

Foundations of Statistics for Data Scientists

With R and Python

Alan Agresti and Maria Kateri



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

Contents

Preface	xv
1 Introduction to Statistical Science	1
1.1 Statistical Science: Description and Inference	1
1.1.1 Design, Descriptive Statistics, and Inferential Statistics	2
1.1.2 Populations and Samples	3
1.1.3 Parameters: Numerical Summaries of the Population	3
1.1.4 Defining Populations: Actual and Conceptual	4
1.2 Types of Data and Variables	4
1.2.1 Data Files	4
1.2.2 Example: The General Social Survey (GSS)	5
1.2.3 Variables	6
1.2.4 Quantitative Variables and Categorical Variables	6
1.2.5 Discrete Variables and Continuous Variables	7
1.2.6 Associations: Response Variables and Explanatory Variables	7
1.3 Data Collection and Randomization	8
1.3.1 Randomization	8
1.3.2 Collecting Data with a Sample Survey	9
1.3.3 Collecting Data with an Experiment	9
1.3.4 Collecting Data with an Observational Study	10
1.3.5 Establishing Cause and Effect: Observational versus Experimental Studies	10
1.4 Descriptive Statistics: Summarizing Data	11
1.4.1 Example: Carbon Dioxide Emissions in European Nations	11
1.4.2 Frequency Distribution and Histogram Graphic	11
1.4.3 Describing the Center of the Data: Mean and Median	13
1.4.4 Describing Data Variability: Standard Deviation and Variance	14
1.4.5 Describing Position: Percentiles, Quartiles, and Box Plots	15
1.5 Descriptive Statistics: Summarizing Multivariate Data	17
1.5.1 Bivariate Quantitative Data: The Scatterplot, Correlation, and Regression	17
1.5.2 Bivariate Categorical Data: Contingency Tables	18
1.5.3 Descriptive Statistics for Samples and for Populations	19
1.6 Chapter Summary	20
2 Probability Distributions	29
2.1 Introduction to Probability	29
2.1.1 Probabilities and Long-Run Relative Frequencies	29
2.1.2 Sample Spaces and Events	31
2.1.3 Probability Axioms and Implied Probability Rules	32
2.1.4 Example: Diagnostics for Disease Screening	33
2.1.5 Bayes' Theorem	34

2.1.6	Multiplicative Law of Probability and Independent Events	35
2.2	Random Variables and Probability Distributions	36
2.2.1	Probability Distributions for Discrete Random Variables	36
2.2.2	Example: Geometric Probability Distribution	37
2.2.3	Probability Distributions for Continuous Random Variables	38
2.2.4	Example: Uniform Distribution	38
2.2.5	Probability Functions (<i>pdf</i> , <i>pmf</i>) and Cumulative Distribution Function (<i>cdf</i>)	39
2.2.6	Example: Exponential Random Variable	40
2.2.7	Families of Probability Distributions Indexed by Parameters	41
2.3	Expectations of Random Variables	42
2.3.1	Expected Value and Variability of a Discrete Random Variable	42
2.3.2	Expected Values for Continuous Random Variables	43
2.3.3	Example: Mean and Variability for Uniform Random Variable	44
2.3.4	Higher Moments: Skewness	44
2.3.5	Expectations of Linear Functions of Random Variables	45
2.3.6	Standardizing a Random Variable	46
2.4	Discrete Probability Distributions	46
2.4.1	Binomial Distribution	46
2.4.2	Example: Hispanic Composition of Jury List	47
2.4.3	Mean, Variability, and Skewness of Binomial Distribution	48
2.4.4	Example: Predicting Results of a Sample Survey	49
2.4.5	The Sample Proportion as a Scaled Binomial Random Variable	50
2.4.6	Poisson Distribution	50
2.4.7	Poisson Variability and Overdispersion	51
2.5	Continuous Probability Distributions	52
2.5.1	The Normal Distribution	53
2.5.2	The Standard Normal Distribution	53
2.5.3	Examples: Finding Normal Probabilities and Percentiles	54
2.5.4	The Gamma Distribution	55
2.5.5	The Exponential Distribution and Poisson Processes	57
2.5.6	Quantiles of a Probability Distribution	57
2.5.7	Using the Uniform to Randomly Generate a Continuous Random Variable	58
2.6	Joint and Conditional Distributions and Independence	59
2.6.1	Joint and Marginal Probability Distributions	59
2.6.2	Example: Joint and Marginal Distributions of Happiness and Family Income	60
2.6.3	Conditional Probability Distributions	60
2.6.4	Trials with Multiple Categories: The Multinomial Distribution	61
2.6.5	Expectations of Sums of Random Variables	62
2.6.6	Independence of Random Variables	63
2.6.7	Markov Chain Dependence and Conditional Independence	64
2.7	Correlation between Random Variables	64
2.7.1	Covariance and Correlation	64
2.7.2	Example: Correlation between Income and Happiness	65
2.7.3	Independence Implies Zero Correlation, but Not Converse	66
2.7.4	Bivariate Normal Distribution *	66
2.8	Chapter Summary	69

