STUMBLING INTO DATA SCIENCE

STUMBLING INTO DATA SCIENCE

A game-based introduction

by Haim Bar, HaiYing Wang, and Jun Yan

no starch press

San Francisco

Printed in the United States of America

25 24 23 22 21 1 2 3 4 5 6 7 8 9

ISBN-13: 123-4-5678-9012-3 (print)

ISBN-13: 123-4-5678-9012-3 (ebook)

Publisher: Bill Pollock

Production Manager: Rachel Monaghan

Production Editor: E. D. Itor

Developmental Editor: D. E. Itor

Cover Illustrator: Illustrator

Interior Design: Calamari Studios

Technical Reviewer: R. E. Viewer

Copyeditor: C. O. Editor

Proofreader: P. R. Øfreader

Indexer: I. N. Dexer

For information on book distributors or translations, please contact No Starch Press, Inc. directly:

No Starch Press, Inc.

245 8th Street, San Francisco, CA 94103

phone: 415.863.9900; fax: 415.863.9950; info@nostarch.com; www.nostarch.com

Library of Congress Cataloging-in-Publication Data

```
Uthor, A. U. and Respondent, C. O. R. and Iter, W. R.
  Pellentesque habitant morbi: Cum sociis natoque penatibus/
  A. U. Thor, C. O. R. Respondent and W. R. Iter
       p. cm.
  Includes index.
  ISBN-10: 1-23456-789-0
  ISBN-13: 123-4-56789-012-3
  1.  Morbi (habitant).  2. Penatibus. I. Title.
XX2303.5.T324 2008
```

The information in this book is distributed on an “As Is” basis, without warranty. While every precaution has been taken in the preparation of this work, neither the author nor No Starch Press, Inc. shall have any liability to any person or entity with respect to any loss or damage caused or alleged to be caused directly or indirectly by the information contained in it.

[S]

Haim: To my family … …

HaiYing: To my family … …

Jun: To Jo, Bo, and Jiafeng, without whom the book would have been finished much sooner, but life would be unimaginable.

About the Authors

Haim Bar is an Associate Professor in the Department of Statistics, University of Connecticut. Before joining UConn, he worked as a software engineer for Motorola, a director of software development in MicroPatent, LLC, a Principal Scientist at ATC-NY, and a statistician at Cornell's Statistical Consulting Unit. His research interests include Bayesian methods, statistical modeling and variable selection in high-dimensional data, especially in applications to genomics.

HaiYing Wang is an Associate Professor in the Department of Statistics, University of Connecticut. He was an Assistant Professor in the Department of Mathematics and Statistics at the University of New Hampshire and an cooling system engineer in the Midea Group. His research interests include informative subdata selection for big data, model selection, model averaging, measurement error models, optimal design, and semi-parametric regression.

Jun Yan is a Professor in the Department of Statistics at the University of Connecticut. a Research Fellow in the Center for Population Health at UConn Health. He received his PhD in Statistics from University of Wisconsin–Madison in 2003. After four years on the faculty of the Department of Statistics and Actuarial Science at the University of Iowa, he joined UConn in 2007. His methodological research interests include survival analysis, clustered data analysis, spatial extremes, and statistical computing. His application domains are public health, environmental sciences, and sports. With a special interest in making his statistical methods available via open source software, he and his coauthors developed and maintain a collection of R packages in the public domain. He is an Elected Member of the International Statistical Institute and a Fellow of the American Statistical Association.

All three authors have been members of the Computer Committee of the Department for years, during which time the idea of the book emerged.

BRIEF CONTENTS

PART I: MI ALIQUAM DICTUM

CONTENTS IN DETAIL

PART I
MI ALIQUAM DICTUM

FOREWORD

Morbi justo. Aenean nec dolor. In hac habitasse platea dictumst. Proin nonummy porttitor velit. Sed sit amet leo nec metus rhoncus varius. Cras ante. Vestibulum commodo sem tincidunt massa. Nam justo. Aenean luctus, felis et condimentum lacinia, lectus enim pulvinar purus, non porta velit nisl sed eros. Suspendisse consequat. Mauris a dui et tortor mattis pretium. Sed nulla metus, volutpat id, aliquam eget, ullamcorper ut, ipsum. Morbi eu nunc. Praesent pretium. Duis aliquam pulvinar ligula. Ut blandit egestas justo. Quisque posuere metus viverra pede.

Vivamus sodales elementum neque. Vivamus dignissim accumsan neque. Sed at enim. Vestibulum nonummy interdum purus. Mauris ornare velit id nibh pretium ultricies. Fusce tempor pellentesque odio. Vivamus augue purus, laoreet in, scelerisque vel, commodo id, wisi. Duis enim. Nulla interdum, nunc eu semper eleifend, enim dolor pretium elit, ut commodo ligula nisl a est. Vivamus ante. Nulla leo massa, posuere nec, volutpat vitae, rhoncus eu, magna.

Quisque facilisis auctor sapien. Pellentesque gravida hendrerit lectus. Mauris rutrum sodales sapien. Fusce hendrerit sem vel lorem. Integer pellentesque massa vel augue. Integer elit tortor, feugiat quis, sagittis et, ornare non, lacus. Vestibulum posuere pellentesque eros. Quisque venenatis ipsum dictum nulla. Aliquam quis quam non metus eleifend interdum. Nam eget sapien ac mauris malesuada adipiscing. Etiam eleifend neque sed quam. Nulla facilisi. Proin a ligula. Sed id dui eu nibh egestas tincidunt. Suspendisse arcu.

