Scientific Research and Methodology

PETER K DUNN



Peter K. Dunn

Scientific Research and Methodology

An introduction to quantitative research and statistics in science, engineering and health

Contents

Preface **xvii**

1 Research: An introduction **1**
1.1 How do we know what we know? . . . 1
1.2 The purpose of research . . . 2
1.3 Evidence-based research . . . 2
1.4 Using software in research . . . 3
1.5 An example: Research in action . . . 4
1.6 The components of research . . . 4
1.7 Types of research . . . 5
1.8 Quick review questions . . . 7
1.9 Exercises . . . 7

I Asking research questions **9**

2 Research questions **11**
2.1 Introduction . . . 11
2.2 Conceptual and operational definitions . . . 12
2.3 Elements of RQs . . . 15
2.3.1 The Population . . . 15
2.3.2 The Outcome . . . 18
2.3.3 The Comparison or Connection . . . 18
2.3.4 The Intervention . . . 19
2.4 Types of RQs . . . 21
2.4.1 Descriptive RQs (PO) . . . 21
2.4.2 Relational RQs (POC) . . . 22
2.4.3 Interventional RQs (POCI) . . . 23
2.4.4 Comparing the three levels of RQs . . . 24
2.5 Two approaches to RQs . . . 25
2.5.1 Estimation: Confidence intervals questions . . . 26
2.5.2 Making decisions: Hypothesis testing questions . . . 26
2.6 Writing good RQs . . . 27
2.7 Writing RQs: An example . . . 29
2.8 Variables: From populations to individuals . . . 30
2.9 Units of observation and units of analysis . . . 33
2.10 Preparing software for data entry . . . 36
2.11 Summary . . . 37
2.12 Quick review questions . . . 38
2.13 Exercises . . . 38

II Research design 41

3 Types of study designs 43
3.1 Three types of study designs . . . 43
3.2 Descriptive studies . . . 44
3.3 Observational studies . . . 45
3.3.1 Retrospective studies . . . 46
3.3.2 Prospective studies . . . 46
3.3.3 Cross-sectional studies . . . 46
3.4 Experimental studies . . . 47
3.4.1 True experimental studies . . . 48
3.4.2 Quasi-experimental studies . . . 49
3.5 Comparing study designs . . . 49
3.6 External and internal validity . . . 50
3.7 The importance of design . . . 51
3.8 Summary . . . 52
3.9 Quick review questions . . . 52
3.10 Exercises . . . 52

4 Ethics in research 55
4.1 Ethical guidelines . . . 55
4.2 Common ethical issues . . . 56
4.3 Academic integrity . . . 57
4.3.1 Collusion . . . 57
4.3.2 Fraud . . . 57
4.3.3 Reproducible research . . . 58
4.4 Summary . . . 59
4.5 Quick review questions . . . 59
4.6 Exercises . . . 60

5 External validity: Sampling 61
5.1 The idea of sampling . . . 61
5.2 Precision and accuracy . . . 63
5.3 Types of sampling . . . 64
5.3.1 Random sampling methods . . . 65
5.3.2 Non-random sampling methods . . . 65
5.4 Simple random sampling . . . 67
5.5 Systematic sampling . . . 68
5.6 Stratified sampling . . . 70
5.7 Cluster sampling . . . 71
5.8 Multistage sampling . . . 72
5.9 Representative sampling . . . 73
5.10 Bias in selecting samples . . . 75
5.11 Final example . . . 77
5.12 Summary . . . 79
5.13 Quick review questions . . . 79
5.14 Exercises . . . 79

6 Overview of internal validity **81**
6.1 Introduction . . . 81
6.2 The explanatory variable and variation in the response . . . 82
6.3 Extraneous variables and variation in the response . . . 83
6.4 Study design and variation in the response . . . 86
6.5 Chance and variation in the response . . . 86
6.6 Summary . . . 87
6.7 Quick review questions . . . 87
6.8 Exercises . . . 88

7 Internal validity and experimental studies **91**
7.1 Introduction . . . 91
7.2 Managing confounding . . . 93
7.2.1 Restrictions . . . 95
7.2.2 Blocking . . . 95
7.2.3 Analysis . . . 95
7.2.4 Random allocation . . . 96
7.3 Random allocation vs random sampling . . . 97
7.4 Carry-over effect and washout periods . . . 97
7.5 Hawthorne effect and blinding individuals . . . 99
7.6 Observer effect and blinding researchers . . . 100
7.7 Placebo effect and using controls . . . 101
7.8 Comments on describing blinding . . . 103
7.9 Design issues: Overview . . . 104
7.10 Summary . . . 106
7.11 Quick review questions . . . 106
7.12 Exercises . . . 107

8 Internal validity and observational studies **109**
8.1 Introduction . . . 109
8.2 Managing confounding . . . 110
8.2.1 Restrictions . . . 110
8.2.2 Blocking . . . 111
8.2.3 Analysis . . . 111
8.2.4 Random allocation . . . 111
8.3 Carry-over effect and washout periods . . . 113
8.4 Hawthorne effect and blinding individuals . . . 114
8.5 Observer effect and blinding researchers . . . 114
8.6 Placebo effect and using controls . . . 115
8.7 Summary . . . 116
8.8 Quick review questions . . . 116
8.9 Exercises . . . 117

9 Identifying study limitations **119**
9.1 Introduction . . . 119
9.2 Limitations: External validity . . . 120
9.3 Limitations: Internal validity . . . 121
9.4 Limitations: Ecological validity . . . 122

9.5 Summary . . . 123
9.6 Quick review questions . . . 123
9.7 Exercises . . . 124

III Collecting data 127

10 Procedures for collecting data 129
10.1 Protocols . . . 129
10.2 Collecting data using surveys . . . 131
10.2.1 Asking survey questions . . . 131
10.2.2 Online and paper surveys . . . 134
10.3 Summary . . . 134
10.4 Quick review questions . . . 134
10.5 Exercises . . . 135

IV Describing and summarising data 137

11 Describing data 139
11.1 Quantitative and qualitative data . . . 139
11.1.1 Quantitative data: Discrete and continuous data . . . 140
11.1.2 Qualitative data: Nominal and ordinal data . . . 141
11.2 Describing data in jamovi and SPSS . . . 143
11.2.1 Using jamovi . . . 144
11.2.2 Using SPSS . . . 144
11.3 Summary . . . 145
11.4 Quick revision questions . . . 146
11.5 Exercises . . . 146

12 Graphical summaries of data 149
12.1 Introduction . . . 149
12.2 One quantitative variable . . . 150
12.2.1 Stem-and-leaf plots . . . 150
12.2.2 Dot charts (quantitative data) . . . 152
12.2.3 Histograms . . . 153
12.2.4 Describing the distribution . . . 155
12.3 One qualitative variable . . . 156
12.3.1 Dot charts (qualitative data) . . . 157
12.3.2 Bar charts . . . 157
12.3.3 Pie charts . . . 158
12.3.4 Comparing pie charts and bar charts . . . 159
12.3.5 Is a graph needed? . . . 161
12.4 One qualitative variable and one quantitative variable . . . 161
12.4.1 Back-to-back stem-and-leaf . . . 161
12.4.2 2-D dot charts . . . 162
12.4.3 Boxplots . . . 162
12.5 Two quantitative variables . . . 166
12.6 Two qualitative variables . . . 167
12.6.1 Stacked bar charts . . . 168

12.6.2 Side-by-side bar charts . . . 168
12.6.3 Dot charts . . . 169
12.6.4 Other variations . . . 169
12.7 Comparing 2-D and 3-D graphs . . . 171
12.8 Other types of graphs . . . 172
12.8.1 Geographic plots . . . 172
12.8.2 Case-profile plots . . . 173
12.8.3 Histogram of differences . . . 174
12.8.4 Time plots . . . 176
12.9 Notes on constructing graphs . . . 177
12.10Case Study: The NHANES data . . . 179
12.11Summary . . . 181
12.12Quick review questions . . . 182
12.13Exercises . . . 183

13 Numerical summaries: quantitative data 187
13.1 Introduction . . . 187
13.2 Computing the average value . . . 189
13.2.1 Computing the average: The mean . . . 190
13.2.2 Computing the average: The median . . . 192
13.2.3 Which average to use? . . . 193
13.3 Computing the variation . . . 195
13.3.1 Computing the variation: Range . . . 195
13.3.2 Computing the variation: Standard deviation . . . 196
13.3.3 Computing the variation: IQR . . . 198
13.3.4 Computing the variation: Percentiles . . . 200
13.3.5 Which measure of variation to use? . . . 201
13.4 Describing shape . . . 201
13.5 Identifying outliers . . . 201
13.5.1 Bell-shaped (normal) distributions and the 68–95–99.7 rule . . . 202
13.5.2 The standard deviation rule for identifying outliers . . . 205
13.5.3 The IQR rule for identifying outliers . . . 205
13.5.4 When to use which rule? . . . 206
13.6 Compiling tables of numerical summary information . . . 207
13.7 Observing relationships: The NHANES study . . . 208
13.8 Summary . . . 211
13.9 Quick review questions . . . 212
13.10Exercises . . . 212

14 Numerical summaries: qualitative data 215
14.1 Proportions and percentages . . . 215
14.1.1 Introduction . . . 215
14.1.2 Overall proportions and percentages . . . 216
14.1.3 Row proportions and percentages . . . 217
14.1.4 Column proportions and percentages . . . 217
14.1.5 Example: Large kidney stones . . . 217
14.2 Odds . . . 219
14.3 Odds ratios . . . 221

14.4 Observing relationships . . . 222
14.5 Example: Skipping breakfast . . . 223
14.6 Case Study: The NHANES data . . . 224
14.7 Summary . . . 226
14.8 Quick revision questions . . . 226
14.9 Exercises . . . 227

V Tools for answering RQs 229

15 Making decisions: An introduction 231
15.1 Introduction . . . 231
15.2 The need for making decisions . . . 232
15.3 How decisions are made . . . 234
15.4 Making decisions in research . . . 236
15.4.1 Assumption about the population parameter . . . 236
15.4.2 Expectations of sample statistics . . . 237
15.4.3 Observations about our sample . . . 238
15.5 Tools for describing sampling variation . . . 239
15.6 Summary . . . 239
15.7 Quick review questions . . . 240
15.8 Exercises . . . 240

16 Probability 241
16.1 Introduction . . . 241
16.2 Classical approach . . . 243
16.3 Relative frequency approach . . . 245
16.4 Subjective approach . . . 246
16.5 Independence . . . 247
16.6 Summary . . . 247
16.7 Quick review questions . . . 248
16.8 Exercises . . . 248

17 Distributions and models 251
17.1 Introduction . . . 251
17.2 Distributions: An example . . . 252
17.3 Normal distributions . . . 253
17.4 Standardising (z-scores) . . . 254
17.5 Approximating areas using the 68–95–99.7 rule . . . 256
17.6 Exact areas from normal distributions . . . 257
17.6.1 Using the online tables . . . 257
17.6.2 Using the hard-copy tables . . . 258
17.7 Comparing exact and approximate areas . . . 258
17.8 Examples using z-scores . . . 259
17.9 Unstandardising: Working backwards . . . 262
17.9.1 Using the hard-copy tables . . . 263
17.9.2 Using the online tables . . . 263
17.10Summary . . . 265
17.11Quick revision questions . . . 265

17.12Exercises 266

18 Sampling variation **269**
18.1 Introduction 269
18.2 Sample proportions have a distribution 270
18.3 Sample means have a distribution 272
18.4 Standard errors 273
18.5 Standard deviation vs. standard error 274
18.6 Summary 275
18.7 Quick review questions 275
18.8 Exercises 276

VI Analysis: Confidence intervals **277**

19 Introducing confidence intervals **279**

20 Confidence intervals for one proportion **281**
20.1 Sampling distribution: Known proportion 281
20.2 Sampling intervals: Known proportion 285
20.3 Sampling distribution: Unknown proportion 286
20.4 Confidence intervals: Unknown proportion 288
20.5 Interpretation of a CI 291
20.6 Statistical validity conditions 291
20.7 Summary: Finding a CI for p 294
20.8 Estimating sample sizes: one proportion 295
20.9 Example: Female coffee drinkers 295
20.10Quick review questions 296
20.11Exercises 296

21 More about forming CIs **299**
21.1 General comments 299
21.2 Interpretation of a CI 300
21.3 Validity and confidence intervals 301
21.4 Quick revision exercises 302
21.5 Exercises 303

22 Confidence intervals for one mean **305**
22.1 Sampling distribution: One mean with population standard deviation known 305
22.2 Sampling distribution: One mean with population standard deviation unknown 307
22.3 Confidence intervals: One mean 308
22.4 Statistical validity conditions: One mean 310
22.5 Example: NHANES 312
22.6 Example: Cadmium in peanuts 312
22.7 Estimating sample sizes: one mean 313
22.8 Quick review questions 314
22.9 Exercises 314

23 Confidence intervals for mean differences (paired data) **317**
23.1 Mean differences 317

23.2 Mean differences: An example . . . 318
23.3 Notation: Mean differences . . . 320
23.4 Graphical summaries: Mean differences . . . 320
23.5 Numerical summaries: Mean differences . . . 320
23.6 Sampling distribution: Means differences . . . 321
23.7 Confidence intervals: Mean differences . . . 322
23.8 Using software: CIs for mean differences . . . 324
23.9 Statistical validity conditions: Mean differences . . . 325
23.10 Example: Blood pressure . . . 326
23.11 Quick review questions . . . 328
23.12 Exercises . . . 329

24 Confidence intervals for two independent means 331
24.1 Means of two independent samples . . . 331
24.2 Graphical summary: Two independent means . . . 332
24.3 Notation: Two independent means . . . 332
24.4 Numerical summary: Two independent means . . . 333
24.5 Sampling distribution: Two independent means . . . 335
24.6 Confidence intervals: Two independent means . . . 336
24.7 Using software: CIs for two independent means . . . 338
24.8 Statistical validity conditions: Two independent means . . . 339
24.9 Error bar charts . . . 339
24.10 Example: Health Promotion services . . . 342
24.11 Example: Face-plant study . . . 343
24.12 Quick review questions . . . 345
24.13 Exercises . . . 345

25 Confidence intervals for odds ratios 349
25.1 Introduction: Odds ratios . . . 349
25.2 Numerical and graphical summaries: Comparing odds . . . 351
25.3 Sampling distribution: Comparing odds . . . 352
25.4 Confidence intervals: Comparing odds . . . 352
25.5 Statistical validity conditions: Comparing odds . . . 355
25.6 Example: Pet birds . . . 356
25.7 Example: B12 deficiency . . . 358
25.8 Quick review questions . . . 359
25.9 Exercises . . . 361

VII Analysis: Hypothesis testing 365

26 Introducing hypothesis tests 367

27 Hypothesis tests for one mean 369
27.1 Introduction: Body temperatures . . . 369
27.2 Hypotheses and notation: One mean . . . 370
27.3 Sampling distribution: One mean . . . 370
27.4 The test statistic and t-scores: One mean . . . 372
27.5 P-values: One mean . . . 374

27.5.1 Approximating P-values using the 68–95–99.7 rule 374
27.5.2 Finding P-values using sofware 374
27.6 Making decisions with P-values 376
27.7 Communicating results: One mean 377
27.8 Hypothesis testing for one mean: A summary 378
27.9 Statistical validity conditions: One mean 379
27.10Example: Recovery times 380
27.11Summary 383
27.12Quick review questions 383
27.13Exercises 384

28 More about hypothesis testing 387
28.1 Introduction 387
28.2 About hypotheses and assumptions 388
28.2.1 Null hypotheses 388
28.2.2 Alternative hypotheses 389
28.3 About sampling distributions and expectations 390
28.4 About observations and the test statistic 390
28.5 About finding P-values 391
28.6 About interpreting P-values 391
28.7 About writing conclusions 393
28.8 About practical importance and statistical significance 393
28.9 Validity and hypothesis testing 394
28.10Summary 395
28.11Quick review questions 396
28.12Exercises 396

