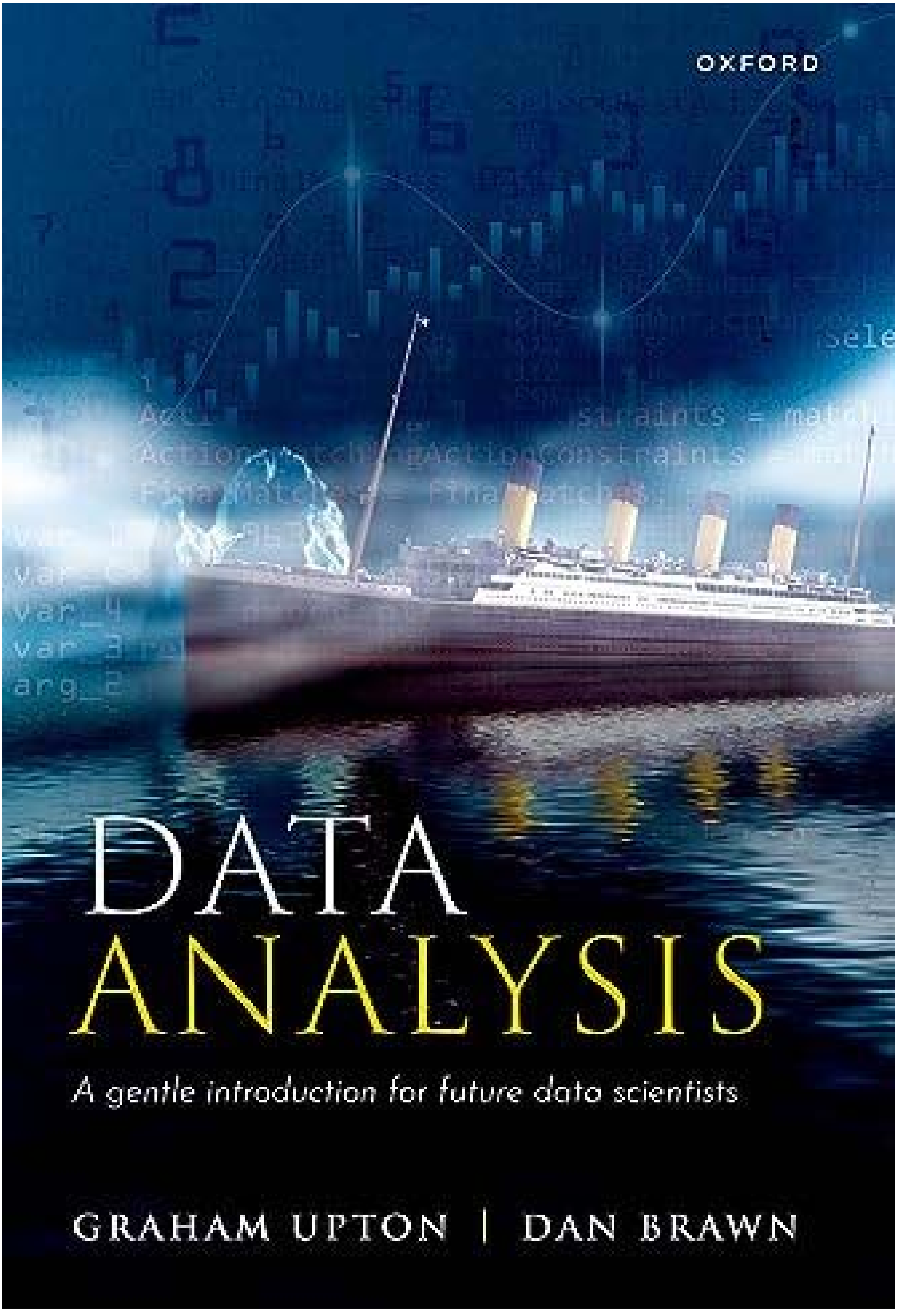

OXFORD

DATA ANALYSIS

A gentle introduction for future data scientists

GRAHAM UPTON | DAN BRAWN

Data Analysis

A gentle introduction for future data scientists

Graham Upton
Former Professor of Applied Statistics, University of Essex

and

Dan Brawn
Lecturer, Department of Mathematical Sciences, University of Essex

OXFORD
UNIVERSITY PRESS

Contents

Preface ix

1. First steps **1**

1.1 Types of data 1

1.2 Sample and population 2

1.2.1 Observations and random variables 3

1.2.2 Sampling variation 3

1.3 Methods for sampling a population 4

1.3.1 The simple random sample 5

1.3.2 Cluster sampling 5

1.3.3 Stratified sampling 6

1.3.4 Systematic sampling 7

1.4 Oversampling and the use of weights 7

2. Summarizing data **9**

2.1 Measures of location 9

2.1.1 The mode 9

2.1.2 The mean 10

2.1.3 The trimmed mean 11

2.1.4 The Winsorized mean 11

2.1.5 The median 12

2.2 Measures of spread 12

2.2.1 The range 12

2.2.2 The interquartile range 12

2.3 Boxplot 13

2.4 Histograms 14

2.5 Cumulative frequency diagrams 16

2.6 Step diagrams 17

2.7 The variance and standard deviation 18

2.8 Symmetric and skewed data 20

3. Probability **21**

3.1 Probability 21

3.2 The rules of probability 24

3.3 Conditional probability and independence 25

3.4 The total probability theorem 28

3.5 Bayes’ theorem 30

4. **Probability distributions** 33
4.1 Notation 33
4.2 Mean and variance of a probability distribution 33
4.3 The relation between sample and population 35
4.4 Combining means and variances 37
4.5 Discrete uniform distribution 39
4.6 Probability density function 39
4.7 The continuous uniform distribution 41

5. **Estimation and confidence** 43
5.1 Point estimates 43
5.1.1 Maximum likelihood estimation (mle) 43
5.2 Confidence intervals 43
5.3 Confidence interval for the population mean 44
5.3.1 The normal distribution 44
5.3.2 The Central Limit Theorem 45
5.3.3 Construction of the confidence interval 48
5.4 Confidence interval for a proportion 48
5.4.1 The binomial distribution 49
5.4.2 Confidence interval for a proportion (large sample case) 50
5.4.3 Confidence interval for a proportion (small sample) 50
5.5 Confidence bounds for other summary statistics 50
5.5.1 The bootstrap 50
