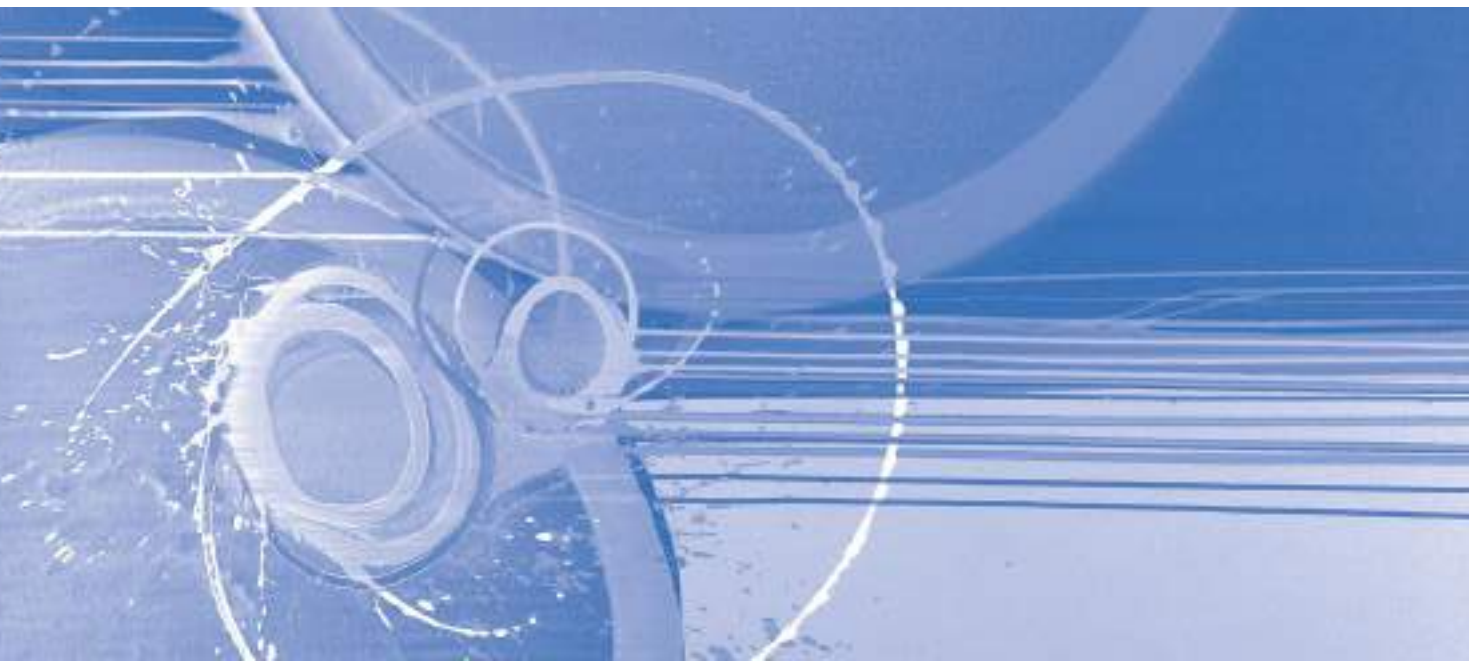
DISCOVERING STATISTICS USING R



ANDY FIELD | JEREMY MILES | ZOË FIELD

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

CONTENTS

Preface xxi
How to use this book xxv
Acknowledgements xxix
Dedication xxxi
Symbols used in this book xxxii
Some maths revision xxxiv

1 Why is my evil lecturer forcing me to learn statistics? 1

1.1. What will this chapter tell me? ① 1
1.2. What the hell am I doing here? I don't belong here ① 2
1.3. Initial observation: finding something that needs explaining ① 4
1.4. Generating theories and testing them ① 4
1.5. Data collection 1: what to measure ① 7
1.5.1. Variables ① 7
1.5.2. Measurement error ① 11
1.5.3. Validity and reliability ① 12
1.6. Data collection 2: how to measure ① 13
1.6.1. Correlational research methods ① 13
1.6.2. Experimental research methods ① 13
1.6.3. Randomization ① 17
1.7. Analysing data ① 19
1.7.1. Frequency distributions ① 19
1.7.2. The centre of a distribution ① 21
1.7.3. The dispersion in a distribution ① 24
1.7.4. Using a frequency distribution to go beyond the data ① 25
1.7.5. Fitting statistical models to the data ① 28
What have I discovered about statistics? ① 29
Key terms that I've discovered 29
Smart Alex's tasks 30
Further reading 31
Interesting real research 31

2 Everything you ever wanted to know about statistics (well, sort of) 32

2.1. What will this chapter tell me? ① 32
2.2. Building statistical models ① 33

2.3. Populations and samples ① 36
2.4. Simple statistical models ① 36
2.4.1. The mean: a very simple statistical model ① 36
2.4.2. Assessing the fit of the mean: sums of squares, variance and standard deviations ① 37
2.4.3. Expressing the mean as a model ② 40
2.5. Going beyond the data ① 41
2.5.1. The standard error ① 42
2.5.2. Confidence intervals ② 43
2.6. Using statistical models to test research questions ① 49
2.6.1. Test statistics ① 53
2.6.2. One- and two-tailed tests ① 55
2.6.3. Type I and Type II errors ① 56
2.6.4. Effect sizes ② 57
2.6.5. Statistical power ② 58
What have I discovered about statistics? ① 59
Key terms that I've discovered 60
Smart Alex's tasks 60
Further reading 60
Interesting real research 61

3 The R environment 62

3.1. What will this chapter tell me? ① 62
3.2. Before you start ① 63
3.2.1. The R-chitecture ① 63
3.2.2. Pros and cons of R ① 64
3.2.3. Downloading and installing R ① 65
3.2.4. Versions of R ① 66
3.3. Getting started ① 66
3.3.1. The main windows in R ① 67
3.3.2. Menus in R ① 67
3.4. Using R ① 71
3.4.1. Commands, objects and functions ① 71
3.4.2. Using scripts ① 75
3.4.3. The R workspace ① 76
3.4.4. Setting a working directory ② 77
3.4.5. Installing packages ① 78
3.4.6. Getting help ① 80
3.5. Getting data into R ① 81
3.5.1. Creating variables ① 81
3.5.2. Creating dataframes ① 81
3.5.3. Calculating new variables from exisiting ones ① 83
3.5.4. Organizing your data ① 85
3.5.5. Missing values ① 92
3.6. Entering data with R Commander ① 92
3.6.1. Creating variables and entering data with R Commander ① 94
3.6.2. Creating coding variables with R Commander ① 95
3.7. Using other software to enter and edit data ① 95
3.7.1. Importing data ① 97
3.7.2. Importing SPSS data files directly ① 99

3.7.3. Importing data with R Commander ① 101
3.7.4. Things that can go wrong ① 102
3.8. Saving data ① 103
3.9. Manipulating data ③ 103
3.9.1. Selecting parts of a dataframe ② 103
3.9.2. Selecting data with the subset() function ② 105
3.9.3. Dataframes and matrices ② 106
3.9.4. Reshaping data ③ 107
What have I discovered about statistics? ① 113
R packages used in this chapter 113
R functions used in this chapter 113
Key terms that I've discovered 114
Smart Alex's tasks 114
Further reading 115

