In statistics, R is the way of the future... I have been waiting for this book for some time, it offers not just the step-by-step guidance needed to complete a particular task, but it also offers the chance to reach the Zen state of total statistical understanding.

Professor Neil Stewart, Warwick University

Field’s Discovering Statistics is popular with students for making a sometimes deemed inaccessible topic accessible, the plain way in Discovering Statistics Using R the authors have managed to do this using a statistics package that is known to be powerful, but sometimes deemed just as inaccessible to the uninitiated, all the while staying true to Field’s off-kilter approach.

Dr Marcel van Egmond, University of Amsterdam

Hot on the heels of the award-winning and best-selling Discovering Statistics Using SPSS, 3rd Edition, Andy Field has teamed up with Jeremy Miles (co-author of Discovering Statistics Using SAS) and Zoe Field to write Discovering Statistics Using R. Keeping the uniquely humorous and self-deprecating style that has made students across the world fall in love with Andy Field’s books, Discovering Statistics Using R takes students on a journey of statistical discovery using R, a free, flexible and dynamically changing software tool for data analysis that is becoming increasingly popular across the social and behavioural sciences throughout the world.

The journey begins by explaining basic statistical and research concepts before a guided tour of the R software environment. Next you discover the importance of exploring and graphing data, before moving onto statistical tests that are the foundations of the rest of the book (for example, correlation and regression). You will then stride confidently into intermediate level analyses such as ANOVA, before ending your journey with advanced techniques such as MANOVA and multilevel models. Although there is enough theory to help you gain the necessary conceptual understanding of what you’re doing, the emphasis is on applying what you learn to real-world examples that should make the experience more fun than you might expect.

Like its sister textbooks, Discovering Statistics Using R is written in an irreverent style and follows the same groundbreaking structure and pedagogical approach. The core material is augmented by a cast of characters to help the reader on their way, together with hundreds of examples, self-assessment tests to consolidate knowledge, and additional website material for those wanting to learn more at: www.sagepub.co.uk/dsur.

Given this book’s accessibility, fun spirit, and use of bizarre real-world research it should be essential for anyone wanting to learn about statistics using the freely-available R software.

ANDY FIELD is Professor of Child Psychopathology at the University of Sussex. He adores cats, and loves to listen to and play very heavy music. His ability to make statistics accessible and fun has been recognized with local and national teaching awards (University of Sussex, 2001; the British Psychological Society, 2007), a prestigious UK National Teaching Fellowship (2010), and the British Psychological Society book award (2006). He lives in Brighton with his wonderful wife Zoe and Fuzzy the cat.

JEREMY MILES works as Behavioral Scientist at the RAND Corporation in Santa Monica, California and as a Professor of Quantitative Methods at the Pardee RAND Graduate School. He thinks dogs are better than cats but is wrong. He lives in Los Angeles with his twin boys, twin dogs, twin chickens but not twin wives. He has published several titles including Understanding and Using Statistics in Psychology (SAGE, 2007).

