

Principles of Managerial Statistics and Data Science

Roberto Rivera

*College of Business
University of Puerto Rico-Mayaguez*

WILEY

Contents

	Preface	<i>xv</i>
	Acknowledgments	<i>xvii</i>
	Acronyms	<i>xix</i>
	About the Companion Site	<i>xxi</i>
	Principles of Managerial Statistics and Data Science	<i>xxiii</i>
1	Statistics Suck; So Why Do I Need to Learn About It?	1
1.1	Introduction	1
	Practice Problems	4
1.2	Data-Based Decision Making: Some Applications	5
1.3	Statistics Defined	9
1.4	Use of Technology and the New Buzzwords: Data Science, Data Analytics, and Big Data	11
1.4.1	A Quick Look at Data Science: Some Definitions	11
	Chapter Problems	14
	Further Reading	14
2	Concepts in Statistics	15
2.1	Introduction	15
	Practice Problems	17
2.2	Type of Data	19
	Practice Problems	20
2.3	Four Important Notions in Statistics	22
	Practice Problems	24
2.4	Sampling Methods	25
2.4.1	Probability Sampling	25
2.4.2	Nonprobability Sampling	27
	Practice Problems	30
2.5	Data Management	31
2.5.1	A Quick Look at Data Science: Data Wrangling Baltimore Housing Variables	34

- 2.6 Proposing a Statistical Study 36
 - Chapter Problems 37
 - Further Reading 39

- 3 Data Visualization 41**
 - 3.1 Introduction 41
 - 3.2 Visualization Methods for Categorical Variables 41
 - Practice Problems 46
 - 3.3 Visualization Methods for Numerical Variables 50
 - Practice Problems 56
 - 3.4 Visualizing Summaries of More than Two Variables Simultaneously 59
 - 3.4.1 A Quick Look at Data Science: Does Race Affect the Chances of a Driver Being Searched During a Vehicle Stop in San Diego? 66
 - Practice Problems 69
 - 3.5 Novel Data Visualization 75
 - 3.5.1 A Quick Look at Data Science: Visualizing Association Between Baltimore Housing Variables Over 14 Years 78
 - Chapter Problems 81
 - Further Reading 96

- 4 Descriptive Statistics 97**
 - 4.1 Introduction 97
 - 4.2 Measures of Centrality 99
 - Practice Problems 108
 - 4.3 Measures of Dispersion 111
 - Practice Problems 115
 - 4.4 Percentiles 116
 - 4.4.1 Quartiles 117
 - Practice Problems 122
 - 4.5 Measuring the Association Between Two Variables 124
 - Practice Problems 128
 - 4.6 Sample Proportion and Other Numerical Statistics 130
 - 4.6.1 A Quick Look at Data Science: Murder Rates in Los Angeles 131
 - 4.7 How to Use Descriptive Statistics 132
 - Chapter Problems 133
 - Further Reading 139

- 5 Introduction to Probability 141**
 - 5.1 Introduction 141
 - 5.2 Preliminaries 142
 - Practice Problems 144

- 5.3 The Probability of an Event 145
 - Practice Problems 148
- 5.4 Rules and Properties of Probabilities 149
 - Practice Problems 152
- 5.5 Conditional Probability and Independent Events 154
 - Practice Problems 159
- 5.6 Empirical Probabilities 161
- 5.6.1 A Quick Look at Data Science: Missing People Reports in Boston by Day of Week 164
 - Practice Problems 165
- 5.7 Counting Outcomes 168
 - Practice Problems 171
 - Chapter Problems 171
 - Further Reading 175

- 6 Discrete Random Variables 177**
- 6.1 Introduction 177
- 6.2 General Properties 178
 - 6.2.1 A Quick Look at Data Science: Number of Stroke Emergency Calls in Manhattan 183
 - Practice Problems 184
- 6.3 Properties of Expected Value and Variance 186
 - Practice Problems 189
- 6.4 Bernoulli and Binomial Random Variables 190
 - Practice Problems 197
- 6.5 Poisson Distribution 198
 - Practice Problems 201
- 6.6 Optional: Other Useful Probability Distributions 203
 - Chapter Problems 205
 - Further Reading 208

