Chester Ismay, Albert Y. Kim and Hendrik Feddersen

HR Analytics in R

Common tasks achieved with the power of R

Contents

List of Tables

List of Figures

1

Introduction

Please note that you are currently looking at the latest version of "HR Analytics in R". However, since work is still in progress it is subject to frequent change.

The intention of this book is to encourage more 'data driven' decisions in HR. HR Analytics is not anymore a nice-to-have addon but rather the way HR practitioners should conduct HR decision making in the future. Where applicable, human judgement is 'added' onto a rigorous analysis of the data done in the first place.

To achieve this ideal world, I need to equip you with some fundamental knowledge of R and RStudio, which are open-source tools for data scientists. I am well aware that on one side you want to do something for your career in HR, however you are most likely completely new to coding.

Help! I'm new to R and RStudio. I'm completely new to coding! What shall I do?

If you're asking yourself this question, then you have come to the right place! There is no better moment to ride the wave of disruptions taking place now in HR.

- *Are you looking to learn about HR Analytics utilising the power of R"? Then start with our Introduction for Students.*
- *Are you looking to contribute to "HR Analytics in R"? Then click here for information on how.*
- *Are you curious about the publishing of this book? Then click here for more information on the open-source technology, in particular R Markdown and the bookdown package.*

This is version 0.5.0 of "HR Analytics in R" published on May 16, 2019. While a PDF version of this book can be found here[1], this is very much a work in progress with many things that still need to be fixed. I appreciate your patience.

[1] hendrikfeddersen.pdf

1.1 Introduction for students

This book assumes no prerequisites: no algebra, no calculus, and no prior programming/coding experience. This is intended to be a gentle introduction to the practice of analyzing data and answering questions using data the way data scientists, statisticians and other researchers would.

Working with the material

You can work your way through the materials by clicking on the arrows to the left and right at the bottom of each page. Alternatively, there is a collapsible contents bar on the left hand side.

If you need to find something specific, you can use the search icon. Typing in a word or phrase will filter the contents bar to relevant sections.

The book by default renders black sans serif on a white background. You can use the A to amend the appearance of the book to make it easier to process, whether that's a larger font, a serif font, or a different colour scheme.

The edit button takes you straight to github, where you can propose editorial changes.

Conventions

Throughout this book various conventions will be used.

In terms of basic formatting:

- This is standard text.
- `This is code or a symbol`
- *This is a Keyboard Key!*
- **This is the first time I mention something important**

This is a book about coding, so expect code blocks. Code blocks will typically look like this:

```
"this is a code block"
```

```
[1] "this is a code block"
```

Directly underneath it, normally starting with two hash symbols (##) is the result of the code executing.

```
## [1] 'this is a code block'`
```

There will also be callouts throughout the book. Some are for information, some expect you to do things.

Anything written here should be read carefully before proceeding.

This is a tip relating to what I've just said.

This is kind of like a tip but is for when you're getting into trouble and need help.

This is something I recommend you do as you're reading.

In Figure 1.1 I present a flowchart of what you'll cover in this book. You'll first get started with data in Chapter 2, where you'll learn about the difference between R and RStudio, start coding in R, understand what R packages are, and explore your first dataset: all domestic departure flights from a New York City airport in 2013. Then

1. **Data science**: You'll assemble your data science toolbox using `tidyverse` packages. In particular:
    - Ch.3: Visualizing data via the `ggplot2` package.
    - Ch.5: Understanding the concept of "tidy" data as a standardized data input format for all packages in the `tidyverse`
    - Ch.4: Wrangling data via the `dplyr` package.
2. **Data modeling**: Using these data science tools and helper functions from the `moderndive` package, you'll start performing data modeling. In particular:
    - Ch.6: Constructing basic regression models.
    - Ch.7: Constructing multiple regression models.
3. **Statistical inference**: Once again using your newly acquired data science tools, I'll unpack statistical inference using the `infer` package. In particular:
    - Ch.8: Understanding the role that sampling variability plays in statistical inference using both tactile and virtual simulations of sampling from a "bowl" with an unknown proportion of red balls.
    - Ch.9: Building confidence intervals.
    - Ch.10: Conducting hypothesis tests.
4. **Data modeling revisited**: Armed with your new understanding of statistical inference, you'll revisit and review the models you constructed in Ch.6 & Ch.7. In particular:

- Ch.11: Interpreting both the statistical and practice significance of the results of the models.
- Ch.12: I'll end the introductory chapters with a discussion on what it means to "think with data" and present an example case study data analysis of house prices in Seattle.

