Table of Contents

Preface ix

1. Exploratory Data Analysis 1
A Statistical Approach 2
The National Survey of Family Growth 2
Importing the Data 3
DataFrames 4
Variables 6
Transformation 7
Validation 8
Interpretation 9
Exercises 11
Glossary 12

2. Distributions 15
Representing Histograms 16
Plotting Histograms 16
NSFG Variables 17
Outliers 19
First Babies 20
Summarizing Distributions 22
Variance 23
Effect Size 23
Reporting Results 24
Exercises 25
Glossary 25

3. Probability Mass Functions 27
Pmfs 27

Plotting PMFs 28
Other Visualizations 30
The Class Size Paradox 30
DataFrame Indexing 34
Exercises 35
Glossary 37

4. Cumulative Distribution Functions. 39
The Limits of PMFs 39
Percentiles 40
CDFs 41
Representing CDFs 42
Comparing CDFs 44
Percentile-Based Statistics 44
Random Numbers 45
Comparing Percentile Ranks 47
Exercises 47
Glossary 48

5. Modeling Distributions. 49
The Exponential Distribution 49
The Normal Distribution 52
Normal Probability Plot 54
The lognormal Distribution 56
The Pareto Distribution 58
Generating Random Numbers 61
Why Model? 62
Exercises 62
Glossary 64

6. Probability Density Functions. 65
PDFs 65
Kernel Density Estimation 67
The Distribution Framework 69
Hist Implementation 69
Pmf Implementation 70
Cdf Implementation 71
Moments 72
Skewness 73
Exercises 75
Glossary 77

7. Relationships Between Variables **79**
Scatter Plots 79
Characterizing Relationships 82
Correlation 83
Covariance 84
Pearson's Correlation 85
Nonlinear Relationships 86
Spearman's Rank Correlation 87
Correlation and Causation 88
Exercises 89
Glossary 89

8. Estimation **91**
The Estimation Game 91
Guess the Variance 93
Sampling Distributions 94
Sampling Bias 97
Exponential Distributions 98
Exercises 99
Glossary 100

9. Hypothesis Testing **101**
Classical Hypothesis Testing 101
HypothesisTest 102
Testing a Difference in Means 104
Other Test Statistics 105
Testing a Correlation 107
Testing Proportions 108
Chi-Squared Tests 109
First Babies Again 110
Errors 111
Power 112
Replication 113
Exercises 114
Glossary 114

10. Linear Least Squares **117**
Least Squares Fit 117
Implementation 118
Residuals 119
Estimation 120
Goodness of Fit 122

Testing a Linear Model 124
Weighted Resampling 126
Exercises 127
Glossary 128

11. Regression..... 129
StatsModels 130
Multiple Regression 131
Nonlinear Relationships 133
Data Mining 134
Prediction 135
Logistic Regression 137
Estimating Parameters 139
Implementation 140
Accuracy 141
Exercises 142
Glossary 143

12. Time Series Analysis..... 145
Importing and Cleaning 145
Plotting 147
Linear Regression 148
Moving Averages 151
Missing Values 153
Serial Correlation 153
Autocorrelation 155
Prediction 157
Further Reading 161
Exercises 161
Glossary 162

13. Survival Analysis..... 165
Survival Curves 165
Hazard Function 167
Estimating Survival Curves 168
Kaplan-Meier Estimation 169
The Marriage Curve 170
Estimating the Survival Function 171
Confidence Intervals 172
Cohort Effects 173
Extrapolation 176
Expected Remaining Lifetime 178

Exercises 180
Glossary 181

14. Analytic Methods..... 183
Normal Distributions 183
Sampling Distributions 184
Representing Normal Distributions 185
Central Limit Theorem 186
Testing the CLT 187
Applying the CLT 190
Correlation Test 191
Chi-Squared Test 193
Discussion 194
Exercises 195

Index..... 197