ZOË FIELD works at the University of Sussex. She has published several research papers but this is her first book. She loves living in Statistics Hell with her beloved husband Andy and their cat Fuzzy.
1 Why is my evil lecturer forcing me to learn statistics? 1
  1.1. What will this chapter tell me? 1
  1.2. What the hell am I doing here? I don’t belong here 2
  1.3. Initial observation: finding something that needs explaining 4
  1.4. Generating theories and testing them 4
  1.5. Data collection 1: what to measure 7
      1.5.1. Variables 7
      1.5.2. Measurement error 11
      1.5.3. Validity and reliability 12
  1.6. Data collection 2: how to measure 13
      1.6.1. Correlational research methods 13
      1.6.2. Experimental research methods 13
      1.6.3. Randomization 17
  1.7. Analysing data 19
      1.7.1. Frequency distributions 19
      1.7.2. The centre of a distribution 21
      1.7.3. The dispersion in a distribution 24
      1.7.4. Using a frequency distribution to go beyond the data 25
      1.7.5. Fitting statistical models to the data 28
  What have I discovered about statistics? 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? 32
  2.2. Building statistical models 33
2.3. Populations and samples 36
2.4. Simple statistical models 36
  2.4.1. The mean: a very simple statistical model 36
  2.4.2. Assessing the fit of the mean: sums of squares, variance and standard deviations 37
  2.4.3. Expressing the mean as a model 40
2.5. Going beyond the data 41
  2.5.1. The standard error 42
  2.5.2. Confidence intervals 43
2.6. Using statistical models to test research questions 49
  2.6.1. Test statistics 53
  2.6.2. One- and two-tailed tests 55
  2.6.3. Type I and Type II errors 56
  2.6.4. Effect sizes 57
  2.6.5. Statistical power 58
What have I discovered about statistics? 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? 62
  3.2. Before you start 63
    3.2.1. The R-chitecture 63
    3.2.2. Pros and cons of R 64
    3.2.3. Downloading and installing R 65
    3.2.4. Versions of R 66
  3.3. Getting started 66
    3.3.1. The main windows in R 67
    3.3.2. Menus in R 67
  3.4. Using R 71
    3.4.1. Commands, objects and functions 71
    3.4.2. Using scripts 75
    3.4.3. The R workspace 76
    3.4.4. Setting a working directory 77
    3.4.5. Installing packages 78
    3.4.6. Getting help 80
  3.5. Getting data into R 81
    3.5.1. Creating variables 81
    3.5.2. Creating dataframes 81
    3.5.3. Calculating new variables from existing ones 83
    3.5.4. Organizing your data 85
    3.5.5. Missing values 92
  3.6. Entering data with R Commander 92
    3.6.1. Creating variables and entering data with R Commander 94
    3.6.2. Creating coding variables with R Commander 95
  3.7. Using other software to enter and edit data 95
    3.7.1. Importing data 97
    3.7.2. Importing SPSS data files directly 99
CONTENTS

3.7.3. Importing data with R Commander 101
3.7.4. Things that can go wrong 102
3.8. Saving data 103
3.9. Manipulating data 103
3.9.1. Selecting parts of a dataframe 103
3.9.2. Selecting data with the subset() function 105
3.9.3. Dataframes and matrices 106
3.9.4. Reshaping data 107
What have I discovered about statistics? 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? 116
4.2. The art of presenting data 117
4.2.1. Why do we need graphs 117
4.2.2. What makes a good graph? 117
4.2.3. Lies, damned lies, and … erm … graphs 120
4.3. Packages used in this chapter 121
4.4. Introducing ggplot2 121
4.4.1. The anatomy of a plot 121
4.4.2. Geometric objects (geoms) 123
4.4.3. Aesthetics 125
4.4.4. The anatomy of the ggplot() function 127
4.4.5. Stats and geoms 128
4.4.6. Avoiding overplotting 130
4.4.7. Saving graphs 131
4.4.8. Putting it all together: a quick tutorial 132
4.5. Graphing relationships: the scatterplot 136
4.5.1. Simple scatterplot 136
4.5.2. Adding a funky line 138
4.5.3. Grouped scatterplot 140
4.6. Histograms: a good way to spot obvious problems 142
4.7. Boxplots (box–whisker diagrams) 144
4.8. Density plots 148
4.9. Graphing means 149
4.9.1. Bar charts and error bars 149
4.9.2. Line graphs 155
4.10. Themes and options 161
What have I discovered about statistics? 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? 166
5.2. What are assumptions? 167
5.3. Assumptions of parametric data 167
5.4. Packages used in this chapter 169
5.5. The assumption of normality 169
  5.5.1. Oh no, it’s that pesky frequency distribution again: checking normality visually 169
  5.5.2. Quantifying normality with numbers 173
  5.5.3. Exploring groups of data 177
5.6. Testing whether a distribution is normal 182
  5.6.1. Doing the Shapiro–Wilk test in R 182
  5.6.2. Reporting the Shapiro–Wilk test 185
5.7. Testing for homogeneity of variance 185
  5.7.1. Levene’s test 186
  5.7.2. Reporting Levene’s test 188
  5.7.3. Hartley’s F_max; the variance ratio 189
5.8. Correcting problems in the data 190
  5.8.1. Dealing with outliers 190
  5.8.2. Dealing with non-normality and unequal variances 191
  5.8.3. Transforming the data using R 194
  5.8.4. When it all goes horribly wrong 201

