Chester Ismay and Albert Y. Kim

ModernDive

Contents

1
Preamble

1.1 Principles of this Book

These are some principles we keep in mind. If you agree with them, this might be the book for you.

1. **Blur the lines between lecture and lab**
   - Laptops and open source software are rendering the lab/lecture dichotomy ever more archaic.
   - It's much harder for students to understand the importance of using the software if they only use it once a week or less. They forget the syntax in much the same way someone learning a foreign language forgets the rules.
2. **Focus on the entire data/science research pipeline**
   - Grolemund and Wickham's graphic
   - George Cobb argued for "Minimizing prerequisites to research"
3. **It's all about data, data, data**
   - We leverage R packages for rich/complex yet easy-to-load data sets.
   - We've heard it before: "You can't teach `ggplot2` for data visualization in intro stats!" We, like David Robinson, are more optimistic and we've had success doing so.
   - `dplyr` is a game changer for data manipulation: the verb describing your desired data action *is* the command name!
4. **Use simulation/resampling for intro stats, not probability/large sample approximation**
   - Reinforce concepts, not equations, formulas, and probability tables.
   - To this end, we're big fans of the `mosaic` package's `shuffle()`, `resample()`, and `do()` functions for sampling and simulation.
5. **Don't fence off students from the computation pool, throw them in!**
   - Don't teach them coding/programming per se, but computation and algorithmic thinking.

- Drawing Venn diagrams delineating statistics, computer science, and data science is also ever more archaic; embrace computation!

6. **Complete reproducibility**

- We find it frustrating when textbooks give examples but not the source code and the data itself. We not only give you the source code for all examples, but also the source code for the whole book!
- We encourage use of R Markdown to foster notions of reproducible research.
- **Ultimately the best textbook is one you've written yourself**
  - You best know your audience, their background, and their priorities and you know best your own style and types of examples and problems you like best. Customizability is the ultimate end.
  - A new paradigm for textbooks? Versions, not editions? Pull requests, crowd-sourcing, and development versions?

1.2 Contribute

- This book is in beta testing and is currently at Version 0.1.0.9000. If you would like to receive periodic updates on this book and other similar projects, please fill out this Google Form.
- The source code for this book is available for download/forking on GitHub. If you find typos or other errors or have suggestions on how to better word something in the book, please create a pull request too!
- Please feel free to modify the book as you wish for your own needs! All we ask is that you list the authors field above as "Chester Ismay, Albert Y. Kim, and YOU!"
- We'd also appreciate if you let us now what changes you've made and how you've used the textbook. We'd love some data on what's working well and what's not working so well.

1.3 Getting Started

This book was written using the **bookdown** R package from Yihui Xie. In order to follow along and run the code in this book on your own, you'll need to have access to R and RStudio. You can find more information on both of these with a simple Google search for "R" and for "RStudio." An introduction to using R, RStudio, and R Markdown is also available in a free book here (Ismay, 2016). It is recommended that you refer back to this book frequently as it has GIF screen recordings that you can follow along with as you learn.

We will keep a running list of R packages you will need to have installed to complete the analysis as well here in the `needed_pkgs` character vector. You can check if you have all of the needed packages installed by running all of the lines below. The last lines including the `if` will install them as needed (i.e., download their needed files from the internet to your hard drive).

You can run the `library` function on them to load them into your current analysis. Prior to each analysis where a package is needed, you will see the corresponding `library` function in the text. Make sure to check the top of the chapter to see if a package was loaded there.

```
needed_pkgs <- c("nycflights13", "dplyr", "ggplot2", "knitr",
  "okcupiddata", "dygraphs", "rmarkdown", "mosaic", "ggplot2movies")

new.pkgs <- needed_pkgs[!(needed_pkgs %in% installed.packages())]

if(length(new.pkgs)) {
  install.packages(new.pkgs, repos = "http://cran.rstudio.com")
}
```

Colophon

The source of the book is available here and was built with versions of R packages (and their dependent packages) given below. This may not be of importance for initial readers of this book, but the hope is you can reproduce a duplicate of this book by installing these versions of the packages.

