

/// DETAILED CONTENTS

PREFACE	xviii		
ABOUT THE AUTHOR	xxv		
1: Preparing Data for Analysis and Visualization in R: The R-Team and the Pot Policy Problem	1		
1.1 Choosing and learning R	1		
1.2 Learning R with publicly available data	4		
1.3 Achievements to unlock	4		
1.4 The tricky weed problem	4		
1.4.1 <i>Marijuana legalization</i>	4		
1.5 Achievement 1: Observations and variables	8		
1.5.1 <i>Defining observations and variables</i>	8		
1.5.2 <i>Entering and storing variables in R</i>	8		
1.5.3 <i>Achievement 1: Check your understanding</i>	10		
1.6 Achievement 2: Using reproducible research practices	10		
1.6.1 <i>Using comments to organize and explain code</i>	10		
1.6.2 <i>Including a prolog to introduce a script file</i>	12		
1.6.3 <i>Naming objects</i>	14		
1.6.4 <i>Achievement 2: Check your understanding</i>	16		
1.7 Achievement 3: Understanding and changing data types	16		
1.7.1 <i>Numeric data type</i>	16		
1.7.2 <i>Integer data type</i>	17		
1.7.3 <i>Logical data type</i>	17		
1.7.4 <i>Character data type</i>	18		
1.7.5 <i>Factor data type</i>	19		
1.7.6 <i>Achievement 3: Check your understanding</i>	19		
1.8 Achievement 4: Entering or loading data into R	19		
1.8.1 <i>Creating vectors for different data types</i>	19		
1.8.2 <i>Creating a matrix to store data in rows and columns</i>	21		
1.8.3 <i>Creating a data frame</i>	23		
1.8.4 <i>Importing data frames from outside sources</i>	25		
1.8.5 <i>Importing a comma separated values (csv) file</i>	25		
1.8.6 <i>Cleaning data types in an imported file</i>	30		
1.8.7 <i>Achievement 4: Check your understanding</i>	32		
1.9 Achievement 5: Identifying and treating missing values	32		
1.9.1 <i>Recoding missing values to NA</i>	33		
1.9.2 <i>Achievement 5: Check your understanding</i>	38		
1.10 Achievement 6: Building a basic bar chart	38		
1.10.1 <i>Omitting NA from a graph</i>	40		
1.10.2 <i>Working with color in a bar chart</i>	41		
1.10.3 <i>Using special variables in graphs</i>	44		
1.10.4 <i>Achievement 6: Check your understanding</i>	48		
1.11 Chapter summary	49		
1.11.1 <i>Achievements unlocked in this chapter: Recap</i>	49		
1.11.2 <i>Chapter exercises</i>	49		
2: Computing and Reporting Descriptive Statistics: The R-Team and the Troubling Transgender Health Care Problem	52		
2.1 Achievements to unlock	53		
2.2 The transgender health care problem	53		
2.3 Data, codebook, and R packages for learning about descriptive statistics	56		
2.4 Achievement 1: Understanding variable types and data types	57		
2.4.1 <i>Data types for categorical variables</i>	57		
2.4.2 <i>Data types for continuous variables</i>	57		
2.4.3 <i>Achievement 1: Check your understanding</i>	58		
2.5 Achievement 2: Choosing and conducting descriptive analyses for categorical (factor) variables	58		
2.5.1 <i>Computing frequencies and frequency distributions</i>	58		
2.5.2 <i>Making a basic table of frequencies and percentages</i>	61		

2.5.3 Data management	62	3.4.3 Waffle charts	157
2.5.4 Achievement 2: Check your understanding	70	3.4.4 Achievement 1: Check your understanding	160
2.6 Achievement 3: Choosing and conducting descriptive analyses for continuous (numeric) variables	71	3.5 Achievement 2: Choosing and creating graphs for a single continuous variable	160
2.6.1 Why frequency distributions do not work for numeric variables	71	3.5.1 Histograms	161
2.6.2 Defining and calculating central tendency	71	3.5.2 Density plots	167
2.6.3 Messy data and computing measures of central tendency	79	3.5.3 Boxplots	170
2.6.4 Defining and calculating spread	83	3.5.4 Achievement 2: Check your understanding	173
2.6.5 Achievement 3: Check your understanding	91	3.6 Achievement 3: Choosing and creating graphs for two variables at once	173
2.7 Achievement 4: Developing clear tables for reporting descriptive statistics	91	3.6.1 Mosaic plots for two categorical variables	173
2.7.1 Data cleaning before analysis	92	3.6.2 Bar charts for two categorical variables	177
2.7.2 Creating a table from the clean data	103	3.6.3 Bar charts, point charts, boxplots, and violin plots for one categorical and one continuous variable	182
2.7.3 Creating a table from clean data (another way)	114	3.6.4 Line graphs and scatterplots for two continuous variables	202
2.8 Chapter summary	133	3.6.5 Achievement 3: Check your understanding	213
2.8.1 Achievements unlocked in this chapter: Recap	133	3.7 Achievement 4: Ensuring graphs are well-formatted with appropriate and clear titles, labels, colors, and other features	213
2.8.2 Chapter exercises	134	3.8 Chapter summary	214
		3.8.1 Achievements unlocked in this chapter: Recap	214
		3.8.2 Chapter exercises	215
3: Data Visualization: The R-Team and the Tricky Trigger Problem	136	4: Probability Distributions and Inference: The R-Team and the Opioid Overdose Problem	218
3.1 Achievements to unlock	137	4.1 Achievements to unlock	219
3.2 The tricky trigger problem	137	4.2 The awful opioid overdose problem	219
3.2.1 Comparing gun deaths to other causes of death in the United States	138	4.3 Data, codebook, and R packages for learning about distributions	227
3.2.2 Weapons used in homicide	138	4.4 Achievement 1: Defining and using the probability distributions to infer from a sample	227
3.2.3 Types of guns used in homicide	140	4.4.1 Characteristics of probability distributions	228
3.2.4 The role of gun manufacturers in reducing gun deaths	140	4.4.2 Types of probability distributions	228
3.3 Data, codebook, and R packages for graphs	142	4.4.3 Achievement 1: Check your understanding	228
3.4 Achievement 1: Choosing and creating graphs for a single categorical variable	143		
3.4.1 Pie charts	144		
3.4.2 Bar charts	146		

