Praise for the First Edition

Lucid and engaging—this is without doubt the fun way to learn R!
—Amos A. Folarin, University College London

Be prepared to quickly raise the bar with the sheer quality that R can produce.
—Patrick Breen, Rogers Communications Inc.

An excellent introduction and reference on R from the author of the best R website.
—Christopher Williams, University of Idaho

Thorough and readable. A great R companion for the student or researcher.
—Samuel McQuillin, University of South Carolina

Finally, a comprehensive introduction to R for programmers.
—Philipp K. Janert, Author of *Gnuplot in Action*

Essential reading for anybody moving to R for the first time.
—Charles Malpas, University of Melbourne

One of the quickest routes to R proficiency. You can buy the book on Friday and have a working program by Monday.
—Elizabeth Ostrowski, Baylor College of Medicine

One usually buys a book to solve the problems they know they have. This book solves problems you didn't know you had.
—Carles Fenollosa, Barcelona Supercomputing Center

Clear, precise, and comes with a lot of explanations and examples…the book can be used by beginners and professionals alike, and even for teaching R!
—Atef Ouni, Tunisian National Institute of Statistics

A great balance of targeted tutorials and in-depth examples.
—Landon Cox, 360VL Inc.

R in Action

SECOND EDITION

Data analysis and graphics with R

ROBERT I. KABACOFF

MANNING

SHELTER ISLAND

For online information and ordering of this and other Manning books, please visit www.manning.com. The publisher offers discounts on this book when ordered in quantity. For more information, please contact

Special Sales Department
Manning Publications Co.
20 Baldwin Road
PO Box 761
Shelter Island, NY 11964
Email: orders@manning.com

Manning Publications Co.
20 Baldwin Road
PO Box 761
Shelter Island, NY 11964

Development editor: Jennifer Stout
Copyeditor: Tiffany Taylor
Proofreader: Toma Mulligan
Typesetter: Marija Tudor
Cover designer: Marija Tudor

ISBN: 9781617291388
Printed in the United States of America
1 2 3 4 5 6 7 8 9 10 – EBM – 20 19 18 17 16 15

brief contents

contents

preface

What is the use of a book, without pictures or conversations?

—Alice, *Alice's Adventures in Wonderland*

It's wondrous, with treasures to satiate desires both subtle and gross; but it's not for the timid.

—Q, "Q Who?" *Stark Trek: The Next Generation*

When I began writing this book, I spent quite a bit of time searching for a good quote to start things off. I ended up with two. R is a wonderfully flexible platform and language for exploring, visualizing, and understanding data. I chose the quote from *Alice's Adventures in Wonderland* to capture the flavor of statistical analysis today—an interactive process of exploration, visualization, and interpretation.

The second quote reflects the generally held notion that R is difficult to learn. What I hope to show you is that is doesn't have to be. R is broad and powerful, with so many analytic and graphic functions available (more than 50,000 at last count) that it easily intimidates both novice and experienced users alike. But there is rhyme and reason to the apparent madness. With guidelines and instructions, you can navigate the tremendous resources available, selecting the tools you need to accomplish your work with style, elegance, efficiency—and more than a little coolness.

I first encountered R several years ago, when applying for a new statistical consulting position. The prospective employer asked in the pre-interview material if I was conversant in R. Following the standard advice of recruiters, I immediately said yes,

and set off to learn it. I was an experienced statistician and researcher, had 25 years experience as an SAS and SPSS programmer, and was fluent in a half dozen programming languages. How hard could it be? Famous last words.

As I tried to learn the language (as fast as possible, with an interview looming), I found either tomes on the underlying structure of the language or dense treatises on specific advanced statistical methods, written by and for subject-matter experts. The online help was written in a spartan style that was more reference than tutorial. Every time I thought I had a handle on the overall organization and capabilities of R, I found something new that made me feel ignorant and small.

To make sense of it all, I approached R as a data scientist. I thought about what it takes to successfully process, analyze, and understand data, including

- Accessing the data (getting the data into the application from multiple sources)
- Cleaning the data (coding missing data, fixing or deleting miscoded data, transforming variables into more useful formats)
- Annotating the data (in order to remember what each piece represents)
- Summarizing the data (getting descriptive statistics to help characterize the data)
- Visualizing the data (because a picture really is worth a thousand words)
- Modeling the data (uncovering relationships and testing hypotheses)
- Preparing the results (creating publication-quality tables and graphs)

Then I tried to understand how I could use R to accomplish each of these tasks. Because I learn best by teaching, I eventually created a website (www.statmethods.net) to document what I had learned.