3	Sampling Distributions	81
3.1	Sampling Distributions: Probability Distributions for Statistics	81
3.1.1	Example: Predicting an Election Result from an Exit Poll	81
3.1.2	Sampling Distribution: Variability of a Statistic’s Value among Sam- ples	83
3.1.3	Constructing a Sampling Distribution	84
3.1.4	Example: Simulating to Estimate Mean Restaurant Sales	85
3.2	Sampling Distributions of Sample Means	86
3.2.1	Mean and Variance of Sample Mean of Random Variables	86
3.2.2	Standard Error of a Statistic	87
3.2.3	Example: Standard Error of Sample Mean Sales	88
3.2.4	Example: Standard Error of Sample Proportion in Exit Poll	88
3.2.5	Law of Large Numbers: Sample Mean Converges to Population Mean	89
3.2.6	Normal, Binomial, and Poisson Sums of Random Variables Have the Same Distribution	89
3.3	Central Limit Theorem: Normal Sampling Distribution for Large Samples	90
3.3.1	Sampling Distribution of Sample Mean Is Approximately Normal	90
3.3.2	Simulations Illustrate Normal Sampling Distribution in CLT	92
3.3.3	Summary: Population, Sample Data, and Sampling Distributions	93
3.4	Large-Sample Normal Sampling Distributions for Many Statistics*	94
3.4.1	The Delta Method	95
3.4.2	Delta Method Applied to Root Poisson Stabilizes the Variance	96
3.4.3	Simulating Sampling Distributions of Other Statistics	96
3.4.4	The Key Role of Sampling Distributions in Statistical Inference	98
3.5	Chapter Summary	98
4	Statistical Inference: Estimation	105
4.1	Point Estimates and Confidence Intervals	105
4.1.1	Properties of Estimators: Unbiasedness, Consistency, Efficiency	106
4.1.2	Evaluating Properties of Estimators	107
4.1.3	Interval Estimation: Confidence Intervals for Parameters	107
4.2	The Likelihood Function and Maximum Likelihood Estimation	108
4.2.1	The Likelihood Function	108
4.2.2	Maximum Likelihood Method of Estimation	109
4.2.3	Properties of Maximum Likelihood (ML) Estimators	110
4.2.4	Example: Variance of ML Estimator of Binomial Parameter	111
4.2.5	Example: Variance of ML Estimator of Poisson Mean	111
4.2.6	Sufficiency and Invariance for ML Estimates	112
4.3	Constructing Confidence Intervals	113
4.3.1	Using a Pivotal Quantity to Induce a Confidence Interval	113
4.3.2	A Large-Sample Confidence Interval for the Mean	115
4.3.3	Confidence Intervals for Proportions	115
4.3.4	Example: Atheists and Agnostics in Europe	116
4.3.5	Using Simulation to Illustrate Long-Run Performance of CIs	117
4.3.6	Determining the Sample Size before Collecting the Data	117
4.3.7	Example: Sample Size for Evaluating an Advertising Strategy	118
4.4	Confidence Intervals for Means of Normal Populations	120
4.4.1	The t Distribution	120
4.4.2	Confidence Interval for a Mean Using the t Distribution	121
4.4.3	Example: Estimating Mean Weight Change for Anorexic Girls	122
4.4.4	Robustness for Violations of Normal Population Assumption	123