5. **HR Analytics - data driven decision making**: The intention is to provide real tangible examples of the application of data science to HR, to illustrate the data science process in the HR context, and to show that the scope mentioned previously in this article, isn't just theoretical - it's real. The last and most important module shall illustrate current best practices of a structured process of thinking and analysis.

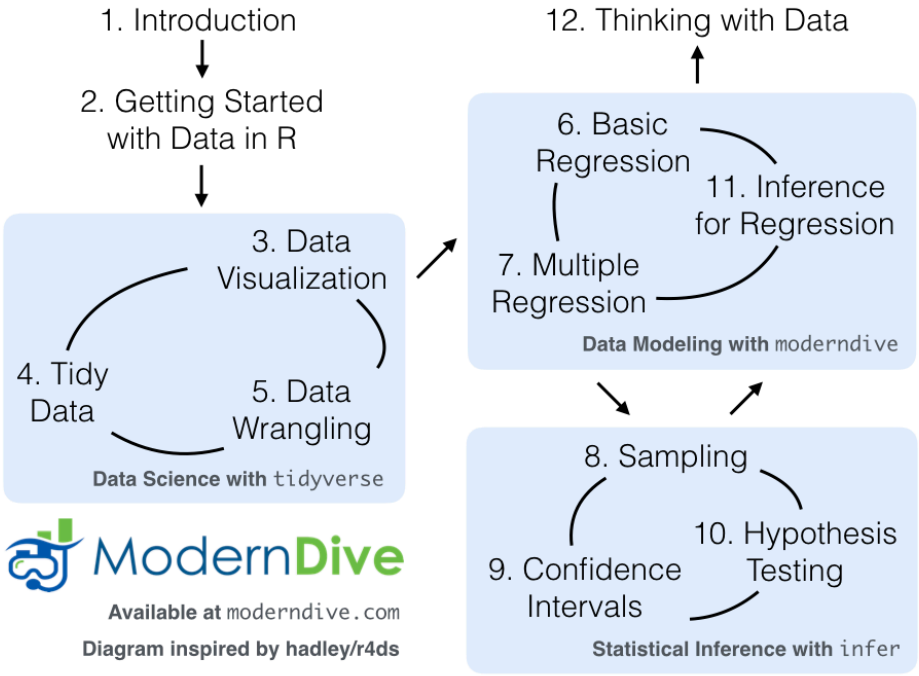

FIGURE 1.1: ModernDive Flowchart

1.1.1 What you will learn from this book

I hope that by the end of this book, you'll have learned

1. How to use R to explore data.

2. How to answer statistical questions using tools like confidence intervals and hypothesis tests.
3. How to effectively create "data stories" using these tools.

What do I mean by data stories? I mean any analysis involving data that engages the reader in answering questions with careful visuals and thoughtful discussion, such as How strong is the relationship between per capita income and crime in Chicago neighborhoods?[2] and How many f**ks does Quentin Tarantino give (as measured by the amount of swearing in his films)?[3]. Further discussions on data stories can be found in this Think With Google article[4].

For other examples of data stories constructed by students like yourselves, look at the final projects for two courses that have previously used ModernDive:

- Middlebury College MATH 116 Introduction to Statistical and Data Sciences[5] using student collected data.
- Pacific University SOC 301 Social Statistics[6] using data from the fivethirtyeight R package[7].

This book will help you develop your "data science toolbox", including tools such as data visualization, data formatting, data wrangling, and data modeling using regression. With these tools, you'll be able to perform the entirety of the "data/science pipeline" while building data communication skills (see Subsection 1.1.2 for more details).

