Contents

AuthorBio xxiii

Preface xxx

ToTheReader 1

I Fundamentals of Probability 1

1 Basic Probability Models 3
  1.1 Example: Bus Ridership 3
  1.2 A “Notebook” View: the Notion of a Repeatable Experiment 4
    1.2.1 Theoretical Approaches 5
    1.2.2 A More Intuitive Approach 5
  1.3 Our Definitions 7
  1.4 “Mailing Tubes” 11
  1.5 Example: Bus Ridership Model (cont’d.) 11
  1.6 Example: ALOHA Network 14
    1.6.1 ALOHA Network Model Summary 16
    1.6.2 ALOHA Network Computations 16
  1.7 ALOHA in the Notebook Context 19
1.8 Example: A Simple Board Game ........................................ 20
1.9 Bayes’ Rule ................................................................. 23
  1.9.1 General Principle ...................................................... 23
  1.9.2 Example: Document Classification .............................. 23
1.10 Random Graph Models ................................................... 24
  1.10.1 Example: Preferential Attachment Model ..................... 25
1.11 Combinatorics-Based Computation .................................... 26
  1.11.1 Which Is More Likely in Five Cards, One King or Two Hearts? ......................................................... 26
  1.11.2 Example: Random Groups of Students ....................... 27
  1.11.3 Example: Lottery Tickets ........................................... 27
  1.11.4 Example: Gaps between Numbers .............................. 28
  1.11.5 Multinomial Coefficients .......................................... 29
  1.11.6 Example: Probability of Getting Four Aces in a Bridge Hand ......................................................... 30

2 Monte Carlo Simulation .................................................... 35
  2.1 Example: Rolling Dice .................................................. 35
    2.1.1 First Improvement .................................................. 36
    2.1.2 Second Improvement ............................................... 37
    2.1.3 Third Improvement ................................................ 38
  2.2 Example: Dice Problem ................................................ 39
  2.3 Use of runif() for Simulating Events .............................. 39
  2.4 Example: Bus Ridership (cont’d.) .................................. 40
  2.5 Example: Board Game (cont’d.) .................................... 40
  2.6 Example: Broken Rod ................................................... 41
  2.7 How Long Should We Run the Simulation? ....................... 42
  2.8 Computational Complements ....................................... 42
2.8.1 More on the replicate() Function 42

3 Discrete Random Variables: Expected Value 45
3.1 Random Variables 45
3.2 Discrete Random Variables 46
3.3 Independent Random Variables 46
3.4 Example: The Monty Hall Problem 47
3.5 Expected Value 50
  3.5.1 Generality — Not Just for Discrete Random Variables 50
  3.5.2 Misnomer 50
  3.5.3 Definition and Notebook View 50
3.6 Properties of Expected Value 51
  3.6.1 Computational Formula 51
  3.6.2 Further Properties of Expected Value 54
3.7 Example: Bus Ridership 58
3.8 Example: Predicting Product Demand 58
3.9 Expected Values via Simulation 59
3.10 Casinos, Insurance Companies and “Sum Users,” Compared to Others 60
3.11 Mathematical Complements 61
  3.11.1 Proof of Property E 61

4 Discrete Random Variables: Variance 65
4.1 Variance 65
  4.1.1 Definition 65
  4.1.2 Central Importance of the Concept of Variance 69
  4.1.3 Intuition Regarding the Size of $\text{Var}(X)$ 69
    4.1.3.1 Chebychev’s Inequality 69
CONTENTS

4.1.3.2 The Coefficient of Variation 70
4.2 A Useful Fact 71
4.3 Covariance 72
4.4 Indicator Random Variables, and Their Means and Variances 74
  4.4.1 Example: Return Time for Library Books, Version I 75
  4.4.2 Example: Return Time for Library Books, Version II 76
  4.4.3 Example: Indicator Variables in a Committee Problem 77
4.5 Skewness 79
4.6 Mathematical Complements 79
  4.6.1 Proof of Chebychev’s Inequality 79

