

**Statistical Inference via Data Science: A ModernDive into R and the Tidyverse** provides a pathway for learning about statistical inference using data science tools widely used in industry, academia, and government. It introduces the tidyverse suite of R packages, including the ggplot2 package for data visualization, and the dplyr package for data wrangling. After equipping readers with just enough of these data science tools to perform effective exploratory data analyses, the book covers traditional introductory statistics topics like confidence intervals, hypothesis testing, and multiple regression modeling, while focusing on visualization throughout.

## Features

- Assumes minimal prerequisites, notably, no prior calculus nor coding experience
- Motivates theory using real-world data, including all domestic flights leaving New York City in 2013, the Gapminder project, and the data journalism website, FiveThirtyEight.com
- Centers on simulation-based approaches to statistical inference rather than mathematical formulas
- Uses the infer package for “tidy” and transparent statistical inference to construct confidence intervals and conduct hypothesis tests via the bootstrap and permutation methods
- Provides all code and output embedded directly in the text; also available in the online version at [moderndive.com](http://moderndive.com)

This book is intended for individuals who would like to simultaneously start developing their data science toolbox and start learning about the inferential and modeling tools used in much of modern-day research. The book can be used in methods and data science courses and first courses in statistics, at both the undergraduate and graduate levels.

**Chester Ismay** is a Data Science Evangelist for DataRobot and is based in Portland, Oregon, USA.

**Albert Y. Kim** is an Assistant Professor of Statistical and Data Sciences at Smith College in Northampton, Massachusetts, USA.



**CRC Press**  
Taylor & Francis Group  
an informa business

[www.crcpress.com](http://www.crcpress.com)

CRC Press titles are available as eBook editions in a range of digital formats

STATISTICS

ISBN: 978-0-367-40982-1



9 780367 409821



|                                                            |           |
|------------------------------------------------------------|-----------|
| Foreword                                                   | xv        |
| Preface                                                    | xvii      |
| About the authors                                          | xxix      |
| <b>1 Getting Started with Data in R</b>                    | <b>1</b>  |
| 1.1 What are R and RStudio? . . . . .                      | 1         |
| 1.1.1 Installing R and RStudio . . . . .                   | 2         |
| 1.1.2 Using R via RStudio . . . . .                        | 3         |
| 1.2 How do I code in R? . . . . .                          | 4         |
| 1.2.1 Basic programming concepts and terminology . . . . . | 4         |
| 1.2.2 Errors, warnings, and messages . . . . .             | 6         |
| 1.2.3 Tips on learning to code . . . . .                   | 7         |
| 1.3 What are R packages? . . . . .                         | 8         |
| 1.3.1 Package installation . . . . .                       | 9         |
| 1.3.2 Package loading . . . . .                            | 11        |
| 1.3.3 Package use . . . . .                                | 11        |
| 1.4 Explore your first datasets . . . . .                  | 12        |
| 1.4.1 nycflights13 package . . . . .                       | 12        |
| 1.4.2 flights data frame . . . . .                         | 13        |
| 1.4.3 Exploring data frames . . . . .                      | 14        |
| 1.4.4 Identification and measurement variables . . . . .   | 17        |
| 1.4.5 Help files . . . . .                                 | 18        |
| 1.5 Conclusion . . . . .                                   | 19        |
| 1.5.1 Additional resources . . . . .                       | 19        |
| 1.5.2 What's to come? . . . . .                            | 20        |
| <b>I Data Science with tidyverse</b>                       | <b>21</b> |
| <b>2 Data Visualization</b>                                | <b>23</b> |
| 2.1 The grammar of graphics . . . . .                      | 24        |
| 2.1.1 Components of the grammar . . . . .                  | 24        |
| 2.1.2 Gapminder data . . . . .                             | 25        |
| 2.1.3 Other components . . . . .                           | 26        |



