

---

# Contents

---

- 1 Data ..... 1
  - 1.1 What is data? 1
    - 1.1.1 Problems 4
  - 1.2 Some R essentials 4
    - 1.2.1 Starting R 5
    - 1.2.2 Using R as a calculator 6
    - 1.2.3 Assignment 7
    - 1.2.4 Using `c()` to enter data 9
    - 1.2.5 Using functions on a data vector 10
    - 1.2.6 Creating structured data 13
    - 1.2.7 Problems 14
  - 1.3 Accessing data by using indices 16
    - 1.3.1 Assigning values to data vector 17
    - 1.3.2 Logical values 18
    - 1.3.3 Missing values 20
    - 1.3.4 Managing the work environment 21
    - 1.3.5 Problems 21
  - 1.4 Reading in other sources of data 23
    - 1.4.1 Using R's built-in libraries and data sets 23
    - 1.4.2 Using the data sets that accompany this book 26
    - 1.4.3 Other methods of data entry 27
    - 1.4.4 Problems 29
- 2 Univariate data ..... 31
  - 2.1 Categorical data 32
    - 2.1.1 Tables 32
    - 2.1.2 Barplots 33
    - 2.1.3 Pie charts 36
    - 2.1.4 Dot charts 37

|          |   |            |
|----------|---|------------|
| 2.1.5    | Factors   | 38         |
| 2.1.6    | Problems  | 39         |
| 2.2      | Numeric data  | 41         |
| 2.2.1    | Stem-and-leaf plots   | 41         |
| 2.2.2    | Strip charts  | 42         |
| 2.2.3    | The center: mean, median, and mode                                    | 43         |
| 2.2.4    | Variation: the variance, standard deviation, and IQR                  | 48         |
| 2.2.5    | Problems  | 53         |
| 2.3      | Shape of a distribution   | 56         |
| 2.3.1    | Histogram   | 56         |
| 2.3.2    | Modes, symmetry, and skew   | 61         |
| 2.3.3    | Boxplots  | 64         |
| 2.3.4    | Problems  | 66         |
| <b>3</b> | <b>Bivariate data .....</b>   | <b>69</b>  |
| 3.1      | Pairs of categorical variables  | 69         |
| 3.1.1    | Making two-way tables from summarized data                            | 70         |
| 3.1.2    | Making two-way tables from unsummarized data                          | 71         |
| 3.1.3    | Marginal distributions of two-way tables                              | 72         |
| 3.1.4    | Conditional distributions of two-way tables                           | 73         |
| 3.1.5    | Graphical summaries of two-way contingency tables                     | 74         |
| 3.1.6    | Problems  | 75         |
| 3.2      | Comparing independent samples   | 77         |
| 3.2.1    | Side-by-side boxplots   | 77         |
| 3.2.2    | Densityplots  | 78         |
| 3.2.3    | Strip charts  | 79         |
| 3.2.4    | Quantile-quantile plots   | 79         |
| 3.2.5    | Problems  | 80         |
| 3.3      | Relationships in numeric data   | 82         |
| 3.3.1    | Using scatterplots to investigate relationships                       | 82         |
| 3.3.2    | The correlation between two variables                                 | 86         |
| 3.3.3    | Problems  | 89         |
| 3.4      | Simple linear regression  | 90         |
| 3.4.1    | Using the regression model for prediction                             | 92         |
| 3.4.2    | Finding the regression coefficients using <code>lm()</code>           | 92         |
| 3.4.3    | Transformations of the data   | 95         |
| 3.4.4    | Interacting with a scatterplot  | 96         |
| 3.4.5    | Outliers in the regression model                                      | 98         |
| 3.4.6    | Resistant regression lines: <code>lqs()</code> and <code>r1m()</code> | 99         |
| 3.4.7    | Trend lines   | 101        |
| 3.4.8    | Problems  | 102        |
| <b>4</b> | <b>Multivariate Data .....</b>  | <b>105</b> |
| 4.1      | Viewing multivariate data   | 105        |

