Contents

Preface xiii

1 Introduction 1
1.1 Data Science and Statistics 2
1.2 More on Statistics 3
1.2.1 Populations and Samples 4
1.2.2 Descriptive versus Inferential Statistics 5
1.3 An Introduction to R 6
1.4 Descriptive Statistics 7
1.4.1 Types of Data 8
1.4.2 Example Data: US Domestic Flights from 1987 to 2008 9
1.5 Cross-Sectional Data 9
1.5.1 Measures of Location 9
1.5.2 Measures of Variation 16
1.5.3 Measures of How Two Variables Co-vary 18
1.5.4 Other Summary Statistics 23
1.6 Tabular Summaries of Data 25
1.7 Chapter Summary 27
1.8 Problems 29

2 Data Visualization 31
2.1 Introduction 31
2.2 Traditional Statistical Graphics 32
2.2.1 Bar Charts 32
2.2.2 Pie Charts 35
2.2.3 Histograms 35
2.2.4 Lattice (or Trellis) Plots 38
2.2.5 Box Plots 40
2.2.6 Scatterplots 43
2.3 Graphics for Longitudinal Data 46
2.3.1 Time Series Plots 47
2.3.2 Repeated Cross-Sectional Plots 48
2.3.3 Autocorrelation Plots 50
2.4 Chapter Summary 50
2.5 Problems 52

3 Basic Probability 54
3.1 Introduction 54
3.2 Events and Sample Spaces 54
3.2.1 Probability Axioms 56
3.2.2 Union of Events 57
3.2.3 Intersection of Independent Events 59
3.2.4 Complementary Events 60
3.2.5 Conditional Probability 62
3.3 Calculating Probabilities 63
3.3.1 Sample Point Method 64
3.3.2 Counting Sample Points 66
3.3.3 Combining Events 70
3.4 Bringing It All Together 71
3.4.1 Law of Total Probability 71
3.4.2 Bayes' Theorem 74
3.5 Chapter Summary 77
3.6 Problems 78

4 Random Variables 82
4.1 Introduction 82
4.2 Discrete Random Variables 82
4.2.1 Probability Mass Function 82
4.2.2 Cumulative Distribution Function 86
4.2.3 Expected Value 88
4.2.4 Variance and Standard Deviation 91
4.3 Continuous Random Variables 93
4.3.1 Probability Density Function 93
4.3.2 Cumulative Distribution Function 97
4.3.3 Expected Value 101
4.3.4 Variance and Standard Deviation 101
4.4 Expected Value and Variance Properties 102
4.5 Joint Distributions for Discrete Random Variables 105
4.6 Conditional Distributions for Discrete Random Variables 111
4.7 Joint Distributions for Continuous Random Variables 113
4.8 Conditional Distributions for Continuous Random Variables 121
4.9 Conditioning on a Random Variable 123
4.10 Chapter Summary 126
4.11 Problems 127

5 Discrete Distributions 132
5.1 Introduction 132
5.2 Binomial Distribution 132
5.3 Geometric Distribution 141
5.4 Negative Binomial 146
5.5 Hypergeometric Distribution 149
5.6 Poisson Distribution 154
5.7 Multinomial Distribution 160
5.8 Chapter Summary 164
5.9 Problems 166

6 Continuous Distributions 170
6.1 Introduction 170
6.2 Uniform Distribution 170
6.3 Exponential Distribution 173
6.4 Normal Distribution 180
6.4.1 Standardizing 183
6.4.2 Bivariate and Multivariate Normal Distributions 186
6.5 Gamma and Weibull Distributions 189
6.5.1 Gamma Distribution 189
6.5.2 Weibull Distribution 192
6.6 Distributions Related to the Normal 196
6.6.1 Chi-square (χ^2) Distribution 196
6.6.2 t Distribution 198
6.6.3 F Distribution 201
6.7 Beta Distribution 202
6.8 Transformations 205
6.8.1 Simulating from Distributions 207
6.9 Moment Generating Functions 209
6.10 Quantile–Quantile Plots 215
6.11 Chapter Summary 220
6.12 Problems 221

7 About Data and Data Collection 226
7.1 Introduction 226
7.2 Data and the Scientific Method 228
7.3 Experimental vs. Observational Data 231
7.3.1 Convenience vs. Probability Sampling 234
7.4 Accuracy vs. Precision 234
7.5 Types of Random Samples 236
7.5.1 Sources of Bias 237
7.6 Types of Error 239
7.7 Historical Gaffes in Data Collection 240
7.8 Chapter Summary 241
7.9 Problems 242