What have I discovered about statistics? 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? 205
6.2. Looking at relationships 206
6.3. How do we measure relationships? 206
  6.3.1. A detour into the murky world of covariance 206
  6.3.2. Standardization and the correlation coefficient 208
  6.3.3. The significance of the correlation coefficient 210
  6.3.4. Confidence intervals for r 211
  6.3.5. A word of warning about interpretation: causality 212
6.4. Data entry for correlation analysis 213
6.5. Bivariate correlation 213
  6.5.1. Packages for correlation analysis in R 214
  6.5.2. General procedure for correlations using R Commander 214
  6.5.3. General procedure for correlations using R 216
  6.5.4. Pearson’s correlation coefficient 219
  6.5.5. Spearman’s correlation coefficient 223
  6.5.6. Kendall’s tau (non-parametric) 225
  6.5.7. Bootstrapping correlations 226
  6.5.8. Biserial and point-biserial correlations 229
6.6. Partial correlation 234
6.6.1. The theory behind part and partial correlation 234
6.6.2. Partial correlation using R 235
6.6.3. Semi-partial (or part) correlations 237
6.7. Comparing correlations 238
6.7.1. Comparing independent rs 238
6.7.2. Comparing dependent rs 239
6.8. Calculating the effect size 240
6.9. How to report correlation coefficients 240

What have I discovered about statistics? 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 243
Further reading 244
Interesting real research 244

7  Regression 245

7.1. What will this chapter tell me? 245
7.2. An introduction to regression 246
7.2.1. Some important information about straight lines 247
7.2.2. The method of least squares 248
7.2.3. Assessing the goodness of fit: sums of squares, R and \(R^2\) 249
7.2.4. Assessing individual predictors 252
7.3. Packages used in this chapter 253
7.4. General procedure for regression in R 254
7.4.1. Doing simple regression using R Commander 254
7.4.2. Regression in R 255
7.5. Interpreting a simple regression 257
7.5.1. Overall fit of the object model 258
7.5.2. Model parameters 259
7.5.3. Using the model 260
7.6. Multiple regression: the basics 261
7.6.1. An example of a multiple regression model 261
7.6.2. Sums of squares, \(R\) and \(R^2\) 262
7.6.3. Parsimony-adjusted measures of fit 263
7.6.4. Methods of regression 263
7.7. How accurate is my regression model? 266
7.7.1. Assessing the regression model I: diagnostics 266
7.7.2. Assessing the regression model II: generalization 271
7.8. How to do multiple regression using R Commander and R 276
7.8.1. Some things to think about before the analysis 276
7.8.2. Multiple regression: running the basic model 277
7.8.3. Interpreting the basic multiple regression 280
7.8.4. Comparing models 284
7.9. Testing the accuracy of your regression model 287
7.9.1. Diagnostic tests using R Commander 287
7.9.2. Outliers and influential cases 288
7.9.3. Assessing the assumption of independence 291
7.9.4. Assessing the assumption of no multicollinearity 292
7.9.5. Checking assumptions about the residuals 294
7.9.6. What if I violate an assumption? 298
7.10. Robust regression: bootstrapping 298
7.11. How to report multiple regression 301
7.12. Categorical predictors and multiple regression 302
7.12.1. Dummy coding 302
7.12.2. Regression with dummy variables 305
What have I discovered about statistics? 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? 312
8.2. Background to logistic regression 313
8.3. What are the principles behind logistic regression? 313
8.3.1. Assessing the model: the log-likelihood statistic 315
8.3.2. Assessing the model: the deviance statistic 316
8.3.3. Assessing the model: $R^2$ 316
8.3.4. Assessing the model: information criteria 318
8.3.5. Assessing the contribution of predictors: the $z$-statistic 318
8.3.6. The odds ratio 319
8.3.7. Methods of logistic regression 320
8.4. Assumptions and things that can go wrong 321
8.4.1. Assumptions 321
8.4.2. Incomplete information from the predictors 322
8.4.3. Complete separation 323
8.5. Packages used in this chapter 325
8.6. Binary logistic regression: an example that will make you feel eel 325
8.6.1. Preparing the data 326
8.6.2. The main logistic regression analysis 327
8.6.3. Basic logistic regression analysis using R 329
8.6.4. Interpreting a basic logistic regression 330
8.6.5. Model 1: Intervention only 330
8.6.6. Model 2: Intervention and Duration as predictors 336
8.6.7. Casewise diagnostics in logistic regression 338
8.6.8. Calculating the effect size 341
8.7. How to report logistic regression 341
8.8. Testing assumptions: another example 342
8.8.1. Testing for multicollinearity 343
8.8.2. Testing for linearity of the logit 344
8.9. Predicting several categories: multinomial logistic regression 346
8.9.1. Running multinomial logistic regression in R 347
8.9.2. Interpreting the multinomial logistic regression output 350
9 Comparing two means