4.4.5	Construction of t Distribution Using Chi-Squared and Standard Normal	124
4.4.6	Why Does the Pivotal Quantity Have the t Distribution?	125
4.4.7	Cauchy Distribution: t Distribution with $df = 1$ Has Unusual Behavior	126
4.5	Comparing Two Population Means or Proportions	126
4.5.1	A Model for Comparing Means: Normality with Common Variability	127
4.5.2	A Standard Error and Confidence Interval for Comparing Means	127
4.5.3	Example: Comparing a Therapy to a Control Group	128
4.5.4	Confidence Interval Comparing Two Proportions	130
4.5.5	Example: Does Prayer Help Coronary Surgery Patients?	130
4.6	The Bootstrap	132
4.6.1	Computational Resampling and Bootstrap Confidence Intervals	132
4.6.2	Example: Bootstrap Confidence Intervals for Library Data	132
4.7	The Bayesian Approach to Statistical Inference	134
4.7.1	Bayesian Prior and Posterior Distributions	135
4.7.2	Bayesian Binomial Inference: Beta Prior Distributions	136
4.7.3	Example: Belief in Hell	137
4.7.4	Interpretation: Bayesian versus Classical Intervals	138
4.7.5	Bayesian Posterior Interval Comparing Proportions	138
4.7.6	Highest Posterior Density (HPD) Posterior Intervals	138
4.8	Bayesian Inference for Means	139
4.8.1	Bayesian Inference for a Normal Mean	139
4.8.2	Example: Bayesian Analysis for Anorexia Therapy	140
4.8.3	Bayesian Inference for Normal Means with Improper Priors	141
4.8.4	Predicting a Future Observation: Bayesian Predictive Distribution	142
4.8.5	The Bayesian Perspective, and Empirical Bayes and Hierarchical Bayes Extensions	142
4.9	Why Maximum Likelihood and Bayes Estimators Perform Well *	143
4.9.1	ML Estimators Have Large-Sample Normal Distributions	143
4.9.2	Asymptotic Efficiency of ML Estimators Same as Best Unbiased Estimators	145
4.9.3	Bayesian Estimators Also Have Good Large-Sample Performance	146
4.9.4	The Likelihood Principle	146
4.10	Chapter Summary	147
5	Statistical Inference: Significance Testing	161
5.1	The Elements of a Significance Test	161
5.1.1	Example: Testing for Bias in Selecting Managers	161
5.1.2	Assumptions, Hypotheses, Test Statistic, P -Value, and Conclusion	162
5.2	Significance Tests for Proportions and Means	164
5.2.1	The Elements of a Significance Test for a Proportion	164
5.2.2	Example: Climate Change a Major Threat?	166
5.2.3	One-Sided Significance Tests	166
5.2.4	The Elements of a Significance Test for a Mean	167
5.2.5	Example: Significance Test about Political Ideology	169
5.3	Significance Tests Comparing Means	170
5.3.1	Significance Tests for the Difference between Two Means	170
5.3.2	Example: Comparing a Therapy to a Control Group	171
5.3.3	Effect Size for Comparison of Two Means	172
5.3.4	Bayesian Inference for Comparing Two Means	173
5.3.5	Example: Bayesian Comparison of Therapy and Control Groups	173

