Thomas W. MacFarland
Jan M. Yates

Using R for Biostatistics

EXTRAS ONLINE

Springer

Using R for Biostatistics

Thomas W. MacFarland • Jan M. Yates

Using R for Biostatistics

Springer

Thomas W. MacFarland
Senior Research Associate, Office of Institutional Effectiveness, Nova Southeastern University
Fort Lauderdale, FL, USA

Associate Professor, College of Computing and Engineering, Nova Southeastern University
Fort Lauderdale, FL, USA

Jan M. Yates
Professor Emerita, Abraham S. Fischler College of Education
Nova Southeastern University
Fort Lauderdale, FL, USA

ISBN 978-3-030-62403-3 ISBN 978-3-030-62404-0 (eBook)
https://doi.org/10.1007/978-3-030-62404-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedication

In appreciation for their patience, this text is dedicated to Andrew, Baylen, Courtney, Henry, and Lauren.

This text is also dedicated to the students and beginning researchers who have struggled with the transition from a graphical approach to statistics to the more empowering, but challenging, use of syntax. We hope you continue to explore the many learning resources available to the R community; with practice, the results gained from learning syntax will be more than worth the effort.

Preface

This text is about the use of R in biostatistics. It was prepared to help beginning students and researchers gradually increase their skills with the use of R syntax as they consider how R is used in the quickly expanding the world of biostatistics.

R has become one of the leading languages used for statistical analyses in the biological sciences, and it is increasingly used for data organization, statistical analyses, and the generation of high-quality publishable graphics. There are currently more than 15,000 R-focused packages freely available to the public, and many focus exclusively on applications in biostatistics. It is our view that those who work in biostatistics should have a good working knowledge of R and an understanding of how R fits into the professional toolkit.

Using R for Biostatistics begins with a brief discussion of biostatistics, with an emphasis on how biostatistics grew out of statistics. There is also a short history of the R language and how R developed from the prior S language. The beginning parts of this text also provide a glimpse of how the lessons have been structured, ranging from learning as much as possible about the data to the eventual development of an easy-to-understand summary of statistical analyses. Attention is also given to the many ways data can be imported into R, focusing on how R can accommodate datasets in various formats.

The major part of this text is presented in the form of lessons that address the leading statistical tests typically encountered early on among those who engage in research associated with biostatistics:

- Data Exploration, Descriptive Statistics, and Measures of Central Tendency
- Student's t-Test for Independent Samples
- Student's t-Test for Matched Pairs
- Oneway Analysis of Variance (Oneway ANOVA)

- Twoway Analysis of Variance (Twoway ANOVA)
- Correlation, Association, Regression, Likelihood, and Prediction

For each of these lessons, emphasis is placed on understanding the data, organizing and then working through the data, and subsequently understanding the outcomes of statistical analyses for each test by using many different and complementary approaches:

- The production of graphics is essential to understanding statistical outcomes, and each lesson provides many examples on how to produce beginning and eventually high-quality graphics associated with variables and the selected statistical test.
- Issues inherent to data distribution patterns are also emphasized in the lessons, where many datasets throughout the lessons are first analyzed by using a parametric approach to statistical analysis and analyses are then repeated by using a nonparametric approach. This approach toward analysis takes into account the complexities of real-world analyses in biostatistics, where data are not always as tidy as desired.
- Whenever possible, multiple R packages and multiple R functions are demonstrated, in an attempt to provide wide exposure to the many ways R can be used to gain the desired output.

A special feature in this text is the rich variety of bonus materials provided in the addenda after each lesson. Multiple approaches at statistical analysis, going beyond what was presented in the preceding lesson, are included in the addenda. Analyses that address parametric v nonparametric issues are further stressed in the addenda. Perhaps most importantly, the addenda often include additional datasets and guidance that provide the opportunity for incremental confidence-building practice activities with R. The complexity of R-based syntax is only gradually introduced as engagement with the text continues.

Using R for Biostatistics ends with a large and complex dataset and presentation of the many issues that need to be considered—an introduction to Big Data and breakout subsets of a large dataset. The ending parts of this text look at the future use of R for biostatistics, including a brief introduction to the increasing use of R Markdown.

All external datasets are available on this book's product Web page at Springer. Although multiple file formats are demonstrated in *Using R for Biostatistics*, most datasets are in comma-separated values (.csv) file format. By having access to the data in original format, it is possible to replicate outcomes by using the syntax presented in this text, but now as a self-guided practice activity where the outcomes are known.

Going beyond what is presented in this text, explore the many learning opportunities available to the R community. Join and review what is discussed in R-based discussion groups. View recorded R conference presentations made available through various media. Scan the long list (currently, more than 15,000) of R packages to see how R fits into the way biostatistics is approached by others. The opportunities to learn R syntax are many.

Fort Lauderdale, FL, USA — Thomas W. MacFarland
Fort Lauderdale, FL, USA — Jan M. Yates
Summer 2020

Acknowledgments

We want to thank the many individuals who believe in the open-source paradigm. This text is only possible because of the tireless and often unrecognized efforts of all who have contributed to R, core R, and the thousands of contributed R packages.

We also want to recognize our editor, Laura Aileen Briskman, and the entire Springer team. Thank you for your many ideas, feedback, help, and supporting our efforts.

Introduction

This text is focused on R, a freely available and open-source language that is among the leading languages used in biostatistics. R is typically dependent on user-generated syntax, not menu-driven point and click selections. However, the challenges of using R's syntax, whether working with an Integrated Development Environment (IDE), working interactively at the command line, or perhaps working offline in a separate text editor, are challenges that become a bridge too far for many.

The motivation for this text can be expressed in one simple word: frustration!