In particular, this book will lean heavily on data visualization. In today's world, we are bombarded with graphics that attempt to convey ideas. I will explore what makes a good graphic and what the standard ways are to convey relationships with data. You'll also see the use of visualization to introduce concepts like mean, median, standard deviation, distributions, etc. In general, I'll use visualization as a way of building almost all of the ideas in this book.

To impart the statistical lessons in this book, I have intentionally minimized the number of mathematical formulas used and instead have focused on developing a conceptual understanding via data visualization, statistical computing, and simulations. I hope this is a more intuitive experience than the way

[2] http://rpubs.com/ry_lisa_elana/chicago

[3] https://ismayc.github.io/soc301_s2017/group_projects/group4.html

[4] https://www.thinkwithgoogle.com/marketing-resources/data-measurement/tell-meaningful-stories-with-data/

[5] https://rudeboybert.github.io/MATH116/PS/final_project/final_project_outline.html#past_examples

[6] https://ismayc.github.io/soc301_s2017/group-projects/index.html

[7] https://cran.r-project.org/web/packages/fivethirtyeight/vignettes/fivethirtyeight.html

statistics has traditionally been taught in the past and how it is commonly perceived.

Finally, you'll learn the importance of literate programming. By this I mean you'll learn how to write code that is useful not just for a computer to execute but also for readers to understand exactly what your analysis is doing and how you did it. This is part of a greater effort to encourage reproducible research (see Subsection 1.1.3 for more details). Hal Abelson coined the phrase that I will follow throughout this book:

> "Programs must be written for people to read, and only incidentally for machines to execute."

I understand that there may be challenging moments as you learn to program. Both of us continue to struggle and find ourselves often using web searches to find answers and reach out to colleagues for help. In the long run though, we all can solve problems faster and more elegantly via programming. I wrote this book as our way to help you get started and you should know that there is a huge community of R users that are always happy to help everyone along as well. This community exists in particular on the internet on various forums and websites such as stackoverflow.com[8].

1.1.2 Data/science pipeline

You may think of statistics as just being a bunch of numbers. I commonly hear the phrase "statistician" when listening to broadcasts of sporting events. Statistics (in particular, data analysis), in addition to describing numbers like with baseball batting averages, plays a vital role in all of the sciences. You'll commonly hear the phrase "statistically significant" thrown around in the media. You'll see articles that say "Science now shows that chocolate is good for you." Underpinning these claims is data analysis. By the end of this book, you'll be able to better understand whether these claims should be trusted or whether we should be wary. Inside data analysis are many sub-fields that I will discuss throughout this book (though not necessarily in this order):

- data collection
- data wrangling
- data visualization

[8] https://stackoverflow.com/

- data modeling
- inference
- correlation and regression
- interpretation of results
- data communication/storytelling

These sub-fields are summarized in what Grolemund and Wickham term the "Data/Science Pipeline"[9] in Figure 1.2.

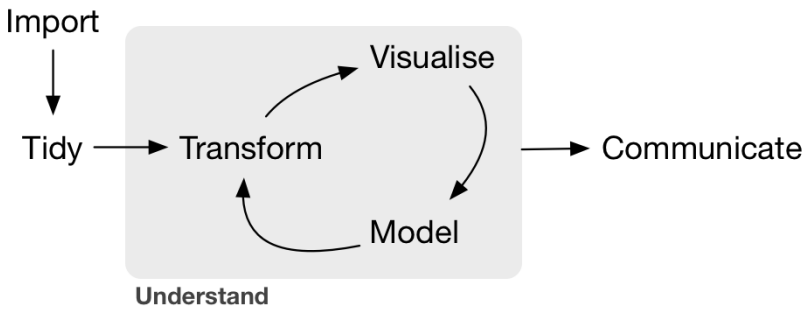

FIGURE 1.2: Data/Science Pipeline

I will begin by digging into the gray **Understand** portion of the cycle with data visualization, then with a discussion on what is meant by tidy data and data wrangling, and then conclude by talking about interpreting and discussing the results of our models via **Communication**. These steps are vital to any statistical analysis. But why should you care about statistics? "Why did they make me take this class?"