4 Exploring data with graphs 116

4.1. What will this chapter tell me? ① 116
4.2. The art of presenting data ① 117
4.2.1. Why do we need graphs ① 117
4.2.2. What makes a good graph? ① 117
4.2.3. Lies, damned lies, and … erm … graphs ① 120
4.3. Packages used in this chapter ① 121
4.4. Introducing ggplot2 ① 121
4.4.1. The anatomy of a plot ① 121
4.3.2. Geometric objects (geoms) ① 123
4.4.3. Aesthetics ① 125
4.4.4. The anatomy of the ggplot() function ① 127
4.4.5. Stats and geoms ③ 128
4.4.6. Avoiding overplotting ② 130
4.4.7. Saving graphs ① 131
4.4.8. Putting it all together: a quick tutorial ② 132
4.5. Graphing relationships: the scatterplot ① 136
4.5.1. Simple scatterplot ① 136
4.5.2. Adding a funky line ① 138
4.5.3. Grouped scatterplot ① 140
4.6. Histograms: a good way to spot obvious problems ① 142
4.7. Boxplots (box–whisker diagrams) ① 144
4.8. Density plots ① 148
4.9. Graphing means ③ 149
4.9.1. Bar charts and error bars ② 149
4.9.2. Line graphs ② 155
4.10. Themes and options ① 161
What have I discovered about statistics? ① 163
R packages used in this chapter 163
R functions used in this chapter 164
Key terms that I've discovered 164
Smart Alex's tasks 164
Further reading 164
Interesting real research 165

5 Exploring assumptions 166

5.1. What will this chapter tell me? ① 166
5.2. What are assumptions? ① 167
5.3. Assumptions of parametric data ① 167
5.4. Packages used in this chapter ① 169
5.5. The assumption of normality ① 169
5.5.1. Oh no, it's that pesky frequency distribution again: checking normality visually ① 169
5.5.2. Quantifying normality with numbers ① 173
5.5.3. Exploring groups of data ① 177
5.6. Testing whether a distribution is normal ① 182
5.6.1. Doing the Shapiro–Wilk test in R ① 182
5.6.2. Reporting the Shapiro–Wilk test ① 185
5.7. Testing for homogeneity of variance ① 185
5.7.1. Levene's test ① 186
5.7.2. Reporting Levene's test ① 188
5.7.3. Hartley's F_{max}: the variance ratio ① 189
5.8. Correcting problems in the data ② 190
5.8.1. Dealing with outliers ② 190
5.8.2. Dealing with non-normality and unequal variances ② 191
5.8.3. Transforming the data using R ② 194
5.8.4. When it all goes horribly wrong ③ 201
What have I discovered about statistics? ① 203
R packages used in this chapter 204
R functions used in this chapter 204
Key terms that I've discovered 204
Smart Alex's tasks 204
Further reading 204

6 Correlation 205

6.1. What will this chapter tell me? ① 205
6.2. Looking at relationships ① 206
6.3. How do we measure relationships? ① 206
6.3.1. A detour into the murky world of covariance ① 206
6.3.2. Standardization and the correlation coefficient ① 208
6.3.3. The significance of the correlation coefficient ③ 210
6.3.4. Confidence intervals for *r* ③ 211
6.3.5. A word of warning about interpretation: causality ① 212
6.4. Data entry for correlation analysis ① 213
6.5. Bivariate correlation ① 213
6.5.1. Packages for correlation analysis in R ① 214
6.5.2. General procedure for correlations using R Commander ① 214
6.5.3. General procedure for correlations using R ① 216
6.5.4. Pearson's correlation coefficient ① 219
6.5.5. Spearman's correlation coefficient ① 223
6.5.6. Kendall's tau (non-parametric) ① 225
6.5.7. Bootstrapping correlations ③ 226
6.5.8. Biserial and point-biserial correlations ③ 229

6.6. Partial correlation ② 234
6.6.1. The theory behind part and partial correlation ② 234
6.6.2. Partial correlation using R ② 235
6.6.3 Semi-partial (or part) correlations ② 237
6.7. Comparing correlations ③ 238
6.7.1. Comparing independent *r*s ③ 238
6.7.2. Comparing dependent *r*s ③ 239
6.8. Calculating the effect size ① 240
6.9. How to report correlation coefficents ① 240
What have I discovered about statistics? ① 242
R packages used in this chapter 243
R functions used in this chapter 243
Key terms that I've discovered 243
Smart Alex's tasks ① 243
Further reading 244
Interesting real research 244

7 Regression 245

7.1. What will this chapter tell me? ① 245
7.2. An introduction to regression ① 246
7.2.1. Some important information about straight lines ① 247
7.2.2. The method of least squares ① 248
7.2.3. Assessing the goodness of fit: sums of squares, *R* and R^2 ① 249
7.2.4. Assessing individual predictors ① 252
7.3. Packages used in this chapter ① 253
7.4. General procedure for regression in R ① 254
7.4.1. Doing simple regression using R Commander ① 254
7.4.2. Regression in R ① 255
7.5. Interpreting a simple regression ① 257
7.5.1. Overall fit of the object model ① 258
7.5.2. Model parameters ① 259
7.5.3. Using the model ① 260
7.6. Multiple regression: the basics ② 261
7.6.1. An example of a multiple regression model ② 261
7.6.2. Sums of squares, *R* and R^2 ② 262
7.6.3. Parsimony-adjusted measures of fit ② 263
7.6.4. Methods of regression ② 263
7.7. How accurate is my regression model? ② 266
7.7.1. Assessing the regression model I: diagnostics ② 266
7.7.2. Assessing the regression model II: generalization ② 271
7.8. How to do multiple regression using R Commander and R ② 276
7.8.1. Some things to think about before the analysis ② 276
7.8.2. Multiple regression: running the basic model ② 277
7.8.3. Interpreting the basic multiple regression ② 280
7.8.4. Comparing models ② 284
7.9. Testing the accuracy of your regression model ② 287
7.9.1. Diagnostic tests using R Commander ② 287
7.9.2. Outliers and influential cases ② 288

7.9.3. Assessing the assumption of independence ② 291
7.9.4. Assessing the assumption of no multicollinearity ② 292
7.9.5. Checking assumptions about the residuals ② 294
7.9.6. What if I violate an assumption? ② 298
7.10. Robust regression: bootstrapping ③ 298
7.11. How to report multiple regression ② 301
7.12. Categorical predictors and multiple regression ③ 302
7.12.1. Dummy coding ③ 302
7.12.2. Regression with dummy variables ③ 305
What have I discovered about statistics? ① 308
R packages used in this chapter 309
R functions used in this chapter 309
Key terms that I've discovered 309
Smart Alex's tasks 310
Further reading 311
Interesting real research 311

8 Logistic regression 312

8.1. What will this chapter tell me? ① 312
8.2. Background to logistic regression ① 313
8.3. What are the principles behind logistic regression? ③ 313
8.3.1. Assessing the model: the log-likelihood statistic ③ 315
8.3.2. Assessing the model: the deviance statistic ③ 316
8.3.3. Assessing the model: R and R^2 ③ 316
8.3.4. Assessing the model: information criteria ③ 318
8.3.5. Assessing the contribution of predictors: the z-statistic ② 318
8.3.6. The odds ratio ③ 319
8.3.7. Methods of logistic regression ② 320
8.4. Assumptions and things that can go wrong ④ 321
8.4.1. Assumptions ② 321
8.4.2. Incomplete information from the predictors ④ 322
8.4.3. Complete separation ④ 323
8.5. Packages used in this chapter ① 325
8.6. Binary logistic regression: an example that will make you feel eel ② 325
8.6.1. Preparing the data 326
8.6.2. The main logistic regression analysis ② 327
8.6.3. Basic logistic regression analysis using R ② 329
8.6.4. Interpreting a basic logistic regression ② 330
8.6.5. Model 1: Intervention only ② 330
8.6.6. Model 2: Intervention and Duration as predictors ② 336
8.6.7. Casewise diagnostics in logistic regression ② 338
8.6.8. Calculating the effect size ② 341
8.7. How to report logistic regression ② 341
8.8. Testing assumptions: another example ② 342
8.8.1. Testing for multicollinearity ③ 343
8.8.2. Testing for linearity of the logit ③ 344
8.9. Predicting several categories: multinomial logistic regression ③ 346
8.9.1. Running multinomial logistic regression in R ③ 347
8.9.2. Interpreting the multinomial logistic regression output ③ 350