5.4	Significance Tests Comparing Proportions	174
5.4.1	Significance Test for the Difference between Two Proportions	174
5.4.2	Example: Comparing Prayer and Non-Prayer Surgery Patients	175
5.4.3	Bayesian Inference for Comparing Two Proportions	176
5.4.4	Chi-Squared Tests for Multiple Proportions in Contingency Tables	177
5.4.5	Example: Happiness and Marital Status	178
5.4.6	Standardized Residuals: Describing the Nature of an Association	179
5.5	Significance Test Decisions and Errors	180
5.5.1	The α -level: Making a Decision Based on the P -Value	181
5.5.2	Never “Accept H_0 ” in a Significance Test	181
5.5.3	Type I and Type II Errors	182
5.5.4	As $P(\text{Type I Error})$ Decreases, $P(\text{Type II Error})$ Increases	182
5.5.5	Example: Testing Whether Astrology Has Some Truth	184
5.5.6	The Power of a Test	185
5.5.7	Making Decisions versus Reporting the P -Value	186
5.6	Duality between Significance Tests and Confidence Intervals	186
5.6.1	Connection between Two-Sided Tests and Confidence Intervals	186
5.6.2	Effect of Sample Size: Statistical versus Practical Significance	187
5.6.3	Significance Tests Are Less Useful than Confidence Intervals	188
5.6.4	Significance Tests and P -Values Can Be Misleading	189
5.7	Likelihood-Ratio Tests and Confidence Intervals *	190
5.7.1	The Likelihood-Ratio and a Chi-Squared Test Statistic	191
5.7.2	Likelihood-Ratio Test and Confidence Interval for a Proportion	191
5.7.3	Likelihood-Ratio, Wald, Score Test Triad	192
5.8	Nonparametric Tests *	194
5.8.1	A Permutation Test to Compare Two Groups	194
5.8.2	Example: Petting versus Praise of Dogs	194
5.8.3	Wilcoxon Test: Comparing Mean Ranks for Two Groups	196
5.8.4	Comparing Survival Time Distributions with Censored Data	197
5.9	Chapter Summary	200
6	Linear Models and Least Squares	211
6.1	The Linear Regression Model and Its Least Squares Fit	211
6.1.1	The Linear Model Describes a Conditional Expectation	211
6.1.2	Describing Variation around the Conditional Expectation	212
6.1.3	Least Squares Model Fitting	213
6.1.4	Example: Linear Model for Scottish Hill Races	214
6.1.5	The Correlation	216
6.1.6	Regression toward the Mean in Linear Regression Models	217
6.1.7	Linear Models and Reality	218
6.2	Multiple Regression: Linear Models with Multiple Explanatory Variables	219
6.2.1	Interpreting Effects in Multiple Regression Models	219
6.2.2	Example: Multiple Regression for Scottish Hill Races	220
6.2.3	Association and Causation	220
6.2.4	Confounding, Spuriousness, and Conditional Independence	221
6.2.5	Example: Modeling the Crime Rate in Florida	222
6.2.6	Equations for Least Squares Estimates in Multiple Regression	223
6.2.7	Interaction between Explanatory Variables in Their Effects	224
6.2.8	Cook’s Distance: Detecting Unusual and Influential Observations	226
6.3	Summarizing Variability in Linear Regression Models	227
6.3.1	The Error Variance and Chi-Squared for Linear Models	228

