A First Course in Statistical Programming with R

This third edition of Braun and Murdoch's bestselling textbook now includes discussion of the use and design principles of the *tidyverse* packages in R, including expanded coverage of *ggplot2*. *R Markdown* is also discussed. The simulation chapter has been expanded to include an introduction to the Box–Müller and Metropolis–Hastings algorithms. New examples and exercises have been added throughout.

This is the only introduction you'll need to start programming in R, the computing standard for analyzing data. Co-written by a retired R Core Team member and an established R author, this book comes with real R code that complies with the standards of the language. Unlike other introductory books on the R system, this book emphasizes portable programming skills that apply to most computing languages and techniques used to develop more complex projects. Solutions, data sets, and any errata are available from the book's website `www.statprogr.science`. Worked examples—all from real applications—hundreds of exercises, and downloadable code, data sets, and solutions make a complete package for anyone working in or learning practical data science.

W. John Braun is Professor of Statistics at UBC's Okanagan campus. His research interests are in the modeling of environmental phenomena, such as wildfire, as well as statistical education, particularly as it relates to the R programming language.

Duncan J. Murdoch is a Professor Emeritus and was a member of the R Core Team of developers and co-president of the R Foundation. He is one of the developers of the rgl package for 3D visualization in R, and has also developed numerous other R packages.

A First Course in Statistical Programming with R

Third Edition

W. John Braun and Duncan J. Murdoch

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, 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
79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.
It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108995146
DOI: 10.1017/9781108993456

First published 2007
Second edition 2016
Third edition 2021

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

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

ISBN 978-1-108-99514-6 Paperback

Contents

Expanded contents

Preface to the third edition

The R community continues to be active, with numerous conferences, user groups, and new packages appearing since we wrote the second edition of this text. In particular, use of the *tidyverse* style of R programming has exploded. In this edition we have included the new Chapter 5 on several *tidyverse* packages, and greatly expanded our coverage of `ggplot2` in Chapter 3 and the `magrittr` pipe operator in Chapter 4. We have also added a section on using R Markdown to Chapter 2.

But this text is about statistical computing *using* R, it's not just a manual on how to use R: so as with other parts of the book, we've written these new parts to emphasize and discuss the underlying ideas. If you work through this book you'll learn a lot about R, but we hope you'll also learn a lot of ideas that will apply to other computing languages and systems.

The later chapters on the more mathematical aspects of statistical computing have also been updated. Chapter 6 now includes expanded coverage of Markov chain Monte Carlo including the Metropolis–Hastings algorithm with an example using it in Bayesian inference.

This edition was built with R version 4.0.2, and once again `knitr` was crucial in putting together the manuscript. Some of the richness of the R environment is indicated by the list of packages and versions used in the production of this book, as shown in the table below. We thank the R Core group and the authors of all of those packages for their work in making R such a rich system. We also thank careful readers Woonchan Cho and Julian Stander for pointing out several errors in the previous edition. All the new errors are ours!

W. John Braun
Duncan Murdoch

October, 2020

Packages used to produce this book.

Package	Version	Package	Version	Package	Version	Package	Version
DBI	1.1.0	evaluate	0.14	leaflet	2.0.3	rgl	0.102.25
KernSmooth	2.23-17	fastmap	1.0.1	lifecycle	0.2.0	rlang	0.4.7
MPV	1.55	forcats	0.5.0	lpSolve	5.6.15	rvest	0.3.6
NLP	0.2-0	fs	1.5.0	lubridate	1.7.9	scales	1.1.1
R6	2.4.1	generics	0.0.2	magrittr	1.5	shiny	1.5.0
Rcpp	1.0.5	ggplot2	3.3.2	manipulateWidget	0.10.1	sos	2.0-0
assertthat	0.2.1	glue	1.4.2	microbenchmark	1.4-7	stringi	1.5.3
backports	1.1.10	grid	4.0.2	mime	0.9	stringr	1.4.0
blob	1.2.1	gtable	0.3.0	miniUI	0.1.1.1	tibble	3.0.3
brew	1.0-6	haven	2.3.1	modelr	0.1.8	tidyr	1.1.2
broom	0.7.1	highr	0.8	munsell	0.5.0	tidyselect	1.1.0
cellranger	1.1.0	hms	0.5.3	patchDVI	1.10.1	tidyverse	1.3.0
colorspace	1.4-1	htmltools	0.5.0	pillar	1.4.6	tools	4.0.2
compiler	4.0.2	htmlwidgets	1.5.2	pkgconfig	2.0.3	vctrs	0.3.4
crayon	1.3.4	httpuv	1.5.4	promises	1.1.1	webshot	0.5.2
crosstalk	1.1.0.1	httr	1.4.2	purrr	0.3.4	xfun	0.18
dbplyr	1.4.4	janeaustenr	0.1.5	quadprog	1.5-8	xml2	1.3.2
digest	0.6.25	jsonlite	1.7.1	readr	1.4.0	xtable	1.8-4
dplyr	1.0.2	later	1.1.0.1	readxl	1.3.1		
ellipsis	0.3.1	lattice	0.20-41	reprex	0.3.0		