8.9.3. Reporting the results 355
What have I discovered about statistics? ① 355
R packages used in this chapter 356
R functions used in this chapter 356
Key terms that I've discovered 356
Smart Alex's tasks 357
Further reading 358
Interesting real research 358

9 Comparing two means 359

9.1. What will this chapter tell me? ① 359
9.2. Packages used in this chapter ① 360
9.3. Looking at differences ① 360
9.3.1. A problem with error bar graphs of repeated-measures designs ① 361
9.3.2. Step 1: calculate the mean for each participant ② 364
9.3.3. Step 2: calculate the grand mean ② 364
9.3.4. Step 3: calculate the adjustment factor ② 364
9.3.5. Step 4: create adjusted values for each variable ② 365
9.4. The *t*-test ① 368
9.4.1. Rationale for the *t*-test ① 369
9.4.2. The *t*-test as a general linear model ② 370
9.4.3. Assumptions of the *t*-test ① 372
9.5. The independent *t*-test ① 372
9.5.1. The independent *t*-test equation explained ① 372
9.5.2. Doing the independent *t*-test ① 375
9.6. The dependent *t*-test ① 386
9.6.1. Sampling distributions and the standard error ① 386
9.6.2. The dependent *t*-test equation explained ① 387
9.6.3. Dependent *t*-tests using *R* ① 388
9.7. Between groups or repeated measures? ① 394
What have I discovered about statistics? ① 395
R packages used in this chapter 396
R functions used in this chapter 396
Key terms that I've discovered 396
Smart Alex's tasks 396
Further reading 397
Interesting real research 397

10 Comparing several means: ANOVA (GLM 1) 398

10.1. What will this chapter tell me? ① 398
10.2. The theory behind ANOVA ② 399
10.2.1 Inflated error rates ② 399
10.2.2. Interpreting *F* ② 400
10.2.3. ANOVA as regression ② 400
10.2.4. Logic of the *F*-ratio ② 405
10.2.5. Total sum of squares (SS_T) ② 407
10.2.6. Model sum of squares (SS_M) ② 409
10.2.7. Residual sum of squares (SS_R) ② 410
10.2.8. Mean squares ② 411

10.2.9. The *F*-ratio ② 411
10.3. Assumptions of ANOVA ③ 412
10.3.1. Homogeneity of variance ② 412
10.3.2. Is ANOVA robust? ③ 412
10.4. Planned contrasts ② 414
10.4.1. Choosing which contrasts to do ② 415
10.4.2. Defining contrasts using weights ② 419
10.4.3. Non-orthogonal comparisons ② 425
10.4.4. Standard contrasts ② 426
10.4.5. Polynomial contrasts: trend analysis ② 427
10.5. *Post hoc* procedures ② 428
10.5.1. *Post hoc* procedures and Type I (α) and Type II error rates ② 431
10.5.2. *Post hoc* procedures and violations of test assumptions ② 431
10.5.3. Summary of *post hoc* procedures ② 432
10.6. One-way ANOVA using R ② 432
10.6.1. Packages for one-way ANOVA in R ① 433
10.6.2. General procedure for one-way ANOVA ① 433
10.6.3. Entering data ① 433
10.6.4. One-way ANOVA using R Commander ② 434
10.6.5. Exploring the data ② 436
10.6.6. The main analysis ② 438
10.6.7. Planned contrasts using R ② 443
10.6.8. *Post hoc* tests using R ② 447
10.7. Calculating the effect size ② 454
10.8. Reporting results from one-way independent ANOVA ② 457
What have I discovered about statistics? ① 458
R packages used in this chapter 459
R functions used in this chapter 459
Key terms that I've discovered 459
Smart Alex's tasks 459
Further reading 461
Interesting real research 461

11 Analysis of covariance, ANCOVA (GLM 2) 462

11.1. What will this chapter tell me? ② 462
11.2. What is ANCOVA? ② 463
11.3. Assumptions and issues in ANCOVA ③ 464
11.3.1. Independence of the covariate and treatment effect ③ 464
11.3.2. Homogeneity of regression slopes ③ 466
11.4. ANCOVA using R ② 467
11.4.1. Packages for ANCOVA in R ① 467
11.4.2. General procedure for ANCOVA ① 468
11.4.3. Entering data ① 468
11.4.4. ANCOVA using R Commander ② 471
11.4.5. Exploring the data ② 471
11.4.6. Are the predictor variable and covariate independent? ② 473
11.4.7. Fitting an ANCOVA model ② 473
11.4.8. Interpreting the main ANCOVA model ② 477

11.4.9. Planned contrasts in ANCOVA ② 479
11.4.10. Interpreting the covariate ② 480
11.4.11. *Post hoc* tests in ANCOVA ② 481
11.4.12. Plots in ANCOVA ② 482
11.4.13. Some final remarks ② 482
11.4.14. Testing for homogeneity of regression slopes ③ 483
11.5. Robust ANCOVA ③ 484
11.6. Calculating the effect size ② 491
11.7. Reporting results ② 494
What have I discovered about statistics? ① 495
R packages used in this chapter 495
R functions used in this chapter 496
Key terms that I've discovered 496
Smart Alex's tasks 496
Further reading 497
Interesting real research 497

12 Factorial ANOVA (GLM 3) 498

12.1. What will this chapter tell me? ② 498
12.2. Theory of factorial ANOVA (independent design) ② 499
12.2.1. Factorial designs ② 499
12.3. Factorial ANOVA as regression ③ 501
12.3.1. An example with two independent variables ② 501
12.3.2. Extending the regression model ③ 501
12.4. Two-way ANOVA: behind the scenes ② 505
12.4.1. Total sums of squares (SS_T) ② 506
12.4.2. The model sum of squares (SS_M) ② 507
12.4.3. The residual sum of squares (SS_R) ② 510
12.4.4. The *F*-ratios ② 511
12.5. Factorial ANOVA using R ② 511
12.5.1. Packages for factorial ANOVA in R ① 511
12.5.2. General procedure for factorial ANOVA ① 512
12.5.3. Factorial ANOVA using R Commander ② 512
12.5.4. Entering the data ② 513
12.5.5. Exploring the data ② 516
12.5.6. Choosing contrasts ② 518
12.5.7. Fitting a factorial ANOVA model ② 520
12.5.8. Interpreting factorial ANOVA ② 520
12.5.9. Interpreting contrasts ② 524
12.5.10. Simple effects analysis ③ 525
12.5.11. *Post hoc* analysis ② 528
12.5.12. Overall conclusions 530
12.5.13. Plots in factorial ANOVA ② 530
12.6. Interpreting interaction graphs ② 530
12.7. Robust factorial ANOVA ③ 534
12.8. Calculating effect sizes ③ 542
12.9. Reporting the results of two-way ANOVA ② 544
What have I discovered about statistics? ① 546

R packages used in this chapter 546
R functions used in this chapter 546
Key terms that I've discovered 547
Smart Alex's tasks 547
Further reading 548
Interesting real research 548