There's a reason so many fields require a statistics course. Scientific knowledge grows through an understanding of statistical significance and data analysis. You needn't be intimidated by statistics. It's not the beast that it used to be and, paired with computation, you'll see how reproducible research in the sciences particularly increases scientific knowledge.

1.1.3 Reproducible research

> "The most important tool is the *mindset*, when starting, that the end product will be reproducible." – Keith Baggerly

---

[9] http://r4ds.had.co.nz/explore-intro.html

Another goal of this book is to help readers understand the importance of reproducible analyses. The hope is to get readers into the habit of making their analyses reproducible from the very beginning. This means I'll be trying to help you build new habits. This will take practice and be difficult at times. You'll see just why it is so important for you to keep track of your code and well-document it to help yourself later and any potential collaborators as well.

Copying and pasting results from one program into a word processor is not the way that efficient and effective scientific research is conducted. It's much more important for time to be spent on data collection and data analysis and not on copying and pasting plots back and forth across a variety of programs.

In a traditional analyses if an error was made with the original data, we'd need to step through the entire process again: recreate the plots and copy and paste all of the new plots and our statistical analysis into your document. This is error prone and a frustrating use of time. I'll see how to use R Markdown to get away from this tedious activity so that we can spend more time doing science.

"We are talking about *computational* reproducibility." - Yihui Xie

---

Reproducibility means a lot of things in terms of different scientific fields. Are experiments conducted in a way that another researcher could follow the steps and get similar results? In this book, I will focus on what is known as **computational reproducibility**. This refers to being able to pass all of one's data analysis, data-sets, and conclusions to someone else and have them get exactly the same results on their machine. This allows for time to be spent interpreting results and considering assumptions instead of the more error prone way of starting from scratch or following a list of steps that may be different from machine to machine.

1.1.4 Final note for students

At this point, if you are interested in instructor perspectives on this book, ways to contribute and collaborate, or the technical details of this book's construction and publishing, then continue with the rest of the chapter below. Otherwise, let's get started with R and RStudio in Chapter 2!

1.2 Introduction for instructors

This book is inspired by the following books:

- "Mathematical Statistics with Resampling and R" (Chihara and Hesterberg, 2011),
- "OpenIntro: Intro Stat with Randomization and Simulation" (Diez et al., 2014), and
- "R for Data Science" (Grolemund and Wickham, 2016).

The first book, while designed for upper-level undergraduates and graduate students, provides an excellent resource on how to use resampling to impart statistical concepts like sampling distributions using computation instead of large-sample approximations and other mathematical formulas. The last two books are free options to learning introductory statistics and data science, providing an alternative to the many traditionally expensive introductory statistics textbooks.

When looking over the large number of introductory statistics textbooks that currently exist, I found that there wasn't one that incorporated many newly developed R packages directly into the text, in particular the many packages included in the `tidyverse`[10] collection of packages, such as `ggplot2`, `dplyr`, `tidyr`, and `broom`. Additionally, there wasn't an open-source and easily reproducible textbook available that exposed new learners all of three of the learning goals listed at the outset of Subsection 1.1.1.

1.2.1 Who is this book for?

This book is intended for instructors of traditional introductory statistics classes using RStudio, either the desktop or server version, who would like to inject more data science topics into their syllabus. I assume that students taking the class will have no prior algebra, calculus, nor programming/coding experience.

Here are some principles and beliefs I kept in mind while writing this text. If you agree with them, this might be the book for you.

1. **Blur the lines between lecture and lab**
    - With increased availability and accessibility of laptops and open-source non-proprietary statistical software, the strict dichotomy between lab and lecture can be loosened.

[10] `http://tidyverse.org/`

- It's much harder for students to understand the importance of using 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. Frequent reinforcement is key.