|          |   |            |
|----------|---|------------|
| 4.1.1    | Summarizing categorical data  | 105        |
| 4.1.2    | Comparing independent samples   | 108        |
| 4.1.3    | Comparing relationships   | 109        |
| 4.1.4    | Problems  | 112        |
| 4.2      | R basics: data frames and lists   | 113        |
| 4.2.1    | Creating a data frame or list   | 114        |
| 4.2.2    | Accessing values in a data frame  | 115        |
| 4.2.3    | Setting values in a data frame or list  | 122        |
| 4.2.4    | Applying functions to a data frame or list  | 123        |
| 4.2.5    | Problems  | 124        |
| 4.3      | Using model formula with multivariate data  | 125        |
| 4.3.1    | Boxplots from a model formula   | 126        |
| 4.3.2    | The <code>pplot()</code> function with model formula                                | 126        |
| 4.3.3    | Creating contingency tables with <code>xtabs()</code>                               | 127        |
| 4.3.4    | Manipulating data frames: <code>split()</code> and <code>stack()</code>             | 129        |
| 4.3.5    | Problems  | 130        |
| 4.4      | Lattice graphics  | 131        |
| 4.4.1    | Problems  | 134        |
| 4.5      | Types of data in R  | 135        |
| 4.5.1    | Factors   | 136        |
| 4.5.2    | Coercion of objects   | 138        |
| <b>5</b> | <b>Describing populations .....</b>   | <b>141</b> |
| 5.1      | Populations   | 141        |
| 5.1.1    | Discrete random variables   | 142        |
| 5.1.2    | Continuous random variables   | 144        |
| 5.1.3    | Sampling from a population  | 146        |
| 5.1.4    | Sampling distributions  | 148        |
| 5.1.5    | Problems  | 148        |
| 5.2      | Families of distributions   | 149        |
| 5.2.1    | The <code>d</code> , <code>p</code> , <code>q</code> , and <code>r</code> functions | 149        |
| 5.2.2    | Binomial, normal, and some other named distributions                                | 150        |
| 5.2.3    | Popular distributions to describe populations                                       | 155        |
| 5.2.4    | Sampling distributions  | 157        |
| 5.2.5    | Problems  | 158        |
| 5.3      | The central limit theorem   | 160        |
| 5.3.1    | Normal parent population  | 160        |
| 5.3.2    | Nonnormal parent population   | 161        |
| 5.3.3    | Problems  | 163        |
| <b>6</b> | <b>Simulation.....</b>  | <b>165</b> |
| 6.1      | The normal approximation for the binomial   | 165        |
| 6.2      | for loops   | 166        |
| 6.3      | Simulations related to the central limit theorem                                    | 168        |