In many classes, and throughout the years, we have seen students and even beginning researchers experience frustration when using data science and information science curricular materials that, though excellent, are presented at a level beyond the capabilities of those who are learning the heuristics of a new programming language for the first time. Introductory texts need to focus on the incoming skills of learners through small confidence-building experiences and present incremental opportunities that build confidence and gradually develop proficiency among enthusiastic learners. Otherwise, frustration, not mastery, will be the main outcome for those learning the new language and its applications to biostatistics.

This text is structured to reduce frustration for learners who will use R to support the research process. To achieve this aim, this text starts with a brief introduction to biostatistics and how biostatistics developed into a distinct science, whether applied for agriculture, medicine, public health, or other subdisciplines. Then, in a set of consistent, structured lessons for many statistical tests the focus moves to a common framework for problem-solving using R. Each lesson is arranged as follows:

- Background
  - Description of the Data
  - Null Hypothesis (Ho)

- Import the Data into R
- Organize the Data and Display the Code Book
- Conduct a Visual Data Check Using Graphics
- Descriptive Statistics for Initial Analysis of the Data
- Quality Assurance, Data Distribution, and Tests for Normality
- Statistical Test(s)
- Summary of Outcomes

This consistent, step-by-step presentation provides a structured and organized introduction to R that supports the following:

- The researcher knows as much about the data as can be reasonably expected.
- The data are well-organized and a descriptive Code Book not only supports a thorough understanding of the data but also aids replication of all analyses, either at a future date or by others.
- Visual presentations improve immediate cognition of trends among the data and an understanding of these trends by others, especially audiences who may not be experienced in biostatistics.
- Descriptive statistics and measures of central tendency further improve understanding of the data.
- Quality assurance is promoted by carefully addressing data distribution patterns, which validates whether parametric, nonparametric, or both approaches should be used for later analyses.
- Statistical tests are conducted, often using multiple approaches in an effort to gain consistency of outcomes.
- The summary for each lesson is prepared, so that beginning biostatisticians can understand not only the initial outcomes but also the potential practical applications of outcomes.

These many precursor activities are often given only marginal attention by those in a rush to complete and then present research results. It is our view, however, that these many activities, although demanding, support appropriate statistical test selection, implementation, and presentation of outcomes in support of the biostatistics research process.

Most lessons in this text are enhanced by addenda, with new skills added to each advancing lesson, which are designed as value-added learning opportunities. The addenda often introduce and/or reinforce specialized packages and functions that go beyond what was previously presented, address parametric

v nonparametric approaches toward the data, and often end with additional practice datasets that support incremental practice with advanced skills.

Each external dataset has been placed at the publisher's Web site for this text, which makes it possible for the learner to practice with the syntax presented in the text and to see if self-generated outcomes match the known correct output.

Contents

1 Biostatistics and R **1**

1.1 Purpose of This Text . . . 2
1.2 Development of Biostatistics . . . 3
1.3 Development of R . . . 6
1.4 How R is Used in This Text . . . 7
1.5 Import Data Into R . . . 10
1.5.1 Import a .csv File of Comma-Separated Values into R . . . 13
1.5.2 Import a .txt File of Tab-Separated Values into R . . . 16
1.5.3 Import a .txt File of Fixed-Width Format Values into R . . . 18
1.5.4 Import a .xlsx Spreadsheet File into R . . . 24
1.5.5 Import a .csv File of Comma-Separated Values from an Online Source into R . . . 27
1.5.6 Import a .csv File of Comma-Separated Values into R by Using Graphical User Interface (GUI) Selections . . . 35
1.5.7 Import by Direct Keyboard *On the Fly* Data Entry into R . . . 38
1.6 Addendum 1: Efficient Programming with R, Project Workflow, and Good Programming Practices (gpp) . . . 41
1.7 Addendum 2: Preview of Descriptive Statistics and Graphics Using R . . . 43
1.8 Addendum 3: R and *Beautiful Graphics* . . . 46
1.9 Addendum 4: Research Designs Used in Biostatistics . . . 51
1.9.1 Case Study and Clinical Trial . . . 52
1.9.2 Pretest–Posttest for One Group . . . 52
1.9.3 Pretest–Posttest for Control Group . . . 52
1.9.4 Posttest Only for Control Group . . . 53
1.9.5 Fixed Group Comparative Analysis of a Single Factor . . . 53
1.9.6 Factorial Data Organization of Multiple Independent Variables . . . 53

1.9.6.1 Goodness of Fit (e.g., Chi-Square) 53
1.9.6.2 Comparison of Group Means (e.g., Analysis of Variance) 54
1.9.7 Correlation, Association, Regression, Likelihood, and Prediction 54
1.10 Prepare to Exit, Save, and Later Retrieve This R Session 55
1.11 External Data and/or Data Resources Used in This Lesson . . . 55

2 Data Exploration, Descriptive Statistics, and Measures of Central Tendency 57
2.1 Background 58
2.1.1 Description of the Data 58
2.1.2 Null Hypothesis 61
2.2 Import Data in Comma-Separated Values (.csv) File Format and/or Self-Generate the Data Using R-Based Functions 61
2.3 Organize the Data and Display the Code Book 63
2.4 Conduct a Visual Data Check Using Graphics (e.g., Figures) 65
2.5 Descriptive Statistics for Initial Analysis of the Data 70
2.6 Quality Assurance, Data Distribution, and Tests for Normality 80
2.7 Statistical Test(s) 86
2.8 Summary 87
2.9 Addendum 1: Specialized External Packages and Functions . . . 88
2.10 Addendum 2: Parametric v Nonparametric 96
2.11 Addendum 3: Additional Practice Datasets for Data with Normal Distribution Patterns and Data That Do Not Exhibit Normal Distribution Patterns 97
2.11.1 Purpose of This Addendum 97
2.11.2 Background 97
2.11.3 Import Data in Comma-Separated Values (.csv) File Format and/or Self-Generate the Data Using R-Based Functions 98
2.11.4 Organize the Data and Display the Code Book 99
2.11.5 Conduct a Visual Data Check Using Graphics (e.g., Figures) 101
2.11.6 Descriptive Statistics for Initial Analysis of the Data . . 104
2.11.7 Quality Assurance, Data Distribution, and Tests for Normality 112
2.11.8 Statistical Test(s) 115
2.11.9 Summary of Outcomes for `SBPNormal` and `SBPNotNormal` 115
2.11.10 Additional Bonus Materials 116

