

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.

Cambridge University Press
978-1-108-99514-6 — A First Course in Statistical Programming with R
W. John Braun , Duncan J. Murdoch
Frontmatter
[More Information](#)

A First Course in Statistical Programming with R

Third Edition

W. John Braun and Duncan J. Murdoch



Cambridge University Press
978-1-108-99514-6 — A First Course in Statistical Programming with R
W. John Braun , Duncan J. Murdoch
Frontmatter
[More Information](#)

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

© W. John Braun and Duncan J. Murdoch 2007, 2016, 2021

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

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

Cambridge University Press has no responsibility for the persistence or accuracy of
URLs for external or third-party internet websites referred to in this publication
and does not guarantee that any content on such websites is, or will remain,
accurate or appropriate.

Contents

	Preface to the third edition	<i>page XV</i>
	Preface to the second edition	xvii
	Preface to the first edition	xix
I	Getting started	1
	1.1 What is statistical programming?	1
	1.2 Outline of this book	2
	1.3 The R package	3
	1.4 Why use a command line?	3
	1.5 Font conventions	4
	1.6 Installation of R and RStudio	4
	1.7 Getting started in RStudio	5
	1.8 Going further	6
2	Introduction to the R language	7
	2.1 First steps	7
	2.2 Basic features of R	12
	2.3 Vectors in R	13
	2.4 Data storage in R	22
	2.5 Packages, libraries, and repositories	27
	2.6 Getting help	29
	2.7 Useful R features	32
	2.8 Logical vectors and relational operators	37
	2.9 Data frames, tibbles, and lists	40
	2.10 Data input and output	46

3	Programming statistical graphics	53
3.1	Simple high level plots	54
3.2	Choosing a high level graphic	67
3.3	Low level graphics functions	68
3.4	Graphics as a language: <code>ggplot2</code>	70
3.5	Other graphics systems	87
4	Programming with R	93
4.1	Flow control	93
4.2	Managing complexity through functions	108
4.3	The <code>replicate()</code> function	114
4.4	Miscellaneous programming tips	115
4.5	Some general programming guidelines	118
4.6	Debugging and maintenance	125
4.7	Efficient programming	132
5	Complex programming in the <i>tidyverse</i>	139
5.1	The <i>tidyverse</i> principles	140
5.2	The <code>tibble</code> package: a data frame improvement	141
5.3	The <code>readr</code> package: reading data in the <i>tidyverse</i>	143
5.4	The <code>stringr</code> package for manipulating strings	144
5.5	The <code>dplyr</code> package for manipulating data sets	146
5.6	Other <i>tidyverse</i> packages	149
6	Simulation	150
6.1	Monte Carlo simulation	150
6.2	Generation of pseudorandom numbers	151
6.3	Simulation of other random variables	156
6.4	Multivariate random number generation	173
6.5	Markov chain simulation	175
6.6	Monte Carlo integration	177
6.7	Advanced simulation methods	179
7	Computational linear algebra	197
7.1	Vectors and matrices in R	198
7.2	Matrix multiplication and inversion	205
7.3	Eigenvalues and eigenvectors	210
7.4	Other matrix decompositions	211
7.5	Other matrix operations	218

8	Numerical optimization	222
8.1	The golden section search method	222
8.2	Newton–Raphson	225
8.3	The Nelder–Mead simplex method	227
8.4	Built-in functions	231
8.5	Linear programming	231
<hr/>		
Appendix A	Review of random variables and distributions	248
<hr/>		
Appendix B	Base graphics details	251
B.1	The plotting region and margins	251
B.2	Adjusting axis tick labels	252
B.3	Setting graphical parameters	255
<hr/>		
	Index	257

Cambridge University Press
978-1-108-99514-6 — A First Course in Statistical Programming with R
W. John Braun , Duncan J. Murdoch
Frontmatter
[More Information](#)