29 Hypothesis tests for the mean difference (paired data) 399
29.1 Introduction: Insulation 399
29.2 Hypotheses and notation: Mean differences 400
29.3 Sampling distribution: Mean differences 401
29.4 The test statistic: Mean differences 402
29.5 P-values: Mean differences 403
29.6 Conclusions: Mean differences 403
29.7 Statistical validity conditions: Mean differences 405
29.8 Example: Endangered species 406
29.9 Example: Blood pressure 407
29.10Summary 410
29.11Quick review questions 410
29.12Exercises 410

30 Hypothesis tests for means of two independent groups 415
30.1 Introduction: Reaction times 415
30.2 Hypotheses and notation: Two independent means 416
30.3 Sampling distribution: Two independent means 417
30.4 The test statistic: Two independent means 418
30.5 P-values: Two independent means 419
30.6 Conclusions: Two independent means 419

30.7 Statistical validity conditions: Two independent means . . . 419
30.8 Example: Health Promotion services . . . 421
30.9 Example: Face-plant study . . . 422
30.10Summary . . . 424
30.11Quick review questions . . . 424
30.12Exercises . . . 425

31 Hypothesis tests for comparing odds 429
31.1 Introduction: Meals on-campus . . . 429
31.2 Hypotheses and notation: Comparing odds . . . 431
31.3 Expected values: Comparing odds . . . 432
31.4 The test statistic: Comparing odds . . . 433
31.5 P-values: Comparing odds . . . 435
31.6 Conclusions: Comparing odds . . . 435
31.7 Statistical validity conditions . . . 436
31.8 Example: Pet birds . . . 437
31.9 Example: B12 deficiency . . . 439
31.10Example: Kerbside dumping . . . 442
31.11Summary . . . 446
31.12Quick review questions . . . 446
31.13Exercises . . . 447

32 Selecting a hypothesis testing 455

VIII Connection RQs: Regression and Correlation 457

33 Relationships between two quantitative variables 459
33.1 Introduction: The red deer data . . . 459
33.2 Two quantitative variables: Graphical summaries . . . 459
33.3 Understanding scatterplots . . . 460
33.4 Summary . . . 461
33.5 Quick review questions . . . 462
33.6 Exercises . . . 464

34 Correlation 467
34.1 Correlation coefficients . . . 467
34.2 Using software . . . 470
34.3 R-squared (R^2) . . . 471
34.4 Hypothesis testing . . . 472
34.4.1 Introduction . . . 472
34.4.2 Hypothesis testing details . . . 472
34.4.3 Statistical validity conditions . . . 473
34.5 Example: Removal efficiency . . . 476
34.6 Summary . . . 477
34.7 Quick review questions . . . 478
34.8 Exercises . . . 478

35 Regression 481

35.1 Introduction . . . 481
35.2 Linear equations: A review . . . 482
35.3 Regression using software . . . 485
35.4 Regression for predictions . . . 486
35.5 Regression for understanding . . . 487
35.5.1 The meaning of b_0 . . . 487
35.5.2 The meaning of b_1 . . . 487
35.6 Hypothesis testing . . . 488
35.6.1 Introduction . . . 488
35.6.2 Hypotheses: Assumption . . . 489
35.6.3 Sampling distribution: Expectation . . . 489
35.6.4 The test statistic: Observation . . . 490
35.6.5 P-value: Consistency with assumption . . . 490
35.7 Confidence intervals . . . 492
35.8 Statistical validity conditions . . . 493
35.9 Example: Obstructive sleep apnoea . . . 494
35.10Example: Food digestibility . . . 496
35.11Summary . . . 497
35.12Quick review questions . . . 499
35.13Exercises . . . 499

IX Reporting, writing and reading research 505

36 Reading research 507
36.1 Introduction . . . 507
36.2 Example 1: Reading research . . . 508
36.3 Example 2: Reading research . . . 509
36.4 Exercises . . . 511

37 Writing research 515
37.1 Introduction . . . 515
37.2 General tips . . . 516
37.3 Article structure . . . 517
37.4 Writing scientifically: Title . . . 518
37.5 Writing scientifically: Abstract . . . 518
37.6 Writing scientifically: Introduction . . . 520
37.7 Writing scientifically: Materials and methods . . . 520
37.8 Writing scientifically: Results . . . 521
37.9 Writing scientifically: Discussion and conclusion . . . 521
37.10Writing carefully: Lexically ambiguous words . . . 521
37.11Constructing tables . . . 523
37.12Constructing figures and graphs . . . 523
37.13Other elements . . . 524
37.14Style . . . 524
37.15Plagiarism . . . 525
37.16Final comments . . . 526
37.17Quick review questions . . . 527
37.18Exercises . . . 527

X Appendices 531

Appendix 533

A Appendix: Data sets 533

B Appendix: Tables 535
B.1 Random numbers . . . 535
B.2 Normal distribution: negative z-values probabilities . . . 537
B.3 Normal distribution: positive z-values probabilities . . . 538

C Appendix: Symbols, formulas, statistics and parameters 539

D Appendix: Answers to end-of-chapter exercises 541
D.1 Answers: Introduction . . . 542
D.2 Answers: RQs . . . 542
D.3 Answers: Research designs . . . 543
D.4 Answers: Ethics . . . 544
D.5 Answers: Sampling . . . 544
D.6 Answers: Overview of internal validity . . . 545
D.7 Answers: Designing experimental studies . . . 546
D.8 Answers: Designing observational studies . . . 546
D.9 Answers: Interpretation . . . 547
D.10 Answers: Data collection . . . 547
D.11 Answers: Describing variables . . . 547
D.12 Answers: Graphs . . . 548
D.13 Answers: Numerical summaries for quantitative data . . . 551
D.14 Answers: Numerical summaries for qualitative data . . . 551
D.15 Answers: Making decisions . . . 552
D.16 Answers: Probability . . . 552
D.17 Answers: Sampling distributions . . . 553
D.18 Answers: Sampling variation . . . 554
D.19 Answers: CIs for one proportion . . . 554
D.20 Answers: More about formings CIs . . . 555
D.21 Answers: CIs for one mean . . . 555
D.22 Answers: CIs for paired data . . . 556
D.23 Answers: CIs for two means . . . 557
D.24 Answers: CIs for odds ratios . . . 558
D.25 Answers: Tests for one mean . . . 559
D.26 Answers: More about hypothesis tests . . . 560
D.27 Answers: Tests for paired means . . . 561
D.28 Answers: Tests for two means . . . 561
D.29 Answers: Tests for odds ratios . . . 563
D.30 Answers: Relationships between two quantitative variables . . . 564
D.31 Answers: Correlation . . . 564
D.32 Answers: Regression . . . 565
D.33 Answers: Reading research . . . 566
D.34 Answers: Writing research . . . 566

E Appendix: Checklists **569**
E.1 A checklist for good scientific graphics 569
E.2 A checklist for good scientific tables 569

F Appendix: Image credits **571**

G Glossary **573**

References **581**

Index **603**

Preface

This book is an introduction to quantitative research in the scientific and health disciplines. The whole research process is introduced, from asking a research question to analysis and reporting of the data. The focus, however, is on the analysis of data.

Supporting documents

To support this textbook, the following are also available:

- A Tutorial Book[1]; and
- A Software Support Book[2] with brief instruction for using jamovi and SPSS to perform *some* of the analyses in this book.

These books are both freely available online.

The data sets used in this book

Almost every data set used in this book is a real data set. Many are available electronically so that you can download them and work with them in statistical software. These are provided in the online version of this book (Appendix A). Some data sets are taken from Smyth (2010).

Statistical software

Most of this book can be read without relying on any specific statistical software. However, some parts explicitly mention and refer to jamovi[3] (The jamovi Project) and/or SPSS[4] (IBM Corp 2016). jamovi is *free* and is like (but not exactly the same as) SPSS. From the jamovi homepage (sic)[5]:

[1] https://bookdown.org/pkaldunn/SRM-tutorials/
[2] https://bookdown.org/pkaldunn/SRM-software/
[3] https://www.jamovi.org/
[4] https://www.ibm.com/products/spss-statistics
[5] https://www.jamovi.org/

> jamovi is a new '3rd generation' statistical spreadsheet. designed from the ground up to be easy to use, jamovi is a compelling alternative to costly statistical products such as SPSS and SAS.
> — https://www.jamovi.org/

Icons used on this book

The icons used in this book have meanings; for example:



These chunks introduce the objectives for the chapters of the book.



These chunks highlight common mistakes or warnings, about a particular concept or about using a formula.



These chunks refer to text that is relevant to using software (jamovi or SPSS) or a calculator.



These chunks offer helpful information.



These chunks indicate how certain symbols and terms are pronounced.

Who can use this book?

This textbooks is **free** for anyone to use: There is no charge for students, instructors or institutions.

Although it is not essential, an email to the author (explaining how the textbook is being used, who is using the textbook, and your thoughts on the textbook) would be appreciated: pdunn2 <at> usc.edu.au.

How this book was made

This book was made using **R** (R Core Team 2018), and the **bookdown** package (Xie 2016), which is based on Markdown[6] syntax, using **knitr** (Xie 2015).

Numerous other **R** packages were used too:

- The diagrams were made in base **R**, or using the **diagram** package (Soetaert 2017).
- The animations in the html version were made using the **animation** package (Xie 2013), and the stills for the PDF version captured by the **webshot** package (Chang 2018).
- The **gifski** package was used to create animations in the online version (Ooms 2018).
- The **kableExtra** package was used for nicer tables (Zhu 2018).
- The maps of Australia were generated using the **oz** package (Venables and Hornik 2016) and plotted using the **ggplot2** package (Wickham 2016).
- The display of some data tables in the online version use the **DT** package (Xie et al. 2018).
- Some plots use the **plotrix** package (Lemon 2006).
- The NHANES data is provided by the **NHANES** package (Pruim 2015).
- Some data are from the **GLMsData** package (Dunn and Smyth 2017).
- The **scales** package is used to rescale data (Wickham 2018).
- The carousels in the online version (for example, Sect. 33.3) are made using the **slickR** package (Sidi 2018).
- The **dygraphs** package (Vanderkam et al. 2018) is used to make interactive graphs (for example, Sect. 12.8.4) for the online version.
- The **plotly** package (Sievert 2018) is used to make some interactive graphs (for example, in Sect. 12.2.3) for the online version.
- The **dplyr** package is used in some data manipulations (Wickham et al. 2019).
- The **viridis** package is used for some colour specifications (Garnier 2018) to make colours easier for those with colour-blindess to distinguish colours, and for better greyscale printing.
- The **webex** package was used to create the interactive web exercises (Barr and DeBruine 2019).
- The **plotfunctions** package was used to add images to existing plots (van Rij 2020)

All of this software is *free* and open source.

Other resources used include:

- The quizzes are embedded using H5P[7] iframes.
- Icons are from **iconmonstr**[8] and are freely available.
- The images of the cards used in Sect. 15.2 are from `https://code.google.com/archive/p/vector-playing-cards/`, and are in the public domain.
- The text folding (in the html version) was implemented by adapting advice from StackOverflow[9].

[6] **https://en.wikipedia.org/wiki/Markdown**
[7] **https://h5p.org**
[8] **https://iconmonstr.com/**
[9] **https://stackoverflow.com/questions/53810294/folding-general-text-in-r-bookdown**

- The images of dice used in, for example, Sect. 16.2, are from https://www.clipartkey.com/, and are free.
- The cover for the book was made using a free image using Canva[10].
- The images used in the online book are free (and used according to their guidelines), as listed in Appendix F.

Learning Outcomes

In this book, you will learn to:

- Develop quantitative research questions and testable hypotheses.
- Design quantitative studies to answer simple quantitative research questions.
- Select and produce appropriate graphical, numerical and statistical analyses.
- Select, apply and interpret the results of the correct statistical technique to analyse data.
- Comprehend, apply and communicate in the language of research and statistics.
- Demonstrate professional integrity in planning, interpreting and reporting the results of quantitative studies.

How to cite this book

Peter K. Dunn (2021). *Scientific Research and Methodology: An introduction to quantitative research in science and health*. https://bookdown.org/pkaldunn/Book

The CC BY-NC-SA 4.0[11] licence is applied to this textbook.

Peter K. Dunn Sippy Downs, Australia

[10] **https://www.canva.com/**
[11] **https://creativecommons.org/licenses/by-nc-sa/4.0/**

1

Research: An introduction



In this chapter, you will learn to:

- identify quantitative and qualitative research.
- identify the steps in the quantitative research process.

1.1 How do we know what we know?

Scientists once believed that all life commonly and regularly arose spontaneously from non-living matter ('spontaneous generation'); that is, any life form could be created on-demand from non-living things by following *recipes*.

J. B. van Helmont (van Helmont 1671; Latour 1989) gave this recipe:

> If a soiled shirt is placed in the opening of a vessel containing grains of wheat, the reaction of the leaven in the shirt with fumes from the wheat will, after approximately twenty-one days, transform the wheat into mice.
> — Translation of van Helmont (1671)

We now know that this isn't true... But how did van Helmont reach this conclusion?

Through *observation*. This is what he saw happen, even though his explanation was incorrect. And why was this idea rejected? Because of the *scientific process*.

Spontaneous generation was proposed after making *observations*. van Helmont then proposed a possible explanation (a *hypothesis*). This hypothesis was then rejected when evidence contradicted the hypothesis. So, a a new hypothesis was proposed and tested, based on further *evidence*. Briefly, this is the *evidence-based, scientific process*.

A more recent example of the scientific process in action is declaring cigarette smoking as harmful. As recently as 1978, the verdict on whether smoking is harmful was debated:

> ...many eminent persons, committees and commissions have unanimously concluded that lung cancer 'is almost entirely due to cigarette smoking.' I once shared this view, but having now studied the evidence in more detail and from new angles I feel unable to reach a definitive conclusion...
> — Burch (1978), p. 456

All scientific knowledge emerges in a similar way: Observations lead to hypotheses, which

are tested against the *evidence*, and the hypotheses are either rejected or temporarily accepted based on this evidence. Notice the approach: hypotheses are *rejected* when contradictory evidence emerges, but hypotheses are only ever *temporarily accepted* until contradictory evidence emerges (if ever).

Knowledge in all scientific disciplines is based on a similar process:

- How do we know the gestation length for *Gilbert's Potoroo* (Stead-Richardson et al. 2010)?
- How do we know that paracetamol eases pain (Weil et al. 2007)?
- How do we know that exercise is good for us (Curfman 1993)?
- How do we know if permeable pavement technology is effective in reducing runoff (Mullaney and Lucke 2014)?

1.2 The purpose of research

Every scientific discipline has changed in the last 10 years. Likewise, the next 10 years will also bring change. We need to know how to adapt to this change.

Every discipline changes, develops, improves, and adapts—usually through *research*. To remain current with developments in your discipline, you need to understand research, even if you will not be conducting research yourself.

Everyone in science-based disciplines must know the language, tools, concepts and ideas of research: Research is the foundation of science.

Research seeks to

> ...confirm, refute or extend previous findings, and potentially reveal new findings...
> — Anastasiadis et al. (2015), p. 410.

Scientific research formally answers questions that arise by observing the world using *data*; that is, science requires *evidence-based answers*.

While *analysis* of the data is often viewed as the hardest part of research, sometimes the hardest part is knowing *what* data to collect, and *how* to collect it (that is, the *study design*).

We study both the study design *and* the analysis of data in this book.

1.3 Evidence-based research

'Evidence-based research' refers to research conclusions based on *evidence*, rather than researchers' hunches, feelings, intuition, hopes, or traditional practice.

Research conclusions are based on **evidence**, which comes from analysing the collected **data**.

Definition 1.1 (Data)**.** Data refers to information (observations or measurements) obtained from a study, as numbers, labels, or text.

Data can come from previously-published studies (Chap. 36). Even so, studies usually leave questions unanswered, or technology develops and new ideas need to be tested, or existing ideas need to be adapted to new technologies, situations and knowledge.

In these cases, **data** comes from new evidence, through research.

Definition 1.2 (Data set)**.** A *data set* refers to a *structured* collection of data from a study.

1.4 Using software in research

Many people use spreadsheets (such as Microsoft Excel) for analysis of data in research.

