

SECOND EDITION

---

# Practical Statistics for Data Scientists

*50+ Essential Concepts Using R and Python*

*Peter Bruce, Andrew Bruce, and Peter Gedeck*

Beijing • Boston • Farnham • Sebastopol • Tokyo

**O'REILLY®**

---

# Table of Contents

|  |          |
|--|----------|
| <b>Preface</b> .....                                       | xiii     |
| <b>1. Exploratory Data Analysis</b> .....                  | <b>1</b> |
| Elements of Structured Data                                | 2        |
| Further Reading  | 4        |
| Rectangular Data   | 4        |
| Data Frames and Indexes                                    | 6        |
| Nonrectangular Data Structures                             | 6        |
| Further Reading  | 7        |
| Estimates of Location                                      | 7        |
| 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                   | 14       |
| Estimates Based on Percentiles                             | 16       |
| Example: Variability Estimates of State Population         | 18       |
| Further Reading  | 19       |
| Exploring the Data Distribution                            | 19       |
| Percentiles and Boxplots                                   | 20       |
| Frequency Tables and Histograms                            | 22       |
| Density Plots and Estimates                                | 24       |
| Further Reading  | 26       |
| Exploring Binary and Categorical Data                      | 27       |
| Mode   | 29       |
| Expected Value   | 29       |
| Probability  | 30       |

|   |           |
|---|-----------|
| Further Reading   | 30        |
| Correlation   | 30        |
| Scatterplots  | 34        |
| Further Reading   | 36        |
| Exploring Two or More Variables                                       | 36        |
| Hexagonal Binning and Contours (Plotting Numeric Versus Numeric Data) | 36        |
| Two Categorical Variables   | 39        |
| Categorical and Numeric Data  | 41        |
| Visualizing Multiple Variables  | 43        |
| Further Reading   | 46        |
| Summary   | 46        |
| <b>2. Data and Sampling Distributions. ....</b>                       | <b>47</b> |
| Random Sampling and Sample Bias                                       | 48        |
| Bias  | 50        |
| Random Selection  | 51        |
| Size Versus Quality: When Does Size Matter?                           | 52        |
| Sample Mean Versus Population Mean                                    | 53        |
| Further Reading   | 53        |
| Selection Bias  | 54        |
| Regression to the Mean  | 55        |
| Further Reading   | 57        |
| Sampling Distribution of a Statistic                                  | 57        |
| Central Limit Theorem   | 60        |
| Standard Error  | 60        |
| Further Reading   | 61        |
| The Bootstrap   | 61        |
| Resampling Versus Bootstrapping                                       | 65        |
| Further Reading   | 65        |
| Confidence Intervals  | 65        |
| Further Reading   | 68        |
| Normal Distribution   | 69        |
| Standard Normal and QQ-Plots  | 71        |
| Long-Tailed Distributions   | 73        |
| Further Reading   | 75        |
| Student's t-Distribution  | 75        |
| Further Reading   | 78        |
| Binomial Distribution   | 78        |
| Further Reading   | 80        |
| Chi-Square Distribution   | 80        |
| Further Reading   | 81        |
| F-Distribution  | 82        |

|   |           |
|---|-----------|
| Further Reading   | 82        |
| Poisson and Related Distributions                               | 82        |
| Poisson Distributions   | 83        |
| Exponential Distribution  | 84        |
| Estimating the Failure Rate                                     | 84        |
| Weibull Distribution  | 85        |
| Further Reading   | 86        |
| Summary   | 86        |
| <b>3. Statistical Experiments and Significance Testing.....</b> | <b>87</b> |
| A/B Testing   | 88        |
| Why Have a Control Group?                                       | 90        |
| Why Just A/B? Why Not C, D,...?                                 | 91        |
| Further Reading   | 92        |
| Hypothesis Tests  | 93        |
| The Null Hypothesis   | 94        |
| Alternative Hypothesis  | 95        |
| One-Way Versus Two-Way Hypothesis Tests                         | 95        |
| Further Reading   | 96        |
| Resampling  | 96        |
| Permutation Test  | 97        |
| Example: Web Stickiness   | 98        |
| Exhaustive and Bootstrap Permutation Tests                      | 102       |
| Permutation Tests: The Bottom Line for Data Science             | 102       |
| Further Reading   | 103       |
| Statistical Significance and p-Values                           | 103       |
| p-Value   | 106       |
| Alpha   | 107       |
| Type 1 and Type 2 Errors  | 109       |
| Data Science and p-Values                                       | 109       |
| Further Reading   | 110       |
| t-Tests   | 110       |
| Further Reading   | 112       |
| Multiple Testing  | 112       |
| Further Reading   | 116       |
| Degrees of Freedom  | 116       |
| Further Reading   | 118       |
| ANOVA   | 118       |
| F-Statistic   | 121       |
| Two-Way ANOVA   | 123       |
| Further Reading   | 124       |
| Chi-Square Test   | 124       |