|          |                                              |           |
|----------|----------------------------------------------|-----------|
| 2.1.4    | ggplot2 package . . . . .                    | 27        |
| 2.2      | Five named graphs - the 5NG . . . . .        | 27        |
| 2.3      | 5NG#1: Scatterplots . . . . .                | 28        |
| 2.3.1    | Scatterplots via geom_point . . . . .        | 29        |
| 2.3.2    | Overplotting . . . . .                       | 31        |
| 2.3.3    | Summary . . . . .                            | 35        |
| 2.4      | 5NG#2: Linegraphs . . . . .                  | 35        |
| 2.4.1    | Linegraphs via geom_line . . . . .           | 36        |
| 2.4.2    | Summary . . . . .                            | 38        |
| 2.5      | 5NG#3: Histograms . . . . .                  | 38        |
| 2.5.1    | Histograms via geom_histogram . . . . .      | 40        |
| 2.5.2    | Adjusting the bins . . . . .                 | 41        |
| 2.5.3    | Summary . . . . .                            | 43        |
| 2.6      | Facets . . . . .                             | 43        |
| 2.7      | 5NG#4: Boxplots . . . . .                    | 45        |
| 2.7.1    | Boxplots via geom_boxplot . . . . .          | 47        |
| 2.7.2    | Summary . . . . .                            | 50        |
| 2.8      | 5NG#5: Barplots . . . . .                    | 50        |
| 2.8.1    | Barplots via geom_bar or geom_col . . . . .  | 51        |
| 2.8.2    | Must avoid pie charts! . . . . .             | 54        |
| 2.8.3    | Two categorical variables . . . . .          | 55        |
| 2.8.4    | Summary . . . . .                            | 60        |
| 2.9      | Conclusion . . . . .                         | 60        |
| 2.9.1    | Summary table . . . . .                      | 60        |
| 2.9.2    | Function argument specification . . . . .    | 61        |
| 2.9.3    | Additional resources . . . . .               | 62        |
| 2.9.4    | What's to come . . . . .                     | 62        |
| <b>3</b> | <b>Data Wrangling</b> . . . . .              | <b>65</b> |
| 3.1      | The pipe operator: %>% . . . . .             | 67        |
| 3.2      | filter rows . . . . .                        | 69        |
| 3.3      | summarize variables . . . . .                | 72        |
| 3.4      | group_by rows . . . . .                      | 75        |
| 3.4.1    | Grouping by more than one variable . . . . . | 78        |
| 3.5      | mutate existing variables . . . . .          | 80        |
| 3.6      | arrange and sort rows . . . . .              | 84        |
| 3.7      | join data frames . . . . .                   | 86        |
| 3.7.1    | Matching "key" variable names . . . . .      | 87        |
| 3.7.2    | Different "key" variable names . . . . .     | 88        |
| 3.7.3    | Multiple "key" variables . . . . .           | 89        |
| 3.7.4    | Normal forms . . . . .                       | 90        |
| 3.8      | Other verbs . . . . .                        | 91        |