- 7 Continuous Random Variables 209**
- 7.1 Introduction 209
 - Practice Problems 211
- 7.2 The Uniform Probability Distribution 211
 - Practice Problems 215
- 7.3 The Normal Distribution 216
 - Practice Problems 225
- 7.4 Probabilities for Any Normally Distributed Random Variable 227
- 7.4.1 A Quick Look at Data Science: Normal Distribution, A Good Match for University of Puerto Rico SATs? 229
 - Practice Problems 231

- 7.5 Approximating the Binomial Distribution 234
 - Practice Problems 236
- 7.6 Exponential Distribution 236
 - Practice Problems 238
 - Chapter Problems 239
 - Further Reading 242

- 8 Properties of Sample Statistics 243**
 - 8.1 Introduction 243
 - 8.2 Expected Value and Standard Deviation of \bar{X} 244
 - Practice Problems 246
 - 8.3 Sampling Distribution of \bar{X} When Sample Comes From a Normal Distribution 247
 - Practice Problems 251
 - 8.4 Central Limit Theorem 252
 - 8.4.1 A Quick Look at Data Science: Bacteria at New York City Beaches 257
 - Practice Problems 259
 - 8.5 Other Properties of Estimators 261
 - Chapter Problems 264
 - Further Reading 267

- 9 Interval Estimation for One Population Parameter 269**
 - 9.1 Introduction 269
 - 9.2 Intuition of a Two-Sided Confidence Interval 270
 - 9.3 Confidence Interval for the Population Mean: σ Known 271
 - Practice Problems 276
 - 9.4 Determining Sample Size for a Confidence Interval for μ 278
 - Practice Problems 279
 - 9.5 Confidence Interval for the Population Mean: σ Unknown 279
 - Practice Problems 284
 - 9.6 Confidence Interval for π 286
 - Practice Problems 287
 - 9.7 Determining Sample Size for π Confidence Interval 288
 - Practice Problems 290
 - 9.8 Optional: Confidence Interval for σ 290
 - 9.8.1 A Quick Look at Data Science: A Confidence Interval for the Standard Deviation of Walking Scores in Baltimore 292
 - Chapter Problems 293
 - Further Reading 296

- 10 Hypothesis Testing for One Population 297**
 - 10.1 Introduction 297

- 10.2 Basics of Hypothesis Testing 299
- 10.3 Steps to Perform a Hypothesis Test 304
 - Practice Problems 305
- 10.4 Inference on the Population Mean: Known Standard Deviation 306
 - Practice Problems 318
- 10.5 Hypothesis Testing for the Mean (σ Unknown) 323
 - Practice Problems 327
- 10.6 Hypothesis Testing for the Population Proportion 329
- 10.6.1 A Quick Look at Data Science: Proportion of New York City High Schools with a Mean SAT Score of 1498 or More 333
 - Practice Problems 334
- 10.7 Hypothesis Testing for the Population Variance 337
- 10.8 More on the p -Value and Final Remarks 338
- 10.8.1 Misunderstanding the p -Value 339
 - Chapter Problems 343
 - Further Reading 347

- 11 Statistical Inference to Compare Parameters from Two Populations 349**
- 11.1 Introduction 349
- 11.2 Inference on Two Population Means 350
- 11.3 Inference on Two Population Means – Independent Samples, Variances Known 351
 - Practice Problems 357
- 11.4 Inference on Two Population Means When Two Independent Samples are Used – Unknown Variances 360
- 11.4.1 A Quick Look at Data Science: Suicide Rates Among Asian Men and Women in New York City 364
 - Practice Problems 366
- 11.5 Inference on Two Means Using Two Dependent Samples 368
 - Practice Problems 370
- 11.6 Inference on Two Population Proportions 371
 - Practice Problems 374
 - Chapter Problems 375
 - References 378
 - Further Reading 378

- 12 Analysis of Variance (ANOVA) 379**
- 12.1 Introduction 379
 - Practice Problems 382
- 12.2 ANOVA for One Factor 383
 - Practice Problems 390