2.12 Prepare to Exit, Save, and Later Retrieve This R Session 139
2.13 External Data and/or Data Resources Used in This Lesson . . . 139

3 Student's t-Test for Independent Samples 141
3.1 Background 142
3.1.1 Description of the Data 142
3.1.2 Null Hypothesis 143
3.2 Import Data in Comma-Separated Values (.csv) File Format and/or Self-Generate the Data Using R-Based Functions 144
3.3 Organize the Data and Display the Code Book 146
3.4 Conduct a Visual Data Check Using Graphics (e.g., Figures) . . 150
3.5 Descriptive Statistics for Initial Analysis of the Data 167
3.6 Quality Assurance, Data Distribution, and Tests for Normality 175
3.7 Statistical Test(s) 186
3.8 Summary of Outcomes 193
3.9 Addendum 1: t-Statistic v z-Statistic 196
3.9.1 Create an Enumerated Dataset 197
3.9.2 Calculate the t-Statistic 199
3.9.3 Calculate the z-Statistic 199
3.10 Addendum 2: Parametric v Nonparametric 200
3.11 Addendum 3: Additional Practice Datasets for Data with Normal Distribution Patterns and Data That Do Not Exhibit Normal Distribution Patterns 201
3.11.1 Data with Normal Distribution Patterns 202
3.11.2 Data That Do Not Exhibit Normal Distribution Patterns 227
3.12 Prepare to Exit, Save, and Later Retrieve This R Session 239
3.13 External Data and/or Data Resources Used in This Lesson . . . 240

4 Student's t-Test for Matched Pairs 241
4.1 Background 242
4.1.1 Description of the Data 242
4.1.2 Null Hypothesis 244
4.1.3 Unstacked (e.g., Wide) Data and Stacked (e.g., Long) Data 244
4.2 Import Data in Comma-Separated Values (.csv) File Format and/or Self-Generate the Data Using R-Based Functions 246
4.3 Organize the Data and Display the Code Book 248
4.4 Conduct a Visual Data Check Using Graphics (e.g., Figures) . . 252
4.5 Descriptive Statistics for Initial Analysis of the Data 256
4.6 Quality Assurance, Data Distribution, and Tests for Normality 262

4.7 Statistical Test(s) . . . 267
4.8 Summary of Outcomes . . . 271
4.9 Addendum 1: R-Based Tools for Unstacked (e.g., Wide) Data . . . 272
4.10 Addendum 2: Stacked Data and Student's t-Test for Matched Pairs . . . 275
4.11 Addendum 3: The Impact of N on Student's t-Test . . . 279
4.12 Addendum 4: Parametric v Nonparametric . . . 282
4.13 Addendum 5: Additional Practice Datasets for Data with Normal Distribution Patterns and Data That Do Not Exhibit Normal Distribution Patterns . . . 285
4.14 Prepare to Exit, Save, and Later Retrieve This R Session . . . 290
4.15 External Data and/or Data Resources Used in This Lesson . . . 291

5 Oneway Analysis of Variance (ANOVA) 293
5.1 Background . . . 294
5.1.1 Description of the Data . . . 294
5.1.2 Null Hypothesis . . . 295
5.2 Import Data in Comma-Separated Values (.csv) File Format and/or Self-Generate the Data Using R-Based Functions . . . 295
5.3 Organize the Data and Display the Code Book . . . 298
5.4 Conduct a Visual Data Check Using Graphics (e.g., Figures) . . . 303
5.5 Descriptive Statistics for Initial Analysis of the Data . . . 310
5.6 Quality Assurance, Data Distribution, and Tests for Normality . . . 314
5.7 Statistical Test(s) . . . 320
5.7.1 Exploratory Oneway ANOVA . . . 334
5.7.2 Oneway ANOVA Method 1: lm() and anova() functions . . . 335
5.7.3 Oneway ANOVA Method 2: aov() and TukeyHSD() Functions . . . 336
5.8 Summary of Outcomes . . . 340
5.9 Addendum 1: Other Packages for Display of Oneway ANOVA . . . 346
5.10 Addendum 2: Parametric v Nonparametric . . . 349
5.10.1 Parametric Approach to Oneway ANOVA . . . 349
5.10.2 Nonparametric Alternative to Oneway ANOVA . . . 349
5.11 Addendum 3: Additional Practice Datasets . . . 352
5.11.1 Data with Normal Distribution Patterns . . . 354
5.11.2 Data That Do Not Exhibit Normal Distribution Patterns . . . 357
5.12 Prepare to Exit, Save, and Later Retrieve This R Session . . . 358
5.13 External Data and/or Data Resources Used in This Lesson . . . 359

6 Twoway Analysis of Variance (ANOVA) 361
6.1 Background . . . 362
6.1.1 Description of the Data . . . 363
6.1.2 Null Hypothesis . . . 363
6.2 Import Data in Comma-Separated Values . . . 364
6.3 Organize the Data and Display the Code Book . . . 370
6.4 Conduct a Visual Data Check Using Graphics . . . 371
6.5 Descriptive Statistics for Initial Analysis . . . 383
6.6 Quality Assurance, Data Distribution, and Tests . . . 387
6.7 Statistical Test(s) . . . 394
6.8 Summary of Outcomes . . . 405
6.9 Addendum 1: Other Packages for Display of Twoway ANOVA . . . 408
6.10 Addendum 2: Parametric v Nonparametric . . . 409
6.11 Addendum 3: Additional Practice Datasets . . . 412
6.11.1 Data with Normal Distribution Patterns . . . 413
6.11.2 Data That Do Not Exhibit Normal Distribution Patterns . . . 418
6.12 Prepare to Exit, Save, and Later Retrieve . . . 425
6.13 External Data and/or Data Resources Used in this Lesson . . . 426