6.3.2	Decomposing Variability into Model Explained and Unexplained Parts	228
6.3.3	R -Squared and the Multiple Correlation	229
6.3.4	Example: R -Squared for Modeling Scottish Hill Races	230
6.4	Statistical Inference for Normal Linear Models	231
6.4.1	The F Distribution: Testing That All Effects Equal 0	231
6.4.2	Example: Normal Linear Model for Mental Impairment	232
6.4.3	t Tests and Confidence Intervals for Individual Effects	233
6.4.4	Multicollinearity: Nearly Redundant Explanatory Variables	234
6.4.5	Confidence Interval for $E(Y)$ and Prediction Interval for Y	235
6.4.6	The F Test That All Effects Equal 0 is a Likelihood-Ratio Test *	236
6.5	Categorical Explanatory Variables in Linear Models	238
6.5.1	Indicator Variables for Categories	238
6.5.2	Example: Comparing Mean Incomes of Racial-Ethnic Groups	239
6.5.3	Analysis of Variance (ANOVA): An F Test Comparing Several Means . . .	240
6.5.4	Multiple Comparisons of Means: Bonferroni and Tukey Methods	241
6.5.5	Models with Both Categorical and Quantitative Explanatory Variables	243
6.5.6	Comparing Two Nested Normal Linear Models	244
6.5.7	Interaction with Categorical and Quantitative Explanatory Variables . . .	245
6.6	Bayesian Inference for Normal Linear Models	246
6.6.1	Prior and Posterior Distributions for Normal Linear Models	246
6.6.2	Example: Bayesian Linear Model for Mental Impairment	246
6.6.3	Bayesian Approach to the Normal One-Way Layout	247
6.7	Matrix Formulation of Linear Models *	248
6.7.1	The Model Matrix	248
6.7.2	Least Squares Estimates and Standard Errors	249
6.7.3	The Hat Matrix and the Leverage	250
6.7.4	Alternatives to Least Squares: Robust Regression and Regularization . . .	250
6.7.5	Restricted Optimality of Least Squares: Gauss–Markov Theorem	250
6.7.6	Matrix Formulation of Bayesian Normal Linear Model	251
6.8	Chapter Summary	252
7	Generalized Linear Models	263
7.1	Introduction to Generalized Linear Models	263
7.1.1	The Three Components of a Generalized Linear Model	263
7.1.2	GLMs for Normal, Binomial, and Poisson Responses	264
7.1.3	Example: GLMs for House Selling Prices	265
7.1.4	The Deviance	267
7.1.5	Likelihood-Ratio Model Comparison Uses Deviance Difference	268
7.1.6	Model Selection: AIC and the Bias/Variance Tradeoff	268
7.1.7	Advantages of GLMs versus Transforming the Data	271
7.1.8	Example: Normal and Gamma GLMs for Covid-19 Data	271
7.2	Logistic Regression Model for Binary Data	272
7.2.1	Logistic Regression: Model Expressions	273
7.2.2	Parameter Interpretation: Effects on Probabilities and Odds	273
7.2.3	Example: Dose–Response Study for Flour Beetles	274
7.2.4	Grouped and Ungrouped Binary Data: Effects on Estimates and Deviance	276
7.2.5	Example: Modeling Italian Employment with Logit and Identity Links . . .	278
7.2.6	Complete Separation and Infinite Logistic Parameter Estimates	279

7.3	Bayesian Inference for Generalized Linear Models	281
7.3.1	Normal Prior Distributions for GLM Parameters	281
7.3.2	Example: Logistic Regression for Endometrial Cancer Patients	282
7.4	Poisson Loglinear Models for Count Data	284
7.4.1	Poisson Loglinear Models	284
7.4.2	Example: Modeling Horseshoe Crab Satellite Counts	285
7.4.3	Modeling Rates: Including an Offset in the Model	286
7.4.4	Example: Lung Cancer Survival	287
7.5	Negative Binomial Models for Overdispersed Count Data *	288
7.5.1	Increased Variance Due to Heterogeneity	289
7.5.2	Negative Binomial: Gamma Mixture of Poisson Distributions	289
7.5.3	Example: Negative Binomial Modeling of Horseshoe Crab Data	290
7.6	Iterative GLM Model Fitting *	291
7.6.1	The Newton–Raphson Method	291
7.6.2	Newton–Raphson Fitting of Logistic Regression Model	292
7.6.3	Covariance Matrix of Parameter Estimators and Fisher Scoring	294
7.6.4	Likelihood Equations and Covariance Matrix for Poisson GLMs	294
7.7	Regularization with Large Numbers of Parameters *	295
7.7.1	Penalized Likelihood Methods	296
7.7.2	Penalized Likelihood Methods: The Lasso	297
7.7.3	Example: Predicting Opinions with Student Survey Data	297
7.7.4	Why Shrink ML Estimates toward 0?	299
7.7.5	Dimension Reduction: Principal Component Analysis	300
7.7.6	Bayesian Inference with a Large Number of Parameters	300
7.7.7	Huge n : Handling Big Data	300
7.8	Chapter Summary	301
8	Classification and Clustering	313
8.1	Classification: Linear Discriminant Analysis and Graphical Trees	314
8.1.1	Classification with Fisher’s Linear Discriminant Function	314
8.1.2	Example: Predicting Whether Horseshoe Crabs Have Satellites	314
8.1.3	Summarizing Predictive Power: Classification Tables and ROC Curves	316
8.1.4	Classification Trees: Graphical Prediction	318
8.1.5	Logistic Regression versus Linear Discriminant Analysis and Classification Trees	320
8.1.6	Other Methods for Classification: k -Nearest Neighbors and Neural Networks *	321
8.2	Cluster Analysis	324
8.2.1	Measuring Dissimilarity between Observations on Binary Responses	325
8.2.2	Hierarchical Clustering Algorithm and Its Dendrogram	325
8.2.3	Example: Clustering States on Presidential Election Outcomes	326
8.3	Chapter Summary	328
9	Statistical Science: A Historical Overview	333
9.1	The Evolution of Statistical Science *	333
9.1.1	Evolution of Probability	333
9.1.2	Evolution of Descriptive and Inferential Statistics	334
9.2	Pillars of Statistical Wisdom and Practice	336
9.2.1	Stigler’s Seven Pillars of Statistical Wisdom	336
9.2.2	Seven Pillars of Wisdom for Practicing Data Science	338