Preface to the second edition

A lot of things have happened in the R community since we wrote the first edition of this text. Millions of new users have started to use R, and it is now the premier platform for data analytics. (In fact, the term "data analytics" hardly existed when we wrote the first edition.)

RStudio, a cross-platform integrated development environment for R, has had a large influence on the increase in popularity. In this edition we recommend RStudio as the platform for most new users, and have integrated simple RStudio instructions into the text. In fact, we have used RStudio and the `knitr` package in putting together the manuscript.

We have also added numerous examples and exercises, and cleaned up existing ones when they were unclear. Chapter 2 (Introduction to the R Language) has had extensive revision and reorganization. We have added short discussions of newer graphics systems to Chapter 3 (Programming Statistical Graphics). Reference material on some common error messages has been added to Chapter 4 (Programming with R), and a list of pseudo-random number generators as well as a more extensive discussion of Markov chain Monte Carlo is new in Chapter 5 (Simulation). In Chapter 6 (Computational Linear Algebra), some applications have been added to give students a better idea of why some of the matrix decompositions are so important.

Once again we have a lot of people to thank. Many students have used the first edition, and we are grateful for their comments and criticisms. Some anonymous reviewers also provided some helpful suggestions and pointers so that we could make improvements to the text. We hope our readers find this new edition as interesting and educational as we think it is.

W. John Braun
Duncan Murdoch

November, 2015

Preface to the first edition

This text began as notes for a course in statistical computing for second year actuarial and statistical students at the University of Western Ontario. Both authors are interested in statistical computing, both as support for our other research and for its own sake. However, we have found that our students were not learning the right sort of programming basics before they took our classes. At every level from undergraduate through Ph.D., we found that the students were not able to produce simple, reliable programs; that they didn't understand enough about numerical computation to understand how rounding error could influence their results, and that they didn't know how to begin a difficult computational project.

We looked into service courses from other departments, but we found that they emphasized languages and concepts that our students would not use again. Our students need to be comfortable with simple programming so that they can put together a simulation of a stochastic model; they also need to know enough about numerical analysis so that they can do numerical computations reliably. We were unable to find this mix in an existing course, so we designed our own.

We chose to base this text on R. R is an open source computing package which has seen a huge growth in popularity in the last few years. Being open source, it is easily obtainable by students and economical to install in our computing lab. One of us (Murdoch) is a member of the core R development team, and the other (Braun) is a co-author of a book on data analysis using R. These facts made it easy for us to choose R, but we are both strong believers in the idea that there are certain universals of programming, and in this text we try to emphasize those: it is not a manual about programming in R, it is a course in statistical programming that uses R.

Students starting this course are not assumed to have any programming experience or advanced statistical knowledge. They should be familiar with university-level calculus, and should have had exposure to a course in introductory probability, though that could be taken concurrently: the probabilistic concepts start in Chapter 5. (We include a concise appendix reviewing the probabilistic material.) We include some advanced topics in simulation, linear algebra, and optimization that an instructor may choose to skip in a one-semester course offering.

We have a lot of people to thank for their help in writing this book. The students in Statistical Sciences 259b have provided motivation and feedback, Lutong Zhou drafted several figures, Kristy Alexander, Yiwen Diao, Qiang Fu, and Yu Han went over the exercises and wrote up detailed solutions, and Diana Gillooly of Cambridge University Press, Prof. Brian Ripley of Oxford University, and some anonymous reviewers all provided helpful suggestions. And of course, this book could not exist without R, and R would be far less valuable without the contributions of the worldwide R community.

W. John Braun
Duncan Murdoch

February, 2007