Maecenas dui. Aliquam volutpat auctor lorem. Cras placerat est vitae lectus. Curabitur massa lectus, rutrum euismod, dignissim ut, dapibus a, odio. Ut eros erat, vulputate ut, interdum non, porta eu, erat. Cras fermentum, felis in porta congue, velit leo facilisis odio, vitae consectetuer lorem quam vitae orci. Sed ultrices, pede eu placerat auctor, ante ligula rutrum tellus, vel posuere nibh lacus nec nibh. Maecenas laoreet dolor at enim. Donec molestie dolor nec metus. Vestibulum libero. Sed quis erat. Sed tristique. Duis pede leo, fermentum quis, consectetuer eget, vulputate sit amet, erat.

Donec vitae velit. Suspendisse porta fermentum mauris. Ut vel nunc non mauris pharetra varius. Duis consequat libero quis urna. Maecenas at ante. Vivamus varius, wisi sed egestas tristique, odio wisi luctus nulla, lobortis dictum dolor ligula in lacus. Vivamus aliquam, urna sed interdum porttitor, metus orci interdum odio, sit amet euismod lectus felis et leo. Praesent ac wisi. Nam suscipit vestibulum sem. Praesent eu ipsum vitae pede cursus venenatis. Duis sed odio. Vestibulum eleifend. Nulla ut massa. Proin rutrum mattis sapien. Curabitur dictum gravida ante.

Phasellus placerat vulputate quam. Maecenas at tellus. Pellentesque neque diam, dignissim ac, venenatis vitae, consequat ut, lacus. Nam nibh. Vestibulum fringilla arcu mollis arcu. Sed et turpis. Donec sem tellus, volutpat et, varius eu, commodo sed, lectus. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Quisque enim arcu, suscipit nec, tempus at, imperdiet vel, metus. Morbi volutpat purus at erat. Donec dignissim, sem id semper tempus, nibh massa eleifend turpis, sed pellentesque wisi purus sed libero. Nullam lobortis tortor vel risus. Pellentesque consequat nulla eu tellus. Donec velit. Aliquam fermentum, wisi ac rhoncus iaculis, tellus nunc malesuada orci, quis volutpat dui magna id mi. Nunc vel ante. Duis vitae lacus. Cras nec ipsum.

Morbi nunc. Aliquam consectetuer varius nulla. Phasellus eros. Cras dapibus porttitor risus. Maecenas ultrices mi sed diam. Praesent gravida velit at elit vehicula porttitor. Phasellus nisl mi, sagittis ac, pulvinar id, gravida sit amet, erat. Vestibulum est. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Curabitur id sem elementum leo rutrum hendrerit. Ut at mi. Donec tincidunt faucibus massa. Sed turpis quam, sollicitudin a, hendrerit eget, pretium ut, nisl. Duis hendrerit ligula. Nunc pulvinar congue urna.

Nunc velit. Nullam elit sapien, eleifend eu, commodo nec, semper sit amet, elit. Nulla lectus risus, condimentum ut, laoreet eget, viverra nec, odio. Proin lobortis. Curabitur dictum arcu vel wisi. Cras id nulla venenatis tortor congue ultrices. Pellentesque eget pede. Sed eleifend sagittis elit. Nam sed tellus sit amet lectus ullamcorper tristique. Mauris enim sem, tristique eu, accumsan at, scelerisque vulputate, neque. Quisque lacus. Donec et ipsum sit amet elit nonummy aliquet. Sed viverra nisl at sem. Nam diam. Mauris ut dolor. Curabitur ornare tortor cursus velit.

Morbi tincidunt posuere arcu. Cras venenatis est vitae dolor. Vivamus scelerisque semper mi. Donec ipsum arcu, consequat scelerisque, viverra id, dictum at, metus. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut pede sem, tempus ut, porttitor bibendum, molestie eu, elit. Suspendisse potenti. Sed id lectus sit amet purus faucibus vehicula. Praesent sed sem non dui pharetra interdum. Nam viverra ultrices magna.

Aenean laoreet aliquam orci. Nunc interdum elementum urna. Quisque erat. Nullam tempor neque. Maecenas velit nibh, scelerisque a, consequat ut, viverra in, enim. Duis magna. Donec odio neque, tristique et, tincidunt eu, rhoncus ac, nunc. Mauris malesuada malesuada elit. Etiam lacus mauris, pretium vel, blandit in, ultricies id, libero. Phasellus bibendum erat ut diam. In congue imperdiet lectus.

Aenean scelerisque. Fusce pretium porttitor lorem. In hac habitasse platea dictumst. Nulla sit amet nisl at sapien egestas pretium. Nunc non tellus. Vivamus aliquet. Nam adipiscing euismod dolor. Aliquam erat volutpat. Nulla ut ipsum. Quisque tincidunt auctor augue. Nunc imperdiet ipsum eget elit. Aliquam quam leo, consectetuer non, ornare sit amet, tristique quis, felis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque interdum quam sit amet mi. Pellentesque mauris dui, dictum a, adipiscing ac, fermentum sit amet, lorem.

Ut quis wisi. Praesent quis massa. Vivamus egestas risus eget lacus. Nunc tincidunt, risus quis bibendum facilisis, lorem purus rutrum neque, nec porta tortor urna quis orci. Aenean aliquet, libero semper volutpat luctus, pede erat lacinia augue, quis rutrum sem ipsum sit amet pede. Vestibulum aliquet, nibh sed iaculis sagittis, odio dolor blandit augue, eget mollis urna tellus id tellus. Aenean aliquet aliquam nunc. Nulla ultricies justo eget orci. Phasellus tristique fermentum leo. Sed massa metus, sagittis ut, semper ut, pharetra vel, erat. Aliquam quam turpis, egestas vel, elementum in, egestas sit amet, lorem. Duis convallis, wisi sit amet mollis molestie, libero mauris porta dui, vitae aliquam arcu turpis ac sem. Aliquam aliquet dapibus metus.