Appendix A Using R in Statistical Science	341
A.0 Basics of R	341
A.0.1 Starting a Session, Entering Commands, and Quitting	341
A.0.2 Installing and Loading R Packages	341
A.0.3 R Functions and Data Structures	342
A.0.4 Data Input in R	344
A.0.5 R Control Flows	345
A.1 Chapter 1: R for Descriptive Statistics	345
A.1.1 Data Handling and Wrangling	345
A.1.2 Histograms and Other Graphics	346
A.1.3 Descriptive Statistics	347
A.1.4 Missing Values in Data Files	351
A.1.5 Summarizing Bivariate Quantitative Data	352
A.1.6 Summarizing Bivariate Categorical Data	353
A.2 Chapter 2: R for Probability Distributions	353
A.2.1 R Functions for Probability Distributions	353
A.2.2 Quantiles, <i>Q-Q</i> Plots, and the Normal Quantile Plot	355
A.2.3 Joint and Conditional Probability Distributions	358
A.3 Chapter 3: R for Sampling Distributions	358
A.3.1 Simulating the Sampling Distribution of a Statistic	358
A.3.2 Monte Carlo Simulation	359
A.4 Chapter 4: R for Estimation	361
A.4.1 Confidence Intervals for Proportions	361
A.4.2 Confidence Intervals for Means of Subgroups and Paired Differences	362
A.4.3 The <i>t</i> and Other Probability Distributions for Statistical Inference	362
A.4.4 Empirical Cumulative Distribution Function	363
A.4.5 Nonparametric and Parametric Bootstraps	364
A.4.6 Bayesian HPD Intervals Comparing Proportions	366
A.5 Chapter 5: R for Significance Testing	367
A.5.1 Bayes Factors and a Bayesian <i>t</i> Test	367
A.5.2 Simulating the Exact Distribution of the Likelihood-Ratio Statistic	368
A.5.3 Nonparametric Statistics: Permutation Test and Wilcoxon Test	369
A.6 Chapter 6: R for Linear Models	370
A.6.1 Linear Models with the <code>lm</code> Function	370
A.6.2 Diagnostic Plots for Linear Models	370
A.6.3 Plots for Regression Bands and Posterior Distributions	371
A.7 Chapter 7: R for Generalized Linear Models	373
A.7.1 The <code>glm</code> Function	373
A.7.2 Plotting a Logistic Regression Model Fit	373
A.7.3 Model Selection for GLMs	373
A.7.4 Correlated Responses: Marginal, Random Effects, and Transitional Models	376
A.7.5 Modeling Time Series	377
A.8 Chapter 8: R for Classification and Clustering	379
A.8.1 Visualization of Linear Discriminant Analysis Results	379
A.8.2 Cross-Validation and Model Training	379
A.8.3 Classification and Regression Trees	381
A.8.4 Cluster Analysis with Quantitative Variables	381