5 Discrete Parametric Distribution Families 83
  5.1 Distributions 83
    5.1.1 Example: Toss Coin Until First Head 84
    5.1.2 Example: Sum of Two Dice 85
    5.1.3 Example: Watts-Strogatz Random Graph Model 85
      5.1.3.1 The Model 85
  5.2 Parametric Families of Distributions 86
  5.3 The Case of Importance to Us: Parameteric Families of pmfs 86
  5.4 Distributions Based on Bernoulli Trials 88
    5.4.1 The Geometric Family of Distributions 88
      5.4.1.1 R Functions 91
      5.4.1.2 Example: A Parking Space Problem 92
    5.4.2 The Binomial Family of Distributions 94
      5.4.2.1 R Functions 95
      5.4.2.2 Example: Parking Space Model 96
    5.4.3 The Negative Binomial Family of Distributions 96
6.5.1 Properties of Densities .................................. 120
6.5.2 Intuitive Meaning of Densities ......................... 122
6.5.3 Expected Values ........................................ 122
6.6 A First Example ............................................ 123
6.7 Famous Parametric Families of Continuous Distributions . 124
  6.7.1 The Uniform Distributions .............................. 125
    6.7.1.1 Density and Properties ........................ 125
    6.7.1.2 R Functions .................................... 125
    6.7.1.3 Example: Modeling of Disk Performance ....... 126
    6.7.1.4 Example: Modeling of Denial-of-Service Attack .............. 126
  6.7.2 The Normal (Gaussian) Family of Continuous Distributions ..... 127
    6.7.2.1 Density and Properties ........................ 127
    6.7.2.2 R Functions .................................... 127
    6.7.2.3 Importance in Modeling ........................ 128
  6.7.3 The Exponential Family of Distributions ............... 128
    6.7.3.1 Density and Properties ........................ 128
    6.7.3.2 R Functions .................................... 128
    6.7.3.3 Example: Garage Parking Fees .................. 129
    6.7.3.4 Memoryless Property of Exponential Distributions .......... 130
    6.7.3.5 Importance in Modeling ........................ 131
  6.7.4 The Gamma Family of Distributions ................... 131
    6.7.4.1 Density and Properties ........................ 132
    6.7.4.2 Example: Network Buffer ........................ 133
    6.7.4.3 Importance in Modeling ........................ 133
  6.7.5 The Beta Family of Distributions ..................... 134
### CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.7.5.1</td>
<td>Density Etc.</td>
<td>134</td>
</tr>
<tr>
<td>6.7.5.2</td>
<td>Importance in Modeling</td>
<td>138</td>
</tr>
<tr>
<td>6.8</td>
<td>Mathematical Complements</td>
<td>138</td>
</tr>
<tr>
<td>6.8.1</td>
<td>Hazard Functions</td>
<td>138</td>
</tr>
<tr>
<td>6.8.2</td>
<td>Duality of the Exponential Family with the Poisson Family</td>
<td>139</td>
</tr>
<tr>
<td>6.9</td>
<td>Computational Complements</td>
<td>141</td>
</tr>
<tr>
<td>6.9.1</td>
<td>R’s integrate() Function</td>
<td>141</td>
</tr>
<tr>
<td>6.9.2</td>
<td>Inverse Method for Sampling from a Density</td>
<td>141</td>
</tr>
<tr>
<td>6.9.3</td>
<td>Sampling from a Poisson Distribution</td>
<td>142</td>
</tr>
</tbody>
</table>

### II Fundamentals of Statistics

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Statistics: Prologue</td>
<td>149</td>
</tr>
<tr>
<td>7.1</td>
<td>Importance of This Chapter</td>
<td>150</td>
</tr>
<tr>
<td>7.2</td>
<td>Sampling Distributions</td>
<td>150</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Random Samples</td>
<td>150</td>
</tr>
<tr>
<td>7.3</td>
<td>The Sample Mean — a Random Variable</td>
<td>152</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Toy Population Example</td>
<td>152</td>
</tr>
<tr>
<td>7.3.2</td>
<td>Expected Value and Variance of ( \bar{X} )</td>
<td>153</td>
</tr>
<tr>
<td>7.3.3</td>
<td>Toy Population Example Again</td>
<td>154</td>
</tr>
<tr>
<td>7.3.4</td>
<td>Interpretation</td>
<td>155</td>
</tr>
<tr>
<td>7.3.5</td>
<td>Notebook View</td>
<td>155</td>
</tr>
<tr>
<td>7.4</td>
<td>Simple Random Sample Case</td>
<td>156</td>
</tr>
<tr>
<td>7.5</td>
<td>The Sample Variance</td>
<td>157</td>
</tr>
<tr>
<td>7.5.1</td>
<td>Intuitive Estimation of ( \sigma^2 )</td>
<td>157</td>
</tr>
<tr>
<td>7.5.2</td>
<td>Easier Computation</td>
<td>158</td>
</tr>
</tbody>
</table>
7.5.3 Special Case: X Is an Indicator Variable . . . . . . . . 158
7.6 To Divide by n or n-1? . . . . . . . . . . . . . . . . . . . . . 159
  7.6.1 Statistical Bias . . . . . . . . . . . . . . . . . . . . . 159
7.7 The Concept of a “Standard Error” . . . . . . . . . . . . 161
7.8 Example: Pima Diabetes Study . . . . . . . . . . . . . . 162
7.9 Don’t Forget: Sample ≠ Population! . . . . . . . . . . . . 164
7.10 Simulation Issues . . . . . . . . . . . . . . . . . . . . . . 164
  7.10.1 Sample Estimates . . . . . . . . . . . . . . . . . . . . 164
  7.10.2 Infinite Populations? . . . . . . . . . . . . . . . . . . 164
7.11 Observational Studies . . . . . . . . . . . . . . . . . . . . 165
7.12 Computational Complements . . . . . . . . . . . . . . . . 165
  7.12.1 The *apply() Functions . . . . . . . . . . . . . . . . 165
    7.12.1.1 R’s apply() Function . . . . . . . . . . . . . . . . 166
    7.12.1.2 The lapply() and sapply() Function . . . . . . . 166
    7.12.1.3 The split() and tapply() Functions . . . . . . . . 167
  7.12.2 Outliers/Errors in the Data . . . . . . . . . . . . . . 168