8 Sampling Distributions 244
8.1 Introduction 244
8.2 Linear Combinations of Random Variables 244
8.3 Sampling Distributions for Sums and Means 249
8.4 Sampling Distribution for the Sample Variance 253
8.5 The Central Limit Theorem 255
8.6 Normal Approximation to the Binomial 259
8.7 Sampling Distributions for Proportions 263
8.8 Tchebysheff's Theorem and the Law of Large Numbers 265

CAMBRIDGE
Cambridge University Press & Assessment
978-1-107-11304-6 — Introduction to Probability and Statistics for Data Science
Steven E. Rigdon, Ronald D. Fricker, Jr, Douglas C. Montgomery
Table of Contents
More Information

8.9 Chapter Summary 268
8.10 Problems 269

9 Point Estimation 273
9.1 Introduction and Intuitive Estimators 273
9.2 Estimation Criteria 275
9.2.1 Unbiased Estimators 275
9.2.2 Consistent Estimators 277
9.3 Method of Moments 279
9.4 Maximum Likelihood 283
9.5 Approximating MLEs 289
9.6 Sufficiency 294
9.7 Chapter Summary 298
9.8 Problems 299

10 Confidence Intervals 302
10.1 Introduction 302
10.2 Basic Properties 303
10.3 Large Sample Confidence Intervals 307
10.4 Small Sample Confidence Intervals 312
10.5 Confidence Intervals for Differences 316
10.5.1 Confidence Intervals for Differences of Proportions 317
10.5.2 Confidence Intervals for Differences in Means 319
10.5.3 Confidence Interval for Paired Data 324
10.6 Determining the Sample Size 326
10.7 Confidence Intervals from Complex Survey Data 330
10.7.1 Sampling from a Finite Population 330
10.7.2 Stratified Random Samples 333
10.7.3 Cluster Sampling 338
10.7.4 Secondary Data Sources 339
10.7.5 Software for Analyzing Data from Complex Surveys 341
10.8 Chapter Summary 343
10.9 Problems 344

11 Hypothesis Testing 348
11.1 Introduction 348
11.2 Elements of a Statistical Test 350
11.3 Power 354
11.4 *P*-values 356
11.5 Testing the Mean: Variance Known 357
11.5.1 Hypothesis Tests for the Mean from a Population with Known Variance 357
11.5.2 Power 360
11.6 Testing the Mean: Variance Unknown 364
11.7 Testing a Proportion 370
11.8 Testing the Variance 373
11.9 Likelihood Ratio Tests 374
11.10 Chapter Summary 380
11.11 Problems 381

12 Hypothesis Tests for Two or More Populations 386
12.1 Introduction 386
12.2 Testing Two Independent Samples 386
12.2.1 Comparing Two Means 386
12.2.2 Comparing Two Proportions 398
12.2.3 Comparing Variances 402
12.3 Testing Paired Samples 404
12.4 Single-Factor Analysis of Variance 409
12.5 Two-Factor ANOVA 425
12.6 Other Designs for Experiments 430
12.6.1 Two-Level Factorial Designs 431
12.6.2 Fractional Factorial Designs 436
12.6.3 Block Designs 439
12.6.4 Some Experimental Design Principles 442
12.7 Power 444
12.7.1 Power for Two-Sample *t*-Test 445
12.7.2 Power for One-Factor ANOVA 448
12.8 Chapter Summary 451
12.9 Problems 453

13 Hypothesis Tests for Categorical Data 459
13.1 Introduction 459
13.2 Goodness-of-Fit Tests 460
13.3 Contingency Tables: Testing Independence 465
13.4 Contingency Tables: Homogeneity 471
13.5 Fisher's Exact Test 473

CAMBRIDGE
Cambridge University Press & Assessment
978-1-107-11304-6 — Introduction to Probability and Statistics for Data Science
Steven E. Rigdon, Ronald D. Fricker, Jr, Douglas C. Montgomery
Table of Contents
More Information

13.6 The Continuity Correction and Simulation 479
13.7 McNemar's Test 481
13.8 Higher-Dimensional Tables and Simpson's Paradox 485
13.9 Chapter Summary 488
13.10 Problems 489