9.1. What will this chapter tell me? 359
9.2. Packages used in this chapter 360
9.3. Looking at differences 360
  9.3.1. A problem with error bar graphs of repeated-measures designs 361
  9.3.2. Step 1: calculate the mean for each participant 364
  9.3.3. Step 2: calculate the grand mean 364
  9.3.4. Step 3: calculate the adjustment factor 364
  9.3.5. Step 4: create adjusted values for each variable 365
9.4. The t-test 368
  9.4.1. Rationale for the t-test 369
  9.4.2. The t-test as a general linear model 370
  9.4.3. Assumptions of the t-test 372
9.5. The independent t-test 372
  9.5.1. The independent t-test equation explained 372
  9.5.2. Doing the independent t-test 375
9.6. The dependent t-test 386
  9.6.1. Sampling distributions and the standard error 386
  9.6.2. The dependent t-test equation explained 387
  9.6.3. Dependent t-tests using R 388
9.7. Between groups or repeated measures? 394
  What have I discovered about statistics? 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)

10.1. What will this chapter tell me? 398
10.2. The theory behind ANOVA 399
  10.2.1. Inflated error rates 399
  10.2.2. Interpreting F 400
  10.2.3. ANOVA as regression 400
  10.2.4. Logic of the F-ratio 405
  10.2.5. Total sum of squares (SS) 407
  10.2.6. Model sum of squares (SSM) 409
  10.2.7. Residual sum of squares (SSR) 410
  10.2.8. Mean squares 411
11. Analysis of covariance, ANCOVA (GLM 2) 462