|          |   |            |
|----------|---|------------|
| 6.4      | Defining a function   | 169        |
| 6.4.1    | Editing a function  | 169        |
| 6.4.2    | Function arguments  | 170        |
| 6.4.3    | The function body   | 170        |
| 6.5      | Investigating distributions                                 | 172        |
| 6.5.1    | Script files and <code>source()</code>                      | 173        |
| 6.5.2    | The geometric distribution                                  | 174        |
| 6.6      | Bootstrap samples   | 176        |
| 6.7      | Alternates to for loops                                     | 177        |
| <b>7</b> | <b>Confidence intervals.....</b>                            | <b>181</b> |
| 7.1      | Confidence interval ideas                                   | 181        |
| 7.1.1    | Finding confidence intervals using simulation               | 181        |
| 7.2      | Confidence intervals for a population proportion, $p$       | 184        |
| 7.2.1    | Using <code>prop.test()</code> to find confidence intervals | 187        |
| 7.2.2    | Problems  | 188        |
| 7.3      | Confidence intervals for the population mean, $\mu$         | 190        |
| 7.3.1    | One-sided confidence intervals                              | 193        |
| 7.3.2    | Problems  | 194        |
| 7.4      | Other confidence intervals                                  | 197        |
| 7.4.1    | Confidence interval for $\sigma^2$                          | 197        |
| 7.4.2    | Problems  | 198        |
| 7.5      | Confidence intervals for differences                        | 199        |
| 7.5.1    | Difference of proportions                                   | 199        |
| 7.5.2    | Difference of means   | 201        |
| 7.5.3    | Matched samples   | 204        |
| 7.5.4    | Problems  | 205        |
| 7.6      | Confidence intervals for the median                         | 207        |
| 7.6.1    | Confidence intervals based on the binomial                  | 207        |
| 7.6.2    | Confidence intervals based on signed-rank statistic         | 208        |
| 7.6.3    | Confidence intervals based on the rank-sum statistic        | 209        |
| 7.6.4    | Problems  | 211        |
| <b>8</b> | <b>Significance tests.....</b>                              | <b>213</b> |
| 8.1      | Significance test for a population proportion               | 218        |
| 8.1.1    | Using <code>prop.test()</code> to compute $p$ -values       | 220        |
| 8.1.2    | Problems  | 221        |
| 8.2      | Significance test for the mean ( $t$ -tests)                | 222        |
| 8.2.1    | Problems  | 225        |
| 8.3      | Significance tests and confidence intervals                 | 227        |
| 8.4      | Significance tests for the median                           | 228        |
| 8.4.1    | The sign test   | 228        |
| 8.4.2    | The signed-rank test  | 230        |
| 8.4.3    | Problems  | 232        |

|           |   |            |
|-----------|---|------------|
| 8.5       | Two-sample tests of proportion                                | 233        |
| 8.5.1     | Problems  | 235        |
| 8.6       | Two-sample tests of center                                    | 237        |
| 8.6.1     | Two sample tests of center with normal populations            | 237        |
| 8.6.2     | Matched samples   | 241        |
| 8.6.3     | The Wilcoxon rank-sum test for equality of center             | 244        |
| 8.6.4     | Problems  | 246        |
| <b>9</b>  | <b>Goodness of fit</b>  | <b>249</b> |
| 9.1       | The chi-squared goodness-of-fit test                          | 249        |
| 9.1.1     | The multinomial distribution                                  | 249        |
| 9.1.2     | Pearson's $\chi^2$ statistic                                  | 250        |
| 9.1.3     | Problems  | 254        |
| 9.2       | The chi-squared test of independence                          | 258        |
| 9.2.1     | The chi-squared test of homogeneity                           | 262        |
| 9.2.2     | Problems  | 264        |
| 9.3       | Goodness-of-fit tests for continuous distributions            | 266        |
| 9.3.1     | Kolmogorov-Smirnov test                                       | 266        |
| 9.3.2     | The Shapiro-Wilk test for normality                           | 270        |
| 9.3.3     | Finding parameter values using <code>fitdistr()</code>        | 273        |
| 9.3.4     | Problems  | 275        |
| <b>10</b> | <b>Linear regression</b>                                      | <b>277</b> |
| 10.1      | The simple linear regression model                            | 277        |
| 10.1.1    | Model formulas for linear models                              | 278        |
| 10.1.2    | Examples of the linear model                                  | 278        |
| 10.1.3    | Estimating the parameters in simple linear regression         | 279        |
| 10.1.4    | Using <code>lm()</code> to find the estimates                 | 280        |
| 10.1.5    | Problems  | 283        |
| 10.2      | Statistical inference for simple linear regression            | 284        |
| 10.2.1    | Testing the model assumptions                                 | 284        |
| 10.2.2    | Statistical inferences  | 288        |
| 10.2.3    | Using <code>lm()</code> to find values for a regression model | 292        |
| 10.2.4    | Problems  | 297        |
| 10.3      | Multiple linear regression                                    | 300        |
| 10.3.1    | Types of models   | 301        |
| 10.3.2    | Fitting the multiple regression model using <code>lm()</code> | 302        |
| 10.3.3    | Interpreting the regression parameters                        | 305        |
| 10.3.4    | Statistical inferences  | 306        |
| 10.3.5    | Model selection   | 307        |
| 10.3.6    | Problems  | 310        |
| <b>11</b> | <b>Analysis of variance</b>                                   | <b>313</b> |
| 11.1      | One-way ANOVA   | 313        |