14 Regression 493
14.1 Introduction 493
14.1.1 Prediction vs. Explanation 493
14.1.2 Terminology 495
14.1.3 A Working Example 495
14.2 Simple Linear Regression 496
14.3 Properties of the Least Squares Estimators 503
14.4 Inference for Parameters of the Simple Linear Regression Model 510
14.5 Matrix Formulation of Simple Linear Regression 516
14.6 Joint Confidence Regions 518
14.7 Confidence and Prediction Intervals for Responses 520
14.8 Optimal Selection of Levels of Predictor Variables 526
14.9 The ANOVA Table for Simple Linear Regression 528
14.10 Linear Models in More than One Predictor 531
14.11 Indicator Variables 539
14.12 Polynomial and Nonlinear Regression 542
14.13 Inference for a Linear Combination of Model Parameters 546
14.14 Correlation 551
14.15 R^2 and Adjusted R^2 561
14.16 Model Checking 565
14.16.1 Normal Probability Plots 566
14.16.2 Plot of Residuals against the Fitted Values 567
14.17 Chapter Summary 568
14.18 Problems 570

15 Bayesian Methods 574
15.1 Introduction 574
15.2 Bayes' Theorem 575
15.3 The Bayesian Paradigm 581
15.4 Two Paradoxes 585
15.5 Conjugate Priors 588
15.6 Noninformative Priors 596
15.7 Simulation Methods 599
15.7.1 Metropolis–Hastings Algorithm 601
15.7.2 The Gibbs Sampling Algorithm 604
15.8 Hierarchical Bayes Models 606
15.9 Chapter Summary 613
15.10 Problems 614

16 Time Series Methods 618
16.1 Introduction 618
16.2 Using R for Time Series 622
16.3 Numerical Description of Time Series 623
16.4 Exponential Smoothing Methods 628
16.5 Autoregressive Integrated Moving Average (ARIMA) Models 635
16.6 Chapter Summary 640
16.7 Problems 642

17 Estimating the Standard Error: Analytic Approximations, the Jackknife, and the Bootstrap 645
17.1 Introduction 645
17.2 Analytic Approximations to the Standard Error of an Estimator 646
17.3 The Jackknife 656
17.4 The Bootstrap 662
17.4.1 Bootstrap Confidence Intervals Based on Percentiles 667
17.5 Parametric Bootstrap 670
17.6 Bootstrapping in R 674
17.7 Chapter Summary 681
17.8 Problems 682

18 Generalized Linear Models and Regression Trees 684
18.1 Logistic Regression 684
18.2 Multinomial Logistic Regression 698
18.3 Poisson Regression 703
18.4 Generalized Linear Models 706
18.5 Regression Trees 707
18.6 Discrimination and Classification 714

© in this web service Cambridge University Press & Assessment

Cambridge University Press & Assessment
978-1-107-11304-6 — Introduction to Probability and Statistics for Data Science
Steven E. Rigdon, Ronald D. Fricker, Jr, Douglas C. Montgomery
Table of Contents
More Information

18.6.1 $K = 2$ Groups and $p = 1$ Variable 718
18.6.2 K Groups and $p = 1$ Variable 724
18.6.3 K Groups and p Variables 725
18.6.4 Quadratic Discriminant Analysis 737
18.6.5 Dealing with Estimated Parameters 739
18.6.6 Choosing Between Linear and Quadratic Discriminant Analysis 739
18.7 Logistic Regression for Classification 740
18.8 Chapter Summary 745
18.9 Problems 746

19 Cross-Validation and Estimates of Prediction Error 751
19.1 Overfitting and Underfitting 751
19.2 Cross-Validation 754
19.2.1 Splitting the Data at Random 755
19.3 Leave-One-Out Cross-Validation 760
19.4 *k*-Fold Cross-Validation 762
19.5 Cross-Validation with Classification Data 765
19.6 Chapter Summary 767
19.7 Problems 769

20 Large-Scale Hypothesis Testing 773
20.1 Review of Hypothesis Testing 773
20.2 Testing Multiple Hypotheses 775
20.3 The FWER and the Bonferroni Correction 780
20.4 Holm's Method 783
20.5 The False Discovery Rate 785
20.6 Simultaneous Confidence Intervals 789
20.7 Tukey's Method 791
20.8 Scheffé's Method 794
20.9 Chapter Summary 801
20.10 Problems 802

References 805

Index 809

© in this web service Cambridge University Press & Assessment