Using spreadsheets requires extreme care; many extremely expensive and dangerous errors have been made due to using spreadsheets (AlTarawneh and Thorne 2017), including problems when reporting the 2020 COVID-19 pandemic[1].

Problems may emerge for many different reasons:

- Spreadsheets can *automatically change the entered data* (for example, reformatting entries as dates if the spreadsheet *thinks* the data should be a date), even when not appropriate. This has had dire consequences (Ziemann et al. 2016).
- Spreadsheets may include *formulas with errors* (Panko and Sprague Jr 1998), that are incredibly *difficult to locate* and hence fix (Galletta et al. 1996; Panko 2016; London and Slagter 2021).
- Spreadsheets *do not leave a record* of how the data have been analysed or prepared; for example, formulas can be very difficult to understand and parse. Keeping a record of the analysis, preparation of variables, and other operations with the data are part of what is called **reproducible research** (Simons and Holmes 2019). Reproducibility ensures, among other advantages, that the results can be checked by the researchers and by others.
- Excel has *bugs* (Keeling and Pavur 2004; Mélard 2014) even in very basic operations (Berger 2007; Hargreaves and McWilliams 2010). After trying to fix these bugs, sometimes they are made even worse (McCullough and Wilson 2002).

Spreadsheets can be used for research and analysis… but you must be very careful!

Many of the problems with using spreadsheets are due to human error, but spreadsheets make the errors hard to find. Some errors emerge because Excel is being used for purposes it is not really designed for (i.e., scientific analysis).

[1] https://www.zdnet.com/article/excel-errors-microsofts-spreadsheet-may-be-hazardous-to-your-health/

In this course, we will sometimes show output from the statistical software packages jamovi[2] (The jamovi Project) and SPSS[3] (IBM Corp 2016).

1.5 An example: Research in action

During 1988/1989, an unusually high number of cases of the *Legionella longbeachae* infection were observed in South Australia. Why? What could be done to prevent more instances?

The researchers wanted to identify the source of the infection. They noticed that many of those infected were regular gardeners who had handled potting mix recently. So the researchers wondered:

> ... if *L. longbeachae* infection was associated with handling of commercial potting mix.
> — O'Connor et al. (2007), p. 35

They designed a study, then gathered data (using a survey) from 100 people: 25 people *with* the *L. longbeachae* infection, and 75 similar ('matched') people *without* the infection

The researchers described and summarised their data, analysed the data, and then reached an evidence-based conclusion: the potting mix was partially responsible for the increase in infection numbers, but other factors were also involved.

The researchers then communicated their recommendations to reduce the future risks of people contracting the infection:

> Raising people's awareness of a possible health risk when using potting mix should continue in order to protect against *L. longbeachae* infection.
> — O'Connor et al. (2007), p. 39

1.6 The components of research

The research process typically follows the process in Fig. 1.1. This is not always possible or practical, and the process is not always linear (researchers may jump from step to step as necessary). Nonetheless, following this process is good practice when possible.

All these steps are discussed in this book:

- **Asking** the question: Chap. 2.
- **Designing** the study: Chaps. 3 to 9.
- **Collecting** the data: Chap. 10.
- **Describing** and **summarising** the data: Chaps. 11 to 14.

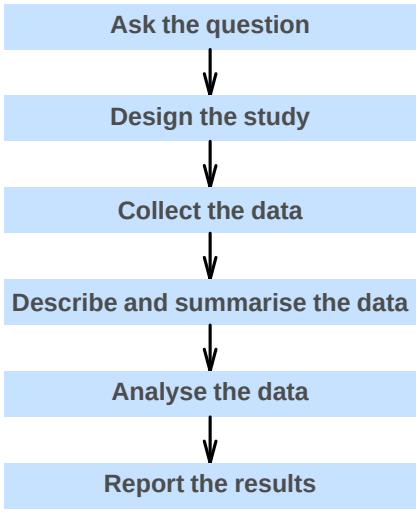


FIGURE 1.1: *The six basic steps in research*

- **Analysing** the data: Chaps. 15 to 35.
- **Reporting** the results: Chaps. 36 and 37.

1.7 Types of research

Research is a formal, evidence-based approach to learning or creating new information. Broadly speaking, the two main types of research are **qualitative** and **quantitative** research.

These two types of research are different and complementary (Table 1.1, which is a *gross generalisation*!).

TABLE 1.1: *Comparing qualitative and quantitative research*

Qualitative	**Aspect**	Quantitative
Feelings, opinions	**What**	**Objective data**
Suggest hypotheses	**Why**	Tests hypotheses
Detailed	**Conclusions**	General
Words, pictures, ...	**Data**	Numbers, statistics
Small number	**Size**	Can be large numbers
Time-consuming	**Time**	More efficient
Rarely generalisable	**Applicability**	Sometimes generalisable
Interviews, focus groups, diaries	**Examples**	Experiments, surveys measurements

Both methods have advantages and disadvantages, and both can be, and often are, used together: using a combination of qualitative and quantitative research is called **mixed methods** research.

The decision to use qualitative, quantitative or mixed methods approaches should depend on the research problem, not the skills of the researchers.

Broadly, *qualitative research* leads to a deeper understanding of what is being studied, usually about a very narrowly-defined group. Meanings, motivations, opinions or themes often emerge from qualitative research.

Broadly, *quantitative research* summarises and analyses data using *numerical* methods, such as averages and percentages. Typically, information from a subset of the population (a *sample*) is used to infer information about a larger group (a *population*) in quantitative research.

> ... quantitative data gets you the numbers to... [support] the broad general points of your research. Qualitative data brings you the details and the depth to understand their full implications.
> — SurveyMonkey website, October 2019.

Definition 1.3 (Quantitative research)**.** *Quantitative research* summarises and analyses data using numerical methods, such as producing averages and percentages.

⚠ This book focuses on *quantitative* research.

Approaches to *qualitative* research (Robson 2002) include grounded theory, action research, and ethnography. Specific examples of *quantitative* research approaches include observational studies and experimental studies.

Example 1.1 (Types of research)**.** Suppose we wish to learn about why people do or do not buy electric cars (for example, see Egbue and Long (2012)).

A *qualitative research study* might:

- Interview a small group of people who have bought *electric* cars,
- Interview another small group of people who have bought *non-electric* cars.

The researchers ask about their reasons for their car purchase.

A *quantitative research* study might survey a large number of buyers of electric and non-electric cars, and ask the buyers' age, sex, and type of car purchased.

The survey may include questions such as 'Which of the following is your *biggest* concern about buying an electric car?' and then list five reasons from which the respondents can select.

The survey responses could be analysed by numerically summarising the ages and sex of car buyers, looking for relationships between age and whether an electric car was purchased, and reporting the percentage of respondents who select each of the five options of concerns about buying electric cars.

A *mixed methods* study may initially use a qualitative approach using small groups (as described above). From this study, any common themes that emerge could be used to create a survey, with these themes as options that respondents can choose between.

The survey may also include open-ended questions such as 'What is the biggest impediment to the uptake of electric cars in Australia?' The survey can be sent to a large number of car buyers, and analysed using qualitative and quantitative methods. This may be followed up by a small focus group (a form of qualitative research) of car owners.

1.8 Quick review questions

Consider the research questions below.

Which are likely to be answered using *quantitative* research studies, and which are likely to be answered using *qualitative* research studies?

1. What percentage of the population experiences minor side-effects from this medication?
2. What is the average number of roof-top solar panels installed on domestic properties?
3. Why do people opt to purchase an electric car?

Answer:

1. The RQ requires numerical summaries of the data: 'What *percentage*...' This RQ would be answered using a **quant**itative RQ.
2. The RQ requires numerical summaries of the data: 'What is the *average*...' This RQ would be answered using a **quant**itative RQ.
3. The RQ *does not* require numerical summaries of the data. This RQ would be answered using a **qual**itative RQ.

1.9 Exercises

Selected answers are available in Sect. D.1.

Exercise 1.1. Consider the research questions 'Which of three different junctional tourniquets are quickest, on average, to apply?'

Is this RQ likely to be answered using a *quantitative* research study, or a *qualitative* research study?

Exercise 1.2. Consider the research questions 'Why do people dump rubbish in mangroves?'

Is this RQ likely to be answered using a *quantitative* research study, or a *qualitative* research study?

Part I

Asking research questions

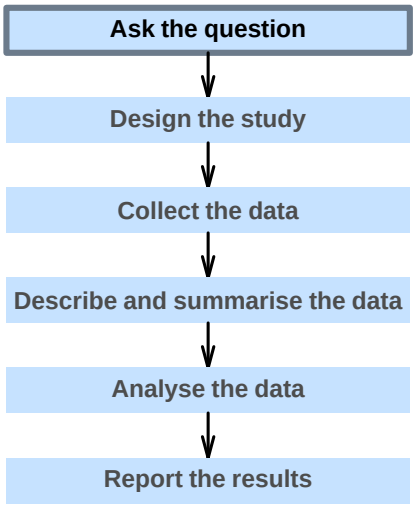
2

Research questions



In this chapter, you will learn to:

- create operational definitions.
- ask quantitative research questions.
- list, explain and give examples of the various types of quantitative research questions.
- identify estimation and decision-making research questions.
- identify the variables implied by a quantitative research question.
- identify observational or experimental studies.
- describe and identify the units of analysis and unit of observations in a study.
- communicate in the language of research and statistics.



2.1 Introduction

In research, asking clear and answerable **research questions** (RQs) is important. The data (evidence) that must be collected depends on the RQ.

In quantitative research, summarising and analysing the data typically uses numerical methods (such as averages or percentages), so the RQs must be appropriate for analysis using these methods.

For this reason, writing the RQ appropriately is important. The RQ drives all other aspects of the research (Fig. 2.1).

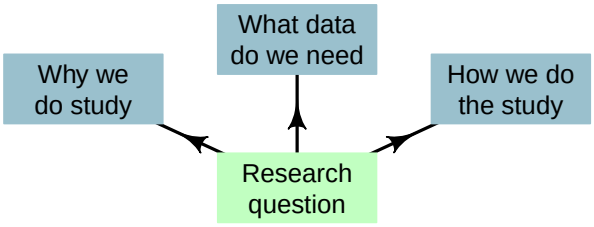


FIGURE 2.1: *The RQ drives all aspects of the study*

Defining the RQ precisely can be challenging. Studies often have an overall, broad research goal with many sub-question (which may be quantitative or qualitative).

Example 2.1 (Research questions)**.** Consider this broad research goal:

> How well are PPs (permeable pavements) working in urban areas?

This goal has many component RQs (Fig. 2.2), and each can be answered using separate studies.

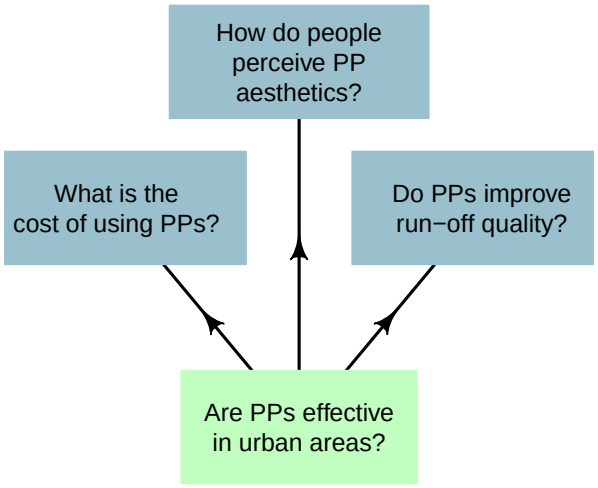


FIGURE 2.2: *A study of permeable pavements (PPs) may have many sub-questions*

2.2 Conceptual and operational definitions

Research studies usually include terms that must be carefully and precisely defined, so that others know *exactly* what has been done and there are no ambiguities. Two types of definitions can be given: *conceptual definitions* and *operational definitions*.

Loosely speaking, a *conceptual definition* explains *what* to measure or observe (what a word or a term *means* for your study), and an *operational definitions* defines exactly *how* to measure or observe it.

For example, in a study of stress in students during a university semester. A *conceptual*

definition would describe what is meant by 'stress.' An *operational definition* would describe how the 'stress' would be measured.

Definition 2.1 (Conceptual definition)**.** A *conceptual definition* articulates *what* exactly is to be measured or observed in a study.

Definition 2.2 (Operational definition)**.** An *operational definition* articulates *how* to capture (identify, create, measure, assess etc.) the value.

Sometimes the definitions themselves aren't important, provided a clear definition is given. Sometimes, commonly-accepted definitions exist, so should be used unless there is a good reason to use a different definition (for example, in criminal law, an 'adult' in Australia is someone aged 18 or over).

Sometimes, a commonly-accepted definition does not exist, so the definition being used should be clearly articulated.

Example 2.2 (Operational and conceptual definitions)**.** Players and fans have become more aware of concussions and head injuries in sport. A Conference on concussion in sport developed this *conceptual definition* (McCrory et al. 2013):

> Concussion is a brain injury and is defined as a complex pathophysiological process affecting the brain, induced by biomechanical forces. Several common features that incorporate clinical, pathologic and biomechanical injury constructs that may be utilised in defining the nature of a concussive head injury include:
>
> 1. Concussion may be caused either by a direct blow to the head, face, neck or elsewhere on the body with an "impulsive" force transmitted to the head.
> 2. Concussion typically results in the rapid onset of short-lived impairment of neurological function that resolves spontaneously. However, in some cases, symptoms and signs may evolve over a number of minutes to hours.
> 3. Concussion may result in neuropathological changes, but the acute clinical symptoms largely reflect a functional disturbance rather than a structural injury and, as such, no abnormality is seen on standard structural neuroimaging studies.
> 4. Concussion results in a graded set of clinical symptoms that may or may not involve loss of consciousness. Resolution of the clinical and cognitive symptoms typically follows a sequential course. However, it is important to note that in some cases symptoms may be prolonged.

While this is all helpful... it does not explain *how* to identify a player with concussion during a game.

Rugby decided on this *operational definition* (Raftery et al. 2016):

> ... a concussion applies with any of the following:
> 1. The presence, pitch side, of any Criteria Set 1 signs or symptoms (table 1)... [**Note**: This table includes symptoms such as 'convulsion,' 'clearly dazed,' etc.];
> 2. An abnormal post game, same day assessment...;
> 3. An abnormal 36–48 h assessment...;
> 4. The presence of clinical suspicion by the treating doctor at any time...

Example 2.3 (Operational and conceptual definitions)**.** Consider a study requiring water temperature to be measured.

An *operational definition* would explain *how* the temperature is measured: the thermometer type, how the thermometer was positioned, how long was it left in the water, and so on.

In contrast, a *conceptual* definition might describe the scientific definition of temperature.

Example 2.4 (Operational definitions)**.** Consider a study measuring stress in first-year university students.

Stress cannot be measured directly, but could be assessed using a survey (like the Perceived Stress Scale (PSS) (Cohen et al. 1983)).

The *operational definition* of stress is the score on the ten-question PSS. Other means of measuring stress are also possible (such as heart rate or blood pressure).

All of these have advantages and disadvantages.

Extra example: Meline (2006) discusses five studies about stuttering, each using a different *operational* definition:

- Study 1: As diagnosed by speech-language pathologist.
- Study 2: Within-word disfluences greater than 5 per 150 words.
- Study 3: Unnatural hesitation, interjections, restarted or incomplete phrases, etc.
- Study 4: More than 3 stuttered words per minute.
- Study 5: State guidelines for fluency disorders.

People may be classified as stutters by some definitions but not others, so it is important to know which definition is used.

Extra example: A study of snacking in Australia (Fayet-Moore et al. 2017) used this operational definition of 'snacking':

> ...an eating occasion that occurred between meals based on time of day.
> — Fayet-Moore et al. (2017) (p. 3)

Extra example: A study examined the possible relationship between the 'pace of life' and the

incidence of heart disease (Levine 1990) in 36 US cities. The researchers used four different *operational* definitions for 'pace of life' (remember the article was published in 1990!):

1. The walking speed of randomly chosen pedestrians.
2. The speed with which bank clerks gave 'change for two $20 bills or [gave] two $20 bills for change.'
3. The talking speed of postal clerks.
4. The proportion of men and women wearing a wristwatch.