Vivamus commodo eros eleifend dui. Vestibulum in leo eu erat tristique mattis. Cras at elit. Cras pellentesque. Nullam id lacus sit amet libero aliquet hendrerit. Proin placerat, mi non elementum laoreet, eros elit tincidunt magna, a rhoncus sem arcu id odio. Nulla eget leo a leo egestas facilisis. Curabitur quis velit. Phasellus aliquam, tortor nec ornare rhoncus, purus urna posuere velit, et commodo risus tellus quis tellus. Vivamus leo turpis, tempus sit amet, tristique vitae, laoreet quis, odio. Proin scelerisque bibendum ipsum. Etiam nisl. Praesent vel dolor. Pellentesque vel magna. Curabitur urna. Vivamus congue urna in velit. Etiam ullamcorper elementum dui. Praesent non urna. Sed placerat quam non mi. Pellentesque diam magna, ultricies eget, ultrices placerat, adipiscing rutrum, sem.

R. E. Viewer
New York
December 2007

ACKNOWLEDGMENTS

Insert acknowledgments here

FOREWORD

Morbi justo. Aenean nec dolor. In hac habitasse platea dictumst. Proin nonummy porttitor velit. Sed sit amet leo nec metus rhoncus varius. Cras ante. Vestibulum commodo sem tincidunt massa. Nam justo. Aenean luctus, felis et condimentum lacinia, lectus enim pulvinar purus, non porta velit nisl sed eros. Suspendisse consequat. Mauris a dui et tortor mattis pretium. Sed nulla metus, volutpat id, aliquam eget, ullamcorper ut, ipsum. Morbi eu nunc. Praesent pretium. Duis aliquam pulvinar ligula. Ut blandit egestas justo. Quisque posuere metus viverra pede.

Vivamus sodales elementum neque. Vivamus dignissim accumsan neque. Sed at enim. Vestibulum nonummy interdum purus. Mauris ornare velit id nibh pretium ultricies. Fusce tempor pellentesque odio. Vivamus augue purus, laoreet in, scelerisque vel, commodo id, wisi. Duis enim. Nulla interdum, nunc eu semper eleifend, enim dolor pretium elit, ut commodo ligula nisl a est. Vivamus ante. Nulla leo massa, posuere nec, volutpat vitae, rhoncus eu, magna.

Quisque facilisis auctor sapien. Pellentesque gravida hendrerit lectus. Mauris rutrum sodales sapien. Fusce hendrerit sem vel lorem. Integer pellentesque massa vel augue. Integer elit tortor, feugiat quis, sagittis et, ornare non, lacus. Vestibulum posuere pellentesque eros. Quisque venenatis ipsum dictum nulla. Aliquam quis quam non metus eleifend interdum. Nam eget sapien ac mauris malesuada adipiscing. Etiam eleifend neque sed quam. Nulla facilisi. Proin a ligula. Sed id dui eu nibh egestas tincidunt. Suspendisse arcu.

Maecenas dui. Aliquam volutpat auctor lorem. Cras placerat est vitae lectus. Curabitur massa lectus, rutrum euismod, dignissim ut, dapibus a, odio. Ut eros erat, vulputate ut, interdum non, porta eu, erat. Cras fermentum, felis in porta congue, velit leo facilisis odio, vitae consectetuer lorem quam vitae orci. Sed ultrices, pede eu placerat auctor, ante ligula rutrum tellus, vel posuere nibh lacus nec nibh. Maecenas laoreet dolor at enim. Donec molestie dolor nec metus. Vestibulum libero. Sed quis erat. Sed tristique. Duis pede leo, fermentum quis, consectetuer eget, vulputate sit amet, erat.

Donec vitae velit. Suspendisse porta fermentum mauris. Ut vel nunc non mauris pharetra varius. Duis consequat libero quis urna. Maecenas at ante. Vivamus varius, wisi sed egestas tristique, odio wisi luctus nulla, lobortis dictum dolor ligula in lacus. Vivamus aliquam, urna sed interdum porttitor, metus orci interdum odio, sit amet euismod lectus felis et leo. Praesent ac wisi. Nam suscipit vestibulum sem. Praesent eu ipsum vitae pede cursus venenatis. Duis sed odio. Vestibulum eleifend. Nulla ut massa. Proin rutrum mattis sapien. Curabitur dictum gravida ante.

Phasellus placerat vulputate quam. Maecenas at tellus. Pellentesque neque diam, dignissim ac, venenatis vitae, consequat ut, lacus. Nam nibh. Vestibulum fringilla arcu mollis arcu. Sed et turpis. Donec sem tellus, volutpat et, varius eu, commodo sed, lectus. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Quisque enim arcu, suscipit nec, tempus at, imperdiet vel, metus. Morbi volutpat purus at erat. Donec dignissim, sem id semper tempus, nibh massa eleifend turpis, sed pellentesque wisi purus sed libero. Nullam lobortis tortor vel risus. Pellentesque consequat nulla eu tellus. Donec velit. Aliquam fermentum, wisi ac rhoncus iaculis, tellus nunc malesuada orci, quis volutpat dui magna id mi. Nunc vel ante. Duis vitae lacus. Cras nec ipsum.

Morbi nunc. Aliquam consectetuer varius nulla. Phasellus eros. Cras dapibus porttitor risus. Maecenas ultrices mi sed diam. Praesent gravida velit at elit vehicula porttitor. Phasellus nisl mi, sagittis ac, pulvinar id, gravida sit amet, erat. Vestibulum est. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Curabitur id sem elementum leo rutrum hendrerit. Ut at mi. Donec tincidunt faucibus massa. Sed turpis quam, sollicitudin a, hendrerit eget, pretium ut, nisl. Duis hendrerit ligula. Nunc pulvinar congue urna.