8 Fitting Continuous Models 171

8.1 Why Fit a Parametric Model? . . . . . . . . . . . . . . . . 171
8.2 Model-Free Estimation of a Density from Sample Data . . . 172
  8.2.1 A Closer Look . . . . . . . . . . . . . . . . . . . . . . 172
  8.2.2 Example: BMI Data . . . . . . . . . . . . . . . . . . . 173
  8.2.3 The Number of Bins . . . . . . . . . . . . . . . . . . . 174
    8.2.3.1 The Bias-Variance Tradeoff . . . . . . . . . . . . . 175
    8.2.3.2 The Bias-Variance Tradeoff in the Histogram Case . . 176
    8.2.3.3 A General Issue: Choosing the Degree of Smoothing . . 178
8.3 Advanced Methods for Model-Free Density Estimation . . . 180
8.4 Parameter Estimation . . . . . . . . . . . . . . . . . . . . . 181
  8.4.1 Method of Moments . . . . . . . . . . . . . . . . . . . 181
  8.4.2 Example: BMI Data . . . . . . . . . . . . . . . . . . . 182
  8.4.3 The Method of Maximum Likelihood . . . . . . . . . 183
  8.4.4 Example: Humidity Data . . . . . . . . . . . . . . . . 185
8.5 MM vs. MLE . . . . . . . . . . . . . . . . . . . . . . . . . 187
8.6 Assessment of Goodness of Fit . . . . . . . . . . . . . . . . 187
8.7 The Bayesian Philosophy . . . . . . . . . . . . . . . . . . . 189
  8.7.1 How Does It Work? . . . . . . . . . . . . . . . . . . . 190
  8.7.2 Arguments For and Against . . . . . . . . . . . . . . 190
8.8 Mathematical Complements . . . . . . . . . . . . . . . . . . 191
  8.8.1 Details of Kernel Density Estimators . . . . . . . . . 191
8.9 Computational Complements . . . . . . . . . . . . . . . . . 192
  8.9.1 Generic Functions . . . . . . . . . . . . . . . . . . . 192
  8.9.2 The gmm Package . . . . . . . . . . . . . . . . . . . . 193
    8.9.2.1 The gmm() Function . . . . . . . . . . . . . . 193
    8.9.2.2 Example: Bodyfat Data . . . . . . . . . . . . 193
9 The Family of Normal Distributions 197
9.1 Density and Properties . . . . . . . . . . . . . . . . . . . . . 197
  9.1.1 Closure under Affine Transformation . . . . . . . . . . 198
  9.1.2 Closure under Independent Summation . . . . . . . . . 199
  9.1.3 A Mystery . . . . . . . . . . . . . . . . . . . . . . . . 200
9.2 R Functions . . . . . . . . . . . . . . . . . . . . . . . . . . 200
9.3 The Standard Normal Distribution . . . . . . . . . . . . . . 200
9.4 Evaluating Normal cdfs . . . . . . . . . . . . . . . . . . . . 201
CONTENTS

9.5 Example: Network Intrusion ......................... 202
9.6 Example: Class Enrollment Size .................. 203
9.7 The Central Limit Theorem ......................... 204
  9.7.1 Example: Cumulative Roundoff Error ........ 205
  9.7.2 Example: Coin Tosses ......................... 205
  9.7.3 Example: Museum Demonstration .............. 206
  9.7.4 A Bit of Insight into the Mystery .......... 207
9.8 $\bar{X}$ Is Approximately Normal ................ 207
  9.8.1 Approximate Distribution of $X$ ............. 207
  9.8.2 Improved Assessment of Accuracy of $X$ .... 208
9.9 Importance in Modeling ............................ 209
9.10 The Chi-Squared Family of Distributions ....... 210
  9.10.1 Density and Properties ....................... 210
  9.10.2 Example: Error in Pin Placement .......... 211
  9.10.3 Importance in Modeling ....................... 211
  9.10.4 Relation to Gamma Family .................... 212
9.11 Mathematical Complements ......................... 212
  9.11.1 Convergence in Distribution, and the Precisely-Stated CLT ........ 212
9.12 Computational Complements ........................ 213
  9.12.1 Example: Generating Normal Random Numbers .. 213