4.5 Achievement 2: Understanding the characteristics and uses of a binomial distribution of a binary variable	228	4.9.2 Confidence intervals for percentages	268
4.5.1 Using distributions to learn about populations from samples	229	4.9.3 Other confidence intervals	271
4.5.2 Statistical properties of a binomial random variable	229	4.9.4 Achievement 6: Check your understanding	272
4.5.3 Visualizing and using the binomial distribution	230	4.10 Chapter summary	272
4.5.4 Achievement 2: Check your understanding	241	4.10.1 Achievements unlocked in this chapter: Recap	272
4.6 Achievement 3: Understanding the characteristics and uses of the normal distribution of a continuous variable	242	4.10.2 Chapter exercises	273
4.6.1 Probability density function	242	5: Computing and Interpreting Chi-Squared: The R-Team and the Vexing Voter Fraud Problem	276
4.6.2 Finding the area under the curve	248	5.1 Achievements to unlock	277
4.6.3 Achievement 3: Check your understanding	249	5.2 The voter fraud problem	278
4.7 Achievement 4: Computing and interpreting z-scores to compare observations to groups	249	5.3 Data, documentation, and R packages for learning about chi-squared	279
4.7.1 Defining the z-score	250	5.4 Achievement 1: Understanding the relationship between two categorical variables using bar charts, frequencies, and percentages	280
4.7.2 Calculating and interpreting z-scores	250	5.4.1 Data cleaning	281
4.7.3 Achievement 4: Check your understanding	250	5.4.2 Using descriptive statistics to examine the relationship between two categorical variables	286
4.8 Achievement 5: Estimating population means from sample means using the normal distribution	251	5.4.3 Using graphs to examine the relationship between two categorical variables	289
4.8.1 Samples and populations	251	5.4.4 Achievement 1: Check your understanding	291
4.8.2 Using the normal distribution and samples to understand populations	251	5.5 Achievement 2: Computing and comparing observed and expected values for the groups	291
4.8.3 Examining a sample from a population	252	5.5.1 Observed values	291
4.8.4 Examining a sample of samples from a population	253	5.5.2 Expected values	291
4.8.5 The Central Limit Theorem	256	5.5.3 Comparing observed and expected values	292
4.8.6 The standard error	257	5.5.4 The assumptions of the chi-squared test of independence	293
4.8.7 Standard deviation versus standard error	259	5.5.5 Achievement 2: Check your understanding	294
4.8.8 Achievement 5: Check your understanding	259	5.6 Achievement 3: Calculating the chi-squared statistic for the test of independence	294
4.9 Achievement 6: Computing and interpreting confidence intervals around means and proportions	259	5.6.1 Summing the differences between observed and expected values	294
4.9.1 Computing and interpreting a 95% confidence interval for a mean	261		

5.6.2 Squaring the summed differences	294	5.10 Achievement 7: Computing and interpreting effect sizes to understand the strength of a significant chi-squared relationship	316
5.6.3 Using R to compute chi-squared	295	5.10.1 Computing the Cramér's V statistic	316
5.6.4 Achievement 3: Check your understanding	295	5.10.2 Interpreting Cramér's V	317
5.7 Achievement 4: Interpreting the chi-squared statistic and making a conclusion about whether or not there is a relationship	296	5.10.3 An example of chi-squared for two binary variables	318
5.7.1 Visualizing the chi-squared distribution	296	5.10.4 Interpreting the Yates continuity correction	320
5.7.2 Area under the curve	297	5.10.5 The phi coefficient effect size statistic	321
5.7.3 Using the chi-squared distribution to determine probability	298	5.10.6 The odds ratio for effect size with two binary variables	322
5.7.4 Selecting the threshold for statistical significance	300	5.10.7 Achievement 7: Check your understanding	324
5.7.5 Achievement 4: Check your understanding	300	5.11 Achievement 8: Understanding the options for failed chi-squared assumptions	324
5.8 Achievement 5: Using Null Hypothesis Significance Testing to organize statistical testing	301	5.11.1 Violating the assumption that the variables must be nominal or ordinal	325
5.8.1 NHST Step 1: Write the null and alternate hypotheses	301	5.11.2 Violating the assumption of expected values of 5 or higher in at least 80% of groups	325
5.8.2 NHST Step 2: Compute the test statistic	301	5.11.3 Violating the independent observations assumption	325
5.8.3 NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)	302	5.12 Chapter summary	326
5.8.4 NHST Step 4: If the probability that the null is true is very small, usually less than 5%, reject the null hypothesis	302	5.12.1 Achievements unlocked in this chapter: Recap	326
5.8.5 NHST Step 5: If the probability that the null is true is not small, usually 5% or greater, retain the null hypothesis	302	5.12.2 Chapter exercises	327
5.8.6 Achievement 5: Check your understanding	302	6: Conducting and Interpreting t-Tests: The R-Team and the Blood Pressure Predicament	330
5.9 Achievement 6: Using standardized residuals to understand which groups contributed to significant relationships	307	6.1 Achievements to unlock	332
5.9.1 Using standardized residuals following chi-squared tests	307	6.2 The blood pressure predicament	332
5.9.2 Interpreting standardized residuals and chi-squared results	316	6.3 Data, codebook, and R packages for learning about t-tests	332
5.9.3 Achievement 6: Check your understanding	316	6.4 Achievement 1: Understanding the relationship between one categorical variable and one continuous variable using histograms, means, and standard deviations	334
		6.4.1 Achievement 1: Check your understanding	338