Nunc velit. Nullam elit sapien, eleifend eu, commodo nec, semper sit amet, elit. Nulla lectus risus, condimentum ut, laoreet eget, viverra nec, odio. Proin lobortis. Curabitur dictum arcu vel wisi. Cras id nulla venenatis tortor congue ultrices. Pellentesque eget pede. Sed eleifend sagittis elit. Nam sed tellus sit amet lectus ullamcorper tristique. Mauris enim sem, tristique eu, accumsan at, scelerisque vulputate, neque. Quisque lacus. Donec et ipsum sit amet elit nonummy aliquet. Sed viverra nisl at sem. Nam diam. Mauris ut dolor. Curabitur ornare tortor cursus velit.

Morbi tincidunt posuere arcu. Cras venenatis est vitae dolor. Vivamus scelerisque semper mi. Donec ipsum arcu, consequat scelerisque, viverra id, dictum at, metus. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut pede sem, tempus ut, porttitor bibendum, molestie eu, elit. Suspendisse potenti. Sed id lectus sit amet purus faucibus vehicula. Praesent sed sem non dui pharetra interdum. Nam viverra ultrices magna.

Aenean laoreet aliquam orci. Nunc interdum elementum urna. Quisque erat. Nullam tempor neque. Maecenas velit nibh, scelerisque a, consequat ut, viverra in, enim. Duis magna. Donec odio neque, tristique et, tincidunt eu, rhoncus ac, nunc. Mauris malesuada malesuada elit. Etiam lacus mauris, pretium vel, blandit in, ultricies id, libero. Phasellus bibendum erat ut diam. In congue imperdiet lectus.

Aenean scelerisque. Fusce pretium porttitor lorem. In hac habitasse platea dictumst. Nulla sit amet nisl at sapien egestas pretium. Nunc non tellus. Vivamus aliquet. Nam adipiscing euismod dolor. Aliquam erat volutpat. Nulla ut ipsum. Quisque tincidunt auctor augue. Nunc imperdiet ipsum eget elit. Aliquam quam leo, consectetuer non, ornare sit amet, tristique quis, felis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque interdum quam sit amet mi. Pellentesque mauris dui, dictum a, adipiscing ac, fermentum sit amet, lorem.

Ut quis wisi. Praesent quis massa. Vivamus egestas risus eget lacus. Nunc tincidunt, risus quis bibendum facilisis, lorem purus rutrum neque, nec porta tortor urna quis orci. Aenean aliquet, libero semper volutpat luctus, pede erat lacinia augue, quis rutrum sem ipsum sit amet pede. Vestibulum aliquet, nibh sed iaculis sagittis, odio dolor blandit augue, eget mollis urna tellus id tellus. Aenean aliquet aliquam nunc. Nulla ultricies justo eget orci. Phasellus tristique fermentum leo. Sed massa metus, sagittis ut, semper ut, pharetra vel, erat. Aliquam quam turpis, egestas vel, elementum in, egestas sit amet, lorem. Duis convallis, wisi sit amet mollis molestie, libero mauris porta dui, vitae aliquam arcu turpis ac sem. Aliquam aliquet dapibus metus.

Vivamus commodo eros eleifend dui. Vestibulum in leo eu erat tristique mattis. Cras at elit. Cras pellentesque. Nullam id lacus sit amet libero aliquet hendrerit. Proin placerat, mi non elementum laoreet, eros elit tincidunt magna, a rhoncus sem arcu id odio. Nulla eget leo a leo egestas facilisis. Curabitur quis velit. Phasellus aliquam, tortor nec ornare rhoncus, purus urna posuere velit, et commodo risus tellus quis tellus. Vivamus leo turpis, tempus sit amet, tristique vitae, laoreet quis, odio. Proin scelerisque bibendum ipsum. Etiam nisl. Praesent vel dolor. Pellentesque vel magna. Curabitur urna. Vivamus congue urna in velit. Etiam ullamcorper elementum dui. Praesent non urna. Sed placerat quam non mi. Pellentesque diam magna, ultricies eget, ultrices placerat, adipiscing rutrum, sem.

R. E. Viewer
New York
December 2007

PART I

MI ALIQUAM DICTUM

Vivamus adipiscing. Curabitur imperdiet tempus turpis. Vivamus sapien dolor, congue venenatis, euismod eget, porta rhoncus, magna. Proin condimentum pretium enim. Fusce fringilla, libero et venenatis facilisis, eros enim cursus arcu, vitae facilisis odio augue vitae orci. Aliquam varius nibh ut odio. Sed condimentum condimentum nunc. Pellentesque eget massa. Pellentesque quis mauris. Donec ut ligula ac pede pulvinar lobortis. Pellentesque euismod. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent elit. Ut laoreet ornare est. Phasellus gravida vulputate nulla. Donec sit amet arcu ut sem tempor malesuada. Praesent hendrerit augue in urna. Proin enim ante, ornare vel, consequat ut, blandit in, justo. Donec felis elit, dignissim sed, sagittis ut, ullamcorper a, nulla. Aenean pharetra vulputate odio.

Quisque enim. Proin velit neque, tristique eu, eleifend eget, vestibulum nec, lacus. Vivamus odio. Duis odio urna, vehicula in, elementum aliquam, aliquet laoreet, tellus. Sed velit. Sed vel mi ac elit aliquet interdum. Etiam sapien neque, convallis et, aliquet vel, auctor non, arcu. Aliquam suscipit aliquam lectus. Proin tincidunt magna sed wisi. Integer blandit lacus ut lorem. Sed luctus justo sed enim.