10 Introduction to Statistical Inference .......... 217
  10.1 The Role of Normal Distributions ............... 217
  10.2 Confidence Intervals for Means ................. 218
    10.2.1 Basic Formulation .......................... 218
  10.3 Example: Pima Diabetes Study ................. 220
  10.4 Example: Humidity Data ......................... 221
10.5 Meaning of Confidence Intervals .......................... 221
  10.5.1 A Weight Survey in Davis ............................. 221
10.6 Confidence Intervals for Proportions ....................... 223
  10.6.1 Example: Machine Classification of Forest Covers . 224
10.7 The Student-t Distribution ................................ 226
10.8 Introduction to Significance Tests ......................... 227
10.9 The Proverbial Fair Coin .................................. 228
10.10 The Basics .................................................. 229
10.11 General Normal Testing .................................... 231
10.12 The Notion of “p-Values” ................................ 231
10.14 Example: The Forest Cover Data ......................... 232
10.15 Problems with Significance Testing ...................... 234
  10.15.1 History of Significance Testing ...................... 234
  10.15.2 The Basic Issues ...................................... 235
  10.15.3 Alternative Approach ................................ 236
10.16 The Problem of “P-hacking” ............................... 237
  10.16.1 A Thought Experiment ................................. 238
  10.16.2 Multiple Inference Methods ......................... 238
10.17 Philosophy of Statistics ................................... 239
  10.17.1 More about Interpretation of CIs ................... 239
    10.17.1.1 The Bayesian View of Confidence Intervals 241

