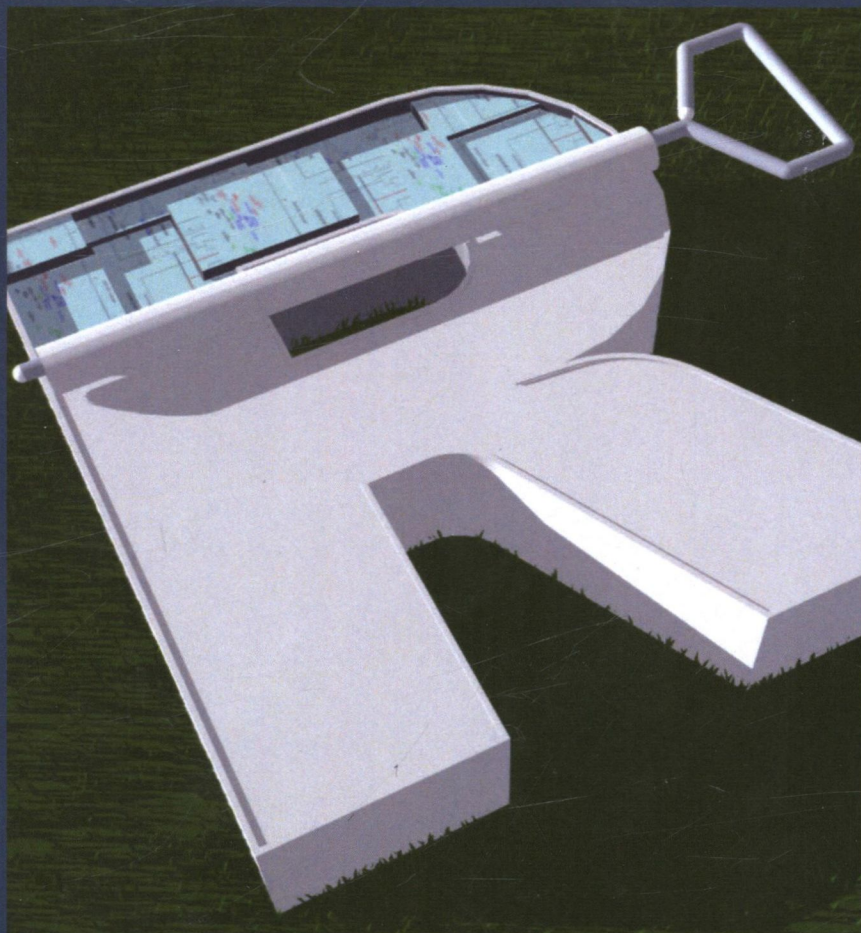


R for **Statistics**



Pierre-André Cornillon

Arnaud Guyader

François Husson

Nicolas Jégou

Julie Josse

Maela Kloareg

Eric Matzner-Løber

Laurent Rouvière



CRC Press

Taylor & Francis Group

A CHAPMAN & HALL BOOK

R for Statistics

Although there are currently a wide variety of software packages suitable for the modern statistician, R has the triple advantage of being comprehensive, widespread, and free. Published in 2008, the second edition of *Statistiques avec R* enjoyed great success as an R guidebook in the French-speaking world. Translated and updated, **R for Statistics** includes a number of expanded and additional worked examples.

Organized into two sections, the book focuses first on the R software, then on the implementation of traditional statistical methods with R.

Focusing on the R software, the first section covers:

- Basic elements of the R software and data processing
- Clear, concise visualization of results, using simple and complex graphs
- Programming basics: pre-defined and user-created functions

The second section of the book presents R methods for a wide range of traditional statistical data processing techniques, including:

- Regression methods
- Analyses of variance and covariance
- Classification methods
- Exploratory multivariate analysis
- Clustering methods
- Hypothesis tests

After a short presentation of the method, the book explicitly details the R command lines and gives commented results. Accessible to novices and experts alike, **R for Statistics** is a clear and enjoyable resource for any scientist.