2. **Focus on the entire data/science research pipeline**
   - I believe that the entirety of Grolemund and Wickham's data/science pipeline[11] should be taught.
   - I believe in "minimizing prerequisites to research"[12]: students should be answering questions with data as soon as possible.
3. **It's all about the data**
   - I leverage R packages for rich, real, and realistic data-sets that at the same time are easy-to-load into R, such as the `nycflights13` and `fivethirtyeight` packages.
   - I believe that data visualization is a gateway drug for statistics[13] and that the Grammar of Graphics as implemented in the `ggplot2` package is the best way to impart such lessons. However, I often hear: "You can't teach `ggplot2` for data visualization in intro stats!" I, like David Robinson[14], are much more optimistic.
   - `dplyr` has made data wrangling much more accessible[15] to novices, and hence much more interesting data-sets can be explored.
4. **Use simulation/resampling to introduce statistical inference, not probability/mathematical formulas**
   - Instead of using formulas, large-sample approximations, and probability tables, statistical concepts using resampling-based inference.
   - This allows for a de-emphasis of traditional probability topics, freeing up room in the syllabus for other topics.
5. **Early exposure to analytics and computing**
   - Computing skills are essential to working with data in the 21st century even for HR managers. Given this fact, I feel that an early exposure to computing can only be of benefit to the whole HR community.
   - I am not teaching a course on coding/programming per se, but rather just enough of the computational and algorithmic thinking necessary for performing a data analysis in HR.
6. **Complete reproducibility and customisability**

[11] http://r4ds.had.co.nz/introduction.html
[12] https://arxiv.org/abs/1507.05346
[13] http://escholarship.org/uc/item/84v3774z
[14] http://varianceexplained.org/r/teach_ggplot2_to_beginners/
[15] http://chance.amstat.org/2015/04/setting-the-stage/

- I am frustrated when people talk about HR Analytics, without giving the source code and the data itself. I give you the source code for all examples as well as the whole book!
- If you want you can even use my book as a starting point and customise for your own non-profit training. For more about how to make this book your own, see About this Book.

1.3 Connect and contribute

If you would like to connect with "HR Analytics in R", check out the following links:

- If you would like to receive periodic updates about HR Analytics, then please sign up for my mailing list[16]. You will receive receive bi-weekly notifications about my new blog posts.
- Please feel free to contact me at `info@hranalytics.live`[17] .
- I am on Twitter at h_feddersen[18].

If you would like to contribute to "HR Analytics in R", there are many ways! Let's all work together to make this book as great as possible for as many students as possible!

- Please let me know if you find any errors, typos, or areas from improvement on my GitHub issue page[19] page. I will fix it as soon as possible.
- If you are familiar with GitHub and would like to contribute even more, please see Section 1.4 below.

I would like to thank Moderndive[20] for their inspirational presentation at a recent R user conference and for their generous example on how to set up a bookdown book and for their introductory pages on how to start using R.

[16] https://hranalytics.live/signup/
[17] mailto:info@hranalytics.live
[18] https://twitter.com/h_feddersen
[19] https://github.com/Hendrik147/HR_Analytics_in_R_book/issues
[20] https://github.com/moderndive/moderndive_book

1.4 About this book

This book was written using RStudio's bookdown[21] package by Yihui Xie (Xie, 2018). This package simplifies the publishing of books by having all content written in R Markdown[22].

- **Latest published version, still in development** The most up-to-date version, which is still in development is available at https://hranalyticslive.netlify.com/
- **Source code** The bookdown/R Markdown source code for the latest version of "HR Analytics in R" is available on Hendrik Feddersen's GitHub repository page[23]
- **Usage** You can share this material with colleagues or for non-commercial purposes but you can't resell or incorporate them into stuff you make money from.
  - As a symbol of gratitude, I would expect you at least to sign up for my mailing list[24].
  - If you think my material is awesome and want to use it for commercial purposes, please contact me at info@hranalytics.live[25]
- **Licence** This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

FIGURE 1.3: Creative Commons License

1.5 About the author

Who am I?