None of these *perfectly* measure 'pace of life,' of course. Nonetheless, the researchers found that, compared to people on the West Coast,

> ... people in the Northeast walk faster, make change faster, talk faster and are more likely to wear a watch...
> — Levine (1990) (p. 455)

Think 2.1 (Definitions)**.** *Define a 'smoker.'*

2.3 Elements of RQs

A RQ must be *written* carefully so it can be properly *answered*. In this section, the four potential components of a RQ are studied:

- The **P**opulation;
- The **O**utcome;
- The **C**omparison or **C**onnection;
- The **I**ntervention.

These form the **POCI** acronym.

2.3.1 The Population

All RQs study some *population*: the larger group of interest in the study.

Definition 2.3 (Population)**.** The *population* is the group of *individuals* (or *cases*; or *subjects* if the individuals are people) from which the total set of observations of interest could be made, and to which the results will (hopefully) generalise.

Individuals or *cases* do not just refer to people., though the words may be commonly used that way.

Similarly, *population* does not just mean people. In this context, a population is *any* group of interest; for example:

- all Australian males between 18 and 35 years of age.
- all bamboo flooring materials manufactured in Queensland.
- all elderly males with glaucoma in Canada.
- all *Pinguicula grandiflora* growing in Europe.

The *population* is not just those individuals from which the data are actually *obtained*. Indeed, *all* these elements of the population may not be accessible in practice.

The population represents all the 'individuals' to which the results are to be generalised. For example, when testing a new drug, the aim is to see if it works on people in general, including people not yet born. The population is 'all people.'

⚠ The **population** in a RQ is *not* just those we end up studying. It is the whole group to which our results would generalise.

In contrast, the *sample* is the *subset* of the population that we actually end up studying, from which data are obtained.

Definition 2.4 (Sample)**.** A *sample* is a subset of the population of interest which is actually studied, and from which data are collected.

Example 2.5 (Samples)**.** Consider a study of American college women, which aimed to:

> . . . assess iron status [. . .] in highly active (>12 hr purposeful physical activity per week) and sedentary (<2 hr purposeful physical activity per week) women. . .
> — Woolf et al. (2009), p. 521.

The *sample* comprises 28 'active' women and 28 'sedentary' American college women, from which data are collected.

The *population* is *all* 'active and sedentary' American college women, not just the 56 in the study. The group of 56 subjects is the *sample*.

Completely defining the population (Banerjee and Chaudhury 2010) sometimes requires *refining* or *clarifying* the population, using *exclusion* and/or *inclusion criteria*.

Exclusion and inclusion criteria clarify which individuals may be explicitly included or excluded from the population.

Exclusion and inclusion criteria should be explained when their purpose is not obvious. Both exclusion and inclusion criteria are not *needed*; none, one or both may be used.

Definition 2.5 (Inclusion criteria)**.** *Inclusion criteria* are characteristics that individuals must meet explicitly to be included in the study.

Example 2.6 (Inclusion criteria)**.** A study of a certain bird species may only included sites where there has been a confirmed sighting within the last two years.

Example 2.7 (Inclusion criteria)**.** A study of weight-loss methods may require people over a certain weight.

Definition 2.6 (Exclusion criteria)**.** *Exclusion criteria* are characteristics that explicitly disqualify potential individuals from being included in the study.

Example 2.8 (Exclusion criteria)**.** Concrete test cylinders with fissure cracks may be excluded from tests of concrete strength.

Example 2.9 (Exclusion criteria)**.** People with severe asthma may be excluded from exercise studies.

Example 2.10 (Population and exclusion criteria)**.** A study on the influenceze vaccine (Kheok et al. 2008) listed the **P**opulation as 'health-care workers' (Kheok et al. 2008p.466), and the sample they studied was:

> All healthcare workers at the National University Hospital (NUH) and KK Women's and Children's Hospital (KKWCH)...
> — Kheok et al. (2008), p. 466

The population was refined by exclusion criteria. The exclusion criteria were:

> ...declining to give consent, a history of egg protein allergy, and neurological or immunological conditions that are contraindications to the influenza vaccine.
> — Kheok et al. (2008), p. 466

Extra example: A study (Guirao et al. 2017) of the walking abilities of amputees used these inclusion and exclusion criteria:

> *Inclusion* criteria were as follows: length of the femur of the amputated limb of at least 15 cm measured from the greater trochanter, use of the prosthesis for at least 12 months prior to enrollment and more than 6 h/day, ability to walk indoors with or without supervision, and with or without ambulation aids and unilateral femoral amputation.
> The criteria for *exclusion* were the presence of cognitive impairment hindering the ability to follow instructions and/or perform the tests, body weight over 100kg, active oncologic pathologies, psychological disorders, previous residuum infection, active infection, residual femur length less than 15cm measured from the greater trochanter, pregnancy, and hip flexion deformity greater than 30^{o}.
> — Guirao et al. (2017), p. 27 (emphasis, line-break added)

2.3.2 The Outcome

All RQs study something *about* the population, called the *outcome*.

Because the RQ concerns a population, the outcome describes a population as a whole; hence, the outcome is usually an *average*, *percentage*, or *general* quantity numerically summarising the population (or subsets of the population).

Definition 2.7 (Outcome)**.** The *outcome* in a RQ is the result, output, consequence or effect of interest in a study, numerically summarising the population (or subsets of the population).

The outcome may be (for example):

- *average* increase in heart rates.
- *average* amount of wear after 1000 hours of use.
- *proportion* of people whose pupils dilate.
- *average* weight loss after three weeks.
- *percentage* of seedlings that die.

⚠ The **outcome** in a RQ summarises a *population*; it does not describe the *individuals* in the population.

2.3.3 The Comparison or Connection

In addition to having a population (P) and an outcome (O), some RQs may *compare* the outcome between a small number of different, distinct subsets of the population (that is, *groups of individuals*), or may explore a *connection* between the outcome and some other quantity that varies.

Definition 2.8 (Comparison)**.** The *comparison* in the RQ identifies the small number of different, distinct subsets of the population between which the outcome is compared. : The *comparison* in the RQ identifies the small number of different, distinct subsets of the population between which the outcome is being compared. The groups being compared have either *imposed* differences, or have *existing* differences.

The outcome may be *compared* between two or more separate subsets of the population:

- Average amount of wear in floor boards (O) could be compared across two groups in the population: standard wooden flooring materials and bamboo flooring.
- Average heart rates (O) could be compared across three subsets of the population: those who received no dose of a drug, those who received a daily dose of the drug, and those who received a dose of the drug twice daily.

⚠ Be careful!

This definition requires that population can be separated into two (or more) subgroups, that have either *imposed* differences (for example, one group is given one dose of fertilizer per day, and another given two doses of fertilizer per day) or have *existing* differences (for example, one group of people aged under 30, and another group of people aged 30 or over).

If all individuals are treated in the same way, there is no *comparison* according to this definition.

Example 2.11 (Comparison)**.** Consider a study to compare the average blood pressure (the Outcome) in Australians (the Population), to see if the average blood pressure in the right arm is the same as the average blood pressure in the left arm.

There is no comparison: the Outcome (average blood pressure) is not compared in two different subsets of the population; every person is treated the same way.

Instead, the blood pressure is measured twice on *every* member of the population. The outcome might be best described as 'the mean *difference* between right- and left-arm blood pressure.'

In contrast, a study comparing the average blood pressure between peoiple aged under 40 and people aged 40 or over *does* have a comparison: two subsets (females and males) of the population (Australians) are compared.

Definition 2.9 (Connection)**.** The *connection* in the RQ identifies another quantity of interest that varies, that may be related to the outcome.

As the value of the *connection* changes, the value of the outcome (potentially) changes:

- The connection between average heart rate (O) and exposure to various doses of caffeine (C) in mg.
- The connection between percentage germination (O) and hours of sunlight per day (C).

2.3.4 The Intervention

In addition to having a population (P), an outcome (O), and possibly a connection or comparison (C), some RQs also have an *intervention*.

Definition 2.10 (Intervention)**.** An *intervention* is a comparison or connection that the researchers have *imposed* upon those in the study, *intending to change the outcome*.

The intervention may be:

- explicitly giving a new drug to patients.
- explicitly applying wear testing loads to two different flooring materials.
- explicitly exposing people to different stimuli.
- explicitly applying a different dose of fertiliser.

Example 2.12 (Interventions)**.** A study comparing the average blood pressure (O) in female and male (C) Australians (P) measured blood pressure using a blood pressure machine (a sphygmomanometer).

The research team needs to interact with the participants and use a machine to measure blood pressure, but there is *no* intervention. Using the sphygmomanometer is just a way to measure blood pressure, to *obtain* the data. The sphygmomanometer is not used *with the intent of changing the outcome*.

There is no intervention, since the *comparison* is between females and males, and this cannot be *imposed* on the individuals by the researchers.

Sometimes, it is not clear from the RQ if an intervention is present or not. If you are writing an interventional RQ, you should try to make it clear when an intervention is used.

Think 2.2 (POCI)**.** *A study of American college women aimed to:*

> *. . . assess iron status [. . .] in highly active (>12 hr purposeful physical activity per week) and sedentary (<2 hr purposeful physical activity per week) women. . .*
> — *Woolf et al. (2009), p. 521.*

In this study, what is the:

- *Outcome?*
- *Comparison or Connection (if any)?*
- *Intervention (if any)?*

Answer: The answer is given in the online book.

Extra example: Researchers examined numerous studies of chest compressions involving paramedics. For their study, they examined research papers in which the **P**opulation was patients who had experienced a cardiac arrest, and where manual chest compressions were compared with another method.

The table below shows the comparison and outcomes of interest:

Interventions	Outcomes
Mechanical chest compression	Mean survival time to hospital discharge
Mechanical CPR	Percentage with a return of spontaneous circulation (ROSC)
Automated chest compressions	
Automated CPR	
Powered chest compressions	
Powered CPR	

The research concluded that:

> Overall, the evidence analysed suggests that mechanical chest compression devices are statistically superior to manual chest compressions of a high quality, when up-to-date protocols and guidelines are followed.
> — Williams et al. (2021), Table 1

2.4 Types of RQs

All RQs have a population (P) and an outcome (O). However, different *types* of RQ emerge depending on whether the RQ also has an comparison/connection (C) or intervention (I).

This section studies different types of research questions:

- Descriptive RQs;
- Relational RQs;
- Interventional RQs.

These are compared in Sect. 2.4.4.

2.4.1 Descriptive RQs (PO)

Descriptive RQs are the most basic RQs, and identify:

- The **P**opulation to be studied.
- The **O**utcome of interest about this population.

Typically, descriptive RQs look like this:

Among {the population}, what is {the outcome}?

⚠ This is not a 'recipe,' but a guideline.

Example 2.13 (Descriptive RQ)**.** Consider this RQ:

> Among Australian males between 18 and 35 years of age, what is the average heart rate?

In this RQ, the *Population* is 'Australians males between 18 and 35 years of age,' and the *Outcome* is '*Average* heart rate.' Notice that the Outcome is a numerical summary of the Outcome across the population (the *average* heart rate).

This is a *descriptive RQ*, as the RQ does not imply studying a connection with, or comparison between, the average heart rate and anything else.

Think 2.3 (POCI)**.** *Consider this RQ:*

> *Among Australian adults, what proportion are coeliacs?*

For this RQ, identify the ***P****opulation and the* ***O****utcome.*

Answer: The answer is given in the online book.

2.4.2 Relational RQs (POC)

Usually, *relationships* are more interesting than just descriptions; relational RQs explore existing relationships. *Relational RQs* identify:

- The **P**opulation.
- The **O**utcome.
- The **C**omparison or **C**onnection.

Relational RQs have no intervention; the connection or comparison is not imposed by the researchers.

Typically, relational RQs based on a *comparison* look like this:

Among {the population}, is {the outcome} the same for {the groups being compared}?

Example 2.14 (Relational RQ)**.** Consider this RQ:

> Among Australians between 18 and 35 years of age, is the average heart rate the same for females and males?

In this RQ, the *Population* is 'Australians between 18 and 35 years of age,' the *Outcome* is 'average heart rate,' and the *Comparison* is 'between females and males.'

This is a *relational RQ* based on a *comparison*. Notice that the average heart rate (Outcome) is a numerical summary across the two population sub-groups being compared (females; males).

The sex of the individual (the C) is not allocated by the researchers, so there is no intervention.

Typically, relational RQs based on a *connection* look like this:

Among {the population}, is {the outcome} related to {something else}?

Example 2.15 (Relational RQ)**.** Consider this RQ:

> Among Australians between 18 and 35 years of age, is the average heart rate related to age?

In this RQ, the *Population* is 'Australians between 18 and 35 years of age,' the *Outcome* is 'average heart rate,' and the *Connection* is with 'age.'

This is a *relational RQ* based on a *connection*. Age (the C) is not allocated by the researchers, so there is no intervention.

Think 2.4 (POCI)**.** *Consider this RQ (based on Brown et al. (2000)):*

> *In the Queensland Ambulance Service last year, what was the difference between the average response time to emergency calls between weekdays and weekends?*

*Identify the **P**opulation, the **O**utcome, and the **C**omparison.*

Answer: The answer is given in the online book.

Think 2.5 (POCI)**.** *Consider this RQ (based on Maron (2007)):*

> *In Queensland state forests, is there a* relationship between *the average number of noisy miners and the number of eucalypts, in general?*

*(A noisy miner is a type of bird.) In this RQ, identify the **P**opulation. the **O**utcome, and the **C**onnection.*

Answer: The answer is given in the online book.

Example 2.16 (Descriptive and relational RQs)**.** Consider a study of blood pressure in Australians (the Population), comparing right- and left-arm blood pressures.

This is a *descriptive* RQ. There is *no* comparison, since there are not two subsets of the population being compared.

The blood pressure is measured twice on each member of the population: every member of the population is treated in the same way. The outcome is 'the average *difference* between right- and left-arm blood pressure.' This is a *descriptive* RQ.

In contrast, a study comparing the average blood pressure between females and males is a *relational* RQ. There *is* a comparison: the two subsets of the population (Australians) being compared are females and males.

2.4.3 Interventional RQs (POCI)

Interventional RQs explore relationships where the comparison/connection is determined or allocated by the researchers. They identify:

- The **P**opulation.
- The **O**utcome.

- The **C**omparison or **C**onnection.
- The **I**ntervention.

Interventional RQs may look like relational RQs, except that the comparison or connection is determined or allocated (i.e., imposed) by the researchers.

Sometimes it is not clear if the comparison or connection has been imposed by the researchers in an interventional RQ. When writing interventional RQs, make efforts to make it clear, if possible, when the RQ is interventional.

Example 2.17 (Interventional RQ)**.** Consider this RQ:

> Among Australians between 18 and 35 years of age, is the average heart rate for people allocated to receive a *new* pill the same as for people allocated to receive an *existing* pill?

In this RQ, the *Population* is 'Australians between 18 and 35 years of age,' the *Outcome* is 'average heart rate,' and the *Comparison* is 'between those taking the new pill, and those taking the existing pill.'

There is an *Intervention*: the researchers *allocate* one of the pills to each subject. This is an *interventional RQ*.

Extra example: Consider this RQ (McLinn et al. 1994):

> In children with acute otitis media, what is the difference in the average duration of symptoms when treated with cefuroxime compared to amoxicillin?

In this RQ, the Population is 'children with acute atitis media,' the Outcome is 'average duration of symptoms,' and the Connection is between the types of drug (comparing 'cefuroxime' and 'amoxicillin').

It is not clear if there is an Intervention.

If the drugs are *given* to the children by the researchers, there is an intervention (giving the drug).

If the researchers just find children who are already taking the two drugs and measure the outcome ('average duration of symptoms'), there is no intervention.

It is probable that there is an intervention.

2.4.4 Comparing the three levels of RQs

Descriptive RQs are the most basic and are usually used when a research topic is in its infancy; descriptive RQs set the platform for asking relational questions.

Relational RQs explore relationships, and provide an understanding of how the outcome of interest is related to certain sub-groups of the population; they may set the platform for asking interventional questions.

