Exploring Data

An Introduction to Data Analysis for Social Scientists

Second Edition

Catherine Marsh and
Jane Elliott

polity
Detailed contents

List of figures xiii
Acknowledgements xxi

Introduction 1

Part I Single Variables

1 Distributions and Variables 7
1.1 Preliminaries 7
1.2 Variables on a household survey 8
1.3 Reducing the number of digits 12
1.4 Bar charts and pie charts 13
1.5 Histograms 13
1.6 From interval level to ordinal level variables – recoding 18
1.7 Household level data 19
1.8 Using SPSS to produce bar charts, pie charts and histograms 24
1.9 Conclusion 32
Exercises 32

2 Numerical Summaries of Level and Spread 35
2.1 Working hours of couples in Britain 35
2.2 Summaries of level 37
2.3 Summaries of spread 40
2.4 Other locational statistics 43
2.5 Choosing between measures 45
2.6 Obtaining summaries of level and spread using SPSS and Excel 47
2.7 Individual and aggregate level data 52
2.8 Hours worked by age 53
2.9 Comparing working hours across Europe: the Working Time Directive 54
2.10 Conclusion 55
Exercises 55
### Detailed contents

#### 3 Scaling and Standardizing
- 3.1 Data are produced not given
- 3.2 Adding or subtracting a constant
- 3.3 Multiplying or dividing by a constant
- 3.4 Standardized variables
- 3.5 The Gaussian distribution
- 3.6 Standardizing with respect to an appropriate base
- 3.7 Has the Equal Pay Act worked?
- 3.8 Conclusion

#### Exercises

#### 4 Inequality
- 4.1 Prosperity and inequality
- 4.2 Income and wealth
- 4.3 Measuring inequality: quantiles and quantile shares
- 4.4 Cumulative income shares and Lorenz curves
- 4.5 Desirable properties in a summary measure of inequality
- 4.6 The Gini coefficient
- 4.7 Alternative measures of inequality
- 4.8 Conclusion

#### Exercises

#### 5 Smoothing Time Series
- 5.1 Time series
- 5.2 Smoothing
- 5.3 The aim of smoothing
- 5.4 Tough on crime, tough on the causes of crime?
- 5.5 Opinion polls
- 5.6 Techniques
- 5.7 Residuals
- 5.8 Refinements
- 5.9 Conclusion

#### Exercises

#### Part II Relationships between Two Variables

#### 6 Percentage Tables
- 6.1 Introduction
- 6.2 Higher education: a question of access?
- 6.3 Proportions, percentages and probabilities
- 6.4 NS-SEC: a new measure of social class?
- 6.5 Contingency tables
- 6.6 Percentage tables
- 6.7 Good table manners
- 6.8 Producing contingency tables using SPSS
- 6.9 Simplifying tables by recoding variables and collapsing categories
- 6.10 Conclusion

#### Exercises
7 Analysing Contingency Tables
7.1 Which way should proportions run? 141
7.2 The base for comparison 142
7.3 Summarizing effects by subtracting proportions 144
7.4 Properties of $d$ as a measure of effect 146
7.5 How large a difference in proportions is a 'significant' difference? 147
7.6 Calculating the chi-square statistic – innocent until proven guilty? 148
7.7 Calculating chi-square using SPSS 156
7.8 Conclusion 157
Exercises 158

8 Handling Several Batches
8.1 Unemployment – an issue of concern 161
8.2 Counting the unemployed 162
8.3 Regional variation in unemployment rates 163
8.4 Boxplots 167
8.5 Outliers 168
8.6 Multiple boxplots 171
8.7 Decomposing the variation in unemployment rates 172
8.8 Moving to the individual as a unit of analysis and using a statistical test 174
8.9 The T-test 176
8.10 Obtaining multiple boxplots using SPSS 182
8.11 One-way analysis of variance 183
8.12 Using SPSS to carry out a one-way analysis of variance 186
8.13 Conclusion 188
Exercises 189

9 Scatterplots and Resistant Lines
9.1 Introduction: lone parents and deprivation 191
9.2 Scatterplots 192
9.3 Lone parents 193
9.4 Linear relationships 194
9.5 Where to draw the line? 195
9.6 Fitting a resistant line 196
9.7 Using SPSS to fit a resistant line 199
9.8 Inspection of residuals 204
9.9 Further examples 206
9.10 Conclusion 209
Exercises 210

10 Transformations
10.1 The Millennium Development Goals 212
10.2 The distribution of GNI per capita 215
10.3 The log transformation 217
10.4 The ladder of powers 219
### Detailed contents

10.5 The goals of transformation .......................... 222
10.6 Promoting equality of spread ......................... 223
10.7 Alternatives to GNI as a measure of welfare ....... 224
10.8 Unbending curvy lines .................................. 225
10.9 Determining the best power for transformation .... 227
10.10 Conclusion .................................................. 228
Exercises ......................................................... 229

Part III  Introducing a Third Variable

11  Causal Explanations ................................. 235
11.1 Why did the chicken cross the road? ............... 236
11.2 Direct and indirect effects ........................... 238
11.3 Controlling the world to learn about causes ....... 239
11.4 Do opinion polls influence people? ................. 241
11.5 Assumptions required to infer causes .............. 242
11.6 Controlling for prior variables ...................... 244
11.7 Controlling for intervening variables ............... 248
11.8 Positive and negative relationships ................. 250
11.9 Conclusion .................................................. 251
Exercises ......................................................... 252

12  Three-Variable Contingency Tables and Beyond .... 254
12.1 Trust and reciprocity .................................... 254
12.2 Controlling for a prior variable ..................... 255
12.3 Causal path models for three variables ............ 258
12.4 More complex models: going beyond three variables 262
12.5 Logistic regression models ........................... 262
12.6 Estimating a more complex model ................. 266
12.7 Interpreting the coefficients in a logistic regression model 268
12.8 Estimating a more complex model – interactions 269
12.9 Age and cohort effects .................................. 273
12.10 Conclusion ................................................ 273
Exercises ......................................................... 274

13  Longitudinal Data ........................................ 277
13.1 Introduction .............................................. 277
13.2 Collecting longitudinal data .......................... 278
13.3 Examining change over time .......................... 282
13.4 Transition tables ......................................... 284
13.5 Poverty dynamics ......................................... 286
13.6 Approaches to the analysis of longitudinal data .... 287
13.7 Event history modelling ............................... 288
13.8 Causality in cross-sectional and longitudinal research 290
13.9 Conclusion ................................................ 290
Exercises ......................................................... 291

References ..................................................... 293
Index ............................................................ 299