[21] https://bookdown.org/
[22] http://rmarkdown.rstudio.com/html_document_format.html
[23] https://github.com/Hendrik147/HR_Analytics_in_R_book
[24] https://hranalytics.live/signup/
[25] mailto:info@hranalytics.live

HF Snappy.bb

- I am Hendrik Feddersen, a long-standing HR practitioner passionate about HR Analytics and living in Amsterdam, Netherlands.
  - Email: info@hranalytics.live[26]
  - Webpage: https://hranalytics.live/
  - Twitter: h_feddersen[27]
  - GitHub: https://github.com/Hendrik147

[26] mailto:info@hranalytics.live

[27] https://twitter.com/h_feddersen

2

Getting Started with Data in R

Placeholder

Part I

Data Science via the tidyverse

3

Data Visualization

Placeholder

Needed packages

3.1 The Grammar of Graphics

3.1.1 Components of the Grammar

3.1.2 Gapminder data

3.1.3 Other components

3.1.4 ggplot2 package

3.2 Five Named Graphs - The 5NG

3.3 5NG#1: Scatterplots

3.3.1 Scatterplots via geom_point

3.3.2 Over-plotting

3.3.3 Summary

3.4 5NG#2: Linegraphs

3.4.1 Linegraphs via geom_line

3.4.2 Summary

3.5 5NG#3: Histograms

3.5.1 Histograms via geom_histogram

3.5.2 Adjusting the bins

3.5.3 Summary

3.6 Facets

3.7 5NG#4: Boxplots

3.7.1 Boxplots via geom_boxplot

3.7.2 Summary

3.8 5NG#5: Barplots

3.8.1 Barplots via geom_bar or geom_col

3.8.2 Must avoid pie charts!

3.8.3 Two categorical variables

3.8.4 Summary

4

Data Wrangling

Placeholder

Needed packages

4.1 The pipe operator: `%>%`

4.2 `filter` rows

4.3 `summarize` variables

4.4 `group_by` rows

4.4.1 Grouping by more than one variable

4.5 `mutate` existing variables

4.6 `arrange` and sort rows

4.7 `join` data frames

4.7.1 Matching "key" variable names

4.7.2 Different "key" variable names

4.7.3 Multiple "key" variables

4.7.4 Normal forms

4.8 Other verbs

4.8.1 `select` variables

4.8.2 `rename` variables

4.8.3 `top_n` values of a variable

4.9 Conclusion

4.9.1 Summary table

4.9.2 Additional resources

4.9.3 What's to come?

5

Data Importing & "Tidy" Data

Placeholder

Needed packages

5.1 Importing data

5.1.1 Using the console

5.1.2 Using RStudio's interface

5.2 Tidy data

5.2.1 Definition of "tidy" data

5.2.2 Converting to "tidy" data

5.2.3 `nycflights13` package

5.3 Case study: Democracy in Guatemala

5.4 Conclusion

5.4.1 `tidyverse` package

5.4.2 Additional resources

5.4.3 What's to come?

Part II

Data Modeling via moderndive

6

Basic Regression

Placeholder

Needed packages

6.1 One numerical explanatory variable

6.1.1 Exploratory data analysis

6.1.2 Simple linear regression

6.1.3 Observed/fitted values and residuals

6.2 One categorical explanatory variable

6.2.1 Exploratory data analysis

6.2.2 Linear regression

6.2.3 Observed/fitted values and residuals

6.3 Related topics

6.3.1 Correlation is not necessarily causation

6.3.2 Best fitting line

6.3.3 `get_regression_x()` functions

6.4 Conclusion

6.4.1 Additional resources

6.4.2 What's to come?

7

Multiple Regression

Placeholder

Needed packages

7.1 One numerical & one categorical explanatory variable

7.1.1 Exploratory data analysis

7.1.2 Interaction model

7.1.3 Parallel slopes model

7.1.4 Observed/fitted values and residuals

7.2 Two numerical explanatory variables

7.2.1 Exploratory data analysis

7.2.2 Regression plane

7.2.3 Observed/fitted values and residuals

7.3 Related topics

7.3.1 Model selection

7.3.2 Correlation coefficient

7.3.3 Simpson's Paradox

7.4 Conclusion

7.4.1 Additional resources

7.4.2 What's to come?

Part III

Statistical inference via infer

8

Sampling

Placeholder

Needed packages

8.1 Sampling activity

8.1.1 What proportion of this bowl's balls are red?

8.1.2 Using the shovel once

8.1.3 Using the shovel 33 times

8.1.4 What are we doing here?

8.2 Computer simulation of sampling

8.2.1 Using the virtual shovel once

8.2.2 Using the virtual shovel 33 times

8.2.3 Using the virtual shovel 1000 times

8.2.4 Using different shovels