6.5 Achievement 2: Comparing a sample mean to a population mean with a one-sample <i>t</i> -test	338	6.8.2 Cohen's <i>d</i> for independent-samples <i>t</i> -tests	356
6.5.1 NHST Step 1: Write the null and alternate hypotheses	339	6.8.3 Cohen's <i>d</i> for dependent-samples <i>t</i> -tests	357
6.5.2 NHST Step 2: Compute the test statistic	339	6.9 Achievement 6: Examining and checking the underlying assumptions for using the <i>t</i> -test	358
6.5.3 NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)	340	6.9.1 Testing normality	358
6.5.4 NHST Steps 4 and 5: Interpret the probability and write a conclusion	342	6.9.2 Achievement 6: Check your understanding	365
6.5.5 Achievement 2: Check your understanding	343	6.10 Achievement 7: Identifying and using alternate tests when <i>t</i> -test assumptions are not met	366
6.6 Achievement 3: Comparing two unrelated sample means with an independent-samples <i>t</i> -test	343	6.10.1 Alternative to one-sample <i>t</i> -test failing assumptions: The sign test	366
6.6.1 NHST Step 1: Write the null and alternate hypotheses	346	6.10.2 Alternative when the dependent-samples <i>t</i> -test fails assumptions: The Wilcoxon signed-ranks test	368
6.6.2 NHST Step 2: Compute the test statistic	346	6.10.3 Alternative when the independent-samples <i>t</i> -test normality assumption fails: The Mann-Whitney U test	370
6.6.3 NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)	349	6.10.4 Effect size for Mann-Whitney U	372
6.6.4 NHST Steps 4 and 5: Interpret the probability and write a conclusion	349	6.10.5 Alternative when the independent-samples <i>t</i> -test variance assumption fails: The Kolmogorov-Smirnov test	373
6.6.5 Achievement 3: Check your understanding	350	6.10.6 Achievement 7: Check your understanding	376
6.7 Achievement 4: Comparing two related sample means with a dependent-samples <i>t</i> -test	350	6.11 Chapter summary	377
6.7.1 NHST Step 1: Write the null and alternate hypotheses	352	6.11.1 Achievements unlocked in this chapter: Recap	377
6.7.2 NHST Step 2: Compute the test statistic	352	6.11.2 Chapter exercises	377
6.7.3 NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)	353	7: Analysis of Variance: The R-Team and the Technical Difficulties Problem	380
6.7.4 NHST Steps 4 and 5: Interpret the probability and write a conclusion	353	7.1 Achievements to unlock	381
6.7.5 Achievement 4: Check your understanding	354	7.2 The technical difficulties problem	382
6.8 Achievement 5: Computing and interpreting an effect size for significant <i>t</i> -tests	354	7.3 Data, codebook, and R packages for learning about ANOVA	382
6.8.1 Cohen's <i>d</i> for one-sample <i>t</i> -tests	355	7.4 Achievement 1: Exploring the data using graphics and descriptive statistics	383
		7.4.1 Data management	383
		7.4.2 Exploratory data analysis	390
		7.4.3 Achievement 1: Check your understanding	392

7.5 Achievement 2: Understanding and conducting one-way ANOVA	392	7.11 Chapter summary	448
7.5.1 The F test statistic for ANOVA	393	7.11.1 Achievements unlocked in this chapter: Recap	448
7.5.2 Achievement 2: Check your understanding	397	7.11.2 Chapter exercises	448
7.6 Achievement 3: Choosing and using post hoc tests and contrasts	398	8: Correlation Coefficients: The R-Team and the Clean Water Conundrum	452
7.6.1 Post hoc tests	398	8.1 Achievements to unlock	454
7.6.2 Planned comparisons	403	8.2 The clean water conundrum	454
7.6.3 Achievement 3: Check your understanding	414	8.3 Data and R packages for learning about correlation	460
7.7 Achievement 4: Computing and interpreting effect sizes for ANOVA	414	8.4 Achievement 1: Exploring the data using graphics and descriptive statistics	461
7.7.1 Achievement 4: Check your understanding	417	8.4.1 Make a scatterplot to examine the relationship	463
7.8 Achievement 5: Testing ANOVA assumptions	418	8.4.2 Achievement 1: Check your understanding	464
7.8.1 Testing normality	418	8.5 Achievement 2: Computing and interpreting Pearson's r correlation coefficient	464
7.8.2 Homogeneity of variances assumption	420	8.5.1 Computing and interpreting the covariance between two variables	464
7.8.3 ANOVA assumptions recap	421	8.5.2 Computing the Pearson's r correlation between two variables	468
7.8.4 Achievement 5: Check your understanding	421	8.5.3 Interpreting the direction of the Pearson's product-moment correlation coefficient, r	468
7.9 Achievement 6: Choosing and using alternative tests when ANOVA assumptions are not met	422	8.5.4 Interpreting the strength of the Pearson's product-moment correlation coefficient	474
7.9.1 Calculating an alternate F-statistic for failing the homogeneity assumption	422	8.5.5 Achievement 2: Check your understanding	475
7.9.2 The Kruskal-Wallis test for failing the normality assumption	429	8.6 Achievement 3: Conducting an inferential statistical test for Pearson's r correlation coefficient	476
7.9.3 Achievement 6: Check your understanding	434	8.6.1 NHST Step 1: Writing the null and alternate hypotheses	476
7.10 Achievement 7: Understanding and conducting two-way ANOVA	434	8.6.2 NHST Step 2: Computing the test statistic	476
7.10.1 Exploratory data analysis for two-way ANOVA	434	8.6.3 NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)	478
7.10.2 Two-way ANOVA NHST	438	8.6.4 NHST Steps 4 and 5: Interpret the probability and write a conclusion	478
7.10.3 Post hoc test for two-way ANOVA	439	8.6.5 Achievement 3: Check your understanding	479
7.10.4 Two-way ANOVA assumptions	443		
7.10.5 Alternatives when two-way ANOVA assumptions fail	445		
7.10.6 Achievement 7: Check your understanding	447		