|       |                                                                  |            |
|-------|------------------------------------------------------------------|------------|
| 3.8.1 | select variables . . . . .                                       | 91         |
| 3.8.2 | rename variables . . . . .                                       | 93         |
| 3.8.3 | top_n values of a variable . . . . .                             | 93         |
| 3.9   | Conclusion . . . . .                                             | 94         |
| 3.9.1 | Summary table . . . . .                                          | 94         |
| 3.9.2 | Additional resources . . . . .                                   | 96         |
| 3.9.3 | What's to come? . . . . .                                        | 96         |
| 4     | <b>Data Importing and "Tidy" Data</b>                            | <b>99</b>  |
| 4.1   | Importing data . . . . .                                         | 100        |
| 4.1.1 | Using the console . . . . .                                      | 101        |
| 4.1.2 | Using RStudio's interface . . . . .                              | 102        |
| 4.2   | "Tidy" data . . . . .                                            | 103        |
| 4.2.1 | Definition of "tidy" data . . . . .                              | 106        |
| 4.2.2 | Converting to "tidy" data . . . . .                              | 108        |
| 4.2.3 | nycflights13 package . . . . .                                   | 112        |
| 4.3   | Case study: Democracy in Guatemala . . . . .                     | 113        |
| 4.4   | tidyverse package . . . . .                                      | 116        |
| 4.5   | Conclusion . . . . .                                             | 117        |
| 4.5.1 | Additional resources . . . . .                                   | 117        |
| 4.5.2 | What's to come? . . . . .                                        | 117        |
| II    | <b>Data Modeling with moderndive</b>                             | <b>119</b> |
| 5     | <b>Basic Regression</b>                                          | <b>121</b> |
| 5.1   | One numerical explanatory variable . . . . .                     | 123        |
| 5.1.1 | Exploratory data analysis . . . . .                              | 124        |
| 5.1.2 | Simple linear regression . . . . .                               | 133        |
| 5.1.3 | Observed/fitted values and residuals . . . . .                   | 137        |
| 5.2   | One categorical explanatory variable . . . . .                   | 139        |
| 5.2.1 | Exploratory data analysis . . . . .                              | 140        |
| 5.2.2 | Linear regression . . . . .                                      | 147        |
| 5.2.3 | Observed/fitted values and residuals . . . . .                   | 151        |
| 5.3   | Related topics . . . . .                                         | 152        |
| 5.3.1 | Correlation is not necessarily causation . . . . .               | 152        |
| 5.3.2 | Best-fitting line . . . . .                                      | 154        |
| 5.3.3 | get_regression_x() functions . . . . .                           | 157        |
| 5.4   | Conclusion . . . . .                                             | 160        |
| 5.4.1 | Additional resources . . . . .                                   | 160        |
| 5.4.2 | What's to come? . . . . .                                        | 160        |
| 6     | <b>Multiple Regression</b>                                       | <b>161</b> |
| 6.1   | One numerical and one categorical explanatory variable . . . . . | 162        |



|            |                                                         |            |
|------------|---------------------------------------------------------|------------|
| 6.1.1      | Exploratory data analysis . . . . .                     | 162        |
| 6.1.2      | Interaction model . . . . .                             | 166        |
| 6.1.3      | Parallel slopes model . . . . .                         | 169        |
| 6.1.4      | Observed/fitted values and residuals . . . . .          | 173        |
| 6.2        | Two numerical explanatory variables . . . . .           | 175        |
| 6.2.1      | Exploratory data analysis . . . . .                     | 176        |
| 6.2.2      | Regression plane . . . . .                              | 181        |
| 6.2.3      | Observed/fitted values and residuals . . . . .          | 183        |
| 6.3        | Related topics . . . . .                                | 184        |
| 6.3.1      | Model selection . . . . .                               | 184        |
| 6.3.2      | Correlation coefficient . . . . .                       | 188        |
| 6.3.3      | Simpson's Paradox . . . . .                             | 188        |
| 6.4        | Conclusion . . . . .                                    | 191        |
| 6.4.1      | Additional resources . . . . .                          | 191        |
| 6.4.2      | What's to come? . . . . .                               | 191        |
| <b>III</b> | <b>Statistical Inference with infer</b>                 | <b>193</b> |
| <b>7</b>   | <b>Sampling</b>                                         | <b>195</b> |
| 7.1        | Sampling bowl activity . . . . .                        | 195        |
| 7.1.1      | What proportion of this bowl's balls are red? . . . . . | 196        |
| 7.1.2      | Using the shovel once . . . . .                         | 196        |
| 7.1.3      | Using the shovel 33 times . . . . .                     | 198        |
| 7.1.4      | What did we just do? . . . . .                          | 201        |
| 7.2        | Virtual sampling . . . . .                              | 202        |
| 7.2.1      | Using the virtual shovel once . . . . .                 | 203        |
| 7.2.2      | Using the virtual shovel 33 times . . . . .             | 206        |
| 7.2.3      | Using the virtual shovel 1000 times . . . . .           | 209        |
| 7.2.4      | Using different shovels . . . . .                       | 212        |
| 7.3        | Sampling framework . . . . .                            | 216        |
| 7.3.1      | Terminology and notation . . . . .                      | 216        |
| 7.3.2      | Statistical definitions . . . . .                       | 219        |
| 7.3.3      | The moral of the story . . . . .                        | 222        |
| 7.4        | Case study: Polls . . . . .                             | 226        |
| 7.5        | Conclusion . . . . .                                    | 230        |
| 7.5.1      | Sampling scenarios . . . . .                            | 230        |
| 7.5.2      | Central Limit Theorem . . . . .                         | 231        |
| 7.5.3      | Additional resources . . . . .                          | 232        |
| 7.5.4      | What's to come? . . . . .                               | 232        |
| <b>8</b>   | <b>Bootstrapping and Confidence Intervals</b>           | <b>233</b> |
| 8.1        | Pennies activity . . . . .                              | 235        |