8.3 Sampling framework

8.3.1 Terminology & notation

8.3.2 Statistical definitions

8.3.3 The moral of the story

8.4 Case study: Polls

8.5 Conclusion

8.5.1 Random sampling vs random assignment

8.5.2 Central Limit Theorem

8.5.3 Summary table

8.5.4 Additional resources

8.5.5 What's to come?

9

Confidence Intervals

Placeholder

Needed packages

9.1 Resampling activity

9.1.1 What is the average year of circulated US pennies in 2019?

Exploratory data analysis on original sample

9.1.2 Using resampling once

Exploratory data analysis on the resample

9.1.3 Using resampling 33 times

9.1.4 What's the plan?

9.2 Computer simulation of resampling

9.2.1 Using the virtual resample once

9.2.2 Using the virtual resample 33 times

9.2.3 Using the virtual resample 1000 times

9.3 Confidence interval build-up

9.3.1 The percentile method

9.3.2 The standard error method

9.4 The bootstrapping framework

9.4.1 The original workflow needed for this

9.4.2 The infer package for statistical inference

Specify variables

Generate replicates

Calculate summary statistics

Observed statistic / point estimate calculations

Visualize the results

9.4.3 Building confidence intervals with the infer package

9.4.4 The percentile method with infer

9.4.5 The standard error method with infer

9.5 Case study: Revisiting the red ball example

9.5.1 Observed statistic

9.5.2 Bootstrap distribution for one proportion

9.6 Interpreting the confidence interval

10

Hypothesis Testing

Placeholder

Needed packages

10.1 Hypothesis testing activity

10.1.1 Question of interest

10.1.2 What did we actually observe?

10.1.3 Using permuting once

10.1.4 Using permuting 33 times

10.2 Hypothesis testing with infer

10.2.1 Revisiting the infer verb framework

10.2.2 The `infer` pipeline for the activity

Choose the variables of interest

Set the model for the null hypothesis

Replicate samples assuming the null hypothesis is true

Compute the statistic for each replicate

10.2.3 The "There Is Only One Test" framework

10.3 The p-value

10.3.1 Corresponding confidence interval

10.3.2 Summary

10.4 Interpretation of hypothesis testing results

10.4.1 Criminal trial analogy

Two possible conclusions

10.4.2 Types of errors in hypothesis testing

Logic of hypothesis testing

10.4.3 Statistical significance

10.5 Case study: comparing two means

10.5.1 Randomization/permutation

10.5.2 Comparing action and romance movies

10.5.3 Sampling $\rightarrow$ randomization

10.5.4 Data

10.5.5 Model of H_0

10.5.6 Test statistic δ

10.5.7 Observed effect δ^*

11

Inference for Regression

Placeholder

Needed packages

11.1 Simulation-based Inference for Regression

11.1.1 Data

11.1.2 Test statistic δ

11.1.3 Observed effect δ^*

11.1.4 Model of H_0

11.1.5 Simulated data

11.1.6 Distribution of δ under H_0

11.1.7 The p-value

11.2 Bootstrapping for the regression slope

11.3 Inference for multiple regression

11.3.1 Refresher: Professor evaluations data

11.3.2 Refresher: Visualizations

11.3.3 Refresher: Regression tables

11.3.4 Script of R code

11.4 Residual analysis

11.4.1 Residual analysis

11.4.2 Residual analysis

11.4.3 Residual analysis

11.4.4 Residual analysis

Part IV

Learnings so far

12

Thinking with Data

Placeholder

Needed packages

12.1 Case study: Seattle house prices

12.1.1 Exploratory data analysis (EDA)

12.1.2 log10 transformations

12.1.3 EDA Part II

12.1.4 Regression modeling

12.1.5 Inference for regression

12.1.6 Making predictions

12.2 Case study: Effective data storytelling

12.2.1 Bechdel test for Hollywood gender representation

12.2.2 US Births in 1999

12.2.3 Other examples

12.2.4 Script of R code

Concluding remarks

Part V

Practical examples about HR Analytics

13

Gender Pay Gap

Placeholder

13.1 Data Cleaning and Prep.

13.2 Summary Statistics by gender

13.3 Avoiding Simpson's Paradox

13.4 Model Estimation: OLS with controls.

13.4.1 Logarithm of Base Pay