8.7 Achievement 4: Examining effect size for Pearson's r with the coefficient of determination	479	8.10.4 NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)	506
8.7.1 Calculating the coefficient of determination	480	8.10.5 NHST Steps 4 and 5: Interpret the probability and write a conclusion	507
8.7.2 Using R to calculate the coefficient of determination	480	8.10.6 Assumption checking for Spearman's rho	507
8.7.3 Achievement 4: Check your understanding	484	8.10.7 Checking the monotonic assumption	508
8.8 Achievement 5: Checking assumptions for Pearson's r correlation analyses	484	8.10.8 Achievement 7: Check your understanding	508
8.8.1 Checking the normality assumption	485	8.11 Achievement 8: Introducing partial correlations	509
8.8.2 Checking the linearity assumption	489	8.11.1 Computing Pearson's r partial correlations	509
8.8.3 Checking the homoscedasticity assumption	491	8.11.2 Computing Spearman's rho partial correlations	512
8.8.4 Interpreting the assumption checking results	493	8.11.3 Significance testing for partial correlations	513
8.8.5 Achievement 5: Check your understanding	493	8.11.4 Checking assumptions for partial correlations	513
8.9 Achievement 6: Transforming the variables as an alternative when Pearson's r correlation assumptions are not met	493	8.11.5 Interpreting results when assumptions are not met	515
8.9.1 NHST Step 1: Write the null and alternate hypotheses	501	8.12 Chapter summary	516
8.9.2 NHST Step 2: Compute the test statistic	501	8.12.1 Achievements unlocked in this chapter: Recap	516
8.9.3 NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)	501	8.12.2 Chapter exercises	517
8.9.4 NHST Steps 4 and 5: Interpret the probability and write a conclusion	501	9: Linear Regression: The R-Team and the Needle Exchange Examination	520
8.9.5 Testing assumptions for Pearson's r between the transformed variables	502	9.1 Achievements to unlock	521
8.9.6 Achievement 6: Check your understanding	504	9.2 The needle exchange examination	522
8.10 Achievement 7: Using Spearman's rho as an alternative when Pearson's r correlation assumptions are not met	504	9.3 Data, codebook, and R packages for linear regression practice	523
8.10.1 Computing Spearman's rho correlation coefficient	504	9.4 Achievement 1: Using exploratory data analysis to learn about the data before developing a linear regression model	524
8.10.2 NHST Step 1: Write the null and alternate hypotheses	504	9.4.1 Importing and merging data sources	524
8.10.3 NHST Step 2: Compute the test statistic	505	9.4.2 Checking the descriptive statistics	528
		9.4.3 Using a scatterplot to explore the relationship	531
		9.4.4 Using a correlation coefficient to explore the relationship	536

9.4.5 Exploring the data by comparing means across groups	537	9.9 Achievement 6: Checking assumptions and conducting diagnostics	563
9.4.6 Exploring the data with boxplots	538	9.9.1 Assumptions of simple linear regression	563
9.4.7 Achievement 1: Check your understanding	542	9.9.2 Checking the independent observations assumption	563
9.5 Achievement 2: Exploring the statistical model for a line	542	9.9.3 Checking the continuous outcome assumption	563
9.5.1 The equation for a line	542	9.9.4 Checking the linearity assumption	563
9.5.2 Using the equation for a line	542	9.9.5 Checking the homoscedasticity assumption	564
9.5.3 The distinction between deterministic and stochastic	545	9.9.6 Testing the independence of residuals assumption	566
9.5.4 Achievement 2: Check your understanding	546	9.9.7 Testing the normality of residuals assumption	566
9.6 Achievement 3: Computing the slope and intercept in a simple linear regression	547	9.9.8 Interpreting the results of the assumption checking	567
9.6.1 Sampling from a data frame	547	9.9.9 Using model diagnostics to find outliers and influential values	569
9.6.2 Exploring the relationship with a scatterplot	547	9.9.10 Summarizing outliers and influential values	577
9.6.3 Computing the slope of the line	547	9.9.11 Achievement 6: Check your understanding	580
9.6.4 Estimating the linear regression model in R	548	9.10 Achievement 7: Adding variables to the model and using transformation	581
9.6.5 Navigating the linear regression output	549	9.10.1 Adding a binary variable to the model	581
9.6.6 Understanding residuals	549	9.10.2 Adding more variables to the model	584
9.6.7 Achievement 3: Check your understanding	551	9.10.3 No multicollinearity assumption for multiple regression	592
9.7 Achievement 4: Slope interpretation and significance (b_1 , p -value, CI)	551	9.10.4 Checking linearity for multiple regression	594
9.7.1 Interpreting the value of the slope	551	9.10.5 Checking the homoscedasticity assumption for multiple regression	596
9.7.2 Interpreting the statistical significance of the slope	552	9.10.6 Testing the independence of residuals assumption	596
9.7.3 Computing confidence intervals for the slope and intercept	554	9.10.7 Testing the normality of residuals assumption	596
9.7.4 Using the model to make predictions	554	9.10.8 Using the Partial-F test to choose a model	598
9.7.5 Achievement 4: Check your understanding	556	9.10.9 Achievement 7: Check your understanding	603
9.8 Achievement 5: Model significance and model fit	556	9.11 Chapter summary	603
9.8.1 Understanding the F-statistic	557	9.11.1 Achievements unlocked in this chapter: Recap	603
9.8.2 Understanding the R^2 measure of model fit	561	9.11.2 Chapter exercises	604
9.8.3 Reporting linear regression results	561		
9.8.4 Achievement 5: Check your understanding	562		