7 Correlation, Association, Regression, Likelihood, and Prediction 427
7.1 Background . . . 428
7.1.1 Description of the Data . . . 429
7.1.2 Null Hypothesis (Ho) . . . 431
7.2 Import Data in Comma-Separated Values (.csv) . . . 431
7.3 Organize the Data and Display the Code Book . . . 435
7.3.1 Conduct a Visual Data Check Using Graphics (e.g., Figures) . . . 443
7.3.2 Descriptive Statistics for Initial Analysis of the Data . . . 449
7.4 Quality Assurance, Data Distribution . . . 465
7.5 Statistical Test(s) . . . 488
7.5.1 Correlation Using Pearson's r and Spearman's rho . . . 489
7.5.2 Linear Regression Using a Single Predictor Variable . . . 504
7.5.3 Linear Regression Using Multiple Predictor Variables . . . 509
7.5.4 Ordinal Logistic Regression . . . 511
7.5.5 Binary Logistic Regression . . . 523
7.6 Summary of Outcomes . . . 529
7.7 Addendum 1: Multiple Regression . . . 532
7.7.1 Hand-Calculate Multiple Regression . . . 534

7.7.2 Minimal Adequate Model (MAM) for Regression 537
7.7.3 Stepwise Regression 541
7.8 Addendum 2: Likelihood and Odds Ratio 546
7.9 Addendum 3: Parametric v Nonparametric 560
7.10 Addendum 4: Additional Practice Datasets 562
7.10.1 Data with Normal Distribution Patterns 569
7.10.2 Data That Do Not Exhibit Normal Distribution Patterns 574
7.11 Prepare to Exit, Save, and Later Retrieve... 583
7.12 External Data and/or Data Resources Used in... 584

8 Working with Large and Complex Datasets 585
8.1 Background 586
8.1.1 Description of the Data 586
8.1.2 First Null Hypothesis (Ho): Mean Comparisons by Breakout Groups 589
8.1.3 Second Null Hypothesis (Ho): Test of Association 589
8.2 Import Data in Comma-Separated Values (.csv)... 589
8.3 Organize the Data and Display the Code Book 591
8.4 Conduct a Visual Data Check Using Graphics... 606
8.5 Descriptive Statistics for Initial Analysis... 625
8.5.1 Analyses of Object Variables in Original Format 630
8.5.2 Analyses of Boolean-Based Breakouts of Object Variables 632
8.5.2.1 Structure the Boolean-Based Data Selection Process 632
8.5.2.2 Create a New Dataset from an Existing Object Variable 632
8.5.2.3 Download the New Dataset 635
8.6 Quality Assurance, Data Distribution, and... 650
8.6.1 Graphics for Normality 650
8.6.1.1 Histogram 651
8.6.1.2 Density Plot 652
8.6.1.3 Quantile-Quantile (Q-Q) Plot 652
8.6.2 Tests for Normality 654
8.6.2.1 Anderson–Darling Test for Normality 654
8.6.2.2 Jarque–Bera Test for Normality 655
8.6.2.3 Lilliefors (Kolmogorov–Smirnov) Test for Normality Null Hypothesis 656
8.6.2.4 Shapiro–Wilk Test for Normality 657

8.7 Statistical Test(s) . . . 666
8.7.1 Null Hypothesis 1: Analysis of Variance . . . 667
8.7.1.1 Parametric Oneway ANOVA and Tukey HSD Approach . . . 667
8.7.1.2 Parametric Oneway ANOVA and Tukey HSD Approach . . . 681
8.7.1.3 Nonparametric Kruskal–Wallis Approach . . . 685
8.7.2 Null Hypothesis 2: Correlation—Association . . . 690
8.8 Summary of Outcomes . . . 724
8.8.1 Outcomes for First Null Hypothesis (Ho): Mean Comparisons by Breakout Groups . . . 724
8.8.2 Outcomes for Second Null Hypothesis (Ho): Test of Association . . . 725
8.9 Addendum 1: Additional Graphics, to Show Relationships Between and Among Data . . . 725
8.9.1 Graphical Presentation of Grouped (e.g., Factor-Type) Data . . . 726
8.9.1.1 Association Plot . . . 726
8.9.1.2 Bar Plot . . . 728
8.9.1.3 Mosaic Plot . . . 729
8.9.1.4 Pie Chart . . . 731
8.9.1.5 Waffle Chart (e.g., Squared Pie Chart) . . . 735
8.9.2 Graphical Presentation of Interval and Other (e.g., Measured) Numeric Data . . . 740
8.9.2.1 Bagplot (e.g., Bivariate Boxplot) . . . 740
8.9.2.2 Beanplot . . . 741
8.9.2.3 Beeswarm Plot . . . 743
8.9.2.4 Boxplot (e.g., Box-Plot, Box-and-Whiskers Diagram, Box-and-Whiskers Plot) . . . 747
8.9.2.5 Box-Percentile Plot . . . 757
8.9.2.6 Density Plot . . . 760
8.9.2.7 Dotplot (e.g., Dotchart) . . . 760
8.9.2.8 Engelmann–Hecker (EH) Plot . . . 762
8.9.2.9 Histogram . . . 766
8.9.2.10 Line Chart (e.g., Line Graph) . . . 772
8.9.2.11 Pirate Plot . . . 778
8.9.2.12 Quantile-Quantile (Q-Q) Plot . . . 779
8.9.2.13 Scatter Plot (e.g., Scatterplot, Scatter Diagram) . . . 783
8.9.2.14 Color Gradient Plot . . . 787
8.9.2.15 Correlogram . . . 790
8.9.2.16 Hexbin Plot . . . 792
8.9.2.17 Scatter Plot Matrix . . . 794
8.9.2.18 Sunflower Scatterplot . . . 796