Expanded contents

	Preface to the third edition	<i>page</i> xv
	Preface to the second edition	xvii
	Preface to the first edition	xix
I	Getting started	1
	1.1 What is statistical programming?	1
	1.2 Outline of this book	2
	1.3 The R package	3
	1.4 Why use a command line?	3
	1.5 Font conventions	4
	1.6 Installation of R and RStudio	4
	1.7 Getting started in RStudio	5
	1.8 Going further	6
2	Introduction to the R language	7
	2.1 First steps	7
	2.1.1 R can be used as a calculator	7
	2.1.2 Named storage	9
	2.1.3 Quitting R	10
	2.2 Basic features of R	12
	2.2.1 Functions	12
	2.2.2 R is case-sensitive	13
	2.2.3 Listing the objects in the workspace	13
	2.3 Vectors in R	13
	2.3.1 Numeric vectors	13
	2.3.2 Extracting elements from vectors	14
	2.3.3 Vector arithmetic	16
	2.3.4 Simple patterned vectors	17
	2.3.5 Vectors with random patterns	17

x | EXPANDED CONTENTS

2.3.6	Character vectors	17
2.3.7	Factors	18
2.3.8	More on extracting elements from vectors	19
2.3.9	Matrices and arrays	20
2.4	Data storage in R	22
2.4.1	Approximate storage of numbers	22
2.4.2	Exact storage of numbers	24
2.4.3	Dates and times	25
2.4.4	Missing values and other special values	25
2.5	Packages, libraries, and repositories	27
2.6	Getting help	29
2.6.1	Built-in help pages	29
2.6.2	Built-in examples	30
2.6.3	Finding help when you don't know the function name	30
2.7	Useful R features	32
2.7.1	Some built-in graphics functions	32
2.7.2	Some elementary built-in functions	33
2.7.3	Presenting results using R Markdown	35
2.8	Logical vectors and relational operators	37
2.8.1	Boolean algebra	37
2.8.2	Logical operations in R	37
2.8.3	Relational operators	38
2.9	Data frames, tibbles, and lists	40
2.9.1	Extracting data frame elements and subsets	42
2.9.2	Taking random samples from populations	43
2.9.3	Constructing data frames	43
2.9.4	Data frames can have non-numeric columns	43
2.9.5	Lists	45
2.10	Data input and output	46
2.10.1	Changing directories	46
2.10.2	<code>dump()</code> and <code>source()</code>	47
2.10.3	Redirecting R output	48
2.10.4	Saving and retrieving image files	48
2.10.5	The <code>read.table</code> function	49

3 | Programming statistical graphics 53

3.1	Simple high level plots	54
3.1.1	Bar charts and dot charts	54
3.1.2	Pie charts	57
3.1.3	Histograms	58
3.1.4	Boxplots	60
3.1.5	Scatterplots	61
3.1.6	Plotting data from data frames	62
3.1.7	QQ plots	64
3.2	Choosing a high level graphic	67
3.3	Low level graphics functions	68
3.3.1	Adding to plots	68

3.4	Graphics as a language: <code>ggplot2</code>	70
3.4.1	Details of the <code>ggplot2</code> grammar	72
3.4.2	Layers in <code>ggplot2</code>	74
3.4.3	Setting colors	77
3.4.4	Customizing the look of a graph	79
3.4.5	Faceting	83
3.4.6	Groups in <code>ggplot2</code>	85
3.5	Other graphics systems	87
3.5.1	The <code>lattice</code> package	87
3.5.2	The <code>grid</code> package	87
3.5.3	Interactive graphics	89
4	Programming with R	93
4.1	Flow control	93
4.1.1	The <code>for()</code> loop	93
4.1.2	The <code>if()</code> statement	99
4.1.3	The <code>while()</code> loop	102
4.1.4	Newton's method for root finding	104
4.1.5	The <code>repeat</code> loop, and the <code>break</code> and <code>next</code> statements	106
4.2	Managing complexity through functions	108
4.2.1	What are functions?	108
4.2.2	Scope of variables	111
4.2.3	Returning multiple objects	112
4.2.4	Using S3 classes to control printing	112
4.2.5	The <code>magrittr</code> pipe operator	113
4.3	The <code>replicate()</code> function	114
4.4	Miscellaneous programming tips	115
4.4.1	Always edit code in the editor, not in the console	115
4.4.2	Documentation using <code>#</code>	115
4.4.3	Neatness counts!	116
4.5	Some general programming guidelines	118
4.5.1	Top-down design	120
4.6	Debugging and maintenance	125
4.6.1	Recognizing that a bug exists	125
4.6.2	Make the bug reproducible	126
4.6.3	Identify the cause of the bug	126
4.6.4	Fixing errors and testing	129
4.6.5	Look for similar errors elsewhere	129
4.6.6	Debugging in RStudio	129
4.6.7	The <code>browser()</code> , <code>debug()</code> , and <code>debugonce()</code> functions	130
4.6.8	Debugging <code>magrittr</code> pipes	131
4.7	Efficient programming	132
4.7.1	Learn your tools	133
4.7.2	Use efficient algorithms	133
4.7.3	Measure the time your program takes	135
4.7.4	Be willing to use different tools	136
4.7.5	Optimize with care	136