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

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

### 12 Factorial ANOVA (GLM 3)

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.1</td>
<td>What will this chapter tell me?</td>
<td>498</td>
</tr>
<tr>
<td>12.2</td>
<td>Theory of factorial ANOVA (independent design)</td>
<td>499</td>
</tr>
<tr>
<td>12.2.1</td>
<td>Factorial designs</td>
<td>499</td>
</tr>
<tr>
<td>12.3</td>
<td>Factorial ANOVA as regression</td>
<td>501</td>
</tr>
<tr>
<td>12.3.1</td>
<td>An example with two independent variables</td>
<td>501</td>
</tr>
<tr>
<td>12.3.2</td>
<td>Extending the regression model</td>
<td>501</td>
</tr>
<tr>
<td>12.4</td>
<td>Two-way ANOVA: behind the scenes</td>
<td>505</td>
</tr>
<tr>
<td>12.4.1</td>
<td>Total sums of squares ($SS_T$)</td>
<td>506</td>
</tr>
<tr>
<td>12.4.2</td>
<td>The model sum of squares ($SS_M$)</td>
<td>507</td>
</tr>
<tr>
<td>12.4.3</td>
<td>The residual sum of squares ($SS_R$)</td>
<td>510</td>
</tr>
<tr>
<td>12.4.4</td>
<td>The $F$-ratios</td>
<td>511</td>
</tr>
<tr>
<td>12.5</td>
<td>Factorial ANOVA using R</td>
<td>511</td>
</tr>
<tr>
<td>12.5.1</td>
<td>Packages for factorial ANOVA in R</td>
<td>511</td>
</tr>
<tr>
<td>12.5.2</td>
<td>General procedure for factorial ANOVA</td>
<td>512</td>
</tr>
<tr>
<td>12.5.3</td>
<td>Factorial ANOVA using R Commander</td>
<td>512</td>
</tr>
<tr>
<td>12.5.4</td>
<td>Entering the data</td>
<td>513</td>
</tr>
<tr>
<td>12.5.5</td>
<td>Exploring the data</td>
<td>516</td>
</tr>
<tr>
<td>12.5.6</td>
<td>Choosing contrasts</td>
<td>518</td>
</tr>
<tr>
<td>12.5.7</td>
<td>Fitting a factorial ANOVA model</td>
<td>520</td>
</tr>
<tr>
<td>12.5.8</td>
<td>Interpreting factorial ANOVA</td>
<td>520</td>
</tr>
<tr>
<td>12.5.9</td>
<td>Interpreting contrasts</td>
<td>524</td>
</tr>
<tr>
<td>12.5.10</td>
<td>Simple effects analysis</td>
<td>525</td>
</tr>
<tr>
<td>12.5.11</td>
<td>Post hoc analysis</td>
<td>528</td>
</tr>
<tr>
<td>12.5.12</td>
<td>Overall conclusions</td>
<td>530</td>
</tr>
<tr>
<td>12.5.13</td>
<td>Plots in factorial ANOVA</td>
<td>530</td>
</tr>
<tr>
<td>12.6</td>
<td>Interpreting interaction graphs</td>
<td>530</td>
</tr>
<tr>
<td>12.7</td>
<td>Robust factorial ANOVA</td>
<td>534</td>
</tr>
<tr>
<td>12.8</td>
<td>Calculating effect sizes</td>
<td>542</td>
</tr>
<tr>
<td>12.9</td>
<td>Reporting the results of two-way ANOVA</td>
<td>544</td>
</tr>
<tr>
<td></td>
<td>What have I discovered about statistics?</td>
<td>546</td>
</tr>
</tbody>
</table>
13 Repeated-measures designs (GLM 4) 549