III Multivariate Analysis ................................. 243

11 Multivariate Distributions ............................ 245
  11.1 Multivariate Distributions: Discrete .................. 245
11.1.1 Example: Marbles in a Bag .......................... 245
11.2 Multivariate Distributions: Continuous .................. 246
  11.2.1 Motivation and Definition .......................... 246
  11.2.2 Use of Multivariate Densities in Finding Probabilities
    and Expected Values ................................. 247
  11.2.3 Example: Train Rendezvous ......................... 247
11.3 Measuring Co-variation ................................. 248
  11.3.1 Covariance ............................ 248
  11.3.2 Example: The Committee Example Again ............. 250
11.4 Correlation ........................................ 251
  11.4.1 Sample Estimates ............................... 252
11.5 Sets of Independent Random Variables .................... 252
  11.5.1 Mailing Tubes ................................. 252
    11.5.1.1 Expected Values Factor ................. 253
    11.5.1.2 Covariance Is 0 .......................... 253
    11.5.1.3 Variances Add ............................ 253
11.6 Matrix Formulations ................................. 254
  11.6.1 Mailing Tubes: Mean Vectors ...................... 254
  11.6.2 Covariance Matrices ............................ 254
  11.6.3 Mailing Tubes: Covariance Matrices ............... 255
11.7 Sample Estimate of Covariance Matrix .................... 256
  11.7.1 Example: Pima Data ............................. 257
11.8 Mathematical Complements ............................. 257
  11.8.1 Convolution .................................. 257
    11.8.1.1 Example: Backup Battery ................. 258
  11.8.2 Transform Methods .............................. 259
    11.8.2.1 Generating Functions ................. 259
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.8.2.2</td>
<td>Sums of Independent Poisson Random Variables Are Poisson Distributed</td>
<td>261</td>
</tr>
<tr>
<td>12</td>
<td>The Multivariate Normal Family of Distributions</td>
<td>265</td>
</tr>
<tr>
<td>12.1</td>
<td>Densities</td>
<td>265</td>
</tr>
<tr>
<td>12.2</td>
<td>Geometric Interpretation</td>
<td>266</td>
</tr>
<tr>
<td>12.3</td>
<td>R Functions</td>
<td>269</td>
</tr>
<tr>
<td>12.4</td>
<td>Special Case: New Variable Is a Single Linear Combination of a Random Vector</td>
<td>270</td>
</tr>
<tr>
<td>12.5</td>
<td>Properties of Multivariate Normal Distributions</td>
<td>270</td>
</tr>
<tr>
<td>12.6</td>
<td>The Multivariate Central Limit Theorem</td>
<td>272</td>
</tr>
<tr>
<td>13</td>
<td>Mixture Distributions</td>
<td>275</td>
</tr>
<tr>
<td>13.1</td>
<td>Iterated Expectations</td>
<td>276</td>
</tr>
<tr>
<td>13.1.1</td>
<td>Conditional Distributions</td>
<td>277</td>
</tr>
<tr>
<td>13.1.2</td>
<td>The Theorem</td>
<td>277</td>
</tr>
<tr>
<td>13.1.3</td>
<td>Example: Flipping Coins with Bonuses</td>
<td>279</td>
</tr>
<tr>
<td>13.1.4</td>
<td>Conditional Expectation as a Random Variable</td>
<td>280</td>
</tr>
<tr>
<td>13.1.5</td>
<td>What about Variance?</td>
<td>280</td>
</tr>
<tr>
<td>13.2</td>
<td>A Closer Look at Mixture Distributions</td>
<td>281</td>
</tr>
<tr>
<td>13.2.1</td>
<td>Derivation of Mean and Variance</td>
<td>281</td>
</tr>
<tr>
<td>13.2.2</td>
<td>Estimation of Parameters</td>
<td>283</td>
</tr>
<tr>
<td>13.2.2.1</td>
<td>Example: Old Faithful Estimation</td>
<td>283</td>
</tr>
<tr>
<td>13.3</td>
<td>Clustering</td>
<td>284</td>
</tr>
<tr>
<td>14</td>
<td>Multivariate Description and Dimension Reduction</td>
<td>287</td>
</tr>
<tr>
<td>14.1</td>
<td>What Is Overfitting Anyway?</td>
<td>288</td>
</tr>
<tr>
<td>14.1.1</td>
<td>“Desperate for Data”</td>
<td>288</td>
</tr>
<tr>
<td>14.1.2</td>
<td>Known Distribution</td>
<td>289</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>14.1.3 Estimated Mean</td>
<td>289</td>
<td></td>
</tr>
<tr>
<td>14.1.4 The Bias/Variance Tradeoff: Concrete Illustration</td>
<td>290</td>
<td></td>
</tr>
<tr>
<td>14.1.5 Implications</td>
<td>292</td>
<td></td>
</tr>
<tr>
<td>14.2 Principal Components Analysis</td>
<td>293</td>
<td></td>
</tr>
<tr>
<td>14.2.1 Intuition</td>
<td>293</td>
<td></td>
</tr>
<tr>
<td>14.2.2 Properties of PCA</td>
<td>295</td>
<td></td>
</tr>
<tr>
<td>14.2.3 Example: Turkish Teaching Evaluations</td>
<td>296</td>
<td></td>
</tr>
<tr>
<td>14.3 The Log-Linear Model</td>
<td>297</td>
<td></td>
</tr>
<tr>
<td>14.3.1 Example: Hair Color, Eye Color and Gender</td>
<td>297</td>
<td></td>
</tr>
<tr>
<td>14.3.2 Dimension of Our Data</td>
<td>299</td>
<td></td>
</tr>
<tr>
<td>14.3.3 Estimating the Parameters</td>
<td>299</td>
<td></td>
</tr>
<tr>
<td>14.4 Mathematical Complements</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>14.4.1 Statistical Derivation of PCA</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>14.5 Computational Complements</td>
<td>302</td>
<td></td>
</tr>
<tr>
<td>14.5.1 R Tables</td>
<td>302</td>
<td></td>
</tr>
<tr>
<td>14.5.2 Some Details on Log-Linear Models</td>
<td>302</td>
<td></td>
</tr>
<tr>
<td>14.5.2.1 Parameter Estimation</td>
<td>303</td>
<td></td>
</tr>
<tr>
<td>14.5.2.2 The loglin() Function</td>
<td>304</td>
<td></td>
</tr>
<tr>
<td>14.5.2.3 Informal Assessment of Fit</td>
<td>305</td>
<td></td>
</tr>
<tr>
<td>15 Predictive Modeling</td>
<td>309</td>
<td></td>
</tr>
<tr>
<td>15.1 Example: Heritage Health Prize</td>
<td>309</td>
<td></td>
</tr>
<tr>
<td>15.2 The Goals: Prediction and Description</td>
<td>310</td>
<td></td>
</tr>
<tr>
<td>15.2.1 Terminology</td>
<td>310</td>
<td></td>
</tr>
<tr>
<td>15.3 What Does “Relationship” Mean?</td>
<td>311</td>
<td></td>
</tr>
<tr>
<td>15.3.1 Precise Definition</td>
<td>311</td>
<td></td>
</tr>
<tr>
<td>15.3.2 Parametric Models for the Regression Function m()</td>
<td>313</td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>15.4</td>
<td>Estimation in Linear Parametric Regression Models</td>
<td>314</td>
</tr>
<tr>
<td>15.5</td>
<td>Example: Baseball Data</td>
<td>315</td>
</tr>
<tr>
<td>15.5.1</td>
<td>R Code</td>
<td>316</td>
</tr>
<tr>
<td>15.6</td>
<td>Multiple Regression</td>
<td>319</td>
</tr>
<tr>
<td>15.7</td>
<td>Example: Baseball Data (cont’d.)</td>
<td>320</td>
</tr>
<tr>
<td>15.8</td>
<td>Interaction Terms</td>
<td>321</td>
</tr>
<tr>
<td>15.9</td>
<td>Parametric Estimation</td>
<td>322</td>
</tr>
<tr>
<td>15.9.1</td>
<td>Meaning of “Linear”</td>
<td>322</td>
</tr>
<tr>
<td>15.9.2</td>
<td>Random-X and Fixed-X Regression</td>
<td>322</td>
</tr>
<tr>
<td>15.9.3</td>
<td>Point Estimates and Matrix Formulation</td>
<td>323</td>
</tr>
<tr>
<td>15.9.4</td>
<td>Approximate Confidence Intervals</td>
<td>326</td>
</tr>
<tr>
<td>15.10</td>
<td>Example: Baseball Data (cont’d.)</td>
<td>328</td>
</tr>
<tr>
<td>15.11</td>
<td>Dummy Variables</td>
<td>329</td>
</tr>
<tr>
<td>15.12</td>
<td>Classification</td>
<td>330</td>
</tr>
<tr>
<td>15.12.1</td>
<td>Classification = Regression</td>
<td>331</td>
</tr>
<tr>
<td>15.12.2</td>
<td>Logistic Regression</td>
<td>332</td>
</tr>
<tr>
<td>15.12.2.1</td>
<td>The Logistic Model: Motivations</td>
<td>332</td>
</tr>
<tr>
<td>15.12.2.2</td>
<td>Estimation and Inference for Logit</td>
<td>334</td>
</tr>
<tr>
<td>15.12.3</td>
<td>Example: Forest Cover Data</td>
<td>334</td>
</tr>
<tr>
<td>15.12.4</td>
<td>R Code</td>
<td>334</td>
</tr>
<tr>
<td>15.12.5</td>
<td>Analysis of the Results</td>
<td>335</td>
</tr>
<tr>
<td>15.12.5.1</td>
<td>Multiclass Case</td>
<td>336</td>
</tr>
<tr>
<td>15.13</td>
<td>Machine Learning: Neural Networks</td>
<td>336</td>
</tr>
<tr>
<td>15.13.1</td>
<td>Example: Predicting Vertebral Abnormalities</td>
<td>336</td>
</tr>
<tr>
<td>15.13.3</td>
<td>R Packages</td>
<td>339</td>
</tr>
<tr>
<td>15.14</td>
<td>Computational Complements</td>
<td>340</td>
</tr>
</tbody>
</table>
16 Model Parsimony and Overfitting 343
16.1 What Is Overfitting? 343
16.1.1 Example: Histograms 343
16.1.2 Example: Polynomial Regression 344
16.2 Can Anything Be Done about It? 345
16.2.1 Cross-Validation 345
16.3 Predictor Subset Selection 346