10: Binary Logistic Regression: The R-Team and the Perplexing Libraries Problem 606

10.1 Achievements to unlock	607	10.8.2 NHST Step 2: Compute the test statistic	646
10.2 The perplexing libraries problem	608	10.8.3 NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)	646
10.3 Data, codebook, and R packages for logistic regression practice	608	10.8.4 NHST Steps 4 and 5: Interpret the probability and write a conclusion	646
10.4 Achievement 1: Using exploratory data analysis before developing a logistic regression model	609	10.8.5 Achievement 5: Check your understanding	646
10.4.1 Exploratory data analysis	623	10.9 Achievement 6: Interpreting the results of a larger logistic regression model	647
10.4.2 Achievement 1: Check your understanding	628	10.9.1 Computing odds ratios	647
10.5 Achievement 2: Understanding the binary logistic regression statistical model	628	10.9.2 Odds ratio statistical significance	648
10.5.1 The statistical form of the model	628	10.9.3 Interpreting odds ratios	648
10.5.2 The logistic function	629	10.9.4 Compute and interpret model fit	650
10.5.3 Achievement 2: Check your understanding	632	10.9.5 Achievement 6: Check your understanding	650
10.6 Achievement 3: Estimating a simple logistic regression model and interpreting predictor significance and interpretation	632	10.10 Achievement 7: Checking logistic regression assumptions and using diagnostics to identify outliers and influential values	650
10.6.1 NHST	632	10.10.1 Assumption: Independence of observations	651
10.6.2 Interpreting predictor significance	637	10.10.2 Assumption: Linearity	651
10.6.3 Computing odds ratios	637	10.10.3 Assumption: No perfect multicollinearity	652
10.6.4 Odds ratio significance	638	10.10.4 Model diagnostics	653
10.6.5 Interpreting significant odds ratios	638	10.10.5 Achievement 7: Check your understanding	663
10.6.6 Using NHST to organize the significance testing of odds ratios	638	10.11 Achievement 8: Using the model to predict probabilities for observations that are outside the data set	663
10.6.7 Achievement 3: Check your understanding	640	10.11.1 Achievement 8: Check your understanding	664
10.7 Achievement 4: Computing and interpreting two measures of model fit	641	10.12 Achievement 9: Adding and interpreting interaction terms in logistic regression	664
10.7.1 Percent correctly predicted or count R^2	642	10.12.1 NHST	668
10.7.2 Sensitivity and specificity	643	10.12.2 Compute and interpret odds ratios	669
10.7.3 Achievement 4: Check your understanding	644	10.12.3 Compute and interpret model fit	669
10.8 Achievement 5: Estimating a larger logistic regression model with categorical and continuous predictors	644	10.12.4 Check assumptions	669
10.8.1 NHST Step 1: Write the null and alternate hypotheses	645	10.12.5 Achievement 9: Check your understanding	670

10.13 Achievement 10: Using the likelihood ratio test to compare two nested logistic regression models	671	11.6.2 Independence of irrelevant alternatives assumption	715
10.13.1 Using NHST to organize and conduct an LR test	671	11.6.3 Full interpretation of model results	718
10.13.2 Complete interpretation of final model	672	11.6.4 Achievement 3: Check your understanding	719
10.13.3 Achievement 10: Check your understanding	678	11.7 Achievement 4: Using exploratory data analysis for ordinal logistic regression	719
10.14 Chapter summary	678	11.7.1 Visualizing satisfaction with salary by job type and age	720
10.14.1 Achievements unlocked in this chapter: Recap	678	11.7.2 Bivariate statistical tests to examine job satisfaction by job type and age	721
10.14.2 Chapter exercises	679	11.7.3 Achievement 4: Check your understanding	724
11: Multinomial and Ordinal Logistic Regression: The R-Team and the Diversity Dilemma in STEM	682	11.8 Achievement 5: Estimating and interpreting an ordinal logistic regression model	724
11.1 Achievements to unlock	683	11.8.1 NHST for the ordinal regression model	724
11.2 The diversity dilemma in STEM	684	11.8.2 Ordinal logistic regression model fit	726
11.3 Data, codebook, and R packages for multinomial and ordinal regression practice	689	11.8.3 Ordinal regression predictor significance and interpretation	727
11.4 Achievement 1: Using exploratory data analysis for multinomial logistic regression	690	11.8.4 Achievement 5: Check your understanding	728
11.4.1 Visualizing employment in computer science, math, and engineering by sex and age	696	11.9 Achievement 6: Checking assumptions for ordinal logistic regression	729
11.4.2 Checking bivariate statistical associations between job type, sex, and age	700	11.9.1 Assumption of independent observations	729
11.4.3 Achievement 1: Check your understanding	701	11.9.2 Assumption of proportional odds	729
11.5 Achievement 2: Estimating and interpreting a multinomial logistic regression model	701	11.9.3 Full interpretation of model results	730
11.5.1 Multinomial model significance	702	11.9.4 Achievement 6: Check your understanding	730
11.5.2 Multinomial model fit	709	11.10 Chapter summary	731
11.5.3 Multinomial model predictor interpretation	711	11.10.1 Achievements unlocked in this chapter: Recap	731
11.5.4 Achievement 2: Check your understanding	715	11.10.2 Chapter exercises	732
11.6 Achievement 3: Checking assumptions for multinomial logistic regression	715	GLOSSARY	G-1
11.6.1 Independence of observations	715	REFERENCES	R-1
		INDEX	I-1