13.1. What will this chapter tell me? 549

13.2. Introduction to repeated-measures designs 550

13.2.1. The assumption of sphericity 551

13.2.2. How is sphericity measured? 551

13.2.3. Assessing the severity of departures from sphericity 552

13.2.4. What is the effect of violating the assumption of sphericity? 552

13.2.5. What do you do if you violate sphericity? 554

13.3. Theory of one-way repeated-measures ANOVA 554

13.3.1. The total sum of squares (SS_T) 557

13.3.2. The within-participant sum of squares (SS_w) 558

13.3.3. The model sum of squares (SS_m) 559

13.3.4. The residual sum of squares (SS_r) 560

13.3.5. The mean squares 560

13.3.6. The F-ratio 560

13.3.7. The between-participant sum of squares 561

13.4. One-way repeated-measures designs using R 561

13.4.1. Packages for repeated measures designs in R 561

13.4.2. General procedure for repeated-measures designs 562

13.4.3. Repeated-measures ANOVA using R Commander 563

13.4.4. Entering the data 563

13.4.5. Exploring the data 565

13.4.6. Choosing contrasts 568

13.4.7. Analysing repeated measures: two ways to skin a .dat 569

13.4.8. Robust one-way repeated-measures ANOVA 576

13.5. Effect sizes for repeated-measures designs 580

13.6. Reporting one-way repeated-measures designs 581

13.7. Factorial repeated-measures designs 583

13.7.1. Entering the data 584

13.7.2. Exploring the data 586

13.7.3. Setting contrasts 588

13.7.4. Factorial repeated-measures ANOVA 589

13.7.5. Factorial repeated-measures designs as a GLM 594

13.7.6. Robust factorial repeated-measures ANOVA 599

13.8. Effect sizes for factorial repeated-measures designs 599

13.9. Reporting the results from factorial repeated-measures designs 600

What have I discovered about statistics? 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
14  Mixed designs (GLM 5) 604
14.1. What will this chapter tell me? 604
14.2. Mixed designs 605
14.3. What do men and women look for in a partner? 606
14.4. Entering and exploring your data 606
  14.4.1. Packages for mixed designs in R 606
  14.4.2. General procedure for mixed designs 608
  14.4.3. Entering the data 608
  14.4.4. Exploring the data 610
14.5. Mixed ANOVA 613
14.6. Mixed designs as a GLM 617
  14.6.1. Setting contrasts 617
  14.6.2. Building the model 619
  14.6.3. The main effect of gender 622
  14.6.4. The main effect of looks 623
  14.6.5. The main effect of personality 624
  14.6.6. The interaction between gender and looks 625
  14.6.7. The interaction between gender and personality 628
  14.6.8. The interaction between looks and personality 630
  14.6.9. The interaction between looks, personality and gender 635
  14.6.10. Conclusions 639
14.7. Calculating effect sizes 640
14.8. Reporting the results of mixed ANOVA 641
14.9. Robust analysis for mixed designs 643

What have I discovered about statistics? 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? 653
15.2. When to use non-parametric tests 654
15.3. Packages used in this chapter 655
15.4. Comparing two independent conditions: the Wilcoxon rank-sum test 655
  15.4.1. Theory of the Wilcoxon rank-sum test 655
  15.4.2. Inputting data and provisional analysis 659
  15.4.3. Running the analysis using R Commander 661
  15.4.4. Running the analysis using R 662
  15.4.5. Output from the Wilcoxon rank-sum test 664
  15.4.6. Calculating an effect size 664
  15.4.7. Writing the results 666
15.5. Comparing two related conditions: the Wilcoxon signed-rank test
  15.5.1. Theory of the Wilcoxon signed-rank test
  15.5.2. Running the analysis with R Commander
  15.5.3. Running the analysis using R
  15.5.4. Wilcoxon signed-rank test output
  15.5.5. Calculating an effect size
  15.5.6. Writing the results

15.6. Differences between several independent groups: the Kruskal–Wallis test
  15.6.1. Theory of the Kruskal–Wallis test
  15.6.2. Inputting data and provisional analysis
  15.6.3. Doing the Kruskal–Wallis test using R Commander
  15.6.4. Doing the Kruskal–Wallis test using R
  15.6.5. Output from the Kruskal–Wallis test
  15.6.6. Post hoc tests for the Kruskal–Wallis test
  15.6.7. Testing for trends: the Jonckheere–Terpstra test
  15.6.8. Calculating an effect size
  15.6.9. Writing and interpreting the results

15.7. Differences between several related groups: Friedman’s ANOVA
  15.7.1. Theory of Friedman’s ANOVA
  15.7.2. Inputting data and provisional analysis
  15.7.3. Doing Friedman’s ANOVA in R Commander
  15.7.4. Friedman’s ANOVA using R
  15.7.5. Output from Friedman’s ANOVA
  15.7.6. Post hoc tests for Friedman’s ANOVA
  15.7.7. Calculating an effect size
  15.7.8. Writing and interpreting the results

What have I discovered about statistics?