17 Introduction to Discrete Time Markov Chains 349
17.1 Matrix Formulation 350
17.2 Example: Die Game 351
17.3 Long-Run State Probabilities 352
17.3.1 Stationary Distribution 353
17.3.2 Calculation of $\pi$ 354
17.3.3 Simulation Calculation of $\pi$ 355
17.4 Example: 3-Heads-in-a-Row Game 356
17.5 Example: Bus Ridership Problem 358
17.6 Hidden Markov Models 359
17.6.1 Example: Bus Ridership 360
17.6.2 Computation 361
17.7 Google PageRank 361
17.8 Computational Complements 361
17.8.1 Initializing a Matrix to All 0s 361
# CONTENTS

IV Appendices 365

A R Quick Start 367

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1</td>
<td>367</td>
</tr>
<tr>
<td>A.2</td>
<td>368</td>
</tr>
<tr>
<td>A.3</td>
<td>369</td>
</tr>
<tr>
<td>A.4</td>
<td>372</td>
</tr>
<tr>
<td>A.5</td>
<td>372</td>
</tr>
<tr>
<td>A.6</td>
<td>374</td>
</tr>
<tr>
<td>A.7</td>
<td>374</td>
</tr>
<tr>
<td>A.8</td>
<td>375</td>
</tr>
<tr>
<td>A.9</td>
<td>376</td>
</tr>
<tr>
<td>A.9.1</td>
<td>376</td>
</tr>
<tr>
<td>A.9.2</td>
<td>377</td>
</tr>
<tr>
<td>A.10</td>
<td>378</td>
</tr>
<tr>
<td>A.11</td>
<td>380</td>
</tr>
<tr>
<td>A.12</td>
<td>380</td>
</tr>
</tbody>
</table>

B Matrix Algebra 383

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.1</td>
<td>383</td>
</tr>
<tr>
<td>B.1.1</td>
<td>383</td>
</tr>
<tr>
<td>B.2</td>
<td>385</td>
</tr>
<tr>
<td>B.3</td>
<td>385</td>
</tr>
<tr>
<td>B.4</td>
<td>385</td>
</tr>
<tr>
<td>B.5</td>
<td>386</td>
</tr>
<tr>
<td>B.5.1</td>
<td>386</td>
</tr>
</tbody>
</table>

References 394
About the Author

Dr. Norm Matloff is a professor of computer science at the University of California at Davis, and was formerly a professor of statistics at that university. He is a former database software developer in Silicon Valley, and has been a statistical consultant for firms such as the Kaiser Permanente Health Plan.

Dr. Matloff was born in Los Angeles, and grew up in East Los Angeles and the San Gabriel Valley. He has a PhD in pure mathematics from UCLA, specializing in probability theory and statistics. He has published numerous papers in computer science and statistics, with current research interests in parallel processing, regression methodology, machine learning and recommender systems.