5.6 Some other probability distributions 52
5.6.1 The Poisson and exponential distributions 52
5.6.2 The Weibull distribution 55
5.6.3 The chi-squared (χ^2) distribution 56

6. **Models, p-values, and hypotheses** 59
6.1 Models 59
6.2 p-values and the null hypothesis 60
6.2.1 Two-sided or one-sided? 60
6.2.2 Interpreting p-values 61
6.2.3 Comparing p-values 62
6.2.4 Link with confidence interval 63
6.3 p-values when comparing two samples 63
6.3.1 Do the two samples come from the same population? 63
6.3.2 Do the two populations have the same mean? 65

7. **Comparing proportions** 67
7.1 The 2 × 2 table 67
7.2 Some terminology 69
7.2.1 Odds, odds ratios, and independence 70

7.2.2 Relative risk 70
7.2.3 Sensitivity, specificity, and related quantities 71
7.3 The $R \times C$ table 72
7.3.1 Residuals 73
7.3.2 Partitioning 74

8. Relations between two continuous variables 77
8.1 Scatter diagrams 78
8.2 Correlation 79
8.2.1 Testing for independence 81
8.3 The equation of a line 84
8.4 The method of least squares 84
8.5 A random dependent variable, Y 87
8.5.1 Estimation of σ^2 88
8.5.2 Confidence interval for the regression line 88
8.5.3 Prediction interval for future values 88
8.6 Departures from linearity 89
8.6.1 Transformations 89
8.6.2 Extrapolation 90
8.6.3 Outliers 91
8.7 Distinguishing x and Y 93
8.8 Why 'regression'? 93

9. Several explanatory variables 97
9.1 *AIC* and related measures 98
9.2 Multiple regression 99
9.2.1 Two variables 99
9.2.2 Collinearity 101
9.2.3 Using a dummy variable 102
9.2.4 The use of multiple dummy variables 104
9.2.5 Model selection 107
9.2.6 Interactions 107
9.2.7 Residuals 109
9.3 Cross-validation 110
9.3.1 k-fold cross-validation 110
9.3.2 Leave-one-out cross-validation (LOOCV) 112
9.4 Reconciling bias and variability 112
9.5 Shrinkage 113
9.5.1 Standardization 114
9.6 Generalized linear models (GLMs) 115
9.6.1 Logistic regression 116
9.6.2 Loglinear models 118

10. Classification 121
10.1 Naive Bayes classification 121
10.2 Classification using logistic regression 124

10.3 Classification trees 125
10.4 The random forest classifier 127
10.5 *k*-nearest neighbours (*k*NN) 128
10.6 Support-vector machines 131
10.7 Ensemble approaches 133
10.8 Combining variables 134

11. Last words 135
Further reading 138

Index 139