1

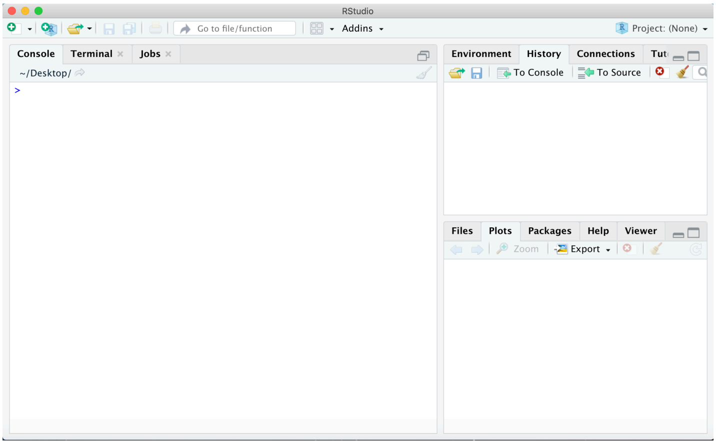
GETTING TO KNOW R

Just like a chef needs a set of tools, a kitchen with a large surface to work on, and a detailed cookbook with different recipes, a data scientist needs a powerful programming language, a convenient development environment, and good documentation. We chose the R language and the RStudio integrated development environment (IDE) for this book, and we hope that this book will be your basic guide into data science. In order to get started, you must first get the necessary software and get acquainted with it, and this is what this chapter is about.

1.1 Installing R and Rstudio

To install R go to the website of the Comprehensive R Archive Network (`https://cran.r-project.org/`) and download the latest version of R for your operating system (Windows, MacOS, or Linux). Follow the installation instructions.

Next, go the Rstudio download website (`https://rstudio.com/products/rstudio/download/`) and get the Desktop version (open source license). Follow the installation instruction. An icon that looks like this R will be on your computer's desktop. Double-click on this icon to start the R session. The Rstudio IDE will open, and look like this:



The left-hand side of the screen contains the Console tab. Notice the > sign (called the 'prompt'). When you see this character, it means that R is ready for the next command. Put the cursor there, and then type

```
2 + 2
```

and hit Enter on the keyboard. You should get the following on the Console:

```
[1] 4
```

Notice the top-right part of the Rstudio window. You should see an 'Environment' and a 'History' tab. Click on History. All your previous input appears there. Try entering another calculation or command and see that they appear in the session's history. For example, try entering the following

```
date()
```

Click on the Environment tab. It should be empty when you start R for the first time. In the Console, type

```
myFirstVariable <- factorial(5)
```

Notice that nothing was printed in the Console, but the Environment tab now contains a table with one row, with 'myFirstVariable' appearing in the cell on the left, and its value (120) on the right. Any object appearing in the Environment tab is available to you throughout your R session, and you don't have to redefine or recalculate it. For example, you can try the following:

```
myFirstVariable/6
```

The Console should now display 20.

The lower-right side of the IDE contains a file browser (the Files tab), information about installed packages (more about it later), and any plot generated during the R session. It also contains a Help tab, to obtain information about built-in functions.

Finally, before we move on to the next section, in the Rstudio top menu, click on File, then on New File, and then on R Script. Alternatively, you can click on little green '+' icon in the top-left part of the IDE. This will split the left side of the Rstudio IDE into two parts – the lower part will contain the Console, and the top part will contain a tab labeled 'Untitled1'. This is where you can enter R code which you will save to a permanent file, and re-use later. For example, enter the following in the blank space in the Untitled1 tab:

```
# This is my first R program
cat("Hello, World!\n")
```

Then, from the main menu in Rstudio, click on File, then on Save, and in the 'Save As' box enter FirstProgram.R and click the Save button. Notice that the tab name is now FirstProgram.R.

In that part of the window, there should now be a small button called Source. Click on it. The program will be executed and the output will be shown in the Console. You can also execute individual lines in the source code. Just put the cursor anywhere in that line, and click on the Run button (which is near the Source button) or click Ctrl+Enter (or Command+Enter on a Mac).

That's it. In the rest of these notes we will see more features of Rstudio, but you are now ready to start learning programming in R.

1.2 Basic Operations in R

1.2.1 Some Useful Functions

R has many built-in functions, and many more in external packages. We will introduce them as we go, but let's get started with some basic ones. The documentation on each function can be obtained by using `?func` or `help(func)` where `func` is some function. For example, to get the documentation about using the `help()` function, try the following:

```
help(help)
?help
```

When we start an R session, it's important to determine the 'working directory'. This is the folder on our computer which R will use to search for data or code files and save results. To find out which folder is currently used, we use the following:

```r
getwd()
```

Try it!

If we want (and we often will), we may change the working directory by using `setwd()`.

```r
setwd("~/Desktop")
```

The `~/` notation is a shortcut to your home directory. Using this shortcut is convenient because if you are using different computers or you share code with others, the home directory may be different on each computer.

Data is stored in variables. We can get the data from a file (Excel, comma-separated values, etc.), the Internet, or we can generate it ourselves. Let's start by generating some data. The simplest function to create data is called `c()`, which stands for 'combine'.

```r
courseNames <- c("Data Science", "Statistics", "Probability")
print(courseNames)
print(class(courseNames))
print(str(courseNames))
print(length(courseNames))
```

```
 [1] "Data Science" "Statistics"   "Probability"
[1] "character"
 chr [1:3] "Data Science" "Statistics" "Probability"
NULL
[1] 3
```

When you want to finish your session, just type `quit()`.

1.2.2 Generating random numbers

1.2.3 Summary statistics

2

SUMMARIZING AND VISUALIZING DATA

People are born with a fantastic ability to detect patterns. However, sometimes the information is hidden and the only way to observe a pattern is to change the viewpoint or rearrange the input. In this chapter we will introduce some tools to summarize data in order to reveal hidden features. We will also discuss principles of graphical excellence. In the words of Edward Tufte, the guru of data visualization, '*Graphical excellence consists of complex idea communicated with clarity, precision, and efficiency.*'

2.1 plots

In this section we will use two Python packages: Matplotlib and Seaborn. Both have excellent features which will help you produce clear and attractive plots, so install them before you read on.