|  |            |
|--|------------|
| Chi-Square Test: A Resampling Approach                   | 124        |
| Chi-Square Test: Statistical Theory                      | 127        |
| Fisher's Exact Test                                      | 128        |
| Relevance for Data Science                               | 130        |
| Further Reading  | 131        |
| Multi-Arm Bandit Algorithm                               | 131        |
| Further Reading  | 134        |
| Power and Sample Size                                    | 135        |
| Sample Size  | 136        |
| Further Reading  | 138        |
| Summary  | 139        |
| <b>4. Regression and Prediction.....</b>                 | <b>141</b> |
| Simple Linear Regression                                 | 141        |
| The Regression Equation                                  | 143        |
| Fitted Values and Residuals                              | 146        |
| Least Squares  | 148        |
| Prediction Versus Explanation (Profiling)                | 149        |
| Further Reading  | 150        |
| Multiple Linear Regression                               | 150        |
| Example: King County Housing Data                        | 151        |
| Assessing the Model                                      | 153        |
| Cross-Validation   | 155        |
| Model Selection and Stepwise Regression                  | 156        |
| Weighted Regression                                      | 159        |
| Further Reading  | 161        |
| Prediction Using Regression                              | 161        |
| The Dangers of Extrapolation                             | 161        |
| Confidence and Prediction Intervals                      | 161        |
| Factor Variables in Regression                           | 163        |
| Dummy Variables Representation                           | 164        |
| Factor Variables with Many Levels                        | 167        |
| Ordered Factor Variables                                 | 169        |
| Interpreting the Regression Equation                     | 169        |
| Correlated Predictors                                    | 170        |
| Multicollinearity  | 172        |
| Confounding Variables                                    | 172        |
| Interactions and Main Effects                            | 174        |
| Regression Diagnostics                                   | 176        |
| Outliers   | 177        |
| Influential Values                                       | 179        |
| Heteroskedasticity, Non-Normality, and Correlated Errors | 182        |

|  |            |
|--|------------|
| Partial Residual Plots and Nonlinearity                      | 185        |
| Polynomial and Spline Regression                             | 187        |
| Polynomial   | 188        |
| Splines  | 189        |
| Generalized Additive Models                                  | 192        |
| Further Reading  | 193        |
| Summary  | 194        |
| <b>5. Classification.....</b>                                | <b>195</b> |
| Naive Bayes  | 196        |
| Why Exact Bayesian Classification Is Impractical             | 197        |
| The Naive Solution   | 198        |
| Numeric Predictor Variables                                  | 200        |
| Further Reading  | 201        |
| Discriminant Analysis  | 201        |
| Covariance Matrix  | 202        |
| Fisher's Linear Discriminant                                 | 203        |
| A Simple Example   | 204        |
| Further Reading  | 207        |
| Logistic Regression  | 208        |
| Logistic Response Function and Logit                         | 208        |
| Logistic Regression and the GLM                              | 210        |
| Generalized Linear Models                                    | 212        |
| Predicted Values from Logistic Regression                    | 212        |
| Interpreting the Coefficients and Odds Ratios                | 213        |
| Linear and Logistic Regression: Similarities and Differences | 214        |
| Assessing the Model  | 216        |
| Further Reading  | 219        |
| Evaluating Classification Models                             | 219        |
| Confusion Matrix   | 221        |
| The Rare Class Problem                                       | 223        |
| Precision, Recall, and Specificity                           | 223        |
| ROC Curve  | 224        |
| AUC  | 226        |
| Lift   | 228        |
| Further Reading  | 229        |
| Strategies for Imbalanced Data                               | 230        |
| Undersampling  | 231        |
| Oversampling and Up/Down Weighting                           | 232        |
| Data Generation  | 233        |
| Cost-Based Classification                                    | 234        |
| Exploring the Predictions                                    | 234        |

|   |            |
|---|------------|
| Further Reading                           | 236        |
| Summary                                   | 236        |
| <b>6. Statistical Machine Learning</b>    | <b>237</b> |
| K-Nearest Neighbors                       | 238        |
| A Small Example: Predicting Loan Default  | 239        |
| Distance Metrics                          | 241        |
| One Hot Encoder                           | 242        |
| Standardization (Normalization, z-Scores) | 243        |
| Choosing K                                | 246        |
| KNN as a Feature Engine                   | 247        |
| Tree Models                               | 249        |
| A Simple Example                          | 250        |
| The Recursive Partitioning Algorithm      | 252        |
| Measuring Homogeneity or Impurity         | 254        |
| Stopping the Tree from Growing            | 256        |
| Predicting a Continuous Value             | 257        |
| How Trees Are Used                        | 258        |
| Further Reading                           | 259        |
| Bagging and the Random Forest             | 259        |
| Bagging                                   | 260        |
| Random Forest                             | 261        |
| Variable Importance                       | 265        |
| Hyperparameters                           | 269        |
| Boosting                                  | 270        |
| The Boosting Algorithm                    | 271        |
| XGBoost                                   | 272        |
| Regularization: Avoiding Overfitting      | 274        |
| Hyperparameters and Cross-Validation      | 279        |
| Summary                                   | 282        |
| <b>7. Unsupervised Learning</b>           | <b>283</b> |
| Principal Components Analysis             | 284        |
| A Simple Example                          | 285        |
| Computing the Principal Components        | 288        |
| Interpreting Principal Components         | 289        |
| Correspondence Analysis                   | 292        |
| Further Reading                           | 294        |
| K-Means Clustering                        | 294        |
| A Simple Example                          | 295        |
| K-Means Algorithm                         | 298        |
| Interpreting the Clusters                 | 299        |

|                                       |            |
|---------------------------------------|------------|
| Selecting the Number of Clusters      | 302        |
| Hierarchical Clustering               | 304        |
| A Simple Example                      | 305        |
| The Dendrogram                        | 306        |
| The Agglomerative Algorithm           | 308        |
| Measures of Dissimilarity             | 309        |
| Model-Based Clustering                | 311        |
| Multivariate Normal Distribution      | 311        |
| Mixtures of Normals                   | 312        |
| Selecting the Number of Clusters      | 315        |
| Further Reading                       | 318        |
| Scaling and Categorical Variables     | 318        |
| Scaling the Variables                 | 319        |
| Dominant Variables                    | 321        |
| Categorical Data and Gower's Distance | 322        |
| Problems with Clustering Mixed Data   | 325        |
| Summary                               | 326        |
| <b>Bibliography.....</b>              | <b>327</b> |
| <b>Index.....</b>                     | <b>329</b> |