

Table of Contents

Preface.....	xiii
1. Exploratory Data Analysis.....	1
Elements of Structured Data	2
Further Reading	4
Rectangular Data	5
Data Frames and Indexes	6
Nonrectangular Data Structures	7
Further Reading	8
Estimates of Location	8
Mean	9
Median and Robust Estimates	10
Example: Location Estimates of Population and Murder Rates	12
Further Reading	13
Estimates of Variability	13
Standard Deviation and Related Estimates	15
Estimates Based on Percentiles	17
Example: Variability Estimates of State Population	18
Further Reading	19
Exploring the Data Distribution	19
Percentiles and Boxplots	20
Frequency Table and Histograms	21
Density Estimates	24
Further Reading	26
Exploring Binary and Categorical Data	26
Mode	28
Expected Value	28
Further Reading	29

Correlation	29
Scatterplots	32
Further Reading	34
Exploring Two or More Variables	34
Hexagonal Binning and Contours (Plotting Numeric versus Numeric Data)	34
Two Categorical Variables	37
Categorical and Numeric Data	38
Visualizing Multiple Variables	40
Further Reading	42
Summary	42
2. Data and Sampling Distributions.....	43
Random Sampling and Sample Bias	44
Bias	46
Random Selection	47
Size versus Quality: When Does Size Matter?	48
Sample Mean versus Population Mean.....	49
Further Reading	49
Selection Bias	50
Regression to the Mean	51
Further Reading	53
Sampling Distribution of a Statistic	53
Central Limit Theorem	55
Standard Error	56
Further Reading	57
The Bootstrap	57
Resampling versus Bootstrapping	60
Further Reading	60
Confidence Intervals	61
Further Reading	63
Normal Distribution	64
Standard Normal and QQ-Plots	65
Long-Tailed Distributions	67
Further Reading	69
Student's t-Distribution	69
Further Reading	72
Binomial Distribution	72
Further Reading	74
Poisson and Related Distributions	74
Poisson Distributions	75
Exponential Distribution	75
Estimating the Failure Rate	76

Weibull Distribution	76
Further Reading	77
Summary	77
3. Statistical Experiments and Significance Testing	79
A/B Testing	80
Why Have a Control Group?	82
Why Just A/B? Why Not C, D...?	83
For Further Reading	84
Hypothesis Tests	85
The Null Hypothesis	86
Alternative Hypothesis	86
One-Way, Two-Way Hypothesis Test	87
Further Reading	88
Resampling	88
Permutation Test	88
Example: Web Stickiness	89
Exhaustive and Bootstrap Permutation Test	92
Permutation Tests: The Bottom Line for Data Science	93
For Further Reading	93
Statistical Significance and P-Values	93
P-Value	96
Alpha	96
Type 1 and Type 2 Errors	98
Data Science and P-Values	98
Further Reading	99
t-Tests	99
Further Reading	101
Multiple Testing	101
Further Reading	104
Degrees of Freedom	104
Further Reading	106
ANOVA	106
F-Statistic	109
Two-Way ANOVA	110
Further Reading	111
Chi-Square Test	111
Chi-Square Test: A Resampling Approach	112
Chi-Square Test: Statistical Theory	114
Fisher's Exact Test	115
Relevance for Data Science	117
Further Reading	118

Multi-Arm Bandit Algorithm	119
Further Reading	121
Power and Sample Size	122
Sample Size	123
Further Reading	125
Summary	125
4. Regression and Prediction.....	127
Simple Linear Regression	127
The Regression Equation	129
Fitted Values and Residuals	131
Least Squares	132
Prediction versus Explanation (Profiling)	133
Further Reading	134
Multiple Linear Regression	134
Example: King County Housing Data	135
Assessing the Model	136
Cross-Validation	138
Model Selection and Stepwise Regression	139
Weighted Regression	141
Further Reading	142
Prediction Using Regression	142
The Dangers of Extrapolation	143
Confidence and Prediction Intervals	143
Factor Variables in Regression	145
Dummy Variables Representation	146
Factor Variables with Many Levels	148
Ordered Factor Variables	149
Interpreting the Regression Equation	150
Correlated Predictors	150
Multicollinearity	152
Confounding Variables	152
Interactions and Main Effects	153
Testing the Assumptions: Regression Diagnostics	155
Outliers	156
Influential Values	158
Heteroskedasticity, Non-Normality and Correlated Errors	161
Partial Residual Plots and Nonlinearity	164
Polynomial and Spline Regression	166
Polynomial	167
Splines	168
Generalized Additive Models	170