8.9.2.19 Stem-and-Leaf Plot . . . 797
8.9.2.20 Stripchart . . . 799
8.9.2.21 Violin Plot . . . 800
8.9.3 Specialized Graphics . . . 801
8.9.3.1 Density Ridge . . . 801
8.9.3.2 Gantt Chart . . . 803
8.9.3.3 Interaction Plot . . . 805
8.9.3.4 Staircase Plot . . . 807
8.9.3.5 Triangular Plot for 3-D Representation . . . 808
8.10 Addendum 2: Graphics Using the lattice Package . . . 810
8.11 Addendum 3: Graphics Using the ggplot2 Package . . . 817
8.12 Addendum 4: Beyond an Introduction to R—Use the tidyverse to Create Subsets of Original Datasets . . . 870
8.12.1 Use the dplyr::filter() Function to Subset a Large Dataset . . . 871
8.12.2 Use the magrittr Package Pipe-Like Operator . . . 877
8.13 Prepare to Exit, Save, and Later Retrieve This R Session . . . 881
8.14 External Data and/or Data Resources Used in This Lesson . . . 881

9 Future Actions and Next Steps 883
9.1 Use of This Text . . . 883
9.2 R and *Beautiful Reporting* with R Markdown . . . 884
9.2.1 Static Reports (e.g., Documents) . . . 885
9.2.2 Dynamic Output (e.g., Presentations) . . . 886
9.3 Future Use of R for Biostatistics . . . 887
9.4 Big Data and BioInformatics . . . 888
9.5 External Resources . . . 888
9.6 Contact the Authors . . . 889

Index 891

List of Figures

1.1 Quality assurance of endurance . . . 15
1.2 Quality assurance of MilkLb365 . . . 17
1.3 Quality assurance of corn yield . . . 21
1.4 Quality assurance of waist . . . 25
1.5 Quality assurance of sorghum yield . . . 27
1.6 Quality assurance of fruit consumption by grade 9–12 students . . . 32
1.7 Quality assurance of highway MPG . . . 35
1.8 Quality assurance of rabbit weights . . . 38
1.9 Quality assurance of systolic blood pressure for adult males . . . 40
1.10 Multiple histograms of birth weight . . . 46
1.11 ggplot2 demonstration 1—simple to complex . . . 49
1.12 ggplot2 demonstration 2—complex . . . 50

2.1 Multiple visualization of weight . . . 66
2.2 Bar plots of section and gender . . . 67
2.3 Multiple visualizations of weight by section and gender . . . 69
2.4 Weight Q-Q plot breakouts by section and gender . . . 85
2.5 Section and gender: frequency distribution overall . . . 90
2.6 Section and gender: frequency distribution breakouts . . . 90
2.7 Weight by section and gender breakouts . . . 92
2.8 Graphical representation of weight by gender . . . 93
2.9 Graphical representation of weight by section . . . 94
2.10 Multiple standard deviations with the same mean . . . 119
2.11 Corn yield by year . . . 123
2.12 Waffle chart of race-ethnicity . . . 131
2.13 Bar plot of race-ethnicity . . . 134
2.14 Dot chart of SBP by race-ethnicity and gender breakouts . . . 138

3.1 Distribution of breed by count—1 . . . 151
3.2 Distribution of percent butterfat—1 . . . 152

3.3 Distribution of percent protein—1 . . . 153
3.4 Distribution of breed by count—2 . . . 155
3.5 Distribution of percent butterfat—2 . . . 159
3.6 Distribution of percent butterfat—3 . . . 160
3.7 Distribution of percent butterfat—4 . . . 162
3.8 Distribution of percent protein—2 . . . 163
3.9 Distribution of percent protein—3 . . . 165
3.10 Distribution of percent protein—4 . . . 166
3.11 Percent butterfat by breed and percent protein by breed—1 . . . 170
3.12 Breakouts of breed by count . . . 175
3.13 Distribution of percent butterfat—5 . . . 180
3.14 Distribution of percent protein—5 . . . 183
3.15 Percent butterfat and percent protein overall and by breed . . . 186
3.16 Percent butterfat by breed and percent protein by breed—2 . . . 196
3.17 Distribution of systolic blood pressure by gender . . . 212
3.18 Distribution of diastolic blood pressure by age breakouts . . . 233

4.1 Comparison of pretest weights to posttest weights using unstacked data—1 . . . 254
4.2 Comparison of pretest weights to posttest weights using unstacked data—2 . . . 255
4.3 Distribution of pretest weights and posttest weights using unstacked data—1 . . . 256
4.4 Distribution of pretest weights and posttest weights using unstacked data—2 . . . 262
4.5 Distribution of pretest weights and posttest weights using unstacked data—3 . . . 267
4.6 Distribution of datapoints by groups using stacked data . . . 274
4.7 Comparison of parametric pre and parametric post datapoints . . . 288

5.1 Distribution of systolic blood pressure by lifestyle breakouts—1 . . . 302
5.2 Distribution of systolic blood pressure by lifestyle breakouts—2 . . . 304
5.3 Distribution of systolic blood pressure by lifestyle breakouts—3 . . . 305
5.4 Distribution of systolic blood pressure by lifestyle breakouts—4 . . . 307
5.5 Distribution of systolic blood pressure by lifestyle breakouts—5 . . . 308

