CONTENTS

PREFACE xiii

1 INTRODUCTION 1

Data Analysis 1
What's in This Book 2
What's with the Workshops? 3
What's with the Math? 4
What You'll Need 5
What's Missing 6

PART I Graphics: Looking at Data

2 A SINGLE VARIABLE: SHAPE AND DISTRIBUTION 11

Dot and Jitter Plots 12
Histograms and Kernel Density Estimates 14
The Cumulative Distribution Function 23
Rank-Order Plots and Lift Charts 30
Only When Appropriate: Summary Statistics and Box Plots 33
Workshop: NumPy 38
Further Reading 45

3 TWO VARIABLES: ESTABLISHING RELATIONSHIPS 47

Scatter Plots 47
Conquering Noise: Smoothing 48
Logarithmic Plots 57
Banking 61
Linear Regression and All That 62
Showing What's Important 66
Graphical Analysis and Presentation Graphics 68
Workshop: matplotlib 69
Further Reading 78

4 TIME AS A VARIABLE: TIME-SERIES ANALYSIS 79

Examples 79
The Task 83
Smoothing 84
Don't Overlook the Obvious! 90
The Correlation Function 91

Optional: Filters and Convolutions 95
Workshop: scipy.signal 96
Further Reading 98

5 MORE THAN TWO VARIABLES: GRAPHICAL MULTIVARIATE ANALYSIS 99

False-Color Plots 100
A Lot at a Glance: Multiplots 105
Composition Problems 110
Novel Plot Types 116
Interactive Explorations 120
Workshop: Tools for Multivariate Graphics 123
Further Reading 125

6 INTERMEZZO: A DATA ANALYSIS SESSION 127

A Data Analysis Session 127
Workshop: gnuplot 136
Further Reading 138

PART II Analytics: Modeling Data

7 GUESSTIMATION AND THE BACK OF THE ENVELOPE 141

Principles of Guesstimation 142
How Good Are Those Numbers? 151
Optional: A Closer Look at Perturbation Theory and Error Propagation 155
Workshop: The Gnu Scientific Library (GSL) 158
Further Reading 161

8 MODELS FROM SCALING ARGUMENTS 163

Models 163
Arguments from Scale 165
Mean-Field Approximations 175
Common Time-Evolution Scenarios 178
Case Study: How Many Servers Are Best? 182
Why Modeling? 184
Workshop: Sage 184
Further Reading 188

9 ARGUMENTS FROM PROBABILITY MODELS 191

The Binomial Distribution and Bernoulli Trials 191
The Gaussian Distribution and the Central Limit Theorem 195
Power-Law Distributions and Non-Normal Statistics 201
Other Distributions 206
Optional: Case Study—Unique Visitors over Time 211
Workshop: Power-Law Distributions 215
Further Reading 218

10 WHAT YOU REALLY NEED TO KNOW ABOUT CLASSICAL STATISTICS 221

Genesis 221
Statistics Defined 223
Statistics Explained 226
Controlled Experiments Versus Observational Studies 230
Optional: Bayesian Statistics—The Other Point of View 235
Workshop: R 243
Further Reading 249

11 INTERMEZZO: MYTHBUSTING—BIGFOOT, LEAST SQUARES, AND ALL THAT 253

How to Average Averages 253
The Standard Deviation 256
Least Squares 260
Further Reading 264

PART III Computation: Mining Data

12 SIMULATIONS 267

A Warm-Up Question 267
Monte Carlo Simulations 270
Resampling Methods 276
Workshop: Discrete Event Simulations with SimPy 280
Further Reading 291

13 FINDING CLUSTERS 293

What Constitutes a Cluster? 293
Distance and Similarity Measures 298
Clustering Methods 304
Pre- and Postprocessing 311
Other Thoughts 314
A Special Case: Market Basket Analysis 316
A Word of Warning 319
Workshop: Pycluster and the C Clustering Library 320
Further Reading 324

14 SEEING THE FOREST FOR THE TREES: FINDING IMPORTANT ATTRIBUTES 327

Principal Component Analysis 328
Visual Techniques 337
Kohonen Maps 339
Workshop: PCA with R 342
Further Reading 348

15 INTERMEZZO: WHEN MORE IS DIFFERENT 351

A Horror Story 353

Some Suggestions 354
What About Map/Reduce? 356
Workshop: Generating Permutations 357
Further Reading 358

PART IV Applications: Using Data

16 REPORTING, BUSINESS INTELLIGENCE, AND DASHBOARDS 361
Business Intelligence 362
Corporate Metrics and Dashboards 369
Data Quality Issues 373
Workshop: Berkeley DB and SQLite 376
Further Reading 381

17 FINANCIAL CALCULATIONS AND MODELING 383
The Time Value of Money 384
Uncertainty in Planning and Opportunity Costs 391
Cost Concepts and Depreciation 394
Should You Care? 398
Is This All That Matters? 399
Workshop: The Newsvendor Problem 400
Further Reading 403

18 PREDICTIVE ANALYTICS 405
Introduction 405
Some Classification Terminology 407
Algorithms for Classification 408
The Process 419
The Secret Sauce 423
The Nature of Statistical Learning 424
Workshop: Two Do-It-Yourself Classifiers 426
Further Reading 431

19 EPILOGUE: FACTS ARE NOT REALITY 433

A PROGRAMMING ENVIRONMENTS FOR SCIENTIFIC COMPUTATION AND DATA ANALYSIS 435
Software Tools 435
A Catalog of Scientific Software 437
Writing Your Own 443
Further Reading 444

B RESULTS FROM CALCULUS 447
Common Functions 448
Calculus 460
Useful Tricks 468

Notation and Basic Math 472
Where to Go from Here 479
Further Reading 481

C WORKING WITH DATA 485

Sources for Data 485
Cleaning and Conditioning 487
Sampling 489
Data File Formats 490
The Care and Feeding of Your Data Zoo 492
Skills 493
Terminology 495
Further Reading 497

INDEX 499