Professor Matloff is an award-winning expositor. He is a recipient of the campuswide Distinguished Teaching Award at his university, and his book, *Statistical Regression and Classification: From Linear Models to Machine Learning*, was selected for the international 2017 Ziegel Award. (He also has been a recipient of the campuswide Distinguished Public Service Award at UC Davis.)
To the Instructor

Statistics is not a discipline like physics, chemistry or biology where we study a subject to solve problems in the same subject. We study statistics with the main aim of solving problems in other disciplines — C.R. Rao, one of the pioneers of modern statistics

The function of education is to teach one to think intensively and to think critically. Intelligence plus character — that is the goal of true education — Dr. Martin Luther King, American civil rights leader

[In spite of] innumerable twists and turns, the Yellow River flows east — Confucius, ancient Chinese philosopher

This text is designed for a junior/senior/graduate-level based course in probability and statistics, aimed specifically at data science students (including computer science). In addition to calculus, the text assumes some knowledge of matrix algebra and rudimentary computer programming.

But why is this book different from all other books on math probability and statistics?

Indeed, it is quite different from the others. Briefly:

• The subtitle of this book, Math + R + Data, immediately signals a difference from other “math stat” books.

• Data Science applications, e.g. random graph models, power law distribution, Hidden Markov models, PCA, Google PageRank, remote sensing, mixture distributions, neural networks, the Curse of Dimensionality, and so on.

• Extensive use of the R language.
The subtitle of this book, *Math + R + Data*, immediately signals that the book follows a very different path. Unlike other “math stat” books, this one has a strong applied emphasis, with lots of real data, facilitated by extensive use of the R language.

The above quotations explain the difference further. First, this book is definitely written from an applications point of view. Second, it pushes the student to think critically about the *how* and *why* of statistics, and to “see the big picture.”

- **Use of real data, and early introduction of statistical issues:**
  
  The Rao quote at the outset of this Preface resonates strongly with me. Though this is a “math stat” book — random variables, density functions, expected values, distribution families, stat estimation and inference, and so on — it takes seriously the Data Science theme claimed in the title, *Probability and Statistics for Data Science*. A book on Data Science, even a mathematical one, should make heavy use of DATA!

  This has implications for the ordering of the chapters. We bring in statistics early, and statistical issues are interspersed throughout the text. Even the introduction to expected value, Chapter 3, includes a simple prediction model, serving as a preview of what will come in Chapter 15. Chapter 5, which covers the famous discrete parametric models, includes an example of fitting the power law distribution to real data. This forms a prelude to Chapter 7, which treats sampling distributions, estimation of mean and variance, bias and so on. Then Chapter 8 covers general point estimation, using MLE and the Method of Moments to fit models to real data. From that point onward, real data is used extensively in every chapter.

  The datasets are all publicly available, so that the instructor can delve further into the data examples.

- **Mathematically correct – yet highly intuitive:**

  The Confucius quote, though made long before the development of formal statistical methods, shows that he had a keen *intuition*, anticipating a fundamental concept in today’s world of data science — data smoothing. Development of such strong intuition in our students is a high priority of this book.

  This is of course a mathematics book. All models, concepts and so on are described precisely in terms of random variables and distributions. In addition to calculus, matrix algebra plays an important role. Optional Mathematical Complements sections at the ends of
many chapters allow inquisitive readers to explore more sophisticated material. The mathematical exercises range from routine to more challenging.

On the other hand, this book is not about “math for math’s sake.” In spite of being mathematically precise in description, it is definitely not a theory book.

For instance, the book does not define probability in terms of sample spaces and set-theoretic terminology. In my experience, defining probability in the classical manner is a major impediment to learning the intuition underlying the concepts, and later to doing good applied work. Instead, I use the intuitive, informal approach of defining probability in terms of long-run frequency, in essence taking the Strong Law of Large Numbers as an axiom.

I believe this approach is especially helpful when explaining conditional probability and expectation, concepts that students notoriously have trouble with. Under the classical approach, students have trouble recognizing when an exercise — and more importantly, an actual application — calls for a conditional probability or expectation if the wording lacks the explicit phrase given that. Instead, I have the reader think in terms of repeated trials, “How often does A occur among those times in which B occurs?”, which is easier to relate to practical settings.

- **Empowering students for real-world applications:**

The word applied can mean different things to different people. Consider for instance the interesting, elegant book for computer science students by Mitzenmacher and Upfal [33]. It focuses on probability, in fact discrete probability, and its intended class of applications is actually the theory of computer science.

I instead focus on the actual use of the material in the real world; which tends to be more continuous than discrete, and more in the realm of statistics than probability. This is especially valuable, as Big Data and Machine Learning now play a significant role in computer and data science.

One sees this philosophy in the book immediately. Instead of starting out with examples involving dice or coins, the book’s very first examples involve a model of a bus transportation system and a model of a computer network. There are indeed also examples using dice, coins and games, but the theme of the late Leo Breiman’s book subtitle [5], “With a View toward Applications,” is never far away.
If I may take the liberty of extending King’s quote, I would note that today statistics is a core intellectual field, affecting virtually everyone’s daily lives. The ability to use, or at the very least understand, statistics is vital to good citizenship, and as an author I take this as a mission.