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

16 Multivariate analysis of variance (MANOVA)

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

16.5.2. Choosing a test statistic 718
16.5.3. Follow-up analysis 719
16.6. MANOVA using R 719
16.6.1. Packages for factorial ANOVA in R 719
16.6.2. General procedure for MANOVA 720
16.6.3. MANOVA using R Commander 720
16.6.4. Entering the data 720
16.6.5. Exploring the data 722
16.6.6. Setting contrasts 728
16.6.7. The MANOVA model 728
16.6.8. Follow-up analysis: univariate test statistics 731
16.6.9. Contrasts 732
16.7. Robust MANOVA 733
16.8. Reporting results from MANOVA 737
16.9. Following up MANOVA with discriminant analysis 738
16.10. Reporting results from discriminant analysis 743
16.11. Some final remarks 743
16.11.1. The final interpretation 743
16.11.2. Univariate ANOVA or discriminant analysis? 745

What have I discovered about statistics? 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? 749
17.2. When to use factor analysis 750
17.3. Factors 751
17.3.1. Graphical representation of factors 752
17.3.2. Mathematical representation of factors 753
17.3.3. Factor scores 755
17.3.4. Choosing a method 758
17.3.5. Communality 759
17.3.6. Factor analysis vs. principal components analysis 760
17.3.7. Theory behind principal components analysis 761
17.3.8. Factor extraction: eigenvalues and the scree plot 762
17.3.9. Improving interpretation: factor rotation 764
17.4. Research example 767
17.4.1. Sample size 769
17.4.2. Correlations between variables 770
17.4.3. The distribution of data 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 772
17.6.2. Initial preparation and analysis 772
17.6.3. Factor extraction using R 778
17.6.4. Rotation 788
17.6.5. Factor scores 793
17.6.6. Summary 795
17.7. How to report factor analysis 795
17.8. Reliability analysis 797
  17.8.1. Measures of reliability 797
  17.8.2. Interpreting Cronbach’s α (some cautionary tales …) 799
  17.8.3. Reliability analysis with R Commander 800
  17.8.4. Reliability analysis using R 800
  17.8.5. Interpreting the output 801
17.9. Reporting reliability analysis 806
What have I discovered about statistics? 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? 812
18.2. Packages used in this chapter 813
18.3. Analysing categorical data 813
18.4. Theory of analysing categorical data 814
  18.4.1. Pearson’s chi-square test 814
  18.4.2. Fisher’s exact test 816
  18.4.3. The likelihood ratio 816
  18.4.4. Yates’s correction 817
18.5. Assumptions of the chi-square test 818
18.6. Doing the chi-square test using R 818
  18.6.1. Entering data: raw scores 818
  18.6.2. Entering data: the contingency table 819
  18.6.3. Running the analysis with R Commander 820
  18.6.4. Running the analysis using R 821
  18.6.5. Output from the CrossTable() function 822
  18.6.6. Breaking down a significant chi-square test with standardized residuals 825
  18.6.7. Calculating an effect size 826
  18.6.8. Reporting the results of chi-square 827
18.7. Several categorical variables: loglinear analysis 829
  18.7.1. Chi-square as regression 829
  18.7.2. Loglinear analysis 835
18.8. Assumptions in loglinear analysis 837
18.9. Loglinear analysis using R 838
  18.9.1. Initial considerations 838
  18.9.2. Loglinear analysis as a chi-square test 840
  18.9.3. Output from loglinear analysis as a chi-square test 843
19.8. How to report a multilevel model

What have I discovered about statistics?

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

Epilogue: life after discovering statistics
Troubleshooting R
Glossary
Appendix
A.1. Table of the standard normal distribution
A.2. Critical values of the t-distribution
A.3. Critical values of the F-distribution
A.4. Critical values of the chi-square distribution

References
Index

Functions in R
Packages in R
Karma Police, arrest this man, he talks in maths, he buzzes like a fridge, he’s like a detuned radio.


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 under-mining 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, occasionally 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