Further Reading	172
Summary	172
5. Classification.....	173
Naive Bayes	174
Why Exact Bayesian Classification Is Impractical	175
The Naive Solution	176
Numeric Predictor Variables	178
Further Reading	178
Discriminant Analysis	179
Covariance Matrix	180
Fisher's Linear Discriminant	180
A Simple Example	181
Further Reading	183
Logistic Regression	184
Logistic Response Function and Logit	184
Logistic Regression and the GLM	186
Generalized Linear Models	187
Predicted Values from Logistic Regression	188
Interpreting the Coefficients and Odds Ratios	188
Linear and Logistic Regression: Similarities and Differences	190
Assessing the Model	191
Further Reading	194
Evaluating Classification Models	194
Confusion Matrix	195
The Rare Class Problem	196
Precision, Recall, and Specificity	197
ROC Curve	198
AUC	200
Lift	201
Further Reading	202
Strategies for Imbalanced Data	203
Undersampling	204
Oversampling and Up/Down Weighting	204
Data Generation	205
Cost-Based Classification	206
Exploring the Predictions	206
Further Reading	208
Summary	208
6. Statistical Machine Learning.....	209
K-Nearest Neighbors	210

172	A Small Example: Predicting Loan Default	211
172	Distance Metrics	213
173	One Hot Encoder	214
173	Standardization (Normalization, Z-Scores)	215
174	Choosing K	217
175	KNN as a Feature Engine	218
176	Tree Models	220
178	A Simple Example	221
178	The Recursive Partitioning Algorithm	223
179	Measuring Homogeneity or Impurity	224
180	Stopping the Tree from Growing	226
180	Predicting a Continuous Value	227
181	How Trees Are Used	228
183	Further Reading	229
184	Bagging and the Random Forest	229
184	Bagging	230
186	Random Forest	231
187	Variable Importance	234
188	Hyperparameters	237
188	Boosting	238
190	The Boosting Algorithm	239
191	XGBoost	240
194	Regularization: Avoiding Overfitting	242
194	Hyperparameters and Cross-Validation	246
195	Summary	248
196	7. Unsupervised Learning	249
198	Principal Components Analysis	250
200	A Simple Example	251
201	Computing the Principal Components	254
202	Interpreting Principal Components	254
203	Further Reading	257
204	K-Means Clustering	257
204	A Simple Example	258
205	K-Means Algorithm	260
206	Interpreting the Clusters	261
206	Selecting the Number of Clusters	263
208	Hierarchical Clustering	265
208	A Simple Example	266
208	The Dendrogram	266
209	The Agglomerative Algorithm	268
210	Measures of Dissimilarity	268

Model-Based Clustering	270
Multivariate Normal Distribution	270
Mixtures of Normals	272
Selecting the Number of Clusters	274
Further Reading	276
Scaling and Categorical Variables	276
Scaling the Variables	277
Dominant Variables	279
Categorical Data and Gower's Distance	280
Problems with Clustering Mixed Data	282
Summary	283

Bibliography	285
---------------------------	------------

Index	287
--------------------	------------

This book is aimed at the data scientist with some familiarity with the R programming language, and with some prior (perhaps spotty or ephemeral) exposure to statistics. Both of us came to the world of data science from the world of statistics, so we have some appreciation of the contribution that statistics can make to the art of data science. At the same time, we are well aware of the limitations of traditional statistics instruction: statistics as a discipline is a century and a half old, and most statistics textbooks and courses are laden with the momentum and inertia of an ocean liner.

Two goals underlie this book:

- To lay out, in digestible, navigable, and easily referenced form, key concepts from statistics that are relevant to data science.
- To explain which concepts are important and useful from a data science perspective, which are less so, and why.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic

Indicates new terms, URLs, email addresses, filenames, and file extensions.

Constant width

Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.