1 The goals of this book

The main goal of this book is to prepare students and other readers for the messy and exciting reality of working with data. As data on many aspects of life accumulate at unprecedented rates, there is an increasing need for people who can clean, manage, and analyze information. Secondly, the book aims to encourage women and other underrepresented groups to consider data science careers. Representation of women has decreased in computer science and math fields since 2006, and gaps persist in data science fields by race and ethnicity, limiting the perspectives contributing to the growth of this field. Finally, this book aims to improve the quality of social science through the promotion of reproducible research practices. Science has been increasingly under scrutiny for use of questionable research practices, some of which can be overcome through well-formatted and documented code for data importing, cleaning, and analysis.

To reach all three of these goals, I employed several strategies, such as a narrative writing style, diversity in the main and supporting characters, a focus on social problems, use of publicly available data and open-source R statistical software, and demonstrations of reproducible research practices.

2 The audience for the book

This book was written with first-year statistics courses in the social sciences in mind. Often, these courses start with descriptive statistics and probability theory and continue through general and generalized linear modeling. Others who may find this book useful include people working with data who are interested in learning R for the first time, learning more R, or reinforcing and improving their statistical skills. General readers interested in data science might find this book to be an accessible introduction to one of the primary software packages and many of the foundational concepts used in data science.

3 The features of the book

3.1 A NARRATIVE APPROACH

The book includes an underlying storyline about three characters working together to learn statistics and R: Nancy, Kiara, and Leslie. Nancy and Kiara are data scientists, and Leslie is a student. The first chapter describes their initial meeting at an R-Ladies community event where they discuss the benefits and challenges of using R and decide to work together to learn R and statistics. The remaining chapters each start in a different setting with a conversation between the three introducing the statistical method of the chapter and the social problem they will address while

learning the method. The use of narrative serves at least two purposes. The first purpose is to provide an accessible and relatable way to introduce statistical topics that are often perceived as some combination of difficult and boring. Students start by reading a casual conversation among friends that gently transitions into more technical concerns. The second purpose is to show women in the roles of experts and learners of coding and statistics. Through dialogue, character development, and images, I portray scenarios of women collaborating to learn and apply data management and analysis skills to important social issues.

Each of the three main characters has expertise or interest in a different aspect of learning and applying R code and statistical concepts. Their expertise and interests are highlighted in three different types of boxed features that appear throughout the chapters:

- **Nancy's Fancy Code:** Nancy is interested in writing code and has expertise in more advanced or quirky R coding, which she shares with the other characters in this feature.
- **Kiara's Reproducibility Resource:** Kiara is interested in improving data science by ensuring her work and the work of her colleagues is reproducible through the use of good coding practice; she shares tips for improving reproducibility in this feature.
- **Leslie's Stats Stuff:** Leslie has interest and expertise in statistical theory and adds detail to their discussions through her explanations of statistical concepts in this feature.

3.2 A FOCUS ON SOCIAL PROBLEMS

Each chapter of the book focuses on a different problem from a social science or related field using publicly available data sources. One reason the social sciences are so interesting is because they help us understand and advocate for the world—the one we live in and the one we want to create. I've tried to choose compelling topics, including several with moral dilemmas that are in the news and in students' lives. Most readers should be able to find at least some of the chapters intriguing and relevant. This approach contrasts with textbooks that focus exclusively or predominantly on statistical theory, use simulated data, or choose data sources specifically to avoid controversial topics.

The topics for the chapters are as follows:

- Chapter 1: Marijuana legalization policy
- Chapter 2: Cancer screening for transgender patients
- Chapter 3: Gun deaths, gun use, gun manufacturing, and funding for gun research
- Chapter 4: Opioid deaths, opioid policy, and opioid treatment facilities
- Chapter 5: Perceptions about voter registration and mandatory voting
- Chapter 6: Blood pressure
- Chapter 7: Time spent using technology
- Chapter 8: Global access to clean water, sanitation, and education
- Chapter 9: Distance to needle exchanges
- Chapter 10: The digital divide and library use
- Chapter 11: Representation of women in data science careers

3.3 USE OF PUBLICLY AVAILABLE DATA SETS

Each chapter uses one or more publicly available data sets, and most chapters include instructions for importing the data directly into R from the original online location to encourage the use of reproducible research practices.