13.4.2 Results by Department

13.4.3 Results by Job Title

14

Stop the formal performance management process

Placeholder

15

HR Service Desk

Placeholder

16

Personality insights

Placeholder

17

Commuting time

Placeholder

18

Module - Organisational network analysis

Placeholder

19

Job classification analysis

Placeholder

19.0.1 Final evaluation of the various models

20

Masking HR data

Placeholder

20.0.1 Whitehouse dataset

20.0.2 Fertility dataset

21

Absenteeism MFG

Placeholder

21.0.1 For expediency we will delete the problem records in the dataset.

22

Absenteeism at work

Placeholder

22.1 Data reading

23

Accidents at work

Placeholder

24

Attrition

Here we introduce attrition.

25

Interview attendance problem

Placeholder

25.1 Data reading

25.2 Choosing a model

26

Ranking Medical Schools

Placeholder

27

Webscraping LinkedIn

Placeholder

28

HR dashboards

Here we introduce flexdashboards

29

HR Analytics product with Shiny

Shiny is a very powerful framework for building web applications based on R. It is out of the scope of this book to make a comprehensive introduction to Shiny (which is too big a topic). We recommend that readers who are not familiar with Shiny learn more about it from the website https://shiny.rstudio.com before reading this chapter.

Unlike the more traditional workflow of creating static reports, you can create documents that allow your readers to change the parameters underlying your analysis and see the results immediately in Shiny R Markdown documents. In the example shown in Figure **?**(fig:shiny), the histogram will be automatically updated to reflect the number of bins selected by the reader.

A picture is worth a thousand words, and a Shiny document can potentially show you a thousand pictures as you interact with it. The readers are no longer tied to the fixed analysis and conclusions in the report. They may explore other possibilities by themselves, and possibly make new discoveries or draw different conclusions.

A

Statistical Background

Placeholder

A.1 Basic statistical terms

A.1.1 Mean

A.1.2 Median

A.1.3 Standard deviation

A.1.4 Five-number summary

A.1.5 Distribution

A.1.6 Outliers

A.2 Normal distribution

B

Inference Examples

Placeholder

Needed packages

B.1 Inference mind map

B.2 One mean

B.2.1 Problem statement

B.2.2 Competing hypotheses

In words

In symbols (with annotations)

Set α

B.2.3 Exploring the sample data

Guess about statistical significance

B.2.4 Non-traditional methods

Bootstrapping for hypothesis test

B.2.4.0.1 Calculate p-value

Bootstrapping for confidence interval

B.2.5 Traditional methods

Check conditions

Test statistic

B.2.5.0.1 Observed test statistic

Compute p-value

State conclusion

Confidence interval

B.2.6 Comparing results

B.3 One proportion

B.3.1 Problem statement

B.3.2 Competing hypotheses

In words

In symbols (with annotations)

Set α

B.3.3 Exploring the sample data

Guess about statistical significance

B.3.4 Non-traditional methods

Simulation for hypothesis test

B.3.4.0.1 Calculate p-value

C

Reach for the Stars

Placeholder

Needed packages

C.1 Sorted barplots

C.2 Interactive graphics

C.2.1 Interactive linegraphs

D

Learning Check Solutions

Placeholder

E

Archive HR datasets

Placeholder

E.1 Gender Pay Gap

E.2 Overhead value analysis

E.3 HR Service Desk

E.4 HR recruitment, selection and performance data

E.5 Job classification

E.6 Job classification

E.7 Absenteeism at work

E.8 Job classification

Bibliography

Chihara, L. M. and Hesterberg, T. C. (2011). *Mathematical Statistics with Resampling and R*. John Wiley and Sons, Hoboken, NJ.

Diez, D. M., Barr, C. D., and Çetinkaya Rundel, M. (2014). *Introductory Statistics with Randomization and Simulation*. First edition edition.

Grolemund, G. and Wickham, H. (2016). *R for Data Science*.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.9.