Datasets and all the results described in this book are available on the book's webpage at
<http://www.agrocampus-ouest.fr/math/RforStat>



CRC Press

Taylor & Francis Group
an informa business

www.crcpress.com



R **for** **S** **t** **a** **t** **i** **s** **t** **i** **c** **s**

Cornillon

Guyader

Husson

Jégou

Josse

Kloareg

Matzner-Løber

Rouvière



R **for** **Statistics**

Pierre-André Cornillon

Arnaud Guyader

François Husson

Nicolas Jégou

Julie Josse

Maela Kloareg

Eric Matzner-Løber

Laurent Rouvière



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

First issued in hardback 2017

© 2012 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

ISBN-13: 978-1-4398-8145-3 (pbk)
ISBN-13: 978-1-1384-6934-1 (hbk)

This book contains information obtained from authentic and highly regarded sources. While all reasonable efforts have been made to publish reliable data and information, neither the author[s] nor the publisher can accept any legal responsibility or liability for any errors or omissions that may be made. The publishers wish to make clear that any views or opinions expressed in this book by individual editors, authors or contributors are personal to them and do not necessarily reflect the views/opinions of the publishers. The information or guidance contained in this book is intended for use by medical, scientific or health-care professionals and is provided strictly as a supplement to the medical or other professional's own judgement, their knowledge of the patient's medical history, relevant manufacturer's instructions and the appropriate best practice guidelines. Because of the rapid advances in medical science, any information or advice on dosages, procedures or diagnoses should be independently verified. The reader is strongly urged to consult the relevant national drug formulary and the drug companies' and device or material manufacturers' printed instructions, and their websites, before administering or utilizing any of the drugs, devices or materials mentioned in this book. This book does not indicate whether a particular treatment is appropriate or suitable for a particular individual. Ultimately it is the sole responsibility of the medical professional to make his or her own professional judgements, so as to advise and treat patients appropriately. The authors and publishers have also attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site
<http://www.crcpress.com>

R for **Statistics**

Preface

This book is the English adaptation of the second edition of the book *Statistiques avec R* which was published in 2008 and was a great success in the French-speaking world. In this version, a number of worked examples have been supplemented and new examples have been added. We hope that readers will enjoy using this book for reference when working with R.

This book is aimed at statisticians in the widest sense, that is to say, all those working with datasets: science students, biologists, economists, etc. All statistical studies depend on vast quantities of information, and computerised tools are therefore becoming more and more essential. There are currently a wide variety of software packages which meet these requirements. Here we have opted for R, which has the triple advantage of being free, comprehensive, and its use is booming. However, no prior experience of the software is required. This work aims to be accessible and useful for both novices and experts alike.

This book is organised into two main parts: Part I focuses on the R software and the way it works, and Part II on the implementation of traditional statistical methods with R. In order to render them as independent as possible, a brief chapter offers extra help getting started (Chapter 5, “A Quick Start with R”) and acts as a transition: It will help those readers who are more interested in statistics than in software to be operational more quickly.

In the first chapter we present the basic elements of the software. The second chapter deals with data processing (reading data from file, merging factor levels, etc.), that is to say, common operations in statistics. As any statistical report depends on clear, concise visualisation of results, in Chapter 3 we detail the variety of possibilities available in this domain with R. We first present the construction of simple graphs with the various available options, and then detail the use of more complex graphs. Some programming basics are then presented in Chapter 4: We explain how to construct your own functions but also present some of the useful pre-defined functions to conduct repetitive analyses automatically. Focusing on the R software itself, this first main part enables the reader to understand the commands used in subsequent statistical studies.

The second part of the book offers solutions for dealing with a wide

range of traditional statistical data processing techniques: hypothesis tests (Chapter 6), regression methods (Chapter 7), analyses of variance and covariance (Chapter 8), classification methods (Chapter 9), exploratory multivariate analysis methods (Chapter 10), and clustering methods (Chapter 11). Each technique is preceded by a concrete example and is dealt with independently in its own specific worked example. Following a short presentation of the method, the R command lines are explicitly detailed and the results commented. Readers are able to download the datasets from the address <http://www.agrocampus-ouest.fr/math/RforStat>, where they can easily find all the results described in this work. R is a command line interface, but another easy-to-use graphical interface is available: At the end of each worked example there is a section called “Rcmdr Corner” which can be used to process the example using this second interface. This section also helps readers get to know the R commands better as, with each click on the drop-down menu, the lines of code corresponding to the chosen action are generated.

To draw this preface to a close, we would like to thank Jérôme Pagès and Dominique Dehay, who enabled us to write the French version under the best possible conditions. We would also like to thank Rebecca Clayton for the English translation and Agrocampus Ouest for financing this translation.