Then, about a year later, Marjan Bace, Manning's publisher, called and asked if I would like to write a book on R. I had already written 50 journal articles, 4 technical manuals, numerous book chapters, and a book on research methodology, so how hard could it be? At the risk of sounding repetitive—famous last words.

A year after the first edition came out in 2011, I started working on the second edition. The R platform is evolving, and I wanted to describe these new developments. I also wanted to expand the coverage of predictive analytics and data mining—important topics in the world of big data. Finally, I wanted to add chapters on advanced data visualization, software development, and dynamic report writing.

The book you're holding is the one that I wished I had so many years ago. I have tried to provide you with a guide to R that will allow you to quickly access the power of this great open source endeavor, without all the frustration and angst. I hope you enjoy it.

P.S. I was offered the job but didn't take it. But learning R has taken my career in directions that I could never have anticipated. Life can be funny.

acknowledgments

A number of people worked hard to make this a better book. They include

- Marjan Bace, Manning's publisher, who asked me to write this book in the first place.
- Sebastian Stirling and Jennifer Stout, development editors on the first and second editions, respectively. Each spent many hours helping me organize the material, clarify concepts, and generally make the text more interesting.
- Pablo Domínguez Vaselli, technical proofreader, who helped uncover areas of confusion and provided an independent and expert eye for testing code. I came to rely on his vast knowledge, careful reviews, and considered judgment.
- Olivia Booth, the review editor, who helped obtain reviewers and coordinate the review process.
- Mary Piergies, who helped shepherd this book through the production process, and her team of Tiffany Taylor, Toma Mulligan, Janet Vail, David Novak, and Marija Tudor.
- The peer reviewers who spent hours of their own time carefully reading through the material, finding typos, and making valuable substantive suggestions: Bryce Darling, Christian Theil Have, Cris Weber, Deepak Vohra, Dwight Barry, George Gaines, Indrajit Sen Gupta, Dr. L. Duleep Kumar Samuel, Mahesh Srinivason, Marc Paradis, Peter Rabinovitch, Ravishankar Rajagopalan, Samuel Dale McQuillin, and Zekai Otles.
- The many Manning Early Access Program (MEAP) participants who bought the book before it was finished, asked great questions, pointed out errors, and made helpful suggestions.

Each contributor has made this a better and more comprehensive book.

I would also like to acknowledge the many software authors who have contributed to making R such a powerful data-analytic platform. They include not only the core developers, but also the selfless individuals who have created and maintain contributed packages, extending R's capabilities greatly. Appendix E provides a list of the authors of contributed packages described in this book. In particular, I would like to mention John Fox, Hadley Wickham, Frank E. Harrell, Jr., Deepayan Sarkar, and William Revelle, whose works I greatly admire. I have tried to represent their contributions accurately, and I remain solely responsible for any errors or distortions inadvertently included in this book.

I really should have started this book by thanking my wife and partner, Carol Lynn. Although she has no intrinsic interest in statistics or programming, she read each chapter multiple times and made countless corrections and suggestions. No greater love has any person than to read multivariate statistics for another. Just as important, she suffered the long nights and weekends that I spent writing this book, with grace, support, and affection. There is no logical explanation why I should be this lucky.

There are two other people I would like to thank. One is my father, whose love of science was inspiring and who gave me an appreciation of the value of data. I miss him dearly. The other is Gary K. Burger, my mentor in graduate school. Gary got me interested in a career in statistics and teaching when I thought I wanted to be a clinician. This is all his fault.

about this book

If you picked up this book, you probably have some data that you need to collect, summarize, transform, explore, model, visualize, or present. If so, then R is for you! R has become the worldwide language for statistics, predictive analytics, and data visualization. It offers the widest range of methodologies for understanding data currently available, from the most basic to the most complex and bleeding edge.

As an open source project it's freely available for a range of platforms, including Windows, Mac OS X, and Linux. It's under constant development, with new procedures added daily. Additionally, R is supported by a large and diverse community of data scientists and programmers who gladly offer their help and advice to users.

Although R is probably best known for its ability to create beautiful and sophisticated graphs, it can handle just about any statistical problem. The base installation provides hundreds of data-management, statistical, and graphical functions out of the box. But some of its most powerful features come from the thousands of extensions (packages) provided by contributing authors.