Appendix B Using Python in Statistical Science 383

 B.0 Basics of Python 383

 B.0.1 Python Preliminaries 383

 B.0.2 Data Structures and Data Input 384

 B.1 Chapter 1: PYTHON for Descriptive Statistics 385

 B.1.1 Random Number Generation 385

 B.1.2 Summary Statistics and Graphs for Quantitative Variables 385

 B.1.3 Descriptive Statistics for Bivariate Quantitative Data 386

 B.1.4 Descriptive Statistics for Bivariate Categorical Data 388

 B.1.5 Simulating Samples from a Bell-Shaped Population 388

 B.2 Chapter 2: PYTHON for Probability Distributions 389

 B.2.1 Simulating a Probability as a Long-Run Relative Frequency 389

 B.2.2 Python Functions for Discrete Probability Distributions 390

 B.2.3 Python Functions for Continuous Probability Distributions 391

 B.2.4 Expectations of Random Variables 393

 B.3 Chapter 3: PYTHON for Sampling Distributions 395

 B.3.1 Simulation to Illustrate a Sampling Distribution 395

 B.3.2 Law of Large Numbers 395

 B.4 Chapter 4: PYTHON for Estimation 396

 B.4.1 Confidence Intervals for Proportions 396

 B.4.2 The *t* Distribution 396

 B.4.3 Confidence Intervals for Means 396

 B.4.4 Confidence Intervals Comparing Means and Comparing Proportions . 397

 B.4.5 Bootstrap Confidence Intervals 398

 B.4.6 Bayesian Posterior Intervals for Proportions and Means 399

 B.5 Chapter 5: PYTHON for Significance Testing 400

 B.5.1 Significance Tests for Proportions 400

 B.5.2 Chi-Squared Tests Comparing Multiple Proportions in Contingency
 Tables 400

 B.5.3 Significance Tests for Means 401

 B.5.4 Significance Tests Comparing Means 401

 B.5.5 The Power of a Significance Test 403

 B.5.6 Nonparametric Statistics: Permutation Test and Wilcoxon Test . . . 403

 B.5.7 Kaplan-Meier Estimation of Survival Functions 404

 B.6 Chapter 6: PYTHON for Linear Models 404

 B.6.1 Fitting Linear Models 404

 B.6.2 The Correlation and R-Squared 406

 B.6.3 Diagnostics: Residuals and Cook’s Distances for Linear Models 407

 B.6.4 Statistical Inference and Prediction for Linear Models 410

 B.6.5 Categorical Explanatory Variables in Linear Models 411

 B.6.6 Bayesian Fitting of Linear Models 412

 B.7 Chapter 7: PYTHON for Generalized Linear Models 413

 B.7.1 GLMs with Identity Link 413

 B.7.2 Logistic Regression: Logit Link with Binary Data 415

 B.7.3 Separation and Bayesian Fitting in Logistic Regression 416

 B.7.4 Poisson Loglinear Model for Counts 417

 B.7.5 Negative Binomial Modeling of Count Data 420

 B.7.6 Regularization: Penalized Logistic Regression Using the Lasso 421

 B.8 Chapter 8: PYTHON for Classification and Clustering 421

 B.8.1 Linear Discriminant Analysis 421

 B.8.2 Classification Trees and Neural Networks for Prediction 423

B.8.3 Cluster Analysis	425
Appendix C Brief Solutions to Exercises	427
C.1 Chapter 1: Solutions to Exercises	427
C.2 Chapter 2: Solutions to Exercises	429
C.3 Chapter 3: Solutions to Exercises	431
C.4 Chapter 4: Solutions to Exercises	433
C.5 Chapter 5: Solutions to Exercises	436
C.6 Chapter 6: Solutions to Exercises	439
C.7 Chapter 7: Solutions to Exercises	443
C.8 Chapter 8: Solutions to Exercises	446
Bibliography	447
Example Index	449
Subject Index	453