Interventional RQs (when possible to answer) are the most interesting: they can be used to test theories or models, or to establish cause-and-effect relationships (Table 2.2).

TABLE 2.2: *The three types of RQs*

RQ type	P	O	C	I
Descriptive (D)	Yes	Yes		
Relational (R)	Yes	Yes	Yes	
Interventional (I)	Yes	Yes	Yes	Yes

Research often develops through these stages of RQs as knowledge grows and develops. For example:

- **Descriptive**: What proportion of Australian adults are coeliacs (Cook et al. 2000)?
- **Relational**: Among Australian adults, is the proportion of females who are coeliacs the same as the proportion of males who are coeliacs (Cook et al. 2000)?
- **Interventional**: Among Australian adult coeliacs, is the percentage with adverse symptoms the same for those given a diet *without* oats and those given a diet *with* oats? (Janatuinen et al. 2002; Lundin et al. 2003)?

Think 2.6 (RQ types)**.** *What type of RQs are the following: Descriptive, Relational, or Interventional?*

1. *Among Australian upper-limb amputees, is the percentage wearing prosthesis 'all the time' the same for transradial and transhumeral amputations? (Davidson 2002)*
2. *In New York, what is the difference between the average height of oaks trees ten weeks after planting, comparing trees planted in a concrete sidewalk and a grassed sidewalk? (Grabosky and Bassuk 2016)*
3. *What is the average response time of paramedics to emergency calls? (Pons et al. 2005)*
4. *Is there a relationship between the average weekly hours of physical activity in children and the weekly maximum temperature? (Edwards et al. 2015)*

Answer: The answer is given in the online book.

2.5 Two approaches to RQs

RQs can be approached in one of two ways:

- For **estimation** (*confidence intervals*): These RQs are concerned with, for example, estimating a value in a population. This value may be the size of a difference (probably a RQ with a Comparison), or strength of a relationship (probably a RQ with a Connection).
- For **making decisions** (*hypothesis testing*): These RQs are concerned with making a decision about an unknown population value: for example, is the percentage the same in two different groups of the population?

Think 2.7 (RQ type)**.** *What approach do these RQs take: Decision-making, or Estimation?*

1. *Among Australian upper-limb amputees, is the percentage wearing prosthesis 'all the time' the same for transradial and transhumeral amputations? (Davidson 2002)*
2. *In New York, what is the difference between the average height of oaks trees (ten weeks after planting) comparing trees planted in a concrete sidewalk and a grassed sidewalk? (Grabosky and Bassuk 2016)*
3. *What is the average response time of paramedics to emergency calls? (Pons et al. 2005)*
4. *Is there a relationship between the average weekly hours of physical activity in children and the weekly maximum temperature? (Edwards et al. 2015)*

Answer: The answer is given in the online book.

2.5.1 Estimation: Confidence intervals questions

Sometimes, the RQ concerns how precisely a *value* in the population is estimated by the sample. This value may measure a *difference*, or the *strength* of a relationship.

These RQs are studied in Chapters 19 to 25, and in Sect. 35.7.

Example 2.18 (Estimation RQs)**.** Consider this RQ (based on Barrett et al. (2010)):

> Among Australian teens with a common cold, *how much shorter* are cold symptoms, on *average*, for teens taking a daily dose of echinacea compared to teens taking no medication?

This RQ asks about size of the *difference* (in the *population*) between the average duration of cold symptoms.

Only sample data are available, and there may be no difference (on average) at all in the population.

2.5.2 Making decisions: Hypothesis testing questions

Sometimes, RQs are not about the precision with which a population value is estimated by the sample, but instead about deciding if a difference or a relationship exists in the *population*.

These RQs often are associated with *hypotheses*: *statements* that suggest possible answers to the RQ. Based on the sample, the hypothesis best supported by the data is to be chosen.

These RQs are studied in Chapters 26 to 31, and in Sect. 35.6.

Example 2.19 (Making decisions with samples)**.** Consider this RQ (based on Barrett et al. (2010)):

> Among Australian teens with a common cold, is the average duration of cold symptoms shorter for teens taking a daily dose of echinacea compared to teens taking no medication?

This is a decision-type RQ, with two possible answers (Fig. 2.3): Either echinacea *does* result in shorter average cold durations, or it *doesn't*. In practice the answer is rarely clear cut, and instead *how much* evidence there is in the sample to support a particular hypothesis about the *population* is reported.

Evidence may *support* or to *contradict* a hypothesis; evidence rarely *proves* a hypothesis (at least, without any other support, such as theortical support). Ultimately, after collecting data from a *sample*, a decision must be made about which explanation about the *population* is more consistent with the data collected.

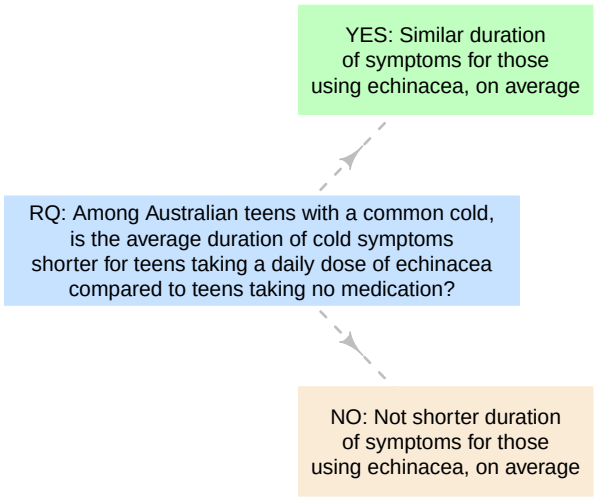


FIGURE 2.3: *Two possible answers to the RQ about echinacea*

2.6 Writing good RQs

Ideally, a well-written RQ (Anastasiadis et al. 2015) should be:

- **F**easible: Answering the RQ should be possible practically; sufficient personnel, time, resources, and money should be available to complete the study properly.
- **I**nteresting: The RQ should be interesting. For example, no-one cares about comparing the percentage of people who prefer drinking tea in blue cups to green cups...
- **N**ovel: The RQ should be original (the RQ should 'seek to confirm, refute or extend previous findings, and potentially reveal new findings' (Anastasiadis et al. (2015), p. 410). Researching something already well known is waste of time and resources.
- **E**thical: The RQ must be able to be answered ethically (Chap. 4). This is not negotiable.
- **R**elevant: The RQ should be relevant and current.

Note the acronym **FINER** to help remember these guidelines.



In most undergraduate university courses, a Project RQ **must** be feasible and ethical. Given the nature of a course, and the short timelines, these RQs don't necessarily *need* to be Interesting, Novel or Relevant. It is great if it is all of these, however.

Example 2.20 (Poor RQ)**.** Here is a RQ submitted by a student group (including typos) at the university where I work:

> Utilising a convenience sample at The University of Sunshine Coast in Sippy Downs, is there a difference in taste perception between students on a Thursday morning and afternoon (, when comparing English and Australian Cadburys milk chocolate ?

This is a poor RQ:

- General poor writing: A round bracket is started but never closed, for example… and for some reason the bracket is followed by a comma. This shows poor attention to detail.
- The RQ starts by describing the sample ('a convenience sample'), but RQs are always about a *population*, not a *sample*.
- The RQ does not have a clear Outcome that numerically summarises the population: a proportion or a mean, for instance.
- It is not clear whether the comparison is between morning and afternoons, or between English and Australian chocolates, or both.

Notice that 'taste perception' is not defined. This is not a criticism: the operational definitions can be provided elsewhere.

This is a far better RQ:

> For USC students at the Sippy Downs campus, is the percentage of people who can correctly identify English or Australians chocolates the same in the mornings as in the afternoons?

Written this way:

- P: USC students at Sippy Downs campus.
- O: The *percentage* correctly identifing English or Australian chocolates.
- C: Between mornings and afternoons.
- I: No intervention: We cannot decide if a particular time of day is morning or afternoon.

This RQ is **F**easible and **E**thical, but probably not really **I**nteresting (except to the project group), **N**ovel or **R**elevant… but that's OK.

Exclusion criteria might exclude people with dairy intolerance, and those who do not eat dairy (such as vegans).

2.7 Writing RQs: An example

RQs emerge from observations, which leads to asking questions, and the need for evidence to answer that question (Tully 2014).

For example, suppose you notice that many people take echinacea when they get a cold; it is reasonable to ask if there is evidence that echinacea helps with a cold in any way. This may lead to an initial RQ (based on Barrett et al. (2010)):

> Is it better to take echinacea when you have a cold?

This RQ is clearly poor, but serves as a starting point.

RQs often start as a basic idea, which can be refined by clarifying the POCI elements. For example, what **population** could we study? Many options exist:

- 'You' is implied by the question... but this is not a useful or practical *population*.
- All Australians.
- Australians adults with a specific "cold."

What **outcome** could be used to to determine echinacea's effectiveness? Again, many options exist:

- *Average* cold duration.
- *Average* severity of cold symptoms.
- *Percentage* of people who take days off work.

The initial RQ cannot be answered because 'better' is ambiguous: better than what? We could decide to **compare** an outcome across different groups, or **connect** it to something else. For example, the *comparison* could be:

- Between taking echinacea and taking no medication.
- Between taking echinacea and taking another medication.
- Between taking different doses of echinacea.

Furthermore, we could decide to **intervene** or not. Whether we decide to include an intervention or not has implications for *how* the study is conducted and how the results are interpreted.

If we decided *not* to intervene, the subjects in the study would decide for themselves how to treat their cold. If we did decide to intervene, various interventions could be used:

- Imposing how *frequently* the dose was taken; and/or
- Imposing what *doses* of echinacea to take.

After making some decisions about P, O, C and I, consider this revised RQ:

> Among Australian teenagers with a common cold, is the duration of cold symptoms shorter for teens taking a daily dose of echinacea compared to teens taking no medication?

The P, O, C and I do not have to be comprehensively described in the RQ; some information could be provided later as operational definitions (e.g., dose).

This RQ is much better, but it is *still not correct*. The outcome is a numerical summary across subsets of the *population*, not of *individuals*. So consider this revised RQ (based on Barrett et al. (2010)):

> Among Australian teenagers with a common cold, is the **average** duration of cold symptoms shorter for teens given a daily dose of echinacea compared to teenagers given no medication?

This is a better RQ.

> **Think 2.8** (POCI)**.** *For this RQ above, identify the* Population*,* Outcome*,* Comparison *or* Connection*, and the* Intervention *(if any).*

Answer: The answer is given in the online book.

2.8 Variables: From populations to individuals

RQs explore relationships in the *population*. The Outcome describes the population in general, and so Outcomes are often worded in terms of averages or percentages or similar. For example, consider this RQ seen above:

> Among Australian teenagers with a common cold, is the average duration of cold symptoms shorter for teens given a daily dose of echinacea compared to teenagers given no medication?

This is an interventional RQ (using a *comparison*) about a *population*.

No relationship could be found with information from just two teenagers. Consider this: suppose a cold lasts for 6 days for a teenager who *does* take echinacea, and a cold lasts for 5 days for a teenager who *does not* take echinacea. Is there a difference between the cold durations *in the population*? We have no way of knowing: Only two teenagers were studied. To explore the relationship using teenagers in general, data from *many* teenagers is needed.

RQs concern numerical summaries about *populations*, but the data to answer the RQ come from *individuals* in the population. (As with the word 'population,' the word 'individual' does not only refer to people.)

Each piece of information that we gather from individuals is called a *variable*, because its values can *vary* from individual to individual.

> **Definition 2.11** (Variable)**.** A *variable* is a single aspect or characteristic associated with each of a group of individuals under consideration, whose values can vary from individual to individual.

The value of a variable can *vary* from one individual to the next. Examples include

- the duration of cold symptoms;
- gender;
- age;
- place of birth;
- amount of tyre wear;
- hair colour.

The RQ identifies the variables *needed* to answer the RQ, though other variables may be (and typically are) measured also (Sect. 6.3).

⚠ A variable is a single aspect that can vary from *individual to individual*.

Your city of birth may not change, but 'city of birth' is still a variable because it can vary from *individual* to *individual*. Your city of birth may not be changing, but that is not relevant.

Example 2.21 (Variables)**.** 'Duration of cold symptoms' is a variable, as it is obtained from individuals, and its value can vary from individual to individual.

The '*average* duration of cold symptoms' is the *outcome*, numerically summarising the individuals cold durations across the population.

While many variables can be measured on individuals, two variables are of greatest importance:

- The **response variable** measures, assesses, describes or records information to determine the outcome; and
- The **explanatory variable** measures, assesses, describes or records information to determine the comparison or connection (Table 2.3).

The RQ cannot be answered without information about these two variables.

TABLE 2.3: *The relationship between the population and the individuals*

Population		Individuals
Outcome:	→	Response variable
Comparison/Connection:	→	Explanatory variable

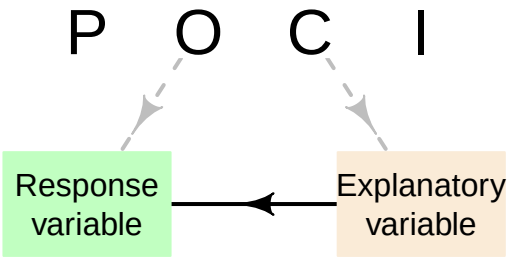


FIGURE 2.4: *The POCI elements*

Definition 2.12 (Response variable)**.** A *response variable* is the variable used to measure, assess or describe the outcome on each individual in the population.

The *outcome* refers to the numerical summary of the values of the response variable (Table 2.4).

TABLE 2.4: *Examples of the Outcome and the corresponding Response variable*

Population	→	**Individuals**
Average increase in diastolic blood pressure	→	Increase in diastolic blood pressure of *individuals* before and after exercise
Percentage of seedings that sprout	→	Whether or not an *individual* seedling sprouts
Proportion owning iPad	→	Whether or not an *individual* owns an iPad
Average cold duration	→	Cold duration for *individuals*
Percentage of concrete cylinders having fissures	→	Whether or not an *individual* cylinder has fissures

Definition 2.13 (Explanatory variable)**.** An *explanatory variable* is a variable of interest from the individuals in the study which (potentially) causes changes in, or is related to, the response variable.

The explanatory variable is a formal description of what *C* measures, observes, assesses or describes in each individual member of the population (Table 2.5).

TABLE 2.5: *Examples of the Comparison and the corresponding Explanatory variable*

Comparison being made		**Explanatory variable in Individuals**
Between males and females	→	The sex of each *individual* person
Between beech, tallowwood, and jarrah floor boards	→	Type of floorboard in each *individual* home
Between 350kg/ha and 400kg/ha fertilizer rates	→	Application rate in each *individual* paddock
Between people in their 20s, 30s and 40s	→	Age group for each *individual* person

In many cases, explanatory variable occurs *before* the response variable, or can be thought of as 'causing' the response variable.

Example 2.22 (Variables)**.** For the final RQ for the echinacea study (Sect. 2.7), the *response variable* would be the length of cold symptoms, and the *explanatory variable* is the type of medication (echinacea; or none).

In this case, the type of medication is taken *before* the cold symptoms disappear, and perhaps even causing them to disappear.

Think 2.9 (Variables)**.** *Consider this RQ:*

> *For carrots grown in Buderim, is the average weight of carrots 8 weeks after planting the same when grown without Thrive, and for carrots grown with weekly applications of Thrive?*

1. *What is the outcome? What is the comparison?*
2. *What data is* needed *from each element of the population to answer this question? That is, what are the response and explanatory variables?*

Answer: The answer is given in the online book.

Example 2.23 (Variables)**.** Consider this RQ:

> For overweight men over 60, is the average weight loss after three weeks the same for a diet high in fresh fruit and a diet high in dried fruit?

The *outcome* is the average weight *loss*; the *response variable* is the weight loss for each individual man. (This would be found by measuring their weight *before* and *after* three weeks on the diets.)

The *comparison* is between the two diets; the *explanatory variable* is which diet each man is on.

2.9 Units of observation and units of analysis

Units of observation and **units of analysis** are important, but similar, concepts that need to be distinguished.