This breadth comes at a price. It can be hard for new users to get a handle on what R is and what it can do. Even the most experienced R user is surprised to learn about features they were unaware of.

R in Action, Second Edition provides you with a guided introduction to R, giving you a 2,000-foot view of the platform and its capabilities. It will introduce you to the most important functions in the base installation and more than 90 of the most useful contributed packages. Throughout the book, the goal is practical application—how you can make sense of your data and communicate that understanding to others. When you finish, you should have a good grasp of how R works and what it can do and where

you can go to learn more. You'll be able to apply a variety of techniques for visualizing data, and you'll have the skills to tackle both basic and advanced data analytic problems.

What's new in the second edition

If you want to delve into the use of R more deeply, the second edition offers more than 200 pages of new material. Concentrated in the second half of the book are new chapters on data mining, predictive analytics, and advanced programming. In particular, chapters 15 (time series), 16 (cluster analysis), 17 (classification), 19 (`ggplot2` graphics), 20 (advanced programming), 21 (creating a package), and 22 (creating dynamic reports) are new. In addition, chapter 2 (creating a dataset) has more detailed information on importing data from text and SAS files, and appendix F (working with large datasets) has been expanded to include new tools for working with big data problems. Finally, numerous updates and corrections have been made throughout the text.

Who should read this book

R in Action, Second Edition should appeal to anyone who deals with data. No background in statistical programming or the R language is assumed. Although the book is accessible to novices, there should be enough new and practical material to satisfy even experienced R mavens.

Users without a statistical background who want to use R to manipulate, summarize, and graph data should find chapters 1–6, 11, and 19 easily accessible. Chapters 7 and 10 assume a one-semester course in statistics; and readers of chapters 8, 9, and 12–18 will benefit from two semesters of statistics. Chapters 20–22 offer a deeper dive into the R language and have no statistical prerequisites. I've tried to write each chapter in such a way that both beginning and expert data analysts will find something interesting and useful.

Roadmap

This book is designed to give you a guided tour of the R platform, with a focus on those methods most immediately applicable for manipulating, visualizing, and understanding data. The book has 22 chapters and is divided into 5 parts: "Getting Started," "Basic Methods," "Intermediate Methods," "Advanced Methods," and "Expanding Your Skills." Additional topics are covered in seven appendices.

Chapter 1 begins with an introduction to R and the features that make it so useful as a data-analysis platform. The chapter covers how to obtain the program and how to enhance the basic installation with extensions that are available online. The remainder of the chapter is spent exploring the user interface and learning how to run programs interactively and in batch.

Chapter 2 covers the many methods available for getting data into R. The first half of the chapter introduces the data structures R uses to hold data, and how to enter

data from the keyboard. The second half discusses methods for importing data into R from text files, web pages, spreadsheets, statistical packages, and databases.

Many users initially approach R because they want to create graphs, so we jump right into that topic in chapter 3. No waiting required. We review methods of creating graphs, modifying them, and saving them in a variety of formats.

Chapter 4 covers basic data management, including sorting, merging, and subsetting datasets, and transforming, recoding, and deleting variables.

Building on the material in chapter 4, chapter 5 covers the use of functions (mathematical, statistical, character) and control structures (looping, conditional execution) for data management. I then discuss how to write your own R functions and how to aggregate data in various ways.

Chapter 6 demonstrates methods for creating common univariate graphs, such as bar plots, pie charts, histograms, density plots, box plots, and dot plots. Each is useful for understanding the distribution of a single variable.

Chapter 7 starts by showing how to summarize data, including the use of descriptive statistics and cross-tabulations. We then look at basic methods for understanding relationships between two variables, including correlations, t-tests, chi-square tests, and nonparametric methods.

Chapter 8 introduces regression methods for modeling the relationship between a numeric outcome variable and a set of one or more numeric predictor variables. Methods for fitting these models, evaluating their appropriateness, and interpreting their meaning are discussed in detail.

Chapter 9 considers the analysis of basic experimental designs through the analysis of variance and its variants. Here we're usually interested in how treatment combinations or conditions affect a numerical outcome. Methods for assessing the appropriateness of the analyses and visualizing the results are also covered.

Chapter 10 provides a detailed treatment of power analysis. Starting with a discussion of hypothesis testing, the chapter focuses on how to determine the sample size necessary to detect a treatment effect of a given size with a given degree of confidence. This can help you to plan experimental and quasi-experimental studies that are likely to yield useful results.