5.6 Distribution of systolic blood pressure by lifestyle breakouts—6 . . . 309
5.7 Distribution of systolic blood pressure by lifestyle breakouts—7 . . . 310
5.8 Distribution of systolic blood pressure by lifestyle breakouts—8 . . . 320
5.9 Distribution of systolic blood pressure . . . 328
5.10 Distribution of systolic blood pressure by lifestyle breakouts—9 . . . 328
5.11 Distribution of systolic blood pressure by lifestyle breakouts—10 . . . 329
5.12 Distribution of systolic blood pressure by lifestyle breakouts—11 . . . 330
5.13 Distribution of systolic blood pressure by lifestyle breakouts—12 . . . 331
5.14 Distribution of systolic blood pressure by lifestyle breakouts—13 . . . 332
5.15 Distribution of systolic blood pressure by lifestyle breakouts—14 . . . 333
5.16 Distribution of systolic blood pressure by lifestyle breakouts—15 . . . 333
5.17 Distribution of systolic blood pressure by lifestyle breakouts—16 . . . 334
5.18 Distribution of systolic blood pressure by lifestyle breakouts—17 . . . 340
5.19 Systolic blood pressure comparative family-wise confidence levels . . . 342
5.20 Distribution of systolic blood pressure by lifestyle breakouts—18 . . . 343
5.21 Distribution of systolic blood pressure by lifestyle breakouts—19 . . . 343
5.22 Distribution of systolic blood pressure by lifestyle breakouts—20 . . . 344
5.23 Distribution of systolic blood pressure by lifestyle breakouts—21 . . . 345
5.24 Distribution of systolic blood pressure by lifestyle breakouts—22 . . . 345
5.25 Differences of systolic blood pressure means by lifestyle breakouts (Tukey—alpha = 0.05) . . . 349

6.1 Distribution of systolic blood pressure overall . . . 372
6.2 Distribution of systolic blood pressure overall and by breakouts—1 . . . 376
6.3 Distribution of systolic blood pressure overall and by breakouts—2 . . . 378

6.4 Distribution of systolic blood pressure overall and by breakouts—3 . . . 379
6.5 Distribution of systolic blood pressure overall and by breakouts—4 . . . 380
6.6 Factorial representation of systolic blood pressure overall and by breakouts . . . 383
6.7 Distribution of systolic blood pressure overall and by breakouts—5 . . . 385
6.8 Distribution of systolic blood pressure overall and by breakouts—6 . . . 393
6.9 Distribution of systolic blood pressure by breakouts . . . 403
6.10 Interaction of gender and drug for systolic blood pressure—1 . . . 403
6.11 Interaction of gender and drug for systolic blood pressure—2 . . . 405
6.12 Interaction of gender and drug for systolic blood pressure—3 . . . 405
6.13 Corn yield by county breakouts . . . 422
6.14 Corn yield by management breakouts . . . 423

7.1 Distribution of factor-type object variables gender, RaceEthnicity, BMIStatus, and obesity . . . 445
7.2 Distribution of numeric-type object variable age . . . 446
7.3 Distribution of numeric-type object variable cholesterol . . . 447
7.4 Distribution of numeric-type object variable systolic blood pressure . . . 448
7.5 Distribution of numeric-type object variable diastolic blood pressure . . . 448
7.6 Distribution of numeric-type object variable body mass index . . . 449
7.7 Distribution of age by factor-type object variables gender, RaceEthnicity, BMIStatus, and obesity . . . 460
7.8 Distribution of cholesterol by factor-type object variables gender, RaceEthnicity, BMIStatus, and obesity . . . 461
7.9 Distribution of systolic blood pressure by factor-type object variables gender, RaceEthnicity, BMIStatus, and obesity . . . 462
7.10 Distribution of diastolic blood pressure by factor-type object variables gender, RaceEthnicity, BMIStatus, and obesity . . . 463
7.11 Distribution of body mass index by factor-type object variables gender, RaceEthnicity, BMIStatus, and obesity . . . 464

7.12 Normality of numeric-type object variables AgeYears, TotalCholesterolmgdL, SBPmmHg, DBPmmHg, and BMIMetric . . . 469
7.13 Normality of age by factor-type object variables gender, RaceEthnicity, BMIStatus, and obesity . . . 480
7.14 Normality of cholesterol by factor-type object variables gender, RaceEthnicity, BMIStatus, and obesity . . . 481
7.15 Normality of systolic blood pressure by factor-type object variables gender, RaceEthnicity, BMIStatus, and obesity . . . 482
7.16 Normality of diastolic blood pressure by factor-type object variables gender, RaceEthnicity, BMIStatus, and obesity . . . 483
7.17 Normality of body mass index by factor-type object variables gender, RaceEthnicity, BMIStatus, and obesity . . . 485
7.18 Normality of body mass index by obesity and accommodation for missing data . . . 487
7.19 Brute force one-by-one display of correlation: SBP by Age, SBP by cholesterol, SBP by DBP, and SBP by BMI . . . 498
7.20 Correlation matrix of numeric-type object variables age, cholesterol, SBP, DBP, and BMI—Pearson's r . . . 499
7.21 Correlation matrix of numeric-type object variables age, cholesterol, SBP, DBP, and BMI—Spearman's rho . . . 500
7.22 Correlation matrix of numeric-type object variables age, cholesterol, SBP, DBP, and BMI—Pearson's r and Spearman's rho—1 . . . 502
7.23 Correlation matrix of numeric-type object variables age, cholesterol, SBP, DBP, and BMI—Pearson's r and Spearman's rho—2 . . . 503
7.24 Correlation matrix of numeric-type object variables age, cholesterol, SBP, DBP, and BMI . . . 504
7.25 Mean systolic blood pressure by body mass index breakout groups and mean diastolic blood pressure by body mass index breakout groups . . . 513
7.26 Probability of body mass index breakout group assignment by systolic blood pressure . . . 521
7.27 Probability of body mass index breakout group assignment by diastolic blood pressure . . . 522
7.28 Mean systolic blood pressure by obesity breakout groups and mean diastolic blood pressure by obesity breakout groups . . . 525
7.29 Probability of obesity by systolic blood pressure and provability of obesity by diastolic blood pressure . . . 529

