A First Course in Statistical Programming with R

This new, color edition of Braun and Murdoch's bestselling textbook integrates use of the RStudio platform and adds discussion of newer graphics systems, extensive exploration of Markov chain Monte Carlo, expert advice on common error messages, motivating applications of matrix decompositions, and numerous new examples and exercises.

This is the only introduction you'll need to start programming in R, the computing standard for analyzing data. Co-written by an 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 programming, including the principles that apply to most computing languages, and techniques used to develop more complex projects. Solutions, datasets, and any errata are available from the book's website. The many examples, all from real applications, make it particularly useful for anyone working in practical data analysis.

W. John Braun is Deputy Director of the Canadian Statistical Sciences Institute. He is also Professor and Head of the Departments of Computer Science, Physics, Mathematics and Statistics at the University of British Columbia Okanagan. 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 member of the R Core Team of developers, and is 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. He is also a professor in the Department of Statistical and Actuarial Sciences at the University of Western Ontario.

A First Course in Statistical Programming with R

Second Edition

W. John Braun and Duncan J. Murdoch

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York NY 10013

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/9781107576469

First published 2007
Second edition 2016

Printed in the United States of America by Sheridan Books, Inc.

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

ISBN 978-1-107-57646-9 Hardback

Additional resources for this publication at www.cambridge.org/9781107576469.

Contents

Expanded contents

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 pseudorandom 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, Professor 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