Chapter 11 expands on the material in chapter 6, covering the creation of graphs that help you to visualize relationships among two or more variables. These include various types of 2D and 3D scatter plots, scatter-plot matrices, line plots, correlograms, and mosaic plots.

Chapter 12 presents analytic methods that work well in cases where data are sampled from unknown or mixed distributions, where sample sizes are small, where outliers are a problem, or where devising an appropriate test based on a theoretical distribution is too complex and mathematically intractable. They include both resampling and bootstrapping approaches—computer-intensive methods that are easily implemented in R.

Chapter 13 expands on the regression methods in chapter 8 to cover data that are not normally distributed. The chapter starts with a discussion of generalized linear

models and then focuses on cases where you're trying to predict an outcome variable that is either categorical (logistic regression) or a count (Poisson regression).

One of the challenges of multivariate data problems is simplification. Chapter 14 describes methods of transforming a large number of correlated variables into a smaller set of uncorrelated variables (principal component analysis), as well as methods for uncovering the latent structure underlying a given set of variables (factor analysis). The many steps involved in an appropriate analysis are covered in detail.

Chapter 15 describes methods for creating, manipulating, and modeling time series data. It covers visualizing and decomposing time series data, as well as exponential and ARIMA approaches to forecasting future values.

Chapter 16 illustrates methods of clustering observations into naturally occurring groups. The chapter begins with a discussion of the common steps in a comprehensive cluster analysis, followed by a presentation of hierarchical clustering and partitioning methods. Several methods for determining the proper number of clusters are presented.

Chapter 17 presents popular supervised machine-learning methods for classifying observations into groups. Decision trees, random forests, and support vector machines are considered in turn. You'll also learn about methods for evaluating the accuracy of each approach.

In keeping with my attempt to present practical methods for analyzing data, chapter 18 considers modern approaches to the ubiquitous problem of missing data values. R supports a number of elegant approaches for analyzing datasets that are incomplete for various reasons. Several of the best are described here, along with guidance for which ones to use when, and which ones to avoid.

Chapter 19 wraps up the discussion of graphics with a presentation of one of R's most useful and advanced approaches to visualizing data: `ggplot2`. The `ggplot2` package implements a grammar of graphics that provides a powerful and consistent set of tools for graphing multivariate data.

Chapter 20 covers advanced programming techniques. You'll learn about object-oriented programming techniques and debugging approaches. The chapter also presents a variety of tips for efficient programming. This chapter will be particularly helpful if you're seeking a greater understanding of how R works, and it's a prerequisite for chapter 21.

Chapter 21 provides a step-by-step guide to creating R packages. This will allow you to create more sophisticated programs, document them efficiently, and share them with others.

Finally, chapter 22 offers several methods for creating attractive reports from within R. You'll learn how to generate web pages, reports, articles, and even books from your R code. The resulting documents can include your code, tables of results, graphs, and commentary.

The afterword points you to many of the best internet sites for learning more about R, joining the R community, getting questions answered, and staying current with this rapidly changing product.

Last, but not least, the seven appendices (A through G) extend the text's coverage to include such useful topics as R graphic user interfaces, customizing and upgrading an R installation, exporting data to other applications, using R for matrix algebra (à la MATLAB), and working with very large datasets.

We also offer a bonus chapter, which is available online only from the publisher's website at manning.com/RinActionSecondEdition. Online chapter 23 covers the `lattice` package, which is introduced in chapter 19.

Advice for data miners

Data mining is a field of analytics concerned with discovering patterns in large data sets. Many data-mining specialists are turning to R for its cutting-edge analytical capabilities. If you're a data miner making the transition to R and want to access the language as quickly as possible, I recommend the following reading sequence: chapter 1 (introduction), chapter 2 (data structures and those portions of importing data that are relevant to your setting), chapter 4 (basic data management), chapter 7 (descriptive statistics), chapter 8 (sections 1, 2, and 6; regression), chapter 13 (section 2; logistic regression), chapter 16 (clustering), chapter 17 (classification), and appendix F (working with large datasets). Then review the other chapters as needed.

Code examples

In order to make this book as broadly applicable as possible, I've chosen examples from a range of disciplines, including psychology, sociology, medicine, biology, business, and engineering. None of these examples require a specialized knowledge of that field.