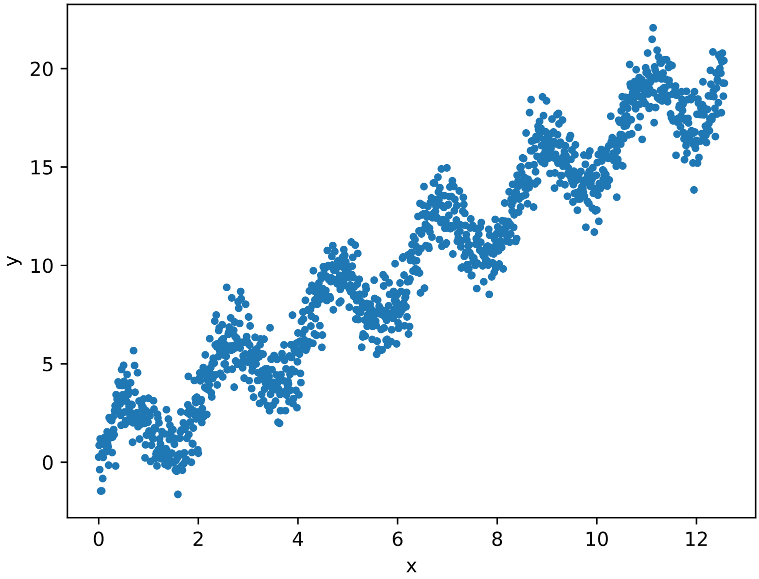


Figure 2-1: Scatterplot

In this book we cover several methods to analyze the data with statistical tool. However, it is very important to always look at the data, and this means creating plots and tables in order to understand it better. We call this process "exploratory data analysis", or EDA, and it is a critical step which should never be skipped. The type of plot which we generate in each case depends on the data, and usually there is more than one option to choose from. In this chapter we cover several visualization methods, and in subsequent chapters we will see additional ones.

2.1.1 One variable

2.1.2 Two variables

To display the relationship between two continuous variables we often use a scatterplot. In Figure 2-1 we see two variables, x and y, and we can see a clear pattern: as x increases, y also increases but we see a cyclical pattern. We can also see that the data is quite noisy. If we draw (in our mind's eye, for now) the line that represents the relationship between x and y, the points will be scattered around that line.

The actual relationship between x and y in this example is $y = 1.6x + 2\sin(3x)$, plus an additional random noise which is drawn from a standard normal distribution (with location 0 and scale 1). The code that was used to generate this plot is as follows:

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3 x = np.arange(0, 4*np.pi, 0.01)
4 y = 1.6*x + 2*np.sin(3*x) + np.random.normal(loc=0.0, scale=1.0, size=len(x))
5 plt.plot(x, y, linestyle="", marker=".")
6 plt.xlabel('x')
7 plt.ylabel('y')
8 plt.savefig('images/ScaPlot.pdf')
9 #plt.show()
```

In line #3 we create a sequence of points between 0 and 4π, in jumps of 0.01. These are our x values. For each x_i in this sequence, we calculate $y = 1.6x + 2sin(3x)$ and add the random (standard normal) noise. We use a shorthand notation for the `matplotlib.pyplot` module and call it `plt`. This is just used to save us extra typing. The, we can create the scatterplot in line #5 by using the `plt.plot` function. Setting the line style to blank means that the points will not be connected to each other with a line. The marker argument allows us to choose the shape of the points. Then, we simply add the axes labels. Finally, note that in order to save the figure to a file we us the `savefig` function, but if we wanted to view it on the screen, we would use the `show` function.

Try running this code, but with the following changes:

1. Without the x term: `y = 2*np.sin(3*x) + np.random.normal(loc=0.0, scale=1.0, size=len(x))`
2. Without the cyclical term: `y = 1.6*x + np.random.normal(loc=0.0, scale=1.0, size=len(x))`
3. Without the random noise: `y = 1.6*x + 2*np.sin(3*x)`
4. With less noise: `y = 1.6*x + 2*np.sin(3*x) + np.random.normal(loc=0.0, scale=0.1, size=len(x))`
5. With more noise: `y = 1.6*x + 2*np.sin(3*x) + np.random.normal(loc=0.0, scale=2, size=len(x))`
6. Change the coefficients (1.6, 2, and 3). Include negative values in your attempts.
7. Use a different marker in the plot.

2.2 Tables

3

INTUITION, PARADOXES, AND PROBABILITY

The world is random and probability is a tool to quantify the randomness. With random events, our intuition may give us high confidence on a wrong conclusion. For example, knowing that A is larger than B and B is larger than C, can we say A is always larger than C? Not really for random events.

3.1 Probability

We will go through examples to see how probability helps interpret phenomena in real life and how probability may counter intuition. We begin by introducing some concepts:

- Experiment: a situations in which the outcome occur randomly.
- Sample Space: the set of all possible outcomes in an experiment.
- Event: a subset of the sample space is called an event; an event occur if the outcome from an experiment belong to the event.

- Probability: assuming that all the elements of the sample space have equal chance to occur, the probability of event A is

$$P(A) = \frac{n}{N} = \frac{\text{number of elements in A}}{\text{total number of outcomes}},$$

where n is the number of elements in A and N is the number of elements in the sample space. Note that this formula holds only if all the outcomes are equally likely.

3.2 Elevator waiting time

Example 1. (Elevator waiting time) Mr. Smith works on the 13th floor of a 15 floor building. The only elevator moves continuously through floors 1, 2, . . . 14, 15, 14, . . . 2, 1, 2, . . . , except that it stops on a floor on which the button has been pressed. Assume that time spent loading and unloading passengers is very small compared to the travelling time. Mr. Smith complains that at 5pm, when he wants to go home, the elevator almost always goes up when it stops on his floor. What is the explanation?

When Mr. Smith gets to the elevator, it may be below the 13th floor or above it. The elevator will goes up if it is below the 13th floor and it will goes down if it is above the 13th floor. There are 12 floors below the 13th floor and 2 floors above it, so the probability that the elevator is below the 13th floor is $12/14 \approx 0.86 > 0.5$. Thus no matter when Mr. Smith wants to go home, it is more likely that the elevator is going up.

We can simulate this situation:

3.3 Intransitive Dice

We know that in general if $A > B$ and $B > C$ then $A > C$. However, this may not be the case in the word of probability.

Consider three dice with different sides numbers:

- Die A has sides 2, 2, 4, 4, 9, 9.
- Die B has sides 1, 1, 6, 6, 8, 8.
- Die C has sides 3, 3, 5, 5, 7, 7.

To play a game using dice A and B, you can chose which die to roll and your opponent rolls the other. The one who tolls a larger number wins. Which die do you want to choose?

Let's table the possible results to see the better die.

		B		
		1	6	8
A	2	$A > B$	$A < B$	$A < B$
	4	$A > B$	$A < B$	$A < B$
	9	$A > B$	$A > B$	$A > B$

Since die A wins five out of the nine possible results and all possible results occur with equal probability, we know that die A has a higher winning probability ($\frac{5}{9}$) than die B. We should choose die A over die B.

Now consider dice B and C.

			B	
		1	6	8
C	3	$C > B$	$C < B$	$C < B$
	5	$C > B$	$C < B$	$C < B$
	7	$C > B$	$C > B$	$C < B$

We see that die B has a higher winning probability ($\frac{5}{9}$) than die C, so we should choose die B over die C.

Since we should choose die A over die B, and choose die B over die C, does this mean that we should choose die A over die C if these two dice are to be selected. Surprisingly, the answer is NO. Here is the table of the possible results.

			C	
		3	5	7
A	2	$A < C$	$A < C$	$A < C$
	4	$A > C$	$A < C$	$A < C$
	9	$A > C$	$A > C$	$A > C$

We should choose die C over die A!

Here is an experiment to simulate the intransitive dice.