Many textbooks in statistics use simulated data or data that have been pre-cleaned. This book takes a different approach in order to provide the audience experience with data situations they are likely to encounter outside of the learning environment. Simulated and pre-cleaned data sources have advantages, including being useful for clearly demonstrating what it means to meet assumptions for statistical models or fail assumptions in specific ways. However, use of pre-cleaned data can set unrealistic expectations of how most data actually look outside the classroom. Likewise, simulated data can reinforce the stereotype that learning and using statistics are not only difficult but also disconnected from the real world. My book tries to overcome these challenges by using real data that address compelling social problems.

Admittedly, there are challenges to my approach. It is decidedly more difficult to demonstrate some concepts and to meet statistical model assumptions with real-world data. That's life. The challenges with this approach mimic the challenges of data science. Moreover, readers will be introduced to strategies for thinking like a data scientist to identify and deal with common obstacles.

The data sets used are as follows:

- 2016 General Social Survey (Chapter 1)
- 2014 Behavioral Risk Factor Surveillance Survey (Chapter 2)
- 2016 Federal Bureau of Investigation homicide data table (Chapter 3)
- 2011–2012 National Health and Nutrition Examination Survey (Chapter 3)
- 2017 Kaiser Family Foundation state opioid policy data (Chapter 4)
- 2017 American Foundation for AIDS Research distance to substance use treatment facility data (Chapter 4)
- 2017 Pew Research Center voting perceptions data (Chapter 5)
- 2015–2016 National Health and Nutrition Examination Survey (Chapter 6)
- 2018 General Social Survey (Chapter 7)
- 2015 World Health Organization data on access to water and sanitation (Chapter 8)
- 2015 United Nations Educational, Scientific, and Cultural Organization data on education (Chapter 8)
- 2017 American Foundation for AIDS Research distance to needle exchange data (Chapter 9)
- 2016 Pew Research Center library use data (Chapter 10)
- 2017 National Science Foundation Scientists and Engineers Statistical Data System (Chapter 11)

3.4 USE OF R STATISTICAL SOFTWARE

R is a coding language used to conduct statistical analyses and to create visual displays of data. The options available in R for analysis and graphing are competitive with, and in some cases surpass, the Statistical Package for the Social Sciences (SPSS) and Statistical Analysis System (SAS) software packages. However, R statistical software is free and open source. Anyone can contribute a “package” to R and, if it is accepted to the Comprehensive R Archive Network (CRAN) (<https://cran.r-project.org/>)

submit.html), it becomes available for all R users worldwide to access and use. As of July 2019, there are approximately 15,000 user-contributed packages available on the CRAN, with most packages being designed to conduct some specific type of analysis. The use of R has been growing and, by some metrics, has caught up to or surpassed SPSS and SAS when it comes to data-related jobs (Muenchen, 2019). Because R is free, it allows people and organizations with fewer resources to access and use a high-quality data science tool (Krishnaswamy & Marinova, 2012; Sullivan, 2011). Consequently, the software is more egalitarian and inclusive, creating opportunities for collaborations and contributions to emerge from less privileged nations and individuals.

Among the unique qualities of R is the community that has formed around its use. On any given day, the #rstats hashtag on Twitter includes several hundred tweets from people making suggestions, asking questions, and posting R memes and jokes. The R community is also highly focused on diversity. For example, groups like R-Ladies Global support and encourage underrepresented voices in R (Daish, Frick, LeDell, de Queiroz, & Vitolo, 2019).

3.5 INCLUSION OF DIVERSE CHARACTERS, AUTHORS, AND ARTWORK

The book emphasizes diversity and inclusion in several ways. First, to address the underrepresentation of women in math and computer science fields, the underlying story features three women. In addition, the final chapter of the book examines data related to job type, job satisfaction, and sex, with a focus on computer science and math. The chapter portrays the student character coming to terms with data on employment and job satisfaction as she ponders a data science career. Third, when relevant and available, I cited women and authors from countries underrepresented in the statistical literature. Specifically, when two equally citable sources supporting the same concept were available, I opted for the underrepresented authors as determined by a commonsense reading of names and affiliations and, in some cases, searching for and reading of online profiles. Finally, the main and supporting characters in the artwork included throughout the text include diverse representation of race, ethnicity, and sex. If women and students of color can literally “see” themselves in the characters, they may find data jobs more appealing and feasible (Herrmann et al., 2016; Johnson, 2011).

3.6 EMPHASIS ON REPRODUCIBLE RESEARCH PRACTICES

Science is increasingly under scrutiny for the use of questionable research practices that threaten the quality of the evidence underlying decisions that impact lives (Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016; Steen, Casadevall, & Fang, 2013). Among the many strategies that could improve the quality of the science is the use of reproducible practices during data management and analysis that allow other scientists to reproduce the work (Harris et al., 2019; Harris, Wondmeneh, Zhao, & Leider, 2019). This text suggests and demonstrates choices that contribute to reproducibility when importing, cleaning, managing, and analyzing data. Reproducibility is mentioned throughout the text by one of the characters, and most chapters have a resource box that offers one or more strategies for ensuring that data management and analysis tasks are performed in a reproducible way. For example, one of the sections in the first chapter describes and demonstrates how to pick a consistent and clear way to name and format constants, variables, and functions. This strategy results in code that is more readable for humans, which improves code comprehension and facilitates the reuse of code to verify results with the same data or new data.