|          |                                                           |            |
|----------|-----------------------------------------------------------|------------|
| 8.1.1    | What is the average year on US pennies in 2019? . . . . . | 235        |
| 8.1.2    | Resampling once . . . . .                                 | 239        |
| 8.1.3    | Resampling 35 times . . . . .                             | 244        |
| 8.1.4    | What did we just do? . . . . .                            | 246        |
| 8.2      | Computer simulation of resampling . . . . .               | 247        |
| 8.2.1    | Virtually resampling once . . . . .                       | 247        |
| 8.2.2    | Virtually resampling 35 times . . . . .                   | 249        |
| 8.2.3    | Virtually resampling 1000 times . . . . .                 | 251        |
| 8.3      | Understanding confidence intervals . . . . .              | 254        |
| 8.3.1    | Percentile method . . . . .                               | 255        |
| 8.3.2    | Standard error method . . . . .                           | 256        |
| 8.4      | Constructing confidence intervals . . . . .               | 258        |
| 8.4.1    | Original workflow . . . . .                               | 259        |
| 8.4.2    | infer package workflow . . . . .                          | 259        |
| 8.4.3    | Percentile method with infer . . . . .                    | 267        |
| 8.4.4    | Standard error method with infer . . . . .                | 269        |
| 8.5      | Interpreting confidence intervals . . . . .               | 271        |
| 8.5.1    | Did the net capture the fish? . . . . .                   | 272        |
| 8.5.2    | Precise and shorthand interpretation . . . . .            | 280        |
| 8.5.3    | Width of confidence intervals . . . . .                   | 281        |
| 8.6      | Case study: Is yawning contagious? . . . . .              | 284        |
| 8.6.1    | <i>Mythbusters</i> study data . . . . .                   | 284        |
| 8.6.2    | Sampling scenario . . . . .                               | 286        |
| 8.6.3    | Constructing the confidence interval . . . . .            | 287        |
| 8.6.4    | Interpreting the confidence interval . . . . .            | 294        |
| 8.7      | Conclusion . . . . .                                      | 295        |
| 8.7.1    | Comparing bootstrap and sampling distributions . . . . .  | 295        |
| 8.7.2    | Theory-based confidence intervals . . . . .               | 300        |
| 8.7.3    | Additional resources . . . . .                            | 305        |
| 8.7.4    | What's to come? . . . . .                                 | 305        |
| <b>9</b> | <b>Hypothesis Testing</b>                                 | <b>307</b> |
| 9.1      | Promotions activity . . . . .                             | 308        |
| 9.1.1    | Does gender affect promotions at a bank? . . . . .        | 308        |
| 9.1.2    | Shuffling once . . . . .                                  | 310        |
| 9.1.3    | Shuffling 16 times . . . . .                              | 314        |
| 9.1.4    | What did we just do? . . . . .                            | 316        |
| 9.2      | Understanding hypothesis tests . . . . .                  | 317        |
| 9.3      | Conducting hypothesis tests . . . . .                     | 320        |
| 9.3.1    | infer package workflow . . . . .                          | 322        |
| 9.3.2    | Comparison with confidence intervals . . . . .            | 328        |
| 9.3.3    | "There is only one test" . . . . .                        | 332        |