Think 2.10 (Designs)**.** *Consider this RQ:*

> *In Australian 20-something men, is the average thickness of head hair strands the same for blonds*[1] *and brunets*[2] *(Vaughn et al. 2009)?*

What is the problem with comparing 100 hair strands from one blond man, to 100 hair strands from one brunet man?

In this study, *only one man of each hair colour is represented*. There are 200 observations, but only two people are compared, so little is learnt about 20-something men *in general*.

We learn a lot about two men specifically. The Population is represented by just two men... so we don't learn much about the population of men in general.

In this study, each individual hair is a *unit of observation*: the hair strands are what must be measured to obtain 'thickness of head hair strands.'

But each blond hair comes from the same man, so each of those hairs have essentially lived their life together: They are washed at the same time, with the same shampoo, exposed to the same amount of sunlight and exercise, share genetics, etc. However, different people do their own thing and have their own genetics.

Definition 2.14 (Unit of observation)**.** *Unit of observation*: The 'who' or 'what' which are observed, from which measurements are taken and data collected.

A similar, but different, concept is the **unit of analysis**.

Definition 2.15 (Unit of analysis)**.** *Unit of analysis*: The 'who' or 'what' about which generalizations and conclusions are made; the smallest *independent* 'who' or 'what' for which information is analysed. Units of analysis should not typically share a common underlying source.

In the hair-thickness study each *person* a *unit of analysis*. Importantly, the size of sample in the study is the number units of analysis; so here, *there are only two examples of the population in the study*. The size of the sample is just two.

⚠ The size of the sample in a study is the number of units of *analysis*.

⚠ All studies have units of analysis, and units of observation.

Example 2.24 (Units of analysis)**.** In the hair-strand study, each hair strand is a unit of *observation*: measurements of hair strand thickness are taken from individual hair strands.

However, the unit of *analysis* is the *person*: the hair strands from each man share a lot in common. The men themselves would share little in common, and we are interested in comparing men.

Example 2.25 (Units of analysis)**.** Consider a study comparing the percentage of females and males wearing sunglasses at a specific beach.

People in a *group* at the beach will probably not be operating 'independently': groups of people tend to behave similarly (but perhaps not identically). For example, a couple will often *both* be either wearing or *not* wearing sunglasses.

The researchers have two options; they could either

- Use the people *groups* as the *unit of analysis* (some of which will be groups of one), and record data from just *one* person in any group.

  Ideally, the researchers would specify before-hand which group member from which to take data (e.g., the person closest to the researchers when the group is spotted).

- Alternatively, the researchers may decide not to use data from groups at all, and only gather data from individuals.

Example 2.26 (Units of analysis)**.** A study compares two brands of car tyres. Four tyres of Brand A are allocated to each of Cars 1–5. Four tyres of Brand B are allocated to each of Cars 6–10.

After 12 months, the amount of wear is recorded on each tyre. The *unit of observation* is the *tyre*: the amount of wear is measured on each tyre.

The *unit of analysis* is the *car*: the brand of tyre is allocated to the car and all wheels on the car get the same tyre. Tyres on any one car 'live their life together': They all are exposed to the same day-to-day use, the same drivers, have driven very similar distances, under the same conditions, etc.

Extra example: A report on the *Spectrum* website reported the following:

> Seven years ago, Peter Kind [...] was reading a study about fragile X syndrome, a developmental condition characterized by severe intellectual disability and, often, autism [...] Kind was surprised when he noticed a potentially serious statistical flaw.
> The research team had looked at 10 neurons from each of the 16 mice in the experiment, a practice that in itself was unproblematic. But in the statistical analysis, the researchers had analyzed each neuron as if it were an independent [individual observation]. That gave them 160 data points to work with, 10 times the number of mice in the experiment.
> "The question is, are two neurons in the brain of the same animal truly independent data points? The answer is no,' ' Kind says.
> — Spectrum report

There were 16 units of analysis (mice) so the sample size is 16 mice, but the authors treated the $16 \times 10 = 160$ neurons as the sample size. The 10 neurons from each mouse have a lot in common: the genetic information was the same for all 10 neurons from each mouse.

A total of 160 neurons from 16 mice is very different to a study of 160 neurons from 160 genetically-different mice.

The units of observation and units of analysis *may* be the same, and often are the same. However, they are sometimes different too, and it is *crucial* to be able to identify these situations. Importantly, studies compare units of analysis, not units of observation.

Example 2.27 (Units of analysis)**.** A study compared two school physical activity (PA) programs. Each of 44 children (whose parents agreed for their children to participate in the study) were allocated to one of two PA program. The improvement in children's fitness was measured for every student in the study after six months.

The *units of observation* are the individual students, as the the fitness measurements are taken from the students. The *units of analysis* are also the individual students, as the PA program was allocated to each student individually.

Think 2.11 (Units of analysis and observation)**.** *A study compared two school physical activity (PA) programs. Program 1 was allocated to be used at School A, while Program 2 was allocated to School B. In each school, 22 children (with parental consent) were observed and the improvement in children's fitness was measured for each student after six months.*

What are the units of analysis *and* unit of observation*?*

Answer: The answer is given in the online book.

2.10 Preparing software for data entry

Most statistical software (including jamovi[3] (The jamovi Project) and SPSS[4] (IBM Corp 2016)) uses the same approach for collating the data[5]:

- Each *row* represent one unit of analysis. Hence, the *number* of rows will equal the *number* of units of analysis.
- Each *column* represents one variable. Hence, the *number* of columns will equal the *number* of variables. (There may also be a column of identifying information (such as the person's name).)



In statistical software, the *names* of the variables are not placed in a separate row (say, in Row 1 above the data itself), which might happen when using a spreadsheet.

The *names* of the variables become the names of the columns.

Example 2.28 (Preparing statistical software)**.** In Sect. 2.8, this RQ was posed:

> Among Australian teenagers with a common cold, is the average duration of cold symptoms shorter for teens given a daily dose of echinacea compared to teenagers given no medication?

For this RQ, the *variables* are (Examples 2.21 and 2.22):

- 'Duration of cold symptoms' (response), and
- 'Type of treatment' (explanatory).

To set up the software for data entry:

- The number of *rows* of data would be the number of people in the study.
- The number of *columns* would be two: one column to record the duration of each

[3] https://www.jamovi.org/
[4] https://www.ibm.com/products/spss-statistics
[5] Though there are exceptions for some types of analyses.

individual's cold symptoms, and the other to record whether the individual received a dose of echinacea or received no medication.

In addition, there may be a column recording the name or ID of each individual.

The variable names (say, `Duration` and `Treatment`) would not be in a row of their own; they would be the columns names (Fig 2.5).



	Duration	Treatment
1	6	Echinacea
2	4	None
3	5	None
4	5	Echinacea
5	3	Echinacea
6	6	Echinacea
7	6	Echinacea
8	5	Echinacea
9	6	Echinacea
10	4	None
11	8	None

	Duration	Treatment
1	6	1
2	4	2
3	5	2
4	5	1
5	3	1
6	6	1
7	6	1
8	5	1
9	6	1
10	4	2
11	8	2

FIGURE 2.5: *jamovi (left) and SPSS (right) prepared for the data, with some data entered, and the variable names as the column headers*

While spreadsheets (such as Excel) can be used for analysing data, significant problems can, and do, emerge with using spreadsheets. Great care is needed when using spreadsheets for data analysis!

2.11 Summary

In this chapter, we have learnt about writing and understanding **research questions**. Research questions (RQs) are always about an **outcome** (O) in some **population**. Some RQs have a **comparison** or **connection** (C), and some also have an **intervention** (I). RQs may be **estimation**-type RQs or **decision**-type RQs.

The outcome numerically summarises the population or subsets of the population (so is usually worded in terms of percentages, averages, etc.), but the data comes from **individuals** in the population by measuring, observing or assessing the **response** (or dependent) variable. Similarly, the data concerning the comparison or connection comes from measuring or observing the **explanatory** (or independent) variables.

The *who* or *what* that observations are made from are called the **units of observation**. The smallest independent units (that is, units with very little in common) are called the **units of analysis**.

2.12 Quick review questions

Consider this RQ:

> Is the average walking speed the same when texting and talking on a mobile phone?

1. What is the *explanatory* variable?
2. What is the *response* variable?
3. What is the *outcome*?

Answer:

1. What *individual* people are doing with their phones probably explains their walking speed: the *explanatory variable* is the way in which the mobile phone is being used.
   Notice that 'talking on the phone' and 'texting on the phone' are not *variables*. They are particular values that the *variable* can take. That is, 'what people are doing on their phone' is the variable, because it can vary: Sometimes people will be talking, sometimes texting, etc.
2. Waking speed probably depends on (or *responds to*) how *individual* people are using their phone: the *response variable* is the walking speed.
3. The *outcome* is how the response *variable* is summarised over a group of individuals. The waling speeds from many individuals could be summarised numerically using the *average walking speed*, which would be the outcome.

2.13 Exercises

Selected answers are available in Sect. D.2.

Exercise 2.1. In a study of public acceptance of alternative water supplies (Hurlimann and Dolnicar 2016), various water sources are defined. In Table 2.6, match the term with the appropriate operational definition.

TABLE 2.6: *Match the term with the operational definition*

Term	Definition
Rainwater	Rainwater from a rainwater collection tank on your property
Bottled	Water you presently use throughout your dwelling (home)
Tap	Highly purified seawater deemed by scientists and public health officials as safe for human consumption.
Recycled	Highly purified wastewater deemed by scientists as safe for human consumption
Desalinated	Water sold in bottles by food companies that is widely available to the public for purchase and consumption

Exercise 2.2. Consider this RQ:

> Among university students, is the average resting diastolic blood pressure the same for students who regularly drive to university and those who regularly ride their bicycles?

1. For this RQ, identify the **P**opulation.
2. For this RQ, identify the **O**utcome.
3. For this RQ, identify the **C**onnection, if any.
4. For this RQ, identify the **I**ntervention, if any.
5. What *type* of RQ is this?
6. What *operational definitions* would be needed?
7. What information *must* be collected from each individual to answer the RQ?
8. What are the units of analysis?
9. What are the units of observation?

Exercise 2.3. Consider this article extract (Checkley et al. 2002):

> We conducted a 4-year (1995–1998) field study in a Peruvian peri-urban community… to examine the relation between diarrohea and nutritional status in 230 children < 3 years of age
> — Checkley et al. (2002), p. 210

For this study:

1. Identify PICO.
2. Infer the primary research question.
3. What *type* of question is used?
4. What *operational definitions* would be needed?
5. What are the *response* and *explanatory* variables?

Exercise 2.4. For the following *response* variables, what would be the corresponding *outcomes*?

1. Whether a vehicle crashes or not.
2. The height at which people can jump.
3. The number of tomatoes per plant.
4. Whether or not a person owns a car.

Exercise 2.5. For the following *comparisons*, what would be the corresponding *explanatory* variables?

1. Between 91 octane, 95 octane, and ethanol-blended car fuel.
2. Between caffeinated and decaffeinated coffee.
3. Between taking zero, one or two iron tablet per day.
4. Between vegans and vegetarians.

Exercise 2.6. For the following studies, determine which have a Comparison and which do not. In each case, identify the Outcome.

1. A study to determine if a higher percentage of people at a particular city park wear hats in winter compared to summer.
2. A study to determine if average cholesterol levels are the same when measured on the same people before and after a diet change.

3. A study to determine if the average balance-time on right legs is the same as on left legs.
4. A study to determine if the average yield of tomato plants is the same when three different fertilisers are applied.

Exercise 2.7. Animals in an experiment are divided into pens (three per pen), and feed is allocated to each pen (Sterndale et al. 2017). Animals in different pens receive different feed; animals in the same pen receive the same feed. The weight gain of each animal is recorded.

1. What is the *unit of observation*? Why?
2. What is the *unit of analysis*? Why?

Exercise 2.8. Consider this actual Project Report RQ from the university where I work. Critique the RQ, and write a better RQ (if necessary).

> Among 10 Australian adults, does the time taken to read a passage of text change when different fonts are used?

Exercise 2.9. Consider this actual Project Report RQ from the university where I work. Critique the RQ, and write a better RQ (if necessary).

> Of students that study at USC, Sippy Downs, do males have a larger lung capacity than females?

Part II

Research design

3

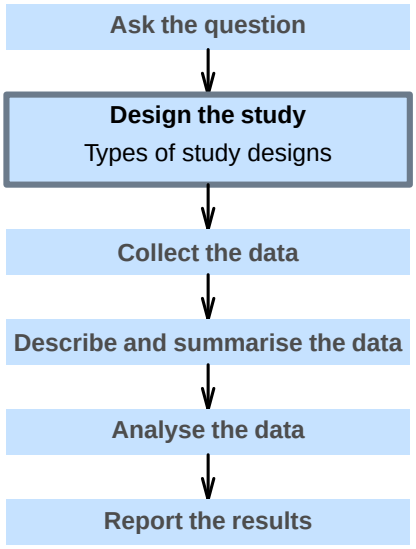
Types of study designs



So far, you have learnt how to ask a RQ.

In this chapter, you will study the details of *how* to collect the data needed to answer the RQ. You will learn to:

- design scientifically sound studies to answer simple quantitative research questions.
- design ethical studies.
- describe the various types of quantitative research studies.
- compare and contrast experimental and observational studies.
- describe and identify retrospective, prospective and cross-sectional observational studies.
- describe and identify true experimental and quasi-experimental studies.



3.1 Three types of study designs

From the RQ, we know what data *must* be collected from the individuals in the study (the response and explanatory variables)... but *how* do we obtain this data? After all, data are important: they are the means by which the RQ is answered.

Three broad methods for obtaining data are to use:

- Descriptive studies (Sect. 3.2), for answering Descriptive RQs;
- Observational studies (Sect. 3.3), for answering Relational RQs; or

- Experimental studies (Sect. 3.4), for answering Interventional RQs.

The type of study depends on the type of RQ.

Example 3.1 (Research design)**.** Suppose we wish to compare the effects of echinacea on the symptoms of the common cold (based on Barrett et al. (2010)).

How would we design such a study to collect the necessary data? What decisions would you need to make?

3.2 Descriptive studies

Descriptive studies *are used to answer descriptive RQs* (Fig. 3.1).

Definition 3.1 (Descriptive study)**.** In a *descriptive study*, researchers only focus on collecting, measuring, assessing or describing an outcome in the population.

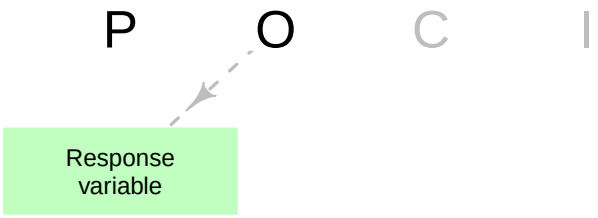


FIGURE 3.1: *A descriptive study, used to answer a descriptive RQ*

Example 3.2 (Descriptive study)**.** Consider this RQ:

> For overweight men over 60, what is the average increase in heart rate after walking 400 metres?

The *outcome* is the average *increase* in heart rate. The *response variable* is the *increase* in heart rate for the individual men.

The increase in heart rate would need to be found by measuring each man's heart rate *before* the walk, then their heart rate *after* the walk, and finding the difference between them. The *increase* in heart rate would be computed as the *after* heart rate minus the *before* heart rate.

Some of these differences might be positive numbers (heart rate went *up*), and some may be negative numbers (heart rate went *down*.

No *comparison* being made: every man in the study is treated in the same way. This is a *descriptive* RQ, which can be answered by a *descriptive* study.

3.3 Observational studies

Observational studies (Fig. 3.2) are used to answer *relational RQs*. They are commonly used, and sometimes are the only possible study design that can be used. These are discussed further in Chap. 3.3.

Definition 3.2 (Observational study)**.** In an *observational study*, researchers do not impose the comparison or connection upon those in the study to (potentially) change the response of the participants.

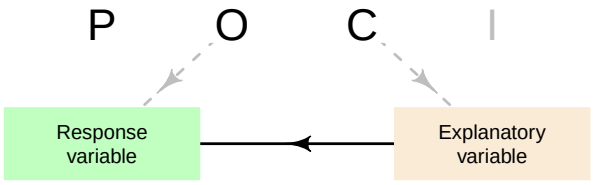


FIGURE 3.2: *An observational study, used to answer a relational RQ*

Definition 3.3 (Condition)**.** *Conditions*: The conditions of interest that those in the observational study are exposed to.