13 Repeated-measures designs (GLM 4) 549

13.1. What will this chapter tell me? ② 549
13.2. Introduction to repeated-measures designs ② 550
13.2.1. The assumption of sphericity ② 551
13.2.2. How is sphericity measured? ② 551
13.2.3. Assessing the severity of departures from sphericity ② 552
13.2.4. What is the effect of violating the assumption of sphericity? ③ 552
13.2.5. What do you do if you violate sphericity? ② 554
13.3. Theory of one-way repeated-measures ANOVA ② 554
13.3.1. The total sum of squares (SS_T) ② 557
13.3.2. The within-participant sum of squares (SS_W) ② 558
13.3.3. The model sum of squares (SS_M) ② 559
13.3.4. The residual sum of squares (SS_R) ② 560
13.3.5. The mean squares ② 560
13.3.6. The *F*-ratio ② 560
13.3.7. The between-participant sum of squares ② 561
13.4. One-way repeated-measures designs using R ② 561
13.4.1. Packages for repeated measures designs in R ① 561
13.4.2. General procedure for repeated-measures designs ① 562
13.4.3. Repeated-measures ANOVA using R Commander ② 563
13.4.4. Entering the data ② 563
13.4.5. Exploring the data ② 565
13.4.6. Choosing contrasts ② 568
13.4.7. Analysing repeated measures: two ways to skin a .dat ② 569
13.4.8. Robust one-way repeated-measures ANOVA ③ 576
13.5. Effect sizes for repeated-measures designs ③ 580
13.6. Reporting one-way repeated-measures designs ② 581
13.7. Factorial repeated-measures designs ② 583
13.7.1. Entering the data ② 584
13.7.2. Exploring the data ② 586
13.7.3. Setting contrasts ② 588
13.7.4. Factorial repeated-measures ANOVA ② 589
13.7.5. Factorial repeated-measures designs as a GLM ③ 594
13.7.6. Robust factorial repeated-measures ANOVA ③ 599
13.8. Effect sizes for factorial repeated-measures designs ③ 599
13.9. Reporting the results from factorial repeated-measures designs ② 600
What have I discovered about statistics? ② 601
R packages used in this chapter 602
R functions used in this chapter 602
Key terms that I've discovered 602
Smart Alex's tasks 602

Further reading 603
Interesting real research 603

14 Mixed designs (GLM 5) 604

14.1. What will this chapter tell me? ① 604
14.2. Mixed designs ② 605
14.3. What do men and women look for in a partner? ② 606
14.4. Entering and exploring your data ② 606
14.4.1. Packages for mixed designs in R ① 606
14.4.2. General procedure for mixed designs ① 608
14.4.3. Entering the data ② 608
14.4.4. Exploring the data ② 610
14.5. Mixed ANOVA ② 613
14.6. Mixed designs as a GLM ③ 617
14.6.1. Setting contrasts ② 617
14.6.2. Building the model ② 619
14.6.3. The main effect of gender ② 622
14.6.4. The main effect of looks ② 623
14.6.5. The main effect of **personality** ② 624
14.6.6. The interaction between **gender** and **looks** ② 625
14.6.7. The interaction between **gender** and **personality** ② 628
14.6.8. The interaction between **looks** and **personality** ② 630
14.6.9. The interaction between **looks, personality** and **gender** ③ 635
14.6.10. Conclusions ③ 639
14.7. Calculating effect sizes ③ 640
14.8. Reporting the results of mixed ANOVA ② 641
14.9. Robust analysis for mixed designs ③ 643
What have I discovered about statistics? ② 650
R packages used in this chapter 650
R functions used in this chapter 651
Key terms that I've discovered 651
Smart Alex's tasks 651
Further reading 652
Interesting real research 652

15 Non-parametric tests 653

15.1. What will this chapter tell me? ① 653
15.2. When to use non-parametric tests ① 654
15.3. Packages used in this chapter ① 655
15.4. Comparing two independent conditions: the Wilcoxon rank-sum test ① 655
15.4.1. Theory of the Wilcoxon rank-sum test ② 655
15.4.2. Inputting data and provisional analysis ① 659
15.4.3. Running the analysis using R Commander ① 661
15.4.4. Running the analysis using R ① 662
15.4.5. Output from the Wilcoxon rank-sum test ① 664
15.4.6. Calculating an effect size ② 664
15.4.7. Writing the results ① 666

15.5. Comparing two related conditions: the Wilcoxon signed-rank test ① 667
15.5.1. Theory of the Wilcoxon signed-rank test ② 668
15.5.2. Running the analysis with R Commander ① 670
15.5.3. Running the analysis using R ① 671
15.5.4. Wilcoxon signed-rank test output ① 672
15.5.5. Calculating an effect size ② 673
15.5.6. Writing the results ① 673
15.6. Differences between several independent groups: the Kruskal–Wallis test ① 674
15.6.1. Theory of the Kruskal–Wallis test ② 675
15.6.2. Inputting data and provisional analysis ① 677
15.6.3. Doing the Kruskal–Wallis test using R Commander ① 679
15.6.4. Doing the Kruskal–Wallis test using R ① 679
15.6.5. Output from the Kruskal–Wallis test ① 680
15.6.6. *Post hoc* tests for the Kruskal–Wallis test ② 681
15.6.7. Testing for trends: the Jonckheere–Terpstra test ② 684
15.6.8. Calculating an effect size ② 685
15.6.9. Writing and interpreting the results ① 686
15.7. Differences between several related groups: Friedman's ANOVA ① 686
15.7.1. Theory of Friedman's ANOVA ② 688
15.7.2. Inputting data and provisional analysis ① 689
15.7.3. Doing Friedman's ANOVA in R Commander ① 690
15.7.4. Friedman's ANOVA using R ① 690
15.7.5. Output from Friedman's ANOVA ① 691
15.7.6. *Post hoc* tests for Friedman's ANOVA ② 691
15.7.7. Calculating an effect size ② 692
15.7.8. Writing and interpreting the results ① 692
What have I discovered about statistics? ① 693
R packages used in this chapter 693
R functions used in this chapter 693
Key terms that I've discovered 694
Smart Alex's tasks 694
Further reading 695
Interesting real research 695

16 Multivariate analysis of variance (MANOVA) 696

16.1. What will this chapter tell me? ② 696
16.2. When to use MANOVA ② 697
16.3. Introduction: similarities to and differences from ANOVA ② 697
16.3.1. Words of warning ② 699
16.3.2. The example for this chapter ② 699
16.4. Theory of MANOVA ③ 700
16.4.1. Introduction to matrices ③ 700
16.4.2. Some important matrices and their functions ③ 702
16.4.3. Calculating MANOVA by hand: a worked example ③ 703
16.4.4. Principle of the MANOVA test statistic ④ 710
16.5. Practical issues when conducting MANOVA ③ 717
16.5.1. Assumptions and how to check them ③ 717

16.5.2. Choosing a test statistic ③ 718
16.5.3. Follow-up analysis ③ 719
16.6. MANOVA using R ② 719
16.6.1. Packages for factorial ANOVA in R ① 719
16.6.2. General procedure for MANOVA ① 720
16.6.3. MANOVA using R Commander ② 720
16.6.4. Entering the data ② 720
16.6.5. Exploring the data ② 722
16.6.6. Setting contrasts ② 728
16.6.7. The MANOVA model ② 728
16.6.8. Follow-up analysis: univariate test statistics ② 731
16.6.9. Contrasts ③ 732
16.7. Robust MANOVA ③ 733
16.8. Reporting results from MANOVA ② 737
16.9. Following up MANOVA with discriminant analysis ③ 738
16.10. Reporting results from discriminant analysis ② 743
16.11. Some final remarks ④ 743
16.11.1. The final interpretation ④ 743
16.11.2. Univariate ANOVA or discriminant analysis? 745
What have I discovered about statistics? ② 745
R packages used in this chapter 746
R functions used in this chapter 746
Key terms that I've discovered 747
Smart Alex's tasks 747
Further reading 748
Interesting real research 748

17 Exploratory factor analysis 749

17.1. What will this chapter tell me? ① 749
17.2. When to use factor analysis ② 750
17.3. Factors ② 751
17.3.1. Graphical representation of factors ② 752
17.3.2. Mathematical representation of factors ② 753
17.3.3. Factor scores ② 755
17.3.4. Choosing a method ② 758
17.3.5. Communality ② 759
17.3.6. Factor analysis vs. principal components analysis ② 760
17.3.7. Theory behind principal components analysis ③ 761
17.3.8. Factor extraction: eigenvalues and the scree plot ② 762
17.3.9. Improving interpretation: factor rotation ③ 764
17.4. Research example ② 767
17.4.1. Sample size ② 769
17.4.2. Correlations between variables ③ 770
17.4.3. The distribution of data ② 772
17.5. Running the analysis with R Commander 772
17.6. Running the analysis with R 772
17.6.1. Packages used in this chapter ① 772
17.6.2. Initial preparation and analysis 772

