisson distribution . . . . .	260
15.1.2 Survival analysis with constant hazard . . . . .	260
15.2 Fitting Poisson models . . . . .	262
15.3 Computing rates . . . . .	266
15.4 Models with piecewise constant intensities . . . . .	270
15.5 Exercises . . . . .	274
<b>16 Nonlinear curve fitting</b>	<b>275</b>
16.1 Basic usage . . . . .	276
16.2 Finding starting values . . . . .	278

16.3	Self-starting models . . . . .	284
16.4	Profiling . . . . .	285
16.5	Finer control of the fitting algorithm . . . . .	287
16.6	Exercises . . . . .	288
<b>A</b>	<b>Obtaining and installing R and the ISwR package</b>	<b>289</b>
<b>B</b>	<b>Data sets in the ISwR package</b>	<b>293</b>
<b>C</b>	<b>Compendium</b>	<b>325</b>
<b>D</b>	<b>Answers to exercises</b>	<b>337</b>
	<b>Bibliography</b>	<b>355</b>
	<b>Index</b>	<b>357</b>