Statistics and Computing

Series Editors:
J. Chambers
D. Hand
W. Härdle

UNCORRECTED PROOF

For further volumes:
http://www.springer.com/series/3022

UNCORRECTED PROOF

Pierre Lafaye de Micheaux • Rémy Drouilhet
Benoit Liquet

The R Software

Fundamentals of Programming and Statistical Analysis

UNCORRECTED PROOF

Springer

Pierre Lafaye de Micheaux
Department of Mathematics and Statistics
Université de Montréal
Montréal, QC, Canada

Rémy Drouilhet
B.S.H.M
Grenoble, France

Benoit Liquet
School of Mathematics and Physics
The University of Queensland
Brisbane, Australia

Series Editors:

J. Chambers
Department of Statistics
Sequoia Hall
390 Serra Mall
Stanford University
Stanford, CA 94305-4065

D. Hand
Department of Mathematics
Imperial College London
South Kensington Campus
London SW7 2AZ
United Kingdom

W. Härdle
C.A.S.E. Centre for Applied Statistics and Economics
School of Business and Economics
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany

ISSN 1431-8784
ISBN 978-1-4614-9019-7
DOI 10.1007/978-1-4614-9020-3
Springer New York Heidelberg Dordrecht London

ISSN 2197-1706 (electronic)
ISBN 978-1-4614-9020-3 (eBook)

Library of Congress Control Number: 2013953875

© Springer Science+Business Media New York 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

UNCORRECTED PROOF

To the open-source community

To all those who have contributed, are contributing and will contribute to the awakening of our consciousness

To our colleagues from Montpellier, Grenoble, Bordeaux and ISPED

UNCORRECTED PROOF

UNCORRECTED PROOF

Foreword

This book has been translated from French by Robin Ryder, who is assistant professor in the CEREMADE (Centre De Recherche en Mathématiques de la Décision) at Université Paris Dauphine (France). We are very pleased that he has agreed to make this translation.

This book is based on notes from a series of lectures given for a few years at the Institut Universitaire de Technologie Grenoble 2 in the Department of Statistics and Business Intelligence (STID, *Statistique et informatique décisionnelle*). It has therefore been "digested" first, in a very imperfect form, by the students of this department, whom we thank here. Had they not shown so much interest, this book would probably never have existed. We also thank our colleague and friend Michel Lejeune, who managed to talk to us about writing a manuscript and submitting it to Springer. It is worth pointing out the role of chance, which made the paths of the three authors cross in the same place for a few years. The human and scientific experience of this encounter was very enriching, and each author provided complementary skills which made it possible to overcome the tremendous amount of work necessary for this book. Finally, we wish to warmly thank our colleague and friend Matthieu Dubois, who is a researcher in experimental psychology and addicted to R and to Macintosh and who was the first to read the French version of the book in its almost finalized version and gave us many ideas for improvement.

The contents of this book were chosen and organized in the best possible way for them to be not only **exhaustive** but also **easy to assimilate** by the reader. This book can be used as support material for lectures on R at any level from beginner to advanced. We have paid particular attention to the form of the book, which we think should aid understanding. It can also be used as a support for self-teaching. Note that most of this book can be useful to users of any operating system. However, a few chapters are mostly meant for users of Microsoft Windows. We have also felt it useful to give, occasionally, complements aimed at users of Linux or Macintosh.

UNCORRECTED PROOF

All chapters follow the same structure. A chapter begins with a small insert listing the prerequisites necessary for the chapter and a short description of the contents. All theoretical notions are explained with numerous examples and include breaks so that the reader can put into practice on a computer the recently introduced notions. Each chapter ends with an assessment section: memorandum of most important terms, followed by a section of theoretical exercises (to be done on paper), which can be used as questions for a test. A practical sheet is also given at the end of each chapter. It can be used to check that the practical aspects of the chapter have been taken in. Note that all exercises and practicals only require the contents of the previous chapters.