17.6.3. Factor extraction using R ② 778
17.6.4. Rotation ② 788
17.6.5. Factor scores ② 793
17.6.6. Summary ② 795
17.7. How to report factor analysis ① 795
17.8. Reliability analysis ② 797
17.8.1. Measures of reliability ③ 797
17.8.2. Interpreting Cronbach's α (some cautionary tales …) ② 799
17.8.3. Reliability analysis with R Commander 800
17.8.4. Reliability analysis using R ② 800
17.8.5. Interpreting the output ② 801
17.9. Reporting reliability analysis ② 806
What have I discovered about statistics? ② 807
R packages used in this chapter 807
R functions used in this chapter 808
Key terms that I've discovered 808
Smart Alex's tasks 808
Further reading 810
Interesting real research 811

18 Categorical data 812

18.1. What will this chapter tell me? ① 812
18.2. Packages used in this chapter ① 813
18.3. Analysing categorical data ① 813
18.4. Theory of analysing categorical data ① 814
18.4.1. Pearson's chi-square test ① 814
18.4.2. Fisher's exact test ① 816
18.4.3. The likelihood ratio ② 816
18.4.4. Yates's correction ② 817
18.5. Assumptions of the chi-square test ① 818
18.6. Doing the chi-square test using R ① 818
18.6.1. Entering data: raw scores ① 818
18.6.2. Entering data: the contingency table ① 819
18.6.3. Running the analysis with R Commander ① 820
18.6.4. Running the analysis using R ① 821
18.6.5. Output from the *CrossTable()* function ① 822
18.6.6. Breaking down a significant chi-square test with standardized residuals ② 825
18.6.7. Calculating an effect size ② 826
18.6.8. Reporting the results of chi-square ① 827
18.7. Several categorical variables: loglinear analysis ③ 829
18.7.1. Chi-square as regression ④ 829
18.7.2. Loglinear analysis ③ 835
18.8. Assumptions in loglinear analysis ② 837
18.9. Loglinear analysis using R ② 838
18.9.1. Initial considerations ② 838
18.9.2. Loglinear analysis as a chi-square test ② 840
18.9.3. Output from loglinear analysis as a chi-square test ② 843

18.9.4. Loglinear analysis ② 845
18.10. Following up loglinear analysis ② 850
18.11. Effect sizes in loglinear analysis ② 851
18.12. Reporting the results of loglinear analysis ② 851
What have I discovered about statistics? ① 852
R packages used in this chapter 853
R functions used in this chapter 853
Key terms that I've discovered 853
Smart Alex's tasks ③ 853
Further reading 854
Interesting real research 854

19 Multilevel linear models 855

19.1. What will this chapter tell me? ① 855
19.2. Hierarchical data ② 856
19.2.1. The intraclass correlation ② 859
19.2.2. Benefits of multilevel models ② 859
19.3. Theory of multilevel linear models ③ 860
19.3.1. An example ② 861
19.3.2. Fixed and random coefficients ③ 862
19.4. The multilevel model ④ 865
19.4.1. Assessing the fit and comparing multilevel models ④ 867
19.4.2. Types of covariance structures ④ 868
19.5. Some practical issues ③ 870
19.5.1. Assumptions ③ 870
19.5.2. Sample size and power ③ 870
19.5.3. Centring variables ④ 871
19.6. Multilevel modelling in R ④ 873
19.6.1. Packages for multilevel modelling in R 873
19.6.2. Entering the data ② 873
19.6.3. Picturing the data ② 874
19.6.4. Ignoring the data structure: ANOVA ② 874
19.6.5. Ignoring the data structure: ANCOVA ② 876
19.6.6. Assessing the need for a multilevel model ③ 878
19.6.7. Adding in fixed effects ③ 881
19.6.8. Introducing random slopes ④ 884
19.6.9. Adding an interaction term to the model ④ 886
19.7. Growth models ④ 892
19.7.1. Growth curves (polynomials) ④ 892
19.7.2. An example: the honeymoon period ② 894
19.7.3. Restructuring the data ③ 895
19.7.4. Setting up the basic model ④ 895
19.7.5. Adding in time as a fixed effect ③ 897
19.7.6. Introducing random slopes ④ 897
19.7.7. Modelling the covariance structure ④ 897
19.7.8. Comparing models ③ 899
19.7.9. Adding higher-order polynomials ③ 901
19.7.10. Further analysis ④ 905

19.8. How to report a multilevel model ③ 906
What have I discovered about statistics? ② 907
R packages used in this chapter 908
R functions used in this chapter 908
Key terms that I've discovered 908
Smart Alex's tasks 908
Further reading 909
Interesting real research 909

Epilogue: life after discovering statistics 910
Troubleshooting R 912
Glossary 913
Appendix 929
A.1. Table of the standard normal distribution 929
A.2. Critical values of the *t*-distribution 935
A.3. Critical values of the *F*-distribution 936
A.4. Critical values of the chi-square distribution 940

References 941
Index 948
Functions in R 956
Packages in R 957

PREFACE

Karma Police, arrest this man, he talks in maths, he buzzes like a fridge, he's like a detuned radio.

Radiohead, 'Karma Police', *OK Computer* (1997)

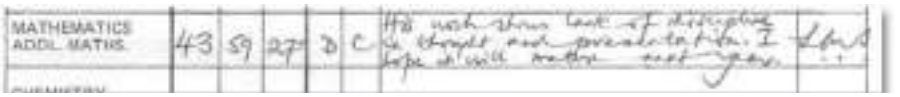
Introduction

Many social science students (and researchers for that matter) despise statistics. For one thing, most of us have a non-mathematical background, which makes understanding complex statistical equations very difficult. Nevertheless, the evil goat-warriors of Satan force our non-mathematical brains to apply themselves to what is, essentially, the very complex task of becoming a statistics expert. The end result, as you might expect, can be quite messy. The one weapon that we have is the computer, which allows us to neatly circumvent the considerable disability that is not understanding mathematics. The advent of computer programs such as SAS, SPSS, R and the like provides a unique opportunity to teach statistics at a conceptual level without getting *too* bogged down in equations. The computer to a goat-warrior of Satan is like catnip to a cat: it makes them rub their heads along the ground and purr and dribble ceaselessly. The only downside of the computer is that it makes it really easy to make a complete idiot of yourself if you don't really understand what you're doing. Using a computer without any statistical knowledge at all can be a dangerous thing. Hence this book. Well, actually, hence a book called *Discovering Statistics Using SPSS*.

