

---

# Contents

---

|                                                  |           |
|--------------------------------------------------|-----------|
| List of Figures                                  | ix        |
| Acknowledgments                                  | xvii      |
| Introduction                                     | xix       |
| <b>1 Getting Started</b>                         | <b>1</b>  |
| 1.1 Installing R                                 | 1         |
| 1.2 Installing RStudio                           | 1         |
| 1.3 Learn R Basics                               | 1         |
| 1.4 Installing Packages                          | 2         |
| 1.5 Importing Data into R                        | 3         |
| 1.6 Exercises                                    | 5         |
| 1.7 Brief Introduction to <code>dplyr</code>     | 6         |
| 1.8 Exercises                                    | 8         |
| 1.9 Mathematical Notation                        | 8         |
| <b>2 Inference</b>                               | <b>13</b> |
| 2.1 Introduction                                 | 13        |
| 2.2 Random Variables                             | 14        |
| 2.3 The Null Hypothesis                          | 15        |
| 2.4 Distributions                                | 16        |
| 2.5 Probability Distribution                     | 17        |
| 2.6 Normal Distribution                          | 19        |
| 2.7 Exercises                                    | 21        |
| 2.8 Populations, Samples and Estimates           | 22        |
| 2.9 Exercises                                    | 23        |
| 2.10 Central Limit Theorem and t-distribution    | 24        |
| 2.11 Exercises                                   | 28        |
| 2.12 Central Limit Theorem in Practice           | 30        |
| 2.13 Exercises                                   | 33        |
| 2.14 t-tests in Practice                         | 36        |
| 2.15 The t-distribution in Practice              | 37        |
| 2.16 Confidence Intervals                        | 39        |
| 2.17 Power Calculations                          | 45        |
| 2.18 Exercises                                   | 51        |
| 2.19 Monte Carlo Simulation                      | 54        |
| 2.20 Parametric Simulations for the Observations | 58        |
| 2.21 Exercises                                   | 58        |
| 2.22 Permutation Tests                           | 60        |
| 2.23 Exercises                                   | 62        |
| 2.24 Association Tests                           | 63        |

|          |                                            |            |
|----------|--------------------------------------------|------------|
| 2.25     | Exercises                                  | 67         |
| <b>3</b> | <b>Exploratory Data Analysis</b>           | <b>69</b>  |
| 3.1      | Quantile Quantile Plots                    | 69         |
| 3.2      | Boxplots                                   | 72         |
| 3.3      | Scatterplots and Correlation               | 74         |
| 3.4      | Stratification                             | 74         |
| 3.5      | Bivariate Normal Distribution              | 75         |
| 3.6      | Plots to Avoid                             | 78         |
| 3.7      | Misunderstanding Correlation (Advanced)    | 91         |
| 3.8      | Exercises                                  | 93         |
| 3.9      | Robust Summaries                           | 94         |
| 3.10     | Wilcoxon Rank Sum Test                     | 99         |
| 3.11     | Exercises                                  | 100        |
| <b>4</b> | <b>Matrix Algebra</b>                      | <b>103</b> |
| 4.1      | Motivating Examples                        | 103        |
| 4.2      | Exercises                                  | 109        |
| 4.3      | Matrix Notation                            | 110        |
| 4.4      | Solving Systems of Equations               | 110        |
| 4.5      | Vectors, Matrices, and Scalars             | 111        |
| 4.6      | Exercises                                  | 113        |
| 4.7      | Matrix Operations                          | 113        |
| 4.8      | Exercises                                  | 117        |
| 4.9      | Examples                                   | 118        |
| 4.10     | Exercises                                  | 122        |
| <b>5</b> | <b>Linear Models</b>                       | <b>125</b> |
| 5.1      | Exercises                                  | 125        |
| 5.2      | The Design Matrix                          | 127        |
| 5.3      | Exercises                                  | 134        |
| 5.4      | The Mathematics Behind <code>lm()</code>   | 135        |
| 5.5      | Exercises                                  | 137        |
| 5.6      | Standard Errors                            | 139        |
| 5.7      | Exercises                                  | 145        |
| 5.8      | Interactions and Contrasts                 | 146        |
| 5.9      | Linear Model with Interactions             | 156        |
| 5.10     | Analysis of Variance                       | 160        |
| 5.11     | Exercises                                  | 166        |
| 5.12     | Collinearity                               | 168        |
| 5.13     | Rank                                       | 169        |
| 5.14     | Removing Confounding                       | 170        |
| 5.15     | Exercises                                  | 170        |
| 5.16     | The QR Factorization (Advanced)            | 172        |
| 5.17     | Going Further                              | 175        |
| <b>6</b> | <b>Inference for High Dimensional Data</b> | <b>177</b> |
| 6.1      | Introduction                               | 177        |
| 6.2      | Exercises                                  | 179        |
| 6.3      | Inference in Practice                      | 180        |
| 6.4      | Exercises                                  | 183        |