Example 3.3 (Observational study)**.** Consider again this RQ (Barrett et al. 2010):

> Among Australian teens with a common cold, is the *average* duration of cold symptoms shorter for teens taking a daily dose of echinacea compared to teens taking no medication?

This would be a relational RQ if the researchers do not impose the echinacea (that is, the individuals make this decision themselves). For this RQ, the *conditions* would be taking echinacea, or not taking echinacea (Fig. 3.3).

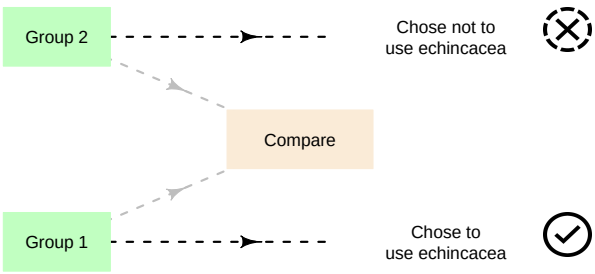


FIGURE 3.3: *Observational studies*

Broadly speaking, three types of observational studies exist (Table 3.1):

- Retrospective: look into the past for the comparison;
- Prospective: look into the future for the outcome;
- Cross-sectional: obtain the outcome in the present.

TABLE 3.1: *The three types of observational studies*

Type	O	C	Reference
Retrospective	Now	Earlier	Sect. 3.3.1
Prospective	Later	Now	Sect. 3.3.2
Cross-sectional	Now	Now	Sect. 3.3.3

These differ in when the Outcome and Comparison/Connection are observed. Many specific types of observational studies exist (case-control studies; cohort studies; etc.), but we will not delve into these.

3.3.1 Retrospective studies

In *retrospective studies*, the Outcome (and response variable) is observed *now*, and the researchers look *back* to see what Comparison/Connection group (and explanatory variable) was in the past (e.g., case-control studies).

Example 3.4 (Retrospective studies)**.** An Australian study (Pamphlett 2012) examined patients with and without sporadic motor neurone disease (SMND), and asked about *past* exposure to metals. The response (whether or not the respondent had SMND) is assessed now, and whether or to they had exposure to metals (explanatory) is assessed from the *past*. This is a *retrospective* observational study.

3.3.2 Prospective studies

In *prospective studies*, the Comparison/Connection (or explanatory variable) is determined *now*, and researchers look *ahead* to assess or measure the Outcome (or response) (e.g., Prospective cohort studies).

Example 3.5 (Prospective studies)**.** A study (Choi and Curhan 2008) measured the softdrink consumption of men, and determined who experienced gout over the following 12 years. The response (whether or not the individuals experience gout) is determined in the future. The explanatory variable (the amount of softdrink consumed) is measured now. This is a *prospective* observational study.

3.3.3 Cross-sectional studies

In *cross-sectional studies*, both the Outcome (response) and Comparison/Connection (explanatory variable) are gathered *now*.

Example 3.6 (Cross-sectional studies)**.** A study (Russell et al. 2014) asked older Australian their opinions of their own food security, and recorded their living arrangements. Individuals' responses to both living arrangements and opinions on food security are obtained *now*. This is a *cross-sectional* observational study.

Think 3.1 (Cross-sectional studies)**.** *In South Australia in 1988–1989, 25 cases of legionella infections (an unusually high number) were investigated (O'Connor et al. 2007). All 25 cases were gardeners, with hanging baskets of ferns.*

Researchers compared 25 cases with legionella infections with 75 non-cases, matching on the basis of age (within 5 years), sex, post codes. The use of potting mix in the previous four weeks was associated with an increase in the risk of contracting illness of about 4.7 times.

What type *of observational study is this?*

Answer: The answer is given in the online book.

3.4 Experimental studies

Experimental studies (Fig. 3.4), or *experiments*, are commonly used (and are discussed in more detail in Chap. 3.4). Well-designed experimental studies can establish a cause-and-effect relationship between the response and explantory variables. However, using experimental studies is not always possible.

In experimental studies, the researcher *creates* changes in the explanatory variable, and *notes* the changes in the response variable. That is, *experimental studies are used to answer interventional RQs*.

Definition 3.4 (Experiment)**.** In an *experimental study* (or an *experiment*), the researchers intervene to control the values of the explanatory variables (C) that are applied to the individuals. The researchers *allocate* treatments (i.e., apply the intervention).

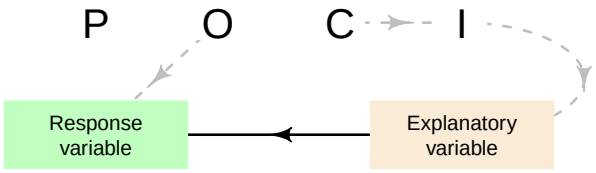


FIGURE 3.4: *An experimental study, used to answer interventional RQs*

Definition 3.5 (Treatments)**.** *Treatments* are the conditions of interest that those in the study can be exposed to (in the comparison/connection). In experiments, treatments are imposed by researchers.

Two types of experimental studies (Table 3.2) are:

- *True experiments*; and
- *Quasi-experiments*.

TABLE 3.2: *Comparing experimental designs (descriptive studies do not have any comparison or connection groups)*

Study type	Researchers allocate who or what to groups	Researchers allocate treatments to groups	Reference
True experiment	Yes	Yes	Sect. 3.4.1
Quasi-experiment	No	Yes	Sect. 3.4.2
Observational	No	No	Sect. 3.3

3.4.1 True experimental studies

True experiment are commonly used, but cannot always be conducted. An example of a true experiment is a *randomised controlled trial*, often used in drug trials.

Definition 3.6 (True experiment)**.** In a *true experiment*, the researchers:

- allocate treatments to groups of individuals (i.e., decide the values of the Comparison/Connection used on the individuals), *and*
- determine who or what individuals are in those groups.

While these may not actually happen explicitly, they can happen conceptually.

Example 3.7 (True experiment)**.** The echinacea study (Barrett et al. 2010) (Sect. 2.7) could be designed as a *true experiment*. The researchers would allocate individuals to one of two groups, and then decide which group took echinacea and which group did not (Fig. 3.5).

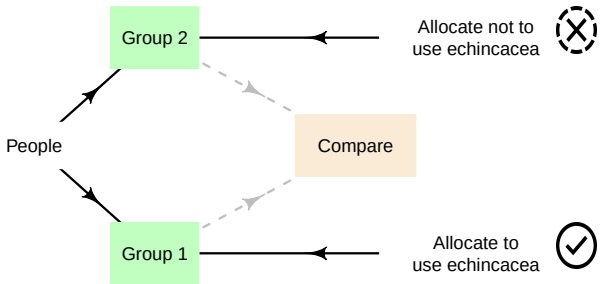


FIGURE 3.5: *True experimental studies*

Think 3.2 (Experimental study)**.** *A researcher wants to examine the effect of an alcohol awareness program (MacDonald 2008) on the amount of alcohol consumed in O-Week. She runs the program at UQ only, then compared the average amount of drinking per person at two universities (A and B). What type of study is this:* observational *or* true experimental*? Answer these questions to help:*

1. *Does the researcher allocate treatments to the groups?*
2. *Does the researcher allocate subjects to groups?*

Answer: The answer is given in the online book.

3.4.2 Quasi-experimental studies

Quasi-experiments are similar to true experiments, but treatment are *allocated* to groups that already exist (as in Example 3.2).

Definition 3.7 (Quasi-experiment)**.** In a *quasi-experiment*, the researchers:

- allocate treatments to groups of individuals (i.e., decide the values of the Comparison/Connection used on the individuals), but
- do **not** determine who or what individuals are in those groups.

Example 3.8 (Quasi-experiments)**.** The echinacea study (based on Barrett et al. (2010)) (Sect. 2.7) could be designed as a quasi-experiment. The researchers would need to *find* (not *create*) two existing groups of people (say, from two different suburbs) then decide which group took echinacea and which group did not (Fig. 3.6).

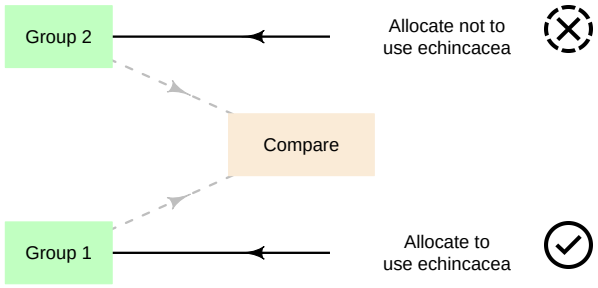


FIGURE 3.6: *Quasi-experimental studies*

3.5 Comparing study designs

In *experimental* studies, researchers *create* differences in the explanatory variable through allocation, and note the effect of this on the response variable. In *observational* studies, researchers *observe* differences in the explanatory variable, and note the values in the response variable. Different RQs require different study designs (Table 3.3).

TABLE 3.3: *Study types and research questions*

RQ type	P	O	C	I	Study type
Descriptive	Yes	Yes			Descriptive
Relational	Yes	Yes	Yes		Observational
Interventional	Yes	Yes	Yes	Yes	Experimental

Importantly, *only well-designed true experiments can show cause-and-effect*. Well-designed true experiments provide stronger evidence than quasi-experiments, which produce stronger evidence than observational studies.

However, experimental studies are often not possible for ethical, financial, practical or logistical reasons.

Well-designed quasi-experiments and observational studies can still produce strong conclusions, but cannot be used by themselves to establish cause-and-effect conclusions.

3.6 External and internal validity

All studies should be designed to be **externally valid** (Chap. 5) and **internally valid** (Chaps. 7 and 8) as far as possible.

A study is *externally valid* if the results are likely to be generalise to *other* groups in the *population*, apart from those studied in the sample.

For a study to be *externally* valid, it first needs to be *internally* valid. Using a *random sample* helps ensure external validity. In addition, the use of *inclusion* and *exclusion criteria* (Sect. 2.3.1) helps clarify to whom or what the results may apply outside of the sample being studied.

Definition 3.8 (External validity)**. Externally validity** refers to the ability to generalise the results to other groups in the *population*, apart from the sample studied.
For a study to be truly externally valid, the sample must be random sample.

A study is *externally valid* if the results from the sample studied are likely to apply to the intended population. It does not mean that the results apply more widely than the intended population.

Example 3.9. Suppose the *population* in a study is *Queensland university students*. The sample would be the students studied. The study is externally valid if the sample is a random sample from the population of students.

The results will not necessarily apply to Queensland residents, but this **has nothing to do with externally validity**. External validity concerns how the *sample* represents the intended population in the RQ, which is *Queensland university students*. The study is not concerned with all Queensland residents.

Internally validity refers to how reasonable and logical it is to draw connections between the outcome and the comparison/connection: that is, the strength of the *inferences* made from the study.

High internal validity means that changes in the response variable can confidently be related to changes in the explanatory variable *in the group that was studied*; the possibility of other explanations for changes in the response variable have been minimised.

Definition 3.9 (Internal validity)**. Internally valid** refers to the strength of the association between the outcome and the comparison/connection.
In a study with high internal validity, the association between the outcome and the comparison/connection can be attributed to that comparison/connection, rather than to other factors.

One of many threats to internal validity might be that the groups being compared are different to begin with (for example, if the group receiving echinacea is younger (on average) than the group receiving no medication).

To check this, the *baseline characteristics* of the individuals in the groups can be compared: the groups being compared should be as similar as possible, so that any differences in the outcome cannot be attributed to pre-existing difference in the two groups being compared.

Example 3.10 (Baseline characteristics)**.** In a study of treating depression in adults (Danielsson et al. 2014), three treatments were compared: exercise, basic body awareness therapy, or advice.

If any differences between the treatments were found, the researchers need to be confident that the differences were due to the treatment.

For this reason, the three groups were compared to ensure the groups were similar in terms of average ages, percentage of women, taking of anti-depressants, and many other aspects.

An *internally valid* study requires studies to be carefully designed; this is discussed at length later (Chaps. 7 and 8). In general, well-designed experimental studies are more likely to be internally valid than observational studies (Fig. 3.7).

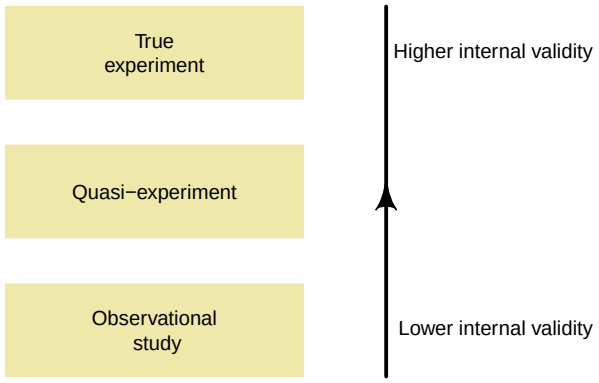


FIGURE 3.7: *Well-designed true experiments are more likely to have high internal validity*

3.7 The importance of design

Choosing the *type* of study is only a small part of research design. Planning the data collection process, and actually collecting the data, is still required. Data may be obtained by:

- Using data already available: This is called *secondary data*.
- Collecting new data: This is called *primary data*.

Either way, knowing *how* the data are obtained is important. The design phase is concerned with planning the best way to obtaining the data to ensure the study is *internally* and *externally* valid, as far as possible.

External validity, considerations include:

- Sampling: Since we can't study the whole population, *who* or *what* do we study in the population (Chap. 5)? And *how many* do we need to study? (We need to learn more before we can answer this critical question.)

Internal validity considerations include:

- *What else* might influence the values of the response variable, apart from the explanatory variable? (Chap. 6)
- Effectiveness: How can the study be designed *effectively* to maximise internal validity? (Chaps. 7 and 8)
- Data collection: How exactly will the data be *collected*? (Chap. 10)

Ethical issues must also be considered (Chap. 4), and the limitations of the study understood when the results are interpreted (Chap. 9).

3.8 Summary

Studies may be **observational** or **experimental**. Observational studies can usually be classified as **retrospective**, **prospective**, or **cross-sectional**. Experimental studies can usually be classified as **true experiments** or **quasi-experiments**. Cause-and-effect conclusions can only be made from well-designed true experiments. Ideally studies should be designed to be **internally** and **externally** valid.

3.9 Quick review questions

1. A study (Fraboni et al. 2018) examined the 'red-light running behaviour of cyclists in Italy.' This study is most likely to be:
2. A study of a sample whose results apply to the wider population of interest would be called
3. In a quasi-experiment, the researchers allocate treatments to groups that they have not organised. True or false?

3.10 Exercises

Selected answers are available in Sect. D.3.

Exercise 3.1. In a study on the shear strength of recycled concrete beams (Gonzalez-Fonteboa and Martinez-Abella 2007), beams were divided into three groups. Different loads were then applied to each group, and the shear strength needed to fracture the beams was measured.

Is this a *quasi-experiment* or a *true experiment*? Answering these questions may help:

1. Do researchers allocate treatments to the groups?
2. Do researchers allocate the who or what to groups?

Exercise 3.2. A study had this aim:

> To compare the effectiveness of alternating pressure air mattresses vs. overlays, to prevent pressure ulcers.
> Manzano et al. (2013) , p. 2099.

Patients were *provided* with either alternating pressure air overlays (in 2001) or alternating pressure air mattresses (in 2006). The number of pressure ulcers were recorded.

This study experimental, because the researchers *provided* the mattresses. Is this a *true* experiment or *quasi*-experiment? Explain.

Exercise 3.3. Consider this initial RQ (based on Friedmann and Thomas (1985)), that clearly requires a lot of refining:

> Are people with pets healthier?

To answer this RQ:

1. Describe a useful and practical definition for P, O and C.
2. Describe an *experimental* study to answer the RQ.
3. Describe an *observational* study to answer the RQ.