• **Use of the R Programming Language:**

The book makes use of some light programming in R, for the purposes of simulation and data analysis. The student is expected to have had some rudimentary prior background in programming, say in one of Python, C, Java or R, but no prior experience with R is assumed. A brief introduction is given in the book’s appendix, and some further R topics are interspersed with the text as Computational Complements.

R is widely used in the world of statistics and data science, with outstanding graphics/visualization capabilities, and a treasure chest of more than 10,000 contributed code packages.

Readers who happen to be in computer science will find R to be of independent interest from a CS perspective. First, R follows the *functional language* and *object-oriented* paradigms: Every action is implemented as a function (even ‘+’); side effects are (almost) always avoided; functions are first-class objects; several different kinds of class structures are offered. R also offers various interesting metaprogramming capabilities. In terms of programming support, there is the extremely popular RStudio IDE, and for the “hard core” coder, the Emacs Speaks Statistics framework. Most chapters in the book have Computational Complements sections, as well as a Computational and Data Problems portion in the exercises.

**Chapter Outline:**

Part I, Chapters 1 through 6: These introduce probability, Monte Carlo simulation, discrete random variables, expected value and variance, and parametric families of discrete distributions.

Part II, Chapters 7 through 10: These then introduce statistics, such as sampling distributions, MLE, bias, Kolmogorov-Smirnov and so on, illustrated by fitting gamma and beta density models to real data. Histograms are viewed as density estimators, and kernel density estimation is briefly covered. This is followed by material on confidence intervals and significance testing.

Part III, Chapters 11 through 17: These cover multivariate analysis in various aspects, such as multivariate distribution, mixture distributions,
PCA/log-linear model, dimension reduction, overfitting and predictive analytics. Again, real data plays a major role.

**Coverage Strategies:**

The book can be comfortably covered in one semester. If a more leisurely pace is desired, or one is teaching under a quarter system, the material has been designed so that some parts can be skipped without loss of continuity. In particular, a more statistics-oriented course might omit the material on Markov chains, while a course focusing more on machine learning may wish to retain this material (e.g. for Hidden Markov models). Individual sections on specialty topics also have been written so as not to create obstacles later on if they are skipped.

The Chapter 11 on multivariate distributions is very useful for data science, e.g. for its relation to clustering. However, instructors who are short on time or whose classes may not have a strong background in matrix algebra may safely skip much of this material.

**A Note on Typography**

In order to help the reader keep track of the various named items, I use math italics for mathematical symbols and expressions, and bold face for program variable and function names. I include R package names for the latter, except for those beginning with a capital letter.

**Thanks:**

The following, among many, provided valuable feedback for which I am very grateful: Ibrahim Ahmed; Ahmed Ahmedin; Stuart Ambler; Earl Barr; Benjamin Beasley; Matthew Butner; Vishal Chakraborti, Michael Clifford; Dipak Ghosal; Noah Gift; Laura Matloff; Nelson Max, Deep Mukhopadhyay, Connie Nguyen, Jack Norman, Richard Oehrle, Michael Rea, Sana Vaziri, Yingkang Xie, and Ivana Zetko. My editor, John Kimmel, is always profoundly helpful. And as always, my books are also inspired tremendously by my wife Gamis and daughter Laura.
To the Reader

I took a course in speed reading, and read War and Peace in 20 minutes. It’s about Russia — comedian Woody Allen

I learned very early the difference between knowing the name of something and knowing something — Richard Feynman, Nobel laureate in physics

Give me six hours to chop down a tree and I will spend the first four sharpening the axe — Abraham Lincoln

This is NOT your ordinary math or programming book.

In order to use this material in real-world applications, it’s crucial to understand what the math means, and what the code actually does.

In this book, you will often find several consecutive paragraphs, maybe even a full page, in which there is no math, no code and no graphs. Don’t skip over these portions of the book! They may actually be the most important ones in the book, in terms of your ability to apply the material in the real world.

And going hand-in-hand with this point, mathematical intuition is key. As you read, stop and think about the intuition underlying those equations.

A closely related point is that the math and code complement each other. Each will give you deeper insight in the other. It may at first seem odd that the book intersperses math and code, but soon you will find their interaction to be quite helpful to your understanding of the material.

The “Plot”

Think of this book as a movie. In order for the “plot” to work well, we will need preparation. This book is aimed at applications to Data Science, so the ultimate destination of the “plot” is statistics and predictive analytics.
The foundation for those fields is probability, so we lay the foundation first in Chapters 1 through 6. We’ll need more probability later — Chapters 9 and 11 — but in order to bring in some “juicy” material into the “movie” as early as possible, we introduce statistics, especially analysis of real DATA, in Chapters 7 and 8 at this early stage.

The final chapter, on Markov chains, is like a “sequel” to the movie. This sets up some exciting Data Science applications such as Hidden Markov Models and Google’s PageRank search engine.