|           |                                                              |            |
|-----------|--------------------------------------------------------------|------------|
| 9.4       | Interpreting hypothesis tests . . . . .                      | 333        |
| 9.4.1     | Two possible outcomes . . . . .                              | 333        |
| 9.4.2     | Types of errors . . . . .                                    | 335        |
| 9.4.3     | How do we choose alpha? . . . . .                            | 336        |
| 9.5       | Case study: Are action or romance movies rated higher? . . . | 337        |
| 9.5.1     | IMDb ratings data . . . . .                                  | 338        |
| 9.5.2     | Sampling scenario . . . . .                                  | 340        |
| 9.5.3     | Conducting the hypothesis test . . . . .                     | 341        |
| 9.6       | Conclusion . . . . .                                         | 347        |
| 9.6.1     | Theory-based hypothesis tests . . . . .                      | 347        |
| 9.6.2     | When inference is not needed . . . . .                       | 356        |
| 9.6.3     | Problems with p-values . . . . .                             | 358        |
| 9.6.4     | Additional resources . . . . .                               | 359        |
| 9.6.5     | What's to come . . . . .                                     | 359        |
| <b>10</b> | <b>Inference for Regression</b> . . . . .                    | <b>361</b> |
| 10.1      | Regression refresher . . . . .                               | 361        |
| 10.1.1    | Teaching evaluations analysis . . . . .                      | 362        |
| 10.1.2    | Sampling scenario . . . . .                                  | 364        |
| 10.2      | Interpreting regression tables . . . . .                     | 365        |
| 10.2.1    | Standard error . . . . .                                     | 366        |
| 10.2.2    | Test statistic . . . . .                                     | 367        |
| 10.2.3    | p-value . . . . .                                            | 368        |
| 10.2.4    | Confidence interval . . . . .                                | 369        |
| 10.2.5    | How does R compute the table? . . . . .                      | 370        |
| 10.3      | Conditions for inference for regression . . . . .            | 371        |
| 10.3.1    | Residuals refresher . . . . .                                | 371        |
| 10.3.2    | Linearity of relationship . . . . .                          | 373        |
| 10.3.3    | Independence of residuals . . . . .                          | 374        |
| 10.3.4    | Normality of residuals . . . . .                             | 375        |
| 10.3.5    | Equality of variance . . . . .                               | 376        |
| 10.3.6    | What's the conclusion? . . . . .                             | 378        |
| 10.4      | Simulation-based inference for regression . . . . .          | 379        |
| 10.4.1    | Confidence interval for slope . . . . .                      | 380        |
| 10.4.2    | Hypothesis test for slope . . . . .                          | 384        |
| 10.5      | Conclusion . . . . .                                         | 386        |
| 10.5.1    | Theory-based inference for regression . . . . .              | 386        |
| 10.5.2    | Summary of statistical inference . . . . .                   | 388        |
| 10.5.3    | Additional resources . . . . .                               | 389        |
| 10.5.4    | What's to come . . . . .                                     | 389        |



# **IV Conclusion** **391**

## **11 Tell Your Story with Data** **393**

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| 11.1 Review . . . . .                                             | 393 |
| 11.2 Case study: Seattle house prices . . . . .                   | 396 |
| 11.2.1 Exploratory data analysis: Part I . . . . .                | 397 |
| 11.2.2 Exploratory data analysis: Part II . . . . .               | 404 |
| 11.2.3 Regression modeling . . . . .                              | 407 |
| 11.2.4 Making predictions . . . . .                               | 409 |
| 11.3 Case study: Effective data storytelling . . . . .            | 410 |
| 11.3.1 Bechdel test for Hollywood gender representation . . . . . | 411 |
| 11.3.2 US Births in 1999 . . . . .                                | 411 |
| 11.3.3 Scripts of R code . . . . .                                | 414 |

## **Appendix A Statistical Background** **417**

|                                       |     |
|---------------------------------------|-----|
| A.1 Basic statistical terms . . . . . | 417 |
| A.1.1 Mean . . . . .                  | 417 |
| A.1.2 Median . . . . .                | 417 |
| A.1.3 Standard deviation . . . . .    | 417 |
| A.1.4 Five-number summary . . . . .   | 418 |
| A.1.5 Distribution . . . . .          | 418 |
| A.1.6 Outliers . . . . .              | 418 |
| A.2 Normal distribution . . . . .     | 418 |
| A.3 log10 transformations . . . . .   | 421 |

## **Appendix B Versions of R Packages Used** **423**

## **Bibliography** **425**

## **Index** **427**