```python
import numpy as np
np.random.seed(2022)
A = [2, 2, 4, 4, 9, 9]
B = [1, 1, 6, 6, 8, 8]
C = [3, 3, 5, 5, 7, 7]
n = 1000
rollA = np.random.choice(A, size=n)
rollB = np.random.choice(B, size=n)
rollC = np.random.choice(C, size=n)
my_list = np.array([rollA, rollB, rollC])
print("Die A gave a value greater than Die B:", np.average(my_list[0,] > my_list[1,]), "\n Die B gave a value greater than Die C:", np.average(my_list[1,] > my_list[2,]), "\n Die C gave a value greater than Die A:", np.average(my_list[2,] > my_list[0,]))
```

Running this code gives the following:

```
2022-05-11 12:00:49,405 : CRITICAL : read_config : Config file .//python.config not found
```

3.4 Two Children

Example 2. (Two Children) Consider the following two problems:

1. Mr. Jones has two children. The older child is a girl. What is the probability that both children are girls?
2. Mr. Smith has two children. At least one of them is a girl. What is the probability that both children are girls?

Here are the possibilities of {older child, younger child}: {Boy, Boy}, {Boy, Girl}, {Girl, Boy}, {Girl,Girl}, with equal probabilities to occur.

For the first question: two elements satisfy the situation (the sample space for the given situation) – {Girl, Boy}, {Girl,Girl}; one element corresponds to the event of two girls (number of elements in the event of interest). Thus the probability of two girls is $\frac{1}{2}$. This question can also be solved this way: Since the older child is a girl, the probability of two girls is the probability that the younger child is a girl which is $\frac{1}{2}$.

For the second question, three elements satisfy the situation (the sample space for the given situation): {Boy, Girl}, {Girl, Boy}, {Girl,Girl}. Thus the probability is $\frac{1}{3}$.

If we simulate many families, we can find the numerical answer quite accurately.

3.5 Fair Division

Example 3. (Fair Division) Tom and Jerry, each put 30 dollar in a jackpot to start a game. Suppose that they have equal chance to win and the one who wins three times takes all the 60 dollars. Now, Tom has won twice and Jerry has won once, but then something happens and the game must be stopped. How should they split the 60 dollars according their probabilities of winning if the game was finished?

It seems Tom has won twice and Jerry has won once so the 60 dollars should be split as $40 : 20$. However, these are not proportional to their probabilities of winning. They need at most another two games to know the final winner. Here are the four possible results of the two games: {Tom, Tom}, {Tom, Jerry}, {Jerry, Tom}, {Jerry, Jerry}. The probabilities that Tom and Jerry would be the final winner are $\frac{3}{4}$ and $\frac{1}{4}$, respectively, so the fair division should be $45 : 15$.

Let's use simulation to solve this problem.

3.6 Birthday Problem

Example 4. (Birthday Problem) Suppose that a room contains 23 people. What is the probability that at least two of them have a common birthday? Assuming that each year has 365 days, this probability seems very small, but it is actually about 0.5. What is the probability that some one in that room has the same birthday as yours? This probability is quite small (≈ 0.061). In order to have the probability that someone's birthday is the same as yours to be 0.5, we need 253 random selected people to be in that room.

Let's use simulation to verify the aforementioned numbers.

3.7 Henry's Choice

Henry has been caught stealing cattle, and is brought into town for justice. The judge is his ex-wife Gretchen, who wants to show him some sympathy, but the law clearly calls for two shots to be taken at Henry from close range. To make things a little better for Henry, Gretchen tells him she will place two bullets into a six-chambered revolver in successive order. She will spin the chamber, close it, and take one shot. If Henry is still alive, she will then either take another shot, or spin the chamber again before shooting.

Henry is a bit incredulous that his own ex-wife would carry out the punishment, and a bit sad that she was always such a rule follower. He steels himself as Gretchen loads the chambers, spins the revolver, and pulls the trigger. Whew! It was blank. Then Gretchen asks, “Do you want me to pull the trigger again, or should I spin the chamber a second time before pulling the trigger?” What should Henry choose?

We know that the first chamber Gretchen fired was one of the four empty chambers. Since the bullets were placed in consecutive order, one of the empty chambers is followed by a bullet, and the other three empty chambers are followed by another empty chamber. So if Henry has Gretchen pull the trigger again, the probability that a bullet will be fired is 1/4.

If Gretchen spins the chamber again, the probability that she shoots Henry would be 2/6, or 1/3, since there are two possible bullets that would be in firing position out of the six possible chambers that would be in position.

3.8 Bertrand’s Box

Bertrand’s box paradox was first posed by Joseph Bertrand 1889. Here is the question: There are three boxes, one contains two gold coins, one contains two silver coins, and one contains a gold coin and a silver coin.

A box is selected at random and a coin is taken from that box at random. If the coin is a gold coin, what is the probability that the other coin in that box is also a gold coin.