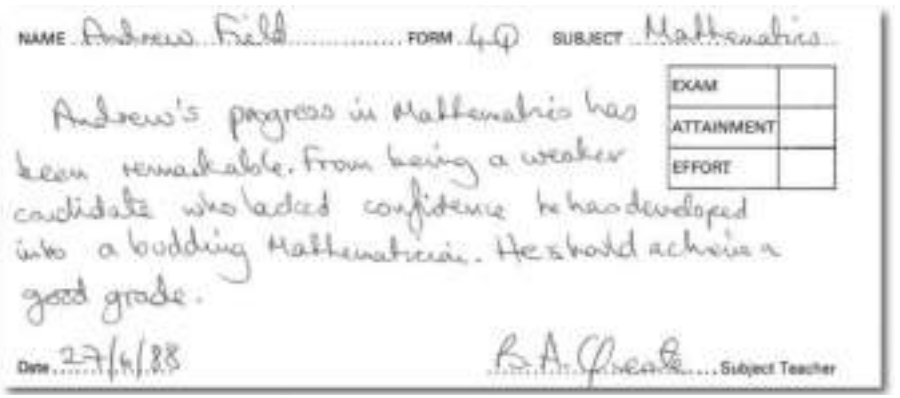
I wrote *Discovering Statistics Using SPSS* just as I was finishing off my Ph.D. in Psychology. My main aim was to write a book that attempted to strike a good balance between theory and practice: I wanted to use the computer as a tool for teaching statistical concepts in the hope that you will gain a better understanding of both theory and practice. If you want theory and you like equations then there are certainly better books: Howell (2006), Stevens (2002) and Tabachnick and Fidell (2007) are peerless as far as I am concerned and have taught me (and continue to teach me) more about statistics than you could possibly imagine. (I have an ambition to be cited in one of these books but I don't think that will ever happen.) However, if you want a book that incorporates digital rectal stimulation then you have just spent your money wisely. (I should probably clarify that the stimulation is in the context of an example, you will not find any devices attached to the inside cover for you to stimulate your rectum while you read. Please feel free to get your own device if you think it will help you to learn.)

A second, not in any way ridiculously ambitious, aim was to make this the only statistics textbook that anyone ever needs to buy. As such, it's a book that I hope will become your friend from first year right through to your professorship. I've tried to write a book that can be read at several levels (see the next section for more guidance). There are chapters for first-year undergraduates (1, 2, 3, 4, 5, 6, 9 and 15), chapters for second-year undergraduates (5, 7, 10, 11, 12, 13 and 14) and chapters on more advanced topics that postgraduates might use (8, 16, 17, 18 and 19). All of these chapters should be accessible to everyone, and I hope to achieve this by flagging the level of each section (see the next section).

My third, final and most important aim is make the learning process fun. I have a sticky history with maths because I used to be terrible at it:



Above is an extract of my school report at the age of 11. The '27=' in the report is to say that I came equal 27th with another student out of a class of 29. That's almost bottom of the class. The 43 is my exam mark as a percentage. Oh dear. Four years later (at 15) this was my school report:



What led to this remarkable change? It was having a good teacher: my brother, Paul. In fact I owe my life as an academic to Paul's ability to do what my maths teachers couldn't: teach me stuff in an engaging way. To this day he still pops up in times of need to teach me things (many tutorials in computer programming spring to mind). Anyway, the reason he's a great teacher is because he's able to make things interesting and relevant to me. He got the 'good teaching' genes in the family, but they're wasted because he doesn't teach for a living; they're a little less wasted though because his approach inspires my lectures and books. One thing that I have learnt is that people appreciate the human touch, and so I tried to inject a lot of my own personality and sense of humour (or lack of) into *Discovering Statistics Using …* books. Many of the examples in this book, although inspired by some of the craziness that you find in the real world, are designed to reflect topics that play on the minds of the average student (i.e., sex, drugs, rock and roll, celebrity, people doing crazy stuff). There are also some examples that are there just because they made me laugh. So, the examples are light-hearted (some have said 'smutty' but I prefer 'light-hearted') and by the end, for better or worse, I think you will have some idea of what goes on in my head on a daily basis. I apologize to those who think it's crass, hate it, or think that I'm undermining the seriousness of science, but, come on, what's not funny about a man putting an eel up his anus?

Did I succeed in these aims? Maybe I did, maybe I didn't, but the SPSS book on which this **R** book is based has certainly been popular and I enjoy the rare luxury of having many complete strangers emailing me to tell me how wonderful I am. (Admittedly, occassionally people email to tell me that they think I'm a pile of gibbon excrement but you have to take the rough with the smooth.) It also won the British Psychological Society book award in 2007. I must have done something right. However, *Discovering Statistics Using SPSS* has one very large flaw: not everybody uses SPSS. Some people use **R**. **R** has one fairly big advantage over other statistical packages in that it is free. That's right, it's free. Completely and utterly free. People say that there's no such thing as a free lunch, but they're wrong:

R is a feast of succulent delights topped off with a baked cheesecake and nothing to pay at the end of it.

It occurred to me that it would be great to have a version of the book that used all of the same theory and examples from the SPSS book but written about **R**. Genius. Genius except that I knew very little about **R**. Six months and quite a few late nights later and I know a lot more about **R** than I did when I started this insane venture. Along the way I have been helped by a very nice guy called Jeremy (a man who likes to put eels in his CD player rather than anywhere else), and an even nicer wife. Both of their contributions have been concealed somewhat by our desire to keep the voice of the book mine, but they have both contributed enormously. (Jeremy's contributions are particularly easy to spot: if it reads like a statistics genius struggling manfully to coerce the words of a moron into something approximating factual accuracy, then Jeremy wrote it.)

What are you getting for your money?

This book takes you on a journey (possibly through a very narrow passage lined with barbed wire) not just of statistics but of the weird and wonderful contents of the world and my brain. In short, it's full of stupid examples, bad jokes, smut and filth. Aside from the smut, I have been forced reluctantly to include some academic content. Over many editions of the SPSS book many people have emailed me with suggestions, so, in theory, what you currently have in your hands should answer any question anyone has asked me over the past ten years. It won't, but it should, and I'm sure you can find some new questions to ask. It has some other unusual features:

- **Everything you'll ever need to know**: I want this to be good value for money so the book guides you from complete ignorance (Chapter 1 tells you the basics of doing research) to being an expert on multilevel modelling (Chapter 19). Of course no book that you can actually lift off the floor will contain everything, but I think this one has a fair crack at taking you from novice to postgraduate level expertise. It's pretty good for developing your biceps also.

- **Stupid faces**: You'll notice that the book is riddled with stupid faces, some of them my own. You can find out more about the pedagogic function of these 'characters' in the next section, but even without any useful function they're still nice to look at.

- **Data sets**: There are about 100 data files associated with this book on the companion website. Not unusual in itself for a statistics book, but my data sets contain more sperm (not literally) than other books. I'll let you judge for yourself whether this is a good thing.

- **My life story**: Each chapter is book-ended by a chronological story from my life. Does this help you to learn about statistics? Probably not, but hopefully it provides some light relief between chapters.

- **R tips**: R does weird things sometimes. In each chapter, there are boxes containing tips, hints and pitfalls related to R.

- **Self-test questions**: Given how much students hate tests, I thought the best way to commit commercial suicide was to liberally scatter tests throughout each chapter. These range from simple questions to test what you have just learned to going back to a technique that you read about several chapters before and applying it in a new context. All of these questions have answers to them on the companion website. They are there so that you can check on your progress.

The book also has some more conventional features:

- **Reporting your analysis**: Every single chapter has a guide to writing up your analysis. Obviously, how one writes up an analysis varies a bit from one discipline to another and, because I'm a psychologist, these sections are quite psychology-based. Nevertheless, they should get you heading in the right direction.
- **Glossary**: Writing the glossary was so horribly painful that it made me stick a vacuum cleaner into my ear to suck out my own brain. You can find my brain in the bottom of the vacuum cleaner in my house.
- **Real-world data**: Students like to have 'real data' to play with. The trouble is that real research can be quite boring. However, just for you, I trawled the world for examples of research on really fascinating topics (in my opinion). I then stalked the authors of the research until they gave me their data. Every chapter has a real research example.

Goodbye

The SPSS version of this book has literally consumed the last 13 years or so of my life, and this **R** version has consumed the last 6 months. I am literally typing this as a withered husk. I have no idea whether people use **R**, and whether this version will sell, but I think they should (use **R**, that is, not necessarily buy the book). The more I have learnt about **R** through writing this book, the more I like it.

This book in its various forms has been a huge part of my adult life; it began as and continues to be a labour of love. The book isn't perfect, and I still love to have feedback (good or bad) from the people who matter most: you.