Contents

Preface	xiii
I An Overview of R	1
1 Main Concepts	3
1.1 Installing R	3
1.2 Work Session	3
1.2.1 Session in Linux	4
1.2.2 Session in Windows	4
1.2.3 Session on a Mac	5
1.3 Help	5
1.3.1 Online Assistance	5
1.3.2 Help on CRAN	6
1.4 R Objects	6
1.4.1 Creating, Displaying and Deleting Objects	6
1.4.2 Type of Object	7
1.4.3 The Missing Value	8
1.4.4 Vectors	9
1.4.4.1 Numeric Vectors	9
1.4.4.2 Character Vectors	10
1.4.4.3 Logical Vectors	11
1.4.4.4 Selecting Part of a Vector	12
1.4.4.5 Selection in Practice	13
1.4.5 Matrices	13
1.4.5.1 Selecting Elements or Part of a Matrix	14
1.4.5.2 Calculating with the Matrices	16
1.4.5.3 Row and Column Operations	17
1.4.6 Factors	18
1.4.7 Lists	20
1.4.7.1 Creating a List	20
1.4.7.2 Extraction	20
1.4.7.3 A Few List Functions	21
1.4.7.4 Dimnames List	21
1.4.8 Data-Frames	22
1.5 Functions	23
1.5.1 Arguments of a Function	23

1.5.2	Output	24
1.6	Packages	24
1.6.1	What Is a Package ?	24
1.6.2	Installing a Package	24
1.6.3	Updating Packages	25
1.6.4	Using Packages	25
1.7	Exercises	26
2	Preparing Data	29
2.1	Reading Data from File	29
2.2	Exporting Results	33
2.3	Manipulating Variables	34
2.3.1	Changing Type	34
2.3.2	Dividing into Classes	35
2.3.3	Working on Factor Levels	36
2.4	Manipulating Individuals	39
2.4.1	Identifying Missing Data	39
2.4.2	Finding Outliers	42
2.5	Concatenating Data Tables	43
2.6	Cross-Tabulation	46
2.7	Exercises	48
3	R Graphics	51
3.1	Conventional Graphical Functions	51
3.1.1	The Plot Function	52
3.1.2	Representing a Distribution	58
3.1.3	Adding to Graphs	60
3.1.4	Graphs with Multiple Dimensions	64
3.1.5	Exporting Graphs	66
3.1.6	Multiple Graphs	67
3.1.7	Multiple Windows	69
3.1.8	Improving and Personalising Graphs	70
3.2	Graphical Functions with lattice	73
3.2.1	Characteristics of a “Lattice” Graph	75
3.2.2	Formulae and Groups	76
3.2.3	Customising Graphs	79
3.2.3.1	Panel Function	79
3.2.3.2	Controlling Size	80
3.2.3.3	Panel Layout	81
3.2.4	Exportation	81
3.2.5	Other Packages	82
3.3	Exercises	82

<i>Contents</i>	vii
4 Making Programs with R	87
4.1 Control Flows	87
4.1.1 Grouped Expressions	87
4.1.2 Loops (for or while)	87
4.1.3 Conditions (if, else)	89
4.2 Predefined Functions	90
4.3 Creating a Function	97
4.4 Exercises	100
II Statistical Methods	101
Introduction to the Statistical Methods Part	103
5 A Quick Start with R	105
5.1 Installing R	105
5.2 Opening and Closing R	105
5.3 The Command Prompt	105
5.4 Attribution, Objects, and Function	106
5.5 Selection	106
5.6 Other	107
5.7 Rcmdr Package	107
5.8 Importing (or Inputting) Data	107
5.9 Graphs	108
5.10 Statistical Analysis	108
6 Hypothesis Test	109
6.1 Confidence Intervals for a Mean	109
6.1.1 Objective	109
6.1.2 Example	109
6.1.3 Steps	109
6.1.4 Processing the Example	110
6.1.5 Rcmdr Corner	112
6.1.6 Taking Things Further	112
6.2 Chi-Square Test of Independence	113
6.2.1 Objective	113
6.2.2 Example	113
6.2.3 Steps	113
6.2.4 Processing the Example	114
6.2.5 Rcmdr Corner	116
6.2.6 Taking Things Further	117
6.3 Comparison of Two Means	118
6.3.1 Objective	118
6.3.2 Example	118
6.3.3 Steps	119
6.3.4 Processing the Example	119
6.3.5 Rcmdr Corner	123

6.3.6	Taking Things Further	124
6.4	Testing Conformity of a Proportion	125
6.4.1	Objective	125
6.4.2	Example	125
6.4.3	Step	125
6.4.4	Processing the Example	125
6.4.5	Rcmdr Corner	126
6.4.6	Taking Things Further	126
6.5	Comparing Several Proportions	127
6.5.1	Objective	127
6.5.2	Example	127
6.5.3	Step	127
6.5.4	Processing the Example	127
6.5.5	Rcmdr Corner	128
6.5.6	Taking Things Further	128
6.6	The Power of a Test	129
6.6.1	Objective	129
6.6.2	Example	129
6.6.3	Steps	129
6.6.4	Processing the Example	130
6.6.5	Rcmdr Corner	130
6.6.6	Taking Things Further	131
7	Regression	133
7.1	Simple Linear Regression	133
7.1.1	Objective	133
7.1.2	Example	134
7.1.3	Steps	134
7.1.4	Processing the Example	134
7.1.5	Rcmdr Corner	138
7.1.6	Taking Things Further	139
7.2	Multiple Linear Regression	140
7.2.1	Objective	140
7.2.2	Example	141
7.2.3	Steps	141
7.2.4	Processing the Example	141
7.2.5	Rcmdr Corner	145
7.2.6	Taking Things Further	146
7.3	Partial Least Squares (PLS) Regression	147
7.3.1	Objective	147
7.3.2	Example	147
7.3.3	Steps	148
7.3.4	Processing the Example	148
7.3.5	Rcmdr Corner	153
7.3.6	Taking Things Further	153