package	*	version	date	source
assertthat		0.1	2013-12-06	CRAN (R 3.3.0)
backports		1.0.4	2016-10-24	cran (@1.0.4)
base64enc		0.1-3	2015-07-28	CRAN (R 3.3.0)
BH		1.62.0-1	2016-11-19	CRAN (R 3.3.2)
bitops		1.0-6	2013-08-17	CRAN (R 3.3.0)
caTools		1.17.1	2014-09-10	CRAN (R 3.3.0)
colorspace		1.3-2	2016-12-14	CRAN (R 3.3.2)
curl		2.3	2016-11-24	CRAN (R 3.3.2)
DBI		0.5-1	2016-09-10	CRAN (R 3.3.0)
dichromat		2.0-0	2013-01-24	CRAN (R 3.3.0)
digest		0.6.11	2017-01-03	CRAN (R 3.3.2)
dplyr		0.5.0	2016-06-24	CRAN (R 3.3.0)
dygraphs		1.1.1.4	2017-01-04	CRAN (R 3.3.2)
evaluate		0.10	2016-10-11	CRAN (R 3.3.0)
ggdendro		0.1-20	2016-04-27	cran (@0.1-20)
ggplot2		2.2.1	2016-12-30	CRAN (R 3.3.2)
ggplot2movies		0.0.1	2015-08-25	CRAN (R 3.3.0)
gridExtra		2.2.1	2016-02-29	CRAN (R 3.3.0)
gtable		0.2.0	2016-02-26	CRAN (R 3.3.0)
highr		0.6	2016-05-09	CRAN (R 3.3.0)
hms		0.3	2016-11-22	CRAN (R 3.3.2)
htmltools		0.3.5	2016-03-21	CRAN (R 3.3.0)
htmlwidgets		0.8	2016-11-09	CRAN (R 3.3.2)
jsonlite		1.2	2016-12-31	CRAN (R 3.3.2)
knitr		1.15.1	2016-11-22	CRAN (R 3.3.2)

labeling	0.3	2014-08-23	CRAN (R 3.3.0)
lattice	0.20-34	2016-09-06	CRAN (R 3.3.2)
latticeExtra	0.6-28	2016-02-09	cran (@0.6-28)
lazyeval	0.2.0	2016-06-12	CRAN (R 3.3.0)
magrittr	1.5	2014-11-22	CRAN (R 3.3.0)
markdown	0.7.7	2015-04-22	CRAN (R 3.3.0)
MASS	7.3-45	2016-04-21	CRAN (R 3.3.2)
Matrix	1.2-7.1	2016-09-01	CRAN (R 3.3.2)
mime	0.5	2016-07-07	CRAN (R 3.3.0)
mosaic	0.14.4	2016-11-05	Github (ProjectMOSAIC/mosaic@c0b1f10)
mosaicData	0.14.0	2016-06-17	cran (@0.14.0)
munsell	0.4.3	2016-02-13	CRAN (R 3.3.0)
nycflights13	0.2.1	2016-12-30	CRAN (R 3.3.2)
okcupiddata	0.1.0	2016-08-19	local
plyr	1.8.4	2016-06-08	CRAN (R 3.3.0)
R6	2.2.0	2016-10-05	CRAN (R 3.3.0)
RColorBrewer	1.1-2	2014-12-07	CRAN (R 3.3.0)
Rcpp	0.12.8	2016-11-17	CRAN (R 3.3.2)
readr	1.0.0	2016-08-03	CRAN (R 3.3.0)
reshape2	1.4.2	2016-10-22	CRAN (R 3.3.0)
rmarkdown	1.3	2016-12-21	CRAN (R 3.3.2)
rprojroot	1.1	2016-10-29	cran (@1.1)
scales	0.4.1	2016-11-09	CRAN (R 3.3.2)
stringi	1.1.2	2016-10-01	CRAN (R 3.3.0)
stringr	1.1.0	2016-08-19	CRAN (R 3.3.0)
tibble	1.2	2016-08-26	CRAN (R 3.3.0)
tidyr	0.6.0	2016-08-12	CRAN (R 3.3.0)
xts	0.9-7	2014-01-02	CRAN (R 3.3.0)
yaml	2.1.14	2016-11-12	CRAN (R 3.3.2)
zoo	1.7-14	2016-12-16	CRAN (R 3.3.2)

Book was last updated:

```
## [1] "By aykim on Monday, January 09, 2017 00:30:34 PST"
```

2
Introduction

Part I

Data Exploration

3
Data Visualization via `ggplot2`

4
Data Manipulation via `dplyr`

Part II

Inference

5
Hypothesis Testing

6
Confidence Intervals

7

Regression via `broom`

Part III

Conclusion

A
Inference Examples

B

Reach for the Starts

C

Placeholder

D
References

Ismay, C. (2016). *Getting used to R, RStudio, and R Markdown.*