The structure of the book is sequential. After a short introduction (see the first part), aimed at getting the reader interested, and a description of a few data sets which will be used throughout the book to illustrate how to use R, the second part of the book is dedicated to the fundamental concepts of R: data organization, import and export, various manipulations, documentation, plots, programming and maintenance. This part should help you "learn the ropes" of R.

The third part of the book is dedicated to using R in a few mathematical and statistical settings. You should read the second part before moving on to this part, although it can be understood by users who already have a few notions in R. It covers R instructions for some of the main statistics and mathematics courses up to third-year undergraduate (e.g., it covers the baccalaureate in statistics and actuarial sciences curriculum at Université de Montréal, as well as the French IUT curriculum in statistics and business intelligence): matrix operations, integration, optimization, descriptive statistics, simulations, confidence intervals and hypothesis testing, simple and multiple linear regression and analysis of variance.

Finally, note that each statistical chapter in the third part relies on one or several real data sets, kindly made available by ISPED (*Institut de santé publique, d'épidémiologie et de développement* in Bordeaux) and described at the beginning of the book. These make learning more concrete and more attractive. We take this opportunity to thank all the teaching staff from the Public Health School of ISPED. These data, as well as several functions developed specially for this book and which are described or used here, are available in an R package associated with this book, called `TheRSoftware`. We also thank Mohamed El Methni and Taghi Barumandzadeh for the material they gave us for the chapter on ANOVA and Hubert Raymondaud for many comments he has made on our French version which allowed us to significantly improve several sections of this book.

UNCORRECTED PROOF

Alternative Order of Reading

We have used the symbol † to make explicit more difficult or less fundamental sections, which you can skip at first read without prejudicing your understanding or mastering of R.

Note that this book was conceived for students from a mathematical or statistical background. However, students or scientists from a more "applied" background can also use it: for these readers, we propose a different order of reading, as follows; the difficult sections should also be omitted.

Part A: The Basics of R

(a) Basic concepts; data organization (Chap. 3)
(b) Import–export and data production (Chap. 4)
(c) Data manipulation (Chap. 5)
(d) R and its documentation (Chap. 6)
(e) Techniques for plots (Chap. 7)
(f) Maintaining sessions (Chap. 9)

Part B: Elementary Statistics

(a) Random variables, distributions and simulation (Chap. 12)
(b) Descriptive statistics (Chap. 11)
(c) Confidence intervals and hypothesis testing (Chap. 13)
(d) Simple and multiple linear regression (Chap. 14)
(e) Elementary analysis of variance (Chap. 15)

Part C: Advanced Concepts

(a) Basic mathematics: matrix operations, integration, optimization (Chap. 10)
(b) Programming in R (Chap. 8)

UNCORRECTED PROOF

UNCORRECTED PROOF

Inserts

We have tried to be careful with the presentation of the book (the form), to make the information (the content) more easy going. Inserts are located at strategic points in the book, to help bring out some of the important information and make notions easier to understand. These inserts are distinguished by icons in the margin.

Tip

Additional information about the topic under study.

Warning

Important point you should pay attention to.

Note

Advice and practical tricks.

See also

Refers to another chapter or a website.

UNCORRECTED PROOF

Advanced users

Advanced elements. You can omit these at first.

11

Linux

Information for Linux users.

12

Mac

Information for Macintosh users.

13

UNCORRECTED PROOF

Solutions to Exercises and Practicals

Solutions to exercises and practicals are given on the book's website (`http://www.biostatisticien.eu/springeR`).

Other projects, more ambitious than the practicals, will be made available on the same website.

Font Conventions

- The letter R refers to the R software.
- We use *italics* for words borrowed from Latin or French or for abbreviations.
- We use a `fixed width` font (`Verbatim` environment) for R instructions.
- We use SMALL CAPS for data sets and sans serif for the name of the file including these data sets. This font will be used for all file names and folder names used in this book.

Montréal, Canada — Pierre Lafaye de Micheaux
Grenoble, France — Rémy Drouilhet
Brisbane, Australia — Benoit Liquet

UNCORRECTED PROOF

UNCORRECTED PROOF