7.30 Scatterplot of two variables (vigor at purchase by vigor at 100 pounds) with added regression line . . . 537
7.31 Percentage distribution of coin tosses . . . 549
7.32 Display of odds ratio—1 . . . 553
7.33 Display of odds ratio—2 . . . 556
7.34 Display of odds ratio—3 . . . 559

8.1 Quality assurance exploratory data analysis—1 . . . 607
8.2 Quality assurance exploratory data analysis—2 . . . 608
8.3 Quality assurance exploratory data analysis—3 . . . 608
8.4 Quality assurance exploratory data analysis—4 . . . 609
8.5 Quality assurance exploratory data analysis—5 . . . 609
8.6 Quality assurance exploratory data analysis—6 . . . 610
8.7 Quality assurance exploratory data analysis—7 . . . 611
8.8 Quality assurance exploratory data analysis—8 . . . 611
8.9 Quality assurance exploratory data analysis—9 . . . 612
8.10 Quality assurance exploratory data analysis—10 . . . 612
8.11 Quality assurance exploratory data analysis—11 . . . 612
8.12 Quality assurance exploratory data analysis—12 . . . 613
8.13 Quality assurance exploratory data analysis—13 . . . 613
8.14 Quality assurance exploratory data analysis—14 . . . 614
8.15 Quality assurance exploratory data analysis—15 . . . 615
8.16 Quality assurance exploratory data analysis—16 . . . 615
8.17 Quality assurance exploratory data analysis—17 . . . 616
8.18 Quality assurance exploratory data analysis—18 . . . 617
8.19 Quality assurance exploratory data analysis—19 . . . 617
8.20 Quality assurance exploratory data analysis—20 . . . 618
8.21 Quality assurance exploratory data analysis—21 . . . 618
8.22 Quality assurance exploratory data analysis—22 . . . 619
8.23 Quality assurance exploratory data analysis—23 . . . 619
8.24 Quality assurance exploratory data analysis—24 . . . 620
8.25 Quality assurance exploratory data analysis—25 . . . 621
8.26 Quality assurance exploratory data analysis—26 . . . 621
8.27 Quality assurance exploratory data analysis—27 . . . 622
8.28 Quality assurance exploratory data analysis—28 . . . 624
8.29 Quality assurance review of original dataset and new dataset using systolic blood pressure and age . . . 635
8.30 Quality assurance review of new dataset using race and age . . . 641
8.31 Quality assurance review of new dataset using race, sex, and age . . . 643
8.32 Distribution (histogram) of two variables: XNormal and XNotNormal . . . 651
8.33 Distribution (density plot) of two variables: XNormal and XNotNormal . . . 652

8.34 Distribution (QQ plot) of two variables: XNormal and XNotNormal . . . 653
8.35 Distribution pattern of systolic blood pressure with multiple subsets . . . 662
8.36 Distribution pattern of diastolic blood pressure and age with multiple subsets . . . 669
8.37 Distribution of diastolic blood pressure (age 040–049)—1 . . . 672
8.38 Distribution of diastolic blood pressure (age 040–049)—2 . . . 673
8.39 Distribution of diastolic blood pressure (age 040–049)—3 . . . 673
8.40 Distribution of diastolic blood pressure (age 040–049)—4 . . . 677
8.41 Distribution of diastolic blood pressure (age 040–049) and race—1 . . . 681
8.42 Distribution of diastolic blood pressure (age 040–049) and race—2 . . . 685
8.43 Distribution of systolic blood pressure by increasingly restrictive subsets—1 . . . 695
8.44 Distribution of weight by increasingly restrictive subsets—1 . . . 696
8.45 Distribution of systolic blood pressure by increasingly restrictive subsets—2 . . . 699
8.46 Distribution of systolic blood pressure by increasingly restrictive subsets—3 . . . 700
8.47 Distribution of systolic blood pressure by increasingly restrictive subsets—4 . . . 701
8.48 Distribution of weight by increasingly restrictive subsets—2 . . . 702
8.49 Distribution of weight by increasingly restrictive subsets—3 . . . 703
8.50 Distribution of weight by increasingly restrictive subsets—4 . . . 704
8.51 Distribution of systolic blood pressure from multiple perspectives—1 . . . 706
8.52 Normality of systolic blood pressure for selected subjects—1 . . . 707
8.53 Distribution of weight from multiple perspectives—1 . . . 709
8.54 Normality of weight for selected subjects—1 . . . 709
8.55 Side-by-side QQ plot of systolic blood pressure and weight . . . 713
8.56 Correlation of systolic blood pressure and weight—1 . . . 719
8.57 Correlation of systolic blood pressure and weight—2 . . . 720
8.58 Correlation of systolic blood pressure and weight—3 . . . 721
8.59 Correlation of systolic blood pressure and weight—4 . . . 722
8.60 Association plot . . . 727
8.61 Stacked bar plot and side-by-Sie bar plot . . . 729