4 Book website

The book website at edge.sagepub.com/harris1e includes the following resources:

INSTRUCTOR TEACHING SITE: edge.sagepub.com/harris1e

SAGE **edge for instructors** supports your teaching by making it easy to integrate quality content and create a rich learning environment for students with

- A **password-protected site** for complete and protected access to all text-specific instructor resources;
- **Test banks** that provide a diverse range of ready-to-use options that save you time—you can also easily edit any question and/or insert your own personalized questions;
- **Tutorial videos** produced exclusively for this text that demonstrate **how to use R** to conduct key statistical tests using real-world data;
- **Editable, chapter-specific PowerPoint® slides** that offer complete flexibility for creating a multimedia presentation for your course;
- **Downloadable Coder (beginner/intermediate) and Hacker (advanced) exercises** from the book that can be used as homework or labs—students can take the **multiple choice pre-test questions** electronically first to check their level;
- **Downloadable data files and R code available** for use with the book and exercises;
- **Solutions** to selected in-text exercises;
- **Instructor Ideas for Gamification** compiled by the author, offered for those who want to gamify their course; and
- **Full-color figures** from the book available for download.

STUDENT STUDY SITE: edge.sagepub.com/harris1e

SAGE **edge for students** enhances learning, it's easy to use, and it offers

- An **open-access site** that makes it easy for students to maximize their study time, anywhere, anytime;
- **Tutorial videos** produced exclusively for this text that demonstrate **how to use R** to conduct key statistical tests using real-world data;
- **Downloadable Coder (beginner/intermediate) and Hacker (advanced) exercises** from the book—students can take the **multiple choice pre-test questions** electronically first to check their level; and
- **Downloadable data files and R code available** for use with the book and exercises.

5 Acknowledgments

Many people helped in major and minor and in obvious and subtle ways throughout the development of this text. Leanne Waugh read well over half of the book, found typos, and identified places where an explanation or the storyline was incomplete or unclear. Amy Sklansky suggested major improvements to the narrative and chapter openings and taught me the difference between writing a story and writing stage directions. Shelly Cooper read several chapters and offered suggestions for fixing and improving the code, including suggesting a useful figure that is a combination of a boxplot, a violin plot, and a scatterplot (check it out in Chapter 9). Scott Harris used his expert copyediting skills to help me with grammar, punctuation, and an unusual amount of repeated words, even for me me. Bobbi Carothers offered

helpful suggestions on very early drafts of the first few chapters, providing a more solid foundation for the rest of the work. Leslie Hinyard answered too many text messages to count. Chelsea West, Chris Prener, Angelique Zeringue, Sarah Van Alsten, Bryan Newman, Alexis Duncan, Kristen Ruckdashel, Joe Steensma, Kim Johnson, Robert Singer, Paaige Turner, Doug Luke, Ellen Mrazek, and some people I do not know on Twitter also gave helpful opinions and suggestions. Thank you all!

At SAGE, Helen Salmon and Chelsea Neve were the epitome of patience, enthusiasm, and professionalism. The reviewers in the following list had great ideas and enthusiastic suggestions that improved the work immensely:

Jennifer Bachner, *Johns Hopkins University*
Matthew C. Bell, *Santa Clara University*
Patrick Bolger, *Texas A&M University*
William J. Bosl, *University of San Francisco*
Joseph Nathan Cohen, *City University of New York–Queens College*
Daniel Conroy-Beam, *University of California, Santa Barbara*
Gabriel I. Cook, *Claremont McKenna College*
James J. Cortright, *University of Wisconsin–River Falls*
Jacqueline S. Craven, *Delta State University*
Todd Daniel, *Missouri State University*
Michael Erickson, *Hawaii Pacific University*
Marte Fallshore, *Central Washington University*
Sylvain Fiset, *Université de Moncton, Edmundston*
Jonathan S. Hack, *Harvard Law School*
Johannes Karreth, *Ursinus College*
George Kikuchi, *California State University, Fresno*
Brandon LeBeau, *University of Iowa*
Michael S. Lynch, *University of Georgia*
Michael E. J. Masson, *University of Victoria*
Avery McIntosh, *Boston University*
Matthew R. Miles, *Brigham Young University, Idaho*
Maura J. Mills, *University of Alabama*
Mary Moore, *Colorado State University–Global*
Derek Mueller, *Carleton University*
David A. M. Peterson, *Iowa State University*
Matthew D. Phillips, *UNC Charlotte*
Darrin L. Rogers, *State University of New York at Fredonia*
Samantha Seals, *University of West Florida*

Yi Shao, *Oklahoma City University*

Ches Thurber, *Northern Illinois University*

Drew Tyre, *University of Nebraska–Lincoln*

Mary Beth Zeni, *Ursuline College*

The artists, Rob Schuster and Rose Storey, did an excellent job producing graphics that fit with the story and goals of the work. My thanks to Hadley Wickham for permission to include one of his tweets as a figure, and to Hao Zhu and Matt Dowle for permission to include their `kableExtra` and `data.table` hex stickers, respectively, in the cover art. Thank you to the data teams at the General Social Survey; the National Health and Nutrition Examination Survey; the World Health Organization; the United Nations Educational, Scientific and Cultural Organization; the Federal Bureau of Investigation; the Pew Research Center; the Kaiser Family Foundation; the Foundation for AIDS Research; and the National Science Foundation. Special thanks go to Dr. David Stark and Dr. Nigam Shah, who answered my emails, sent me a spreadsheet of the data from their 2017 article (Stark & Shah, 2017), and sent the GitHub location of the data and R code for the paper that was the inspiration for Chapter 3 (<https://github.com/davidestark/gun-violence-research/>).

Finally, all the hard work in the world would not have resulted in much had I not had the great fortune of having a math-teaching mom with an unfathomably big heart; too many smart, fun, and supportive friends to count (you know who you are, or, if you don't, you'll be receiving a text to confirm!); and a spouse who is an inspiration and true partner.