Andy

- Contact details: http://www. discoveringstatistics.com/html/email.html
- Twitter: @ProfAndyField
- Blog: http://www.methodspace.com/profile/ProfessorAndyField

HOW TO USE THIS BOOK

When the publishers asked me to write a section on 'How to use this book' it was obviously tempting to write 'Buy a large bottle of Olay anti-wrinkle cream (which you'll need to fend off the effects of ageing while you read), find a comfy chair, sit down, fold back the front cover, begin reading and stop when you reach the back cover.' However, I think they wanted something more useful. ☺

What background knowledge do I need?

In essence, I assume you know nothing about statistics, but I do assume you have some very basic grasp of computers (I won't be telling you how to switch them on, for example) and maths (although I have included a quick revision of some very basic concepts so I really don't assume anything).

Do the chapters get more difficult as I go through the book?

In a sense they do (Chapter 16 on MANOVA is more difficult than Chapter 1), but in other ways they don't (Chapter 15 on non-parametric statistics is arguably less complex than Chapter 14, and Chapter 9 on the *t*-test is definitely less complex than Chapter 8 on logistic regression). Why have I done this? Well, I've ordered the chapters to make statistical sense (to me, at least). Many books teach different tests in isolation and never really give you a grip of the similarities between them; this, I think, creates an unnecessary mystery. Most of the tests in this book are the same thing expressed in slightly different ways. So, I wanted the book to tell this story. To do this I have to do certain things such as explain regression fairly early on because it's the foundation on which nearly everything else is built.

However, to help you through I've coded each section with an icon. These icons are designed to give you an idea of the difficulty of the section. It doesn't necessarily mean you can skip the sections (but see Smart Alex in the next section), but it will let you know whether a section is at about your level, or whether it's going to push you. I've based the icons on my own teaching so they may not be entirely accurate for everyone (especially as systems vary in different countries!):

① This means 'level 1' and I equate this to first-year undergraduate in the UK. These are sections that everyone should be able to understand.

② This is the next level and I equate this to second-year undergraduate in the UK. These are topics that I teach my second years and so anyone with a bit of background in statistics should be able to get to grips with them. However, some of these sections will be quite challenging even for second years. These are intermediate sections.

③ This is 'level 3' and represents difficult topics. I'd expect third-year (final-year) UK undergraduates and recent postgraduate students to be able to tackle these sections.
④ This is the highest level and represents very difficult topics. I would expect these sections to be very challenging to undergraduates and recent postgraduates, but postgraduates with a reasonable background in research methods shouldn't find them too much of a problem.

Why do I keep seeing stupid faces everywhere?

Brian Haemorrhage: Brian's job is to pop up to ask questions and look permanently confused. It's no surprise to note, therefore, that he doesn't look entirely different from the author (he has more hair though). As the book progresses he becomes increasingly despondent. Read into that what you will.

Curious Cat: He also pops up and asks questions (because he's curious). Actually the only reason he's here is because I wanted a cat in the book … and preferably one that looks like mine. Of course the educational specialists think he needs a specific role, and so his role is to look cute and make bad cat-related jokes.

Cramming Sam: Samantha hates statistics. In fact, she thinks it's all a boring waste of time and she just wants to pass her exam and forget that she ever had to know anything about normal distributions. So, she appears and gives you a summary of the key points that you need to know. If, like Samantha, you're cramming for an exam, she will tell you the essential information to save you having to trawl through hundreds of pages of my drivel.

Jane Superbrain: Jane is the cleverest person in the whole universe (she makes Smart Alex look like a bit of an imbecile). The reason she is so clever is that she steals the brains of statisticians and eats them. Apparently they taste of sweaty tank tops, but nevertheless she likes them. As it happens she is also able to absorb the contents of brains while she eats them. Having devoured some top statistics brains she knows all the really hard stuff and appears in boxes to tell you really advanced things that are a bit tangential to the main text. (Readers should note that Jane wasn't interested in eating my brain. That tells you all that you need to know about my statistics ability.)

Labcoat Leni: Leni is a budding young scientist and he's fascinated by real research. He says, 'Andy, man, I like an example about using an eel as a cure for constipation as much as the next man, but all of your examples are made up. Real data aren't like that, we need some real examples, dude!' So off Leni went; he walked the globe, a lone data warrior in a thankless quest for real data. He turned up at universities, cornered academics, kidnapped their families and threatened to put them in a bath of crayfish unless he was given real data. The generous ones relented, but others? Well, let's just say their families are sore. So, when you see Leni you know that you will get some real data, from a real research study to analyse. Keep it real.

Oliver Twisted: With apologies to Charles Dickens, Oliver, like the more famous fictional London urchin, is always asking 'Please Sir, can I have some more?' Unlike Master Twist though, our young Master Twisted always wants more statistics information. Of course he does, who wouldn't? Let us not be the ones to disappoint a young, dirty, slightly smelly boy who dines on gruel, so when Oliver appears you can be certain of one thing: there is additional information to be found on the companion website. (Don't be shy; download it and bathe in the warm asp's milk of knowledge.)

R's Souls: People who love statistics are damned to hell for all eternity, people who like **R** even more so. However, **R** and statistics are secretly so much fun that Satan is inundated with new lost souls, converted to the evil of statistical methods. Satan needs a helper to collect up all the souls of those who have been converted to the joy of **R**. While collecting the souls of the statistical undead, they often cry out useful tips to him. He's collected these nuggets of information and spread them through the book like a demonic plague of beetles. When Satan's busy spanking a goat, his helper pops up in a box to tell you some of **R**'s Souls' Tips.

Smart Alex: Alex is a very important character because he appears when things get particularly difficult. He's basically a bit of a smart alec and so whenever you see his face you know that something scary is about to be explained. When the hard stuff is over he reappears to let you know that it's safe to continue. Now, this is not to say that all of the rest of the material in the book is easy, he just lets you know the bits of the book that you can skip if you've got better things to do with your life than read all 1000 pages! So, if you see Smart Alex then you can *skip the section* entirely and still understand what's going on. You'll also find that Alex pops up at the end of each chapter to give you some tasks to do to see whether you're as smart as he is.

What is on the companion website?

In this age of downloading, CD-ROMs are for losers (at least that's what the 'kids' tell me) so I've put my cornucopia of additional funk on that worldwide interweb thing. This has two benefits: 1) the book is *slightly* lighter than it would have been, and 2) rather than being restricted to the size of a CD-ROM, there is no limit to the amount of fascinating extra material that I can give you (although Sage have had to purchase a new server to fit it all on). To enter my world of delights, go to www.sagepub.co.uk/dsur.

How will you know when there are extra goodies on this website? Easy-peasy, Oliver Twisted appears in the book to indicate that there's something you need (or something extra) on the website. The website contains resources for students and lecturers alike:

- **Data files**: You need data files to work through the examples in the book and they are all on the companion website. We did this so that you're forced to go there and once you're there Sage will flash up subliminal messages that make you buy more of their books.
- **R script files**: if you put all of the **R** commands in this book next to each other and printed them out you'd have a piece of paper that stretched from here to the Tarantula Nebula (which actually exists and sounds like a very scary place). If you type all of these commands into **R** you will wear away your fingers to small stumps. I would never forgive myself if you all got stumpy fingers so the website has script files containing every single **R** command in the book (including within chapter questions and activities).