- 12.3 Multiple Comparisons 391
 - Practice Problems 395
- 12.4 Diagnostics of ANOVA Assumptions 395
 - 12.4.1 A Quick Look at Data Science: Emergency Response Time for Cardiac Arrest in New York City 399
 - Practice Problems 403
- 12.5 ANOVA with Two Factors 404
 - Practice Problems 409
- 12.6 Extensions to ANOVA 413
 - Chapter Problems 416
 - Further Reading 419

13 Simple Linear Regression 421

- 13.1 Introduction 421
- 13.2 Basics of Simple Linear Regression 423
 - Practice Problems 425
- 13.3 Fitting the Simple Linear Regression Parameters 426
 - Practice Problems 429
- 13.4 Inference for Simple Linear Regression 431
 - Practice Problems 440
- 13.5 Estimating and Predicting the Response Variable 443
 - Practice Problems 446
- 13.6 A Binary X 448
 - Practice Problems 449
- 13.7 Model Diagnostics (Residual Analysis) 450
 - Practice Problems 456
- 13.8 What Correlation Doesn't Mean 458
 - 13.8.1 A Quick Look at Data Science: Can Rate of College Educated People Help Predict the Rate of Narcotic Problems in Baltimore? 461
 - Chapter Problems 466
 - Further Reading 472

14 Multiple Linear Regression 473

- 14.1 Introduction 473
- 14.2 The Multiple Linear Regression Model 474
 - Practice Problems 477
- 14.3 Inference for Multiple Linear Regression 478
 - Practice Problems 483
- 14.4 Multicollinearity and Other Modeling Aspects 486
 - Practice Problems 490
- 14.5 Variability Around the Regression Line: Residuals and Intervals 492
 - Practice Problems 494
- 14.6 Modifying Predictors 494

- Practice Problems 495
- 14.7 General Linear Model 496
 - Practice Problems 502
- 14.8 Steps to Fit a Multiple Linear Regression Model 505
- 14.9 Other Regression Topics 507
 - 14.9.1 A Quick Look at Data Science: Modeling Taxi Fares in Chicago 510
 - Chapter Problems 513
 - Further Reading 517

- 15 Inference on Association of Categorical Variables 519**
 - 15.1 Introduction 519
 - 15.2 Association Between Two Categorical Variables 520
 - 15.2.1 A Quick Look at Data Science: Affordability and Business Environment in Chattanooga 525
 - Practice Problems 529
 - Chapter Problems 532
 - Further Reading 532

- 16 Nonparametric Testing 533**
 - 16.1 Introduction 533
 - 16.2 Sign Tests and Wilcoxon Sign-Rank Tests: One Sample and Matched Pairs Scenarios 533
 - Practice Problems 537
 - 16.3 Wilcoxon Rank-Sum Test: Two Independent Samples 539
 - 16.3.1 A Quick Look at Data Science: Austin, Texas, as a Place to Live; Do Men Rate It Higher Than Women? 540
 - Practice Problems 543
 - 16.4 Kruskal–Wallis Test: More Than Two Samples 544
 - Practice Problems 546
 - 16.5 Nonparametric Tests Versus Their Parametric Counterparts 547
 - Chapter Problems 548
 - Further Reading 549

- 17 Forecasting 551**
 - 17.1 Introduction 551
 - 17.2 Time Series Components 552
 - Practice Problems 557
 - 17.3 Simple Forecasting Models 558
 - Practice Problems 562
 - 17.4 Forecasting When Data Has Trend, Seasonality 563
 - Practice Problems 569

- 17.5 Assessing Forecasts 572
- 17.5.1 A Quick Look at Data Science: Forecasting Tourism Jobs in Canada 575
- 17.5.2 A Quick Look at Data Science: Forecasting Retail Gross Sales of Marijuana in Denver 577
- Chapter Problems 580
- Further Reading 581

Appendix A Math Notation and Symbols 583

- A.1 Summation 583
- A.2 p th Power 583
- A.3 Inequalities 584
- A.4 Factorials 584
- A.5 Exponential Function 585
- A.6 Greek and Statistics Symbols 585

Appendix B Standard Normal Cumulative Distribution Function 587

Appendix C t Distribution Critical Values 591

Appendix D Solutions to Odd-Numbered Problems 593

Index 643