8	Analysis of Variance and Covariance	157
8.1	One-Way Analysis of Variance	157
8.1.1	Objectives	157
8.1.2	Example	158
8.1.3	Steps	158
8.1.4	Processing the Example	159
8.1.5	Rcmdr Corner	164
8.1.6	Taking Things Further	165
8.2	Multi-Way Analysis of Variance with Interaction	166
8.2.1	Objective	166
8.2.2	Example	167
8.2.3	Steps	167
8.2.4	Processing the Example	168
8.2.5	Rcmdr Corner	171
8.2.6	Taking Things Further	172
8.3	Analysis of Covariance	173
8.3.1	Objective	173
8.3.2	Example	174
8.3.3	Steps	175
8.3.4	Processing the Example	175
8.3.5	Rcmdr Corner	179
8.3.6	Taking Things Further	179
9	Classification	181
9.1	Linear Discriminant Analysis	181
9.1.1	Objective	181
9.1.2	Example	182
9.1.3	Steps	183
9.1.4	Processing the Example	183
9.1.5	Rcmdr Corner	188
9.1.6	Taking Things Further	188
9.2	Logistic Regression	190
9.2.1	Objective	190
9.2.2	Example	190
9.2.3	Steps	191
9.2.4	Processing the Example	191
9.2.5	Rcmdr Corner	197
9.2.6	Taking Things Further	198
9.3	Decision Tree	199
9.3.1	Objective	199
9.3.2	Example	199
9.3.3	Steps	199
9.3.4	Processing the Example	199
9.3.5	Rcmdr Corner	208
9.3.6	Taking Things Further	208

10 Exploratory Multivariate Analysis	209
10.1 Principal Component Analysis	209
10.1.1 Objective	209
10.1.2 Example	209
10.1.3 Steps	210
10.1.4 Processing the Example	210
10.1.5 Rcmdr Corner	218
10.1.6 Taking Things Further	219
10.2 Correspondence Analysis	222
10.2.1 Objective	222
10.2.2 Example	222
10.2.3 Steps	222
10.2.4 Processing the Example	223
10.2.5 Rcmdr Corner	227
10.2.6 Taking Things Further	228
10.3 Multiple Correspondence Analysis	230
10.3.1 Objective	230
10.3.2 Example	230
10.3.3 Steps	231
10.3.4 Processing the Example	231
10.3.5 Rcmdr Corner	239
10.3.6 Taking Things Further	239
11 Clustering	241
11.1 Ascending Hierarchical Clustering	241
11.1.1 Objective	241
11.1.2 Example	241
11.1.3 Steps	241
11.1.4 Processing the Example	242
11.1.5 Rcmdr Corner	246
11.1.6 Taking Things Further	247
11.2 The k -Means Method	252
11.2.1 Objective	252
11.2.2 Example	252
11.2.3 Steps	252
11.2.4 Processing the Example	252
11.2.5 Rcmdr Corner	255
11.2.6 Taking Things Further	256
Appendix	257
A.1 The Most Useful Functions	257
A.1.1 Generic Functions	257
A.1.2 Numerical Functions	257
A.1.3 Data Handling Functions	258
A.1.4 Probability Distributions	260

A.1.5	Basic Statistical Functions	261
A.1.6	Advanced Statistical Functions	262
A.1.7	Graphical Functions	263
A.1.8	Import and Export Functions	264
A.1.9	Text Management	265
A.1.10	Some Other Useful Functions	265
A.2	Writing a Formula for the Models	266
A.3	The Rcmdr Package	267
A.4	The FactoMineR Package	269
A.4.1	The FactoMineR Package	269
A.4.2	The Drop-Down Menu	270
A.5	Answers to the Exercises	270
A.5.1	Exercises: Chapter 1	270
A.5.2	Exercises: Chapter 2	278
A.5.3	Exercises: Chapter 3	287
A.5.4	Exercises: Chapter 4	292
Bibliography		295
Index of the Functions		297
Index		301