5	Complex programming in the <i>tidyverse</i>	139
5.1	The <i>tidyverse</i> principles	140
5.1.1	Discussion	141
5.2	The <code>tibble</code> package: a data frame improvement	141
5.2.1	Discussion	142
5.3	The <code>readr</code> package: reading data in the <i>tidyverse</i>	143
5.3.1	Discussion	144
5.4	The <code>stringr</code> package for manipulating strings	144
5.4.1	Discussion	146
5.5	The <code>dplyr</code> package for manipulating data sets	146
5.5.1	Discussion	148
5.6	Other <i>tidyverse</i> packages	149
6	Simulation	150
6.1	Monte Carlo simulation	150
6.2	Generation of pseudorandom numbers	151
6.3	Simulation of other random variables	156
6.3.1	Bernoulli random variables	156
6.3.2	Binomial random variables	157
6.3.3	Poisson random variables	162
6.3.4	Exponential random numbers	166
6.3.5	Normal random variables	168
6.3.6	Generating normal variates using the Box–Müller transformation	170
6.3.7	All built-in distributions	172
6.4	Multivariate random number generation	173
6.5	Markov chain simulation	175
6.6	Monte Carlo integration	177
6.7	Advanced simulation methods	179
6.7.1	Rejection sampling	180
6.7.2	Rejection sampling for bivariate distributions	184
6.7.3	Importance sampling	186
6.7.4	The Metropolis–Hastings algorithm	187
7	Computational linear algebra	197
7.1	Vectors and matrices in R	198
7.1.1	Constructing matrix objects	198
7.1.2	Accessing matrix elements; row and column names	200
7.1.3	Matrix properties	202
7.1.4	Triangular matrices	203
7.1.5	Matrix arithmetic	204
7.2	Matrix multiplication and inversion	205
7.2.1	Matrix inversion	207
7.2.2	The <i>LU</i> decomposition	207
7.2.3	Matrix inversion in R	209

7.2.4	Solving linear systems	210
7.3	Eigenvalues and eigenvectors	210
7.4	Other matrix decompositions	211
7.4.1	The singular value decomposition of a matrix	211
7.4.2	The Choleski decomposition of a positive definite matrix	212
7.4.3	The QR decomposition of a matrix	214
7.5	Other matrix operations	218
7.5.1	Kronecker products	219
7.5.2	<code>apply()</code>	219
8	Numerical optimization	222
8.1	The golden section search method	222
8.2	Newton–Raphson	225
8.3	The Nelder–Mead simplex method	227
8.4	Built-in functions	231
8.5	Linear programming	231
8.5.1	Solving linear programming problems in R	234
8.5.2	Maximization and other kinds of constraints	234
8.5.3	Special situations	235
8.5.4	Unrestricted variables	239
8.5.5	Integer programming	240
8.5.6	Alternatives to <code>lp()</code>	241
8.5.7	Quadratic programming	241
Appendix A	Review of random variables and distributions	248
Appendix B	Base graphics details	251
B.1	The plotting region and margins	251
B.2	Adjusting axis tick labels	252
B.3	Setting graphical parameters	255
	Index	257

Cambridge University Press
978-1-108-99514-6 — A First Course in Statistical Programming with R
W. John Braun , Duncan J. Murdoch
Frontmatter
[More Information](#)

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

Cambridge University Press
978-1-108-99514-6 — A First Course in Statistical Programming with R
W. John Braun , Duncan J. Murdoch
Frontmatter
[More Information](#)

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