3.9 Simpson’s Paradox

It is a phenomenon in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

Example 5. (An urn example) A black urn contains 5 red and 6 green balls, and a white urn contains 3 red and 4 green balls. You are allowed to choose an urn and then choose a ball at random from the urn. If you choose a red ball, you get a prize. Which urn should you choose to draw from?

- If you draw from the black urn, the probability of choosing a red ball is $5/11 = .455$.
- If you choose to draw from the white urn, the probability of choosing a red ball is $3/7 = .429$.

You should choose to draw from the black urn.

Now consider another game in which a second black urn has 6 red and 3 green balls, and a second white urn has 9 red and 5 green balls.

Men			Women		
Applicants	Admitted		Applicants	Admitted	
2691	1198	(**44.5**%)	1835	557	(30.4%)

	Men			Women		
Department	Applicants	Admitted		Applicants	Admitted	
A	825	512	(62%)	108	89	(82%)
B	560	353	(63%)	25	17	(68%)
C	325	120	(37%)	593	202	(34%)
D	417	138	(33%)	375	131	(35%)
E	191	53	(28%)	393	94	(24%)
F	373	22	(6%)	341	24	(7%)

- If you draw from the black urn, the probability of a red ball is 6/9 = .667.
- If you choose to draw from the white urn, the probability if 9/14 = .643.

Again you should choose to draw from the black urn.

In the final game, the contents of the second black urn are added to the first black urn, and the contents of the second white urn are added to the first white urn. Again, you can choose which urn to draw from. Which should you choose?

Intuition says choose the black urn, but let's calculate the probabilities.

- The black urn now contains 11 red and 9 green balls, so the probability of drawing a red ball from it is 11/20=0.55
- The white urn now contains 12 red and 9 green balls, so the probability of drawing a red ball from it is 12/21= .571.

You should choose the white urn!

Example 6. (UC Berkeley gender bias) A famous example of Simpson's paradox is a study of gender bias among graduate school admissions to University of California, Berkeley. In 1973 UC Berkeley was sued for sex-discrimination. Here are the overall numbers for the six largest departments in fall admission of 1973.

This table shows that men were more likely than women to be admitted, and the difference was so significant. Let's see which departments were mainly responsible for this gender bias. To do this we broke open the data according to each departments.

Things get strange after we divide the data according different departments. For the six departments, four of them accepted women more than men. To explain this, **?** noticed that women tended to apply to more competitive departments with low admission rates even among qualified applicants, whereas men tended to apply to less competitive departments with high admission rates among the qualified applicants.

To use simulation to further illustrate this, we simulate a data set with four groups in the following.

As seen from Figure **??**, for the whole data, y has an increase pattern as x increases, but for each sub group y has a decrease pattern as x increases.

3.10 100 prisoners problem

The probability may help to obtain some changes in a seemingly hopeless situation. Consider the following example modified from **?**.

In a prison, there are 100 death row prisoners who are numbered from 1 to 100, and there is a room with 100 drawers labeled from 1 to 100. The director randomly puts one prisoner's number in each closed drawer and offers a last chance. The prisoners enter the room, one after another. Each prisoner may open and look into 50 drawers in any order. The drawers are closed again afterwards. If, during this search, every prisoner finds his number in one of the drawers, all prisoners are pardoned. If some prisoner does not find his number, all prisoners die. Before the first prisoner enters the room, the prisoners may discuss strategy, but they cannot communicate once the first prisoner enters the room.

The situation is hopeless if every prisoner selects 50 drawers at random. The probability that a single prisoner finds his number is 0.5, so the probability that all prisoners find their numbers is $0.5^{100} = 7.89 \times 10^{-31} \approx 0$. However, a better strategy gives the prisoners more than 0.30 probability to survive [**?**]. The strategy is described below.

1. Each prisoner first opens the drawer with his own number.
2. If this drawer contains his number he is done and was successful.
3. Otherwise, the drawer contains the number of another prisoner and he next opens the drawer with this number.
4. The prisoner repeats steps 2 and 3 until he finds his own number or has opened 50 drawers.

In the following, we define two functions to simulate the method of randomly open 50 drawers and the better strategy, respectively.

3.11 Monty Hall problem

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

UPDATES

Visit `http://borisv.lk.net/latex.html` for updates, errata, and other information.

The fonts used in *Stumbling into data science* are New Baskerville, Futura, The Sans Mono Condensed, and Dogma. The book was typeset with LaTeX 2$_{\varepsilon}$ package `nostarch` by Boris Veytsman *(2008/06/06 v1.3 Typesetting books for No Starch Press).*

The book was produced as an example of the package `nostarch`.