The datasets used in these examples were selected because they pose interesting questions and because they're small. This allows you to focus on the techniques described and quickly understand the processes involved. When you're learning new methods, smaller is better. The datasets are provided with the base installation of R or available through add-on packages that are available online.

The source code for each example is available from www.manning.com/RinAction SecondEdition and at www.github.com/kabacoff/RiA2. To get the most out of this book, I recommend that you try the examples as you read them.

Finally, a common maxim states that if you ask two statisticians how to analyze a dataset, you'll get three answers. The flip side of this assertion is that each answer will move you closer to an understanding of the data. I make no claim that a given analysis is the best or only approach to a given problem. Using the skills taught in this text, I invite you to play with the data and see what you can learn. R is interactive, and the best way to learn is to experiment.

Code conventions

The following typographical conventions are used throughout this book:

- A `monospaced font` is used for code listings that should be typed as is.

- A `monospaced font` is also used within the general text to denote code words or previously defined objects.
- *`Italics`* within code listings indicate placeholders. You should replace them with appropriate text and values for the problem at hand. For example, *`path_to _my_file`* would be replaced with the actual path to a file on your computer.
- R is an interactive language that indicates readiness for the next line of user input with a prompt (> by default). Many of the listings in this book capture interactive sessions. When you see code lines that start with >, don't type the prompt.
- Code annotations are used in place of inline comments (a common convention in Manning books). Additionally, some annotations appear with numbered bullets like ❶ that refer to explanations appearing later in the text.
- To save room or make text more legible, the output from interactive sessions may include additional white space or omit text that is extraneous to the point under discussion.

Author Online

Purchase of *R in Action, Second Edition* includes free access to a private web forum run by Manning Publications where you can make comments about the book, ask technical questions, and receive help from the author and from other users. To access the forum and subscribe to it, point your web browser to www.manning.com/RinActionSecondEdition. This page provides information on how to get on the forum once you're registered, what kind of help is available, and the rules of conduct on the forum.

Manning's commitment to our readers is to provide a venue where a meaningful dialog between individual readers and between readers and the author can take place. It isn't a commitment to any specific amount of participation on the part of the author, whose contribution to the AO forum remains voluntary (and unpaid). We suggest you try asking the author some challenging questions, lest his interest stray!

The AO forum and the archives of previous discussions will be accessible from the publisher's website as long as the book is in print.

About the author

Dr. Robert Kabacoff is Vice President of Research for Management Research Group, an international organizational development and consulting firm. He has more than 20 years of experience providing research and statistical consultation to organizations in health care, financial services, manufacturing, behavioral sciences, government, and academia. Prior to joining MRG, Dr. Kabacoff was a professor of psychology at Nova Southeastern University in Florida, where he taught graduate courses in quantitative methods and statistical programming. For the past five years, he has managed Quick-R (www.statmethods.net), a popular R tutorial website.

about the cover illustration

The figure on the cover of *R in Action, Second Edition* is captioned "A man from Zadar." The illustration is taken from a reproduction of an album of Croatian traditional costumes from the mid-nineteenth century by Nikola Arsenovic, published by the Ethnographic Museum in Split, Croatia, in 2003. The illustrations were obtained from a helpful librarian at the Ethnographic Museum in Split, itself situated in the Roman core of the medieval center of the town: the ruins of Emperor Diocletian's retirement palace from around AD 304. The book includes finely colored illustrations of figures from different regions of Croatia, accompanied by descriptions of the costumes and of everyday life.

Zadar is an old Roman-era town on the northern Dalmatian coast of Croatia. It's over 2,000 years old and served for hundreds of years as an important port on the trading route from Constantinople to the West. Situated on a peninsula framed by small Adriatic islands, the city is picturesque and has become a popular tourist destination with its architectural treasures of Roman ruins, moats, and old stone walls. The figure on the cover wears blue woolen trousers and a white linen shirt, over which he dons a blue vest and jacket trimmed with the colorful embroidery typical for this region. A red woolen belt and cap complete the costume.

Dress codes and lifestyles have changed over the last 200 years, and the diversity by region, so rich at the time, has faded away. It's now hard to tell apart the inhabitants of different continents, let alone of different hamlets or towns separated by only a few miles. Perhaps we have traded this cultural diversity for a more varied personal life—certainly for a more varied and fast-paced technological life.

Manning celebrates the inventiveness and initiative of the computer business with book covers based on the rich diversity of regional life of two centuries ago, brought back to life by illustrations from old books and collections like this one.