- **Webcasts**: My publisher thinks that watching a film of me explaining what this book is all about will get people flocking to the bookshop. I think it will have people flocking to the medicine cabinet. Either way, if you want to see how truly uncharismatic I am, watch and cringe. There are also a few webcasts of lectures given by me relevant to the content of the book.
- **Self-Assessment Multiple-Choice Questions**: Organized by chapter, these will allow you to test whether wasting your life reading this book has paid off so that you can walk confidently into an examination much to the annoyance of your friends. If you fail said exam, you can employ a good lawyer and sue.
- **Additional material**: Enough trees have died in the name of this book, but still it gets longer and still people want to know more. Therefore, we've written nearly 300 pages, yes, three hundred, of additional material for the book. So for some more technical topics and help with tasks in the book the material has been provided electronically so that (1) the planet suffers a little less, and (2) you won't die when the book falls off of your bookshelf onto your head.
- **Answers**: each chapter ends with a set of tasks for you to test your newly acquired expertise. The chapters are also littered with self-test questions and Labcoat Leni's assignments. How will you know if you get these correct? Well, the companion website contains around 300 pages (that's a different 300 pages to the 300 above) of detailed answers. Will we ever stop writing?
- **Powerpoint slides**: I can't come and personally teach you all. Instead I rely on a crack team of highly skilled and super intelligent pan-dimensional beings called 'lecturers'. I have personally grown each and every one of them in a greenhouse in my garden. To assist in their mission to spread the joy of statistics I have provided them with powerpoint slides for each chapter.
- **Links**: every website has to have links to other useful websites and the companion website is no exception.
- **Cyberworms of knowledge**: I have used nanotechnology to create cyberworms that crawl down your broadband connection, pop out of the USB port of your computer then fly through space into your brain. They re-arrange your neurons so that you understand statistics. You don't believe me? Well, you'll never know for sure unless you visit the companion website …

Happy reading, and don't get sidetracked by Facebook and Twitter.

SYMBOLS USED IN THIS BOOK

Mathematical operators

Σ	This symbol (called sigma) means 'add everything up'. So, if you see something like Σx_i it just means 'add up all of the scores you've collected'.
Π	This symbol means 'multiply everything'. So, if you see something like Πx_i it just means 'multiply all of the scores you've collected'.
$\sqrt{x}$	This means 'take the square root of x'.

Greek symbols

α	The probability of making a Type I error
β	The probability of making a Type II error
β_i	Standardized regression coefficient
χ^2	Chi-square test statistic
χ^2_F	Friedman's ANOVA test statistic
ε	Usually stands for 'error'
η^2	Eta-squared
μ	The mean of a population of scores
ρ	The correlation in the population
σ^2	The variance in a population of data
σ	The standard deviation in a population of data
$\sigma_{\bar{x}}$	The standard error of the mean
τ	Kendall's tau (non-parametric correlation coefficient)
ω^2	Omega squared (an effect size measure). This symbol also means 'expel the contents of your intestine immediately into your trousers'; you will understand why in due course.

English symbols

b_i	The regression coefficient (unstandardized)
df	Degrees of freedom
e_i	The error associated with the *i*th person
F	*F*-ratio (test statistic used in ANOVA)
H	Kruskal–Wallis test statistic
k	The number of levels of a variable (i.e. the number of treatment conditions), or the number of predictors in a regression model
ln	Natural logarithm
MS	The mean squared error. The average variability in the data
N, n, n_i	The sample size. *N* usually denotes the total sample size, whereas *n* usually denotes the size of a particular group
P	Probability (the probability value, *p*-value or significance of a test are usually denoted by *p*)
r	Pearson's correlation coefficient
r_s	Spearman's rank correlation coefficient
r_b, r_{pb}	Biserial correlation coefficient and point–biserial correlation coefficient respectively
R	The multiple correlation coefficient
R^2	The coefficient of determination (i.e. the proportion of data explained by the model)
s^2	The variance of a sample of data
s	The standard deviation of a sample of data
SS	The sum of squares, or sum of squared errors to give it its full title
SS_A	The sum of squares for variable *A*
SS_M	The model sum of squares (i.e. the variability explained by the model fitted to the data)
SS_R	The residual sum of squares (i.e. the variability that the model can't explain – the error in the model)
SS_T	The total sum of squares (i.e. the total variability within the data)
t	Test statistic for Student's *t*-test
T	Test statistic for Wilcoxon's matched-pairs signed-rank test
U	Test statistic for the Mann–Whitney test
W_s	Test statistic for the Shapiro–Wilk test and the Wilcoxon's rank-sum test
$\bar{X}$ or $\bar{x}$	The mean of a sample of scores
z	A data point expressed in standard deviation units

Why is my evil lecturer forcing me to learn statistics?

1

FIGURE 1.1
When I grow up, please don't let me be a statistics lecturer

1.1. What will this chapter tell me? ①

I was born on 21 June 1973. Like most people, I don't remember anything about the first few years of life and like most children I did go through a phase of driving my parents mad by asking 'Why?' every five seconds. 'Dad, why is the sky blue?', 'Dad, why doesn't mummy have a willy?', etc. Children are naturally curious about the world. I remember at the age of 3 being at a party of my friend Obe (this was just before he left England to return to Nigeria, much to my distress). It was a hot day, and there was an electric fan blowing cold air around the room. As I said, children are natural scientists and my

1.3. Initial observation: finding something that needs explaining ①

The first step in Figure 1.2 was to come up with a question that needs an answer. I spend rather more time than I should watching reality TV. Every year I swear that I won't get hooked on *Big Brother*, and yet every year I find myself glued to the TV screen waiting for the next contestant's meltdown (I am a psychologist, so really this is just research – honestly). One question I am constantly perplexed by is why every year there are so many contestants with really unpleasant personalities (my money is on narcissistic personality disorder[4]) on the show. A lot of scientific endeavour starts this way: not by watching *Big Brother*, but by observing something in the world and wondering why it happens.

Having made a casual observation about the world (*Big Brother* contestants on the whole have profound personality defects), I need to collect some data to see whether this observation is true (and not just a biased observation). To do this, I need to define one or more **variables** that I would like to measure. There's one variable in this example: the personality of the contestant. I could measure this variable by giving them one of the many well-established questionnaires that measure personality characteristics. Let's say that I did this and I found that 75% of contestants did have narcissistic personality disorder. These data support my observation: a lot of *Big Brother* contestants have extreme personalities.

1.4. Generating theories and testing them ①

The next logical thing to do is to explain these data (Figure 1.2). One explanation could be that people with narcissistic personality disorder are more likely to audition for *Big Brother* than those without. This is a **theory**. Another possibility is that the producers of *Big Brother* are more likely to select people who have narcissistic personality disorder to be contestants than those with less extreme personalities. This is another theory. We verified our original observation by collecting data, and we can collect more data to test our theories. We can make two predictions from these two theories. The first is that the number of people turning up for an audition that have narcissistic personality disorder will be higher than the general level in the population (which is about 1%). A prediction from a theory, like this one, is known as a **hypothesis** (see Jane Superbrain Box 1.1). We could test this hypothesis by getting a team of clinical psychologists to interview each person at the *Big Brother* audition and diagnose them as having narcissistic personality disorder or not. The prediction from our second theory is that if the *Big Brother* selection panel are more likely to choose people with narcissistic personality disorder then the rate of this disorder in the final contestants will be even higher than the rate in the group of people going for auditions. This is another hypothesis. Imagine we collected these data; they are in Table 1.1.

In total, 7662 people turned up for the audition. Our first hypothesis is that the percentage of people with narcissistic personality disorder will be higher at the audition than the general level in the population. We can see in the table that of the 7662 people at the audition, 854 were diagnosed with the disorder; this is about 11% ($854/7662 \times 100$), which is much higher than the 1% we'd expect. Therefore, hypothesis 1 is supported by the data. The second hypothesis was that the *Big Brother* selection panel have a bias to chose people with narcissistic personality disorder. If we look at the 12 contestants that they selected, 9 of them had the disorder (a massive 75%). If the producers did not have a bias we would

[4] This disorder is characterized by (among other things) a grandiose sense of self-importance, arrogance, lack of empathy for others, envy of others and belief that others envy them, excessive fantasies of brilliance or beauty, the need for excessive admiration and exploitation of others.