8.62 Mosaic plot—1 . . . 730
8.63 Mosaic plot—2 . . . 731
8.64 Pie chart—1 . . . 733
8.65 Pie chart—2 . . . 734
8.66 Pie chart—3 . . . 735
8.67 Pie chart—4 . . . 736
8.68 Waffle chart—1 . . . 738
8.69 Waffle chart—2 . . . 739
8.70 Bagplot . . . 741
8.71 Beanplot—1 . . . 742
8.72 Beanplot—2 . . . 743
8.73 Beeswarm plot—1 . . . 745
8.74 Beeswarm plot—2 . . . 746
8.75 Boxplot—1 . . . 749
8.76 Boxplot—2 . . . 749
8.77 Boxplot—3 . . . 752
8.78 Boxplot—4 . . . 752
8.79 Beanplot and boxplot—1 . . . 754
8.80 Beanplot and boxplot—2 . . . 755
8.81 Boxplot—5 . . . 756
8.82 Boxplot—6 . . . 756
8.83 Box-percentile plot—1 . . . 758
8.84 Box-percentile plot—2 . . . 759
8.85 Box-percentile plot—3 . . . 760
8.86 Density plot . . . 761
8.87 Dot plot—1 . . . 762
8.88 Dot plot—2 . . . 762
8.89 Engelmann–Hecker plot and boxplot—1 . . . 763
8.90 Engelmann–Hecker plot and boxplot—2 . . . 764
8.91 Engelmann–Hecker plot and boxplot—3 . . . 765
8.92 Engelmann–Hecker plot and boxplot—4 . . . 765
8.93 Histogram—1 . . . 766
8.94 Histogram—2 . . . 767
8.95 Histogram—3 . . . 768
8.96 Histogram—4 . . . 769
8.97 Histogram—5 . . . 770
8.98 Histogram—6 . . . 771
8.99 Histogram—7 . . . 772
8.100 Histogram—8 . . . 773
8.101 Line chart—1 . . . 775
8.102 Line chart—2 . . . 778
8.103 Pirate plot—1 . . . 779
8.104 Pirate plot—2 . . . 780
8.105 Quantile–quantile plot—1 . . . 782

8.106 Quantile–quantile plot—2 . . . 783
8.107 Scatter plot—1 . . . 785
8.108 Scatter plot—2 . . . 787
8.109 Scatter plot—3 . . . 788
8.110 Color gradient plot—1 . . . 789
8.111 Color gradient plot—2 . . . 790
8.112 Color gradient plot—3 . . . 791
8.113 Color gradient plot—4 . . . 791
8.114 Correlogram . . . 793
8.115 Hexbin plot—1 . . . 794
8.116 Hexbin plot—2 . . . 794
8.117 Scatter plot matrix—1 . . . 795
8.118 Scatter plot matrix—2 . . . 796
8.119 Sunflower scatterplot . . . 797
8.120 Stripchart . . . 800
8.121 Violin plot—1 . . . 801
8.122 Violin plot—2 . . . 801
8.123 Density ridge—1 . . . 802
8.124 Density ridge—2 . . . 803
8.125 Density ridge—3 . . . 803
8.126 Density ridge—4 . . . 804
8.127 Gantt chart . . . 805
8.128 Interaction plot—1 . . . 806
8.129 Interaction plot—2 . . . 807
8.130 Staircase plot . . . 809
8.131 Triangular plot . . . 810
8.132 lattice package—barchart . . . 811
8.133 lattice package—box plot and density plot . . . 813
8.134 lattice package—histogram and QQ plot . . . 815
8.135 lattice package—scatter plot, scatter plot matrix (SPLOM), and 3D scatter plot . . . 817
8.136 ggplot2 package—barchart 1 . . . 818
8.137 ggplot2 package—barchart 2 . . . 820
8.138 ggplot2 package—barchart 3 . . . 822
8.139 ggplot2 package—barchart 4 . . . 823
8.140 ggplot2 package—box plot 1 . . . 824
8.141 ggplot2 package—box plot 2 . . . 825
8.142 ggplot2 package—box plot 3 . . . 826
8.143 ggplot2 package—density plot 1 . . . 826
8.144 ggplot2 package—density plot 2 . . . 827
8.145 ggplot2 package—density plot 3 . . . 828
8.146 ggplot2 package—density plot 4 . . . 829
8.147 ggplot2 package—dot plot 1 . . . 830
8.148 ggplot2 package—dot plot 2 . . . 831

8.149 ggplot2 package—dot plot 3 . . . 832
8.150 ggplot2 package—dot plot 4 . . . 833
8.151 ggplot2 package—scatter plot 1 . . . 836
8.152 ggplot2 package—scatter plot 2 . . . 840
8.153 ggplot2 package and GGally package—scatter plot 1 . . . 842
8.154 ggplot2 package and GGally package—scatter plot 2 . . . 843
8.155 ggplot2 package and GGally package—scatter plot 3 . . . 844
8.156 ggplot2 package—violin plot 1 . . . 844
8.157 ggplot2 package—violin plot 2 . . . 845
8.158 ggplot2 package—violin plot 3 . . . 846
8.159 ggplot2 package—violin plot 4 . . . 847
8.160 ggplot2 package—violin plot 5 . . . 848
8.161 ggplot2 package—histogram 1 . . . 849
8.162 ggplot2 package—histogram 2 . . . 851
8.163 ggplot2 package—histogram 3 . . . 852
8.164 ggplot2 package—histogram 4 . . . 853
8.165 ggplot2 package—histogram 5 . . . 854
8.166 ggplot2 package—frequency polygon . . . 857
8.167 ggplot2 package—stripchart . . . 858
8.168 ggplot2 package—QQ plot 1 . . . 860
8.169 ggplot2 package—QQ plot 2 . . . 862
8.170 ggplot2 package—QQ plot 3 . . . 863
8.171 ggplot2 package—interaction plot . . . 865
8.172 ggplot2 package—line chart . . . 868
8.173 ggplot2 package—tile map . . . 870
8.174 Multiple subsets of systolic blood pressure . . . 877
8.175 Comparison of selected variables—original dataset and dataset after multiple subsets . . . 880