|          |                                                |            |
|----------|------------------------------------------------|------------|
| 6.5      | Procedures                                     | 184        |
| 6.6      | Error Rates                                    | 184        |
| 6.7      | The Bonferroni Correction                      | 187        |
| 6.8      | False Discovery Rate                           | 189        |
| 6.9      | Direct Approach to FDR and q-values (Advanced) | 195        |
| 6.10     | Exercises                                      | 198        |
| 6.11     | Basic Exploratory Data Analysis                | 201        |
| 6.12     | Exercises                                      | 206        |
| <b>7</b> | <b>Statistical Models</b>                      | <b>209</b> |
| 7.1      | The Binomial Distribution                      | 209        |
| 7.2      | The Poisson Distribution                       | 210        |
| 7.3      | Maximum Likelihood Estimation                  | 213        |
| 7.4      | Distributions for Positive Continuous Values   | 215        |
| 7.5      | Exercises                                      | 220        |
| 7.6      | Bayesian Statistics                            | 224        |
| 7.7      | Exercises                                      | 229        |
| 7.8      | Hierarchical Models                            | 230        |
| 7.9      | Exercises                                      | 234        |
| <b>8</b> | <b>Distance and Dimension Reduction</b>        | <b>237</b> |
| 8.1      | Introduction                                   | 237        |
| 8.2      | Euclidean Distance                             | 237        |
| 8.3      | Distance in High Dimensions                    | 239        |
| 8.4      | Exercises                                      | 241        |
| 8.5      | Dimension Reduction Motivation                 | 241        |
| 8.6      | Singular Value Decomposition                   | 246        |
| 8.7      | Exercises                                      | 252        |
| 8.8      | Projections                                    | 254        |
| 8.9      | Rotations                                      | 258        |
| 8.10     | Multi-Dimensional Scaling Plots                | 261        |
| 8.11     | Exercises                                      | 266        |
| 8.12     | Principal Component Analysis                   | 267        |
| <b>9</b> | <b>Basic Machine Learning</b>                  | <b>273</b> |
| 9.1      | Clustering                                     | 273        |
| 9.2      | Exercises                                      | 279        |
| 9.3      | Conditional Probabilities and Expectations     | 281        |
| 9.4      | Exercises                                      | 283        |
| 9.5      | Smoothing                                      | 285        |
| 9.6      | Bin Smoothing                                  | 286        |
| 9.7      | Loess                                          | 288        |
| 9.8      | Exercises                                      | 290        |
| 9.9      | Class Prediction                               | 291        |
| 9.10     | Cross-validation                               | 297        |
| 9.11     | Exercises                                      | 302        |

|                                                     |            |
|-----------------------------------------------------|------------|
| <b>10 Batch Effects</b>                             | <b>305</b> |
| 10.1 Confounding                                    | 307        |
| 10.2 Confounding: High-Throughput Example           | 311        |
| 10.3 Exercises                                      | 312        |
| 10.4 Discovering Batch Effects with EDA             | 313        |
| 10.5 Gene Expression Data                           | 314        |
| 10.6 Exercises                                      | 321        |
| 10.7 Motivation for Statistical Approaches          | 323        |
| 10.8 Adjusting for Batch Effects with Linear Models | 325        |
| 10.9 Exercises                                      | 329        |
| 10.10 Factor Analysis                               | 330        |
| 10.11 Exercises                                     | 333        |
| 10.12 Modeling Batch Effects with Factor Analysis   | 335        |
| 10.13 Exercises                                     | 342        |
| <b>Index</b>                                        | <b>343</b> |