|           |   |            |
|-----------|---|------------|
| 11.1.1    | Using R's model formulas to specify ANOVA models  | 316        |
| 11.1.2    | Using <code>oneway.test()</code> to perform ANOVA | 317        |
| 11.1.3    | Using <code>aov()</code> for ANOVA                | 317        |
| 11.1.4    | The nonparametric Kruskal-Wallis test             | 319        |
| 11.1.5    | Problems  | 322        |
| 11.2      | Using <code>lm()</code> for ANOVA                 | 323        |
| 11.2.1    | Treatment coding for analysis of variance         | 326        |
| 11.2.2    | Comparing multiple differences                    | 328        |
| 11.2.3    | Problems  | 330        |
| 11.3      | ANCOVA  | 332        |
| 11.3.1    | Problems  | 334        |
| 11.4      | Two-way ANOVA                                     | 335        |
| 11.4.1    | Treatment coding for additive two-way ANOVA       | 336        |
| 11.4.2    | Testing for row or column effects                 | 336        |
| 11.4.3    | Testing for interactions                          | 337        |
| 11.4.4    | Problems  | 341        |
| <b>12</b> | <b>Two extensions of the linear model</b>         | <b>343</b> |
| 12.1      | Logistic regression                               | 343        |
| 12.1.1    | Generalized linear models                         | 346        |
| 12.1.2    | Fitting the model using <code>glm()</code>        | 346        |
| 12.2      | Nonlinear models                                  | 351        |
| 12.2.1    | Fitting nonlinear models with <code>nls()</code>  | 352        |
| <b>A</b>  | <b>Getting, installing, and running R</b>         | <b>359</b> |
| A.1       | Installing and starting R                         | 359        |
| A.1.1     | Binary installation under Windows                 | 360        |
| A.1.2     | Binary installation under Linux                   | 360        |
| A.1.3     | Binary installation under Mac OS X                | 361        |
| A.1.4     | Installing from the source code                   | 361        |
| A.1.5     | Startup files                                     | 361        |
| A.2       | Extending R using additional packages             | 362        |
| A.2.1     | Upgrading an existing installation                | 363        |
| <b>B</b>  | <b>Graphical user interfaces and R</b>            | <b>365</b> |
| B.1       | The Windows GUI                                   | 365        |
| B.2       | The Mac OS X GUI                                  | 367        |
| B.3       | Rcmdmr  | 368        |
| <b>C</b>  | <b>Teaching with R</b>                            | <b>371</b> |
| <b>D</b>  | <b>More on graphics with R</b>                    | <b>373</b> |
| D.1       | Low- and high-level graphic functions             | 373        |
| D.1.1     | Setting up a plot figure                          | 374        |
| D.1.2     | Adding to a figure                                | 378        |



|                   |   |            |
|-------------------|---|------------|
| D.1.3             | Printing or saving a figure                 | 381        |
| D.2               | Creating new graphics in R                  | 381        |
| <b>E</b>          | <b>Programming in R .....</b>               | <b>387</b> |
| E.1               | Editing functions                           | 387        |
| E.1.1             | Using <code>edit()</code>                   | 387        |
| E.2               | Using functions                             | 388        |
| E.2.1             | Function arguments                          | 389        |
| E.2.2             | Function body and return values             | 391        |
| E.2.3             | Conditional evaluation                      | 393        |
| E.2.4             | Looping                                     | 394        |
| E.3               | Using files and a better editor             | 395        |
| E.3.1             | Using an external editor                    | 395        |
| E.4               | Object-oriented programming with R          | 396        |
| E.4.1             | Method dispatch                             | 397        |
| E.4.2             | S3 methods, S4 methods, and the OOP package | 399        |
| <b>Index.....</b> |   | <b>409</b> |