Exercise 3.4. Consider this journal extract:

> We randomly assigned 811 overweight adults to one of four diets [...] The diets consisted of similar foods and met guidelines for cardiovascular health [...] The primary outcome was the change in body weight after 2 years in [...] comparisons of low fat versus high fat and average protein versus high protein and in the comparison of highest and lowest carbohydrate content.
> — Sacks et al. (2009), p. 859

1. Define POCI.
2. Is this study observational or experimental? Why?
3. Is this study a quasi-experiment or a true experiment? Why?
4. What are the units of analysis?
5. What are the units of observation?
6. What is the response variable?
7. What is the explanatory variable?

4

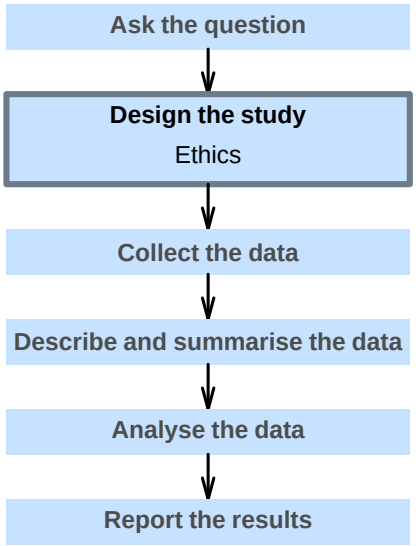
Ethics in research



So far, you have learnt how to ask a RQ, and identify different ways of obtaining data.

In this chapter, you will study how to conduct a study ethically. You will learn to:

- describe and list areas of academic integrity.
- list common ethical issues to be considered in study design.



4.1 Ethical guidelines

Studies *must* be designed to be ethical, and *must* meet ethical guidelines. Every Australian university (and probably every univeristy in the world) is committed to promoting and enforcing responsible research practices (RRPs), for people, animals and the environment.

Studies need to be ethical to minimise risk of harm to the environment and to participants, and to preserve the well-being, dignity, rights and safety of participants (including animals!).

Most research studies require a massive ethics approval process, which must be approved by an ethics committee. This process is necessary for any research conducted at all AUstralian universities and research organisations (such as Queensland Health).

There is insufficient space to cover all ethical issues in detail, but some are obvious and many ethical issues are common-sense.

Example 4.1 (Ethics)**.** Some people think that ethics applies only to studies involving people and/or animals. However, this is **not true**: ethics is important in *all* types of research. For example:

- An evaluation of ethics in *engineering* (Rubbo et al. 2019) found that 238 engineering articles published between 1945 and 2015 were retracted; the most common reason for retraction was unethical research practice.
- A study of 807 researchers in *ecology* (Fraser et al. 2018) found very high rates of Questionable Research Practices (QRPs) by researchers (such as deciding on hypotheses *after* results are known), often above 50% for some types of QRPs; these results were similar to the rates of QRPs in psychology.
- Coudert (2019) documents retractions in the *chemical sciences* in 2017 and 2018, a total of 331 articles. The reasons for the retractions include unethical practices such as falsification of data and plagiarism.

Extra example: A study (Schwitzgebel 2009) found that those who study ethics (called *ethicists*) were more likely to steal library books than other philosophers.

4.2 Common ethical issues

Common ethical issues to consider are:

- **Physical risks**: Participants should not experience physical harm or discomfort.
- **Psychological risks**: Participants should not experience psychological harm or discomfort.
- **Social risks**: Participants should not experience any social harm or discomfort.
- **Environmental risks**: Any damage to the environment should be minimised.
- **Economic risks**: Participants should not experience any significant financial loss. Reimbursements of reasonable costs may need to be considered.
- **Incentives to participate**: If participants are offered incentives to participate (above reimbursement of costs), these should be acknowledged as it may (perhaps unconsciously) cause participants to influence the results.
- **Legal risks**: Participants should not be put in the position of breaking any laws.
- **Acknowledgement**: All those who contributed should be acknowledged.
- **Confidentiality**: Data should be kept confidential.
- **Storage of data**: Data should be stored securely.
- **Consent**: Participants should consent to being in the study, and hence should be told what the study involves. Participants should also be able to withdraw from the study without penalty.
- **Plagiarism**: The work of others should be appropriately acknowledged (Sect. 37.15).
- **Analysis**: The analysis must be approached ethically, and using the appropriate methods.

Example 4.2 (Ethics)**.** In the Tuskegee syphilis experiment (Corbie-Smith 1999) (conducted between 1932 and 1972), effective treatments were witheld from men with syphilis. The men's wives and children often were affected.

The men were lied to about the treatment they were given, and were prevented from seeking treatment elsewhere. This was a highly unethical study, and could never be conducted now.

Extra example: In 1986, the American space shuttle *Challenger* exploded just after launch, killing all seven astronauts on board.

A review of the disaster (Dala et al. 1989) found that part of the cause was that the engineers dismissed some data that they should have used. This was unethical scientific practice.

4.3 Academic integrity

Academic integrity refers to conducting research ethically, honestly and responsibly.

The opposite of academic integrity is *academic misconduct*. You are **strongly encouraged** to read your university's information about academic integrity and academic misconduct, including the information about the *consequences* of academic misconduct.

In the context of research design, academic integrity covers areas such as

- Collusion;
- Fraud;
- Reproducible research; and
- Plagiarism, discussed later.

4.3.1 Collusion

Collusion occurs when people work together to produce a work, but only one gets the credit for it.

In a research context, *collusion* means failing to acknowledge the contributions and ideas of others.

4.3.2 Fraud

In the context of research, **fraud** refers to the intent to deceive. This may happen by falsifying data, inventing data, forgery, fabricating experiemnts of information.

Example 4.3 (Fraud)**.** Microbiologist Keka Sarkar had papers retracted due to fraud, including self-plagiarism and reusing figures that were claimed to be from different studies:

> Two figures in the paper in *Biotechnology Letters* had been taken from [another paper of Sarkar's]: Fig. 2c was identical to Fig. 3 in the *J Nanobiotech* paper and Fig. 4 (inset) was identical to Fig. 7 (left). No acknowledgement was given that these figures were identical. In addition, the two figures illustrated results from apparently different experiments [...] Figure 2a in the *Biotechnology Letters* paper was also used without modification in another publication of this group...
> — Chatterjee and Sarkar (2015), p. 1527

4.3.3 Reproducible research

One way to ensure that the results of research are reliable and trustworthy is to ensure that research is *reproducible*: that someone else can repeat the study (including the analysis):

> Reproducibility involves methods to ensure that independent scientists can reproduce published results by using the same procedures and data as the original investigators. It also requires that the primary investigators share their data and methodological details. These include, at a minimum, the original protocol, the dataset used for the analysis, and the computer code used to produce the results.
> — Laine et al. (2007), p. 452

The means for ensuring that all components of research are reproducible are discipline dependent, are beyond the scope of this book. However, realising the importance of reproducibility is important; for example, it emphasises the importance of describing the protocol. Different journals also have different expectations regarding reproducibility.

Many researchers strongly advise *against* using point-and-click interfaces (such as found in Excel) to analysis since the results are not reproducible:

> ... it is increasingly clear that traditional means of data management, such as storing and manipulating data using spreadsheets, are no longer adequate [...], and new approaches are required for analyzing data.
> — Lewis et al. (2018), p. 172

The importance of reproducibility in the analysis phase is crucial:

> There are serious medical consequences to errors attributable to the effects of spreadsheet programs and software operated through a graphical user interface... Fundamentally, the issue is one of reproducibility. The opacity of graphical user interface–based statistical analysis and the importance of research transparency and reproducibility have been high-lighted by scientific scandals that could have been avoided through a reproducible research paradigm...
> — Simons and Holmes (2019), p. 471

Rather than spreadsheets, which have significant problems when used for analysis, using analysis tools which enable reproducible research (for example, by using scripts all actions are documented), such as R (R Core Team 2018) (on which jamovi is based), are recommended:

We have all had the experience of having performed a laborious calculation in a spreadsheet program only to later be required to redo the analysis because of the availability of additional data, the discovery of an error, or because the analysis is part of a recurring report (e.g., monthly quality indicators). At that point we may have to return and begin the calculation all over, except we may not even remember what we did, or we may inadvertently perform the analysis in a slightly different way each time.
Simons and Holmes (2019), p. 471

Example 4.4 (Unethical reporting and practice)**.** A 'Letter to the Editor' of paramedicine journal (George et al. 2015) questioned an article (Hosseini et al. 2015) in which researchers claimed to have *randomly* allocated participants into two groups. The Letter noted that the initial average weights of the participants in each group were significantly different. The article states:

> It is extraordinarily unlikely that any variable would be that different between two groups if allocation was truly random. Even it was truly random, the stated method of "the samples were randomly divided into two groups"... does not describe the "method used to generate the random allocation sequence" [...] details specified by Consolidated Standards of Reporting Trials (CONSORT)...
> — George et al. (2015)

4.4 Summary

Making studies ethical is not negotiable. Any formal study must obtain ethical approval. Ethics covers issues such as, but not restricted to, physical risks, psychological risks, legal risks, confidentiality, consent and plagiarism.

4.5 Quick review questions

Indicate whether the following are true or false.

1. Ethical considerations apply in *any* type of study.
2. Ethical considerations *only* refer to the interactions of the researchers with participants in the study.
3. Ethical considerations *only* apply when *people* are involved in the study.
4. Ethical considerations only apply when *people* or *animals* are involved in the study.
5. Ethical considerations can extend to storage of data and plagiarism.
6. Ethics only apply to the design of the study.

7. Ethics apply even to the analysis of the data.
8. Ethical clearance is provided by (for example) the University Ethics Office.

4.6 Exercises

Exercise 4.1. Consider this (real) conundrum facing researchers (Crozier and Schulte-Hostedde 2015):

> A research team has an extraordinarily successful long-term study of a population of bighorn sheep (*Ovis canadensis*) on Ram Mountain [...]
> The population contains marked individuals for which the research team has incredibly detailed data on phenotype, pedigree, and life-history. Many graduate students, post-doctoral fellows, and senior scientists have studied this population, and this research has lead to numerous important publications.
> Recently, however, a cougar (*Puma concolor*) that has learned to specialize on these sheep is slowly but surely eating all of them. This is a study of a natural population, which includes predation, but this cougar is drastically reducing the sample size of the study.
> Since it is legal to hunt cougars in the region where this study is taking place, one option is to try to kill the predator; however, even if a cougar were successfully hunted, this would not ensure that it was the correct one.

What action would you recommend? Explain your reasoning, including from an ethics point-of-view.

Exercise 4.2. Suppose a deadly disease breaks out. Is it ethical to begin the use of a new drug to treat those affected, even though the drug is still experimental and the potentially harmful side effects are unknown? Discuss your point-of-view.

Exercise 4.3. Is it ethical to lie to the subjects in a study? *Deception* is common in some disciplines, and may be approved by ethics committees under certain circumstances (such as the potential benefits of the study, and whether the deception is likely to cause physical or psychological discomfort to the participants).

Do you think it is ethical to tell participants that they are taking an active medication, when it is actually ineffective (a 'placebo') (Waber et al. 2008)? Discuss the advantages and disadvantages, including what we can learn from such a study that may be beneficial.

5

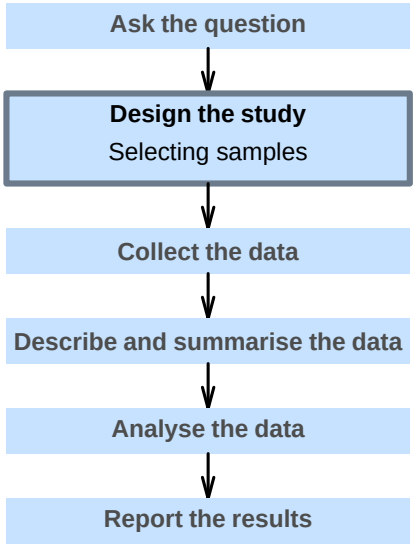
External validity: Sampling



So far, you have learnt to ask a RQ, identify different ways of obtaining data, and design the study.

In this chapter, you will learn how to obtain the sample to study. You will learn to:

- distinguish, and explain, precision and accuracy.
- distinguish between random and non-random sampling.
- select random samples.
- identify, describe why random samples are preferred over non-random samples.
- identify, describe and use simple random sampling.
- identify, describe and use systematic sampling.
- identify, describe and use stratified sampling.
- identify, describe and use cluster sampling.
- identify, describe and use multistage sampling.
- identify ways of obtaining samples that are more likely to be representative samples.



5.1 The idea of sampling

A RQ implies that every member of the population should be studied (the **P** in **P**OCI stand for 'population'). However, being able to do so is very rare because of cost, time, ethics, logistics or practicality. Hence, a subset of the population (a *sample*) is almost always studied.

A sample consists of some *individuals* (or *cases*, or (if the individuals are people) *subjects*) from the population. The *purpose* of a sample is to *approximate* the population.

A study is *externally valid* if the results can be generalised to other groups in the population, apart from the sample studied. This is only possible if the sample is chosen to well-represent the whole population.

However, since a sample doesn't include every member of the population, conclusions made from samples cannot be certain to apply to the whole population.

⚠ In research, the goal is to learn about the *population*, but only a *sample* can be studied.

This book is essentially about how to learn about a population based on an imperfect sample.

Example 5.1 (Samples)**.** A study (based on Lipton et al. (1998)) of the effect of aspirin in treating headaches cannot possible use every single human alive who might one day wish to take aspirin.

Not only would this be prohibitively expensive, time-consuming, and impractical, but such a study would not even use those humans who had not been born yet who might use aspirin. (That is, using the whole target population is *impossible*.) A *sample* must be used.

Having seen that using a sample is necessary, other issues are raised:

- *How* can we learn something useful about the *whole* population if only *some* of that population is studied?
- *Which* individuals should be included in the sample?
- *How many* individuals should be included in the sample be?

The last issue must be left until later, after learning more about the implications of studying a sample rather than a population.

Using a sample instead of the entire population presents challenges. *Every sample is likely to be a bit different*, so what is learnt from a sample depends on which individuals happen to be present in the sample being used. This is called *sampling variation*. That is, each sample produced different data, and so may lead to different answers to the RQ.

This is the challenge of research: *How to make decisions about populations, using an imperfect sample information.* Perhaps surprisingly, *lots* can be learnt about the population if we approach the task of selecting a sample correctly.

⚠ Almost always, *samples* are studied, not *populations*.

Every sample is likely to be different, and hence the *results from every sample are likely to be different*. This is called *sampling variation*.

While *we can never be certain* about the conclusions from the sample, special tools can be used to help make decisions about the *population* from a *sample*.

Consider a fair pack of cards, where 50% of cards are red. Figure 5.1 shows five samples of 10

cards, and the percentage of red cards is not the same in every sample (and not necessarily 50%).

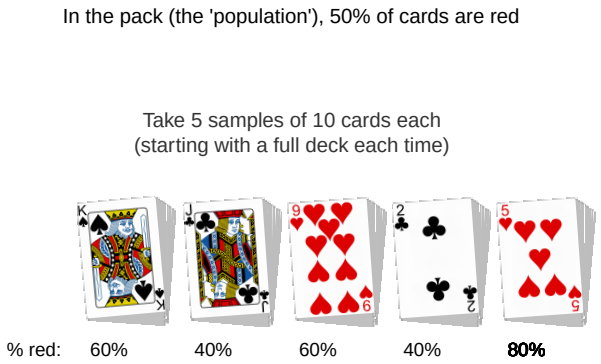


FIGURE 5.1: *The sample of 10 cards do not always produce the same sample percentage of red cards*

Think 5.1. *Two surveys were conducted before the 1936 presidential election in the USA to predict the winner (Bryson 1976), summarised in Table 5.1. Which do you think predicted correctly the winner of the election? Why?*

TABLE 5.1: *Two surveys about the USA presidential election*

Study	Number in sample	Population	Method
A	10 000 000	Specific groups	Voluntary survey
B	50 000	All Americans	Random survey

5.2 Precision and accuracy

Two issues concerning sampling, raised in Sect. 5.1, were: *which* individuals should be in the sample, and *how many* individuals should be in the sample be. These two issues address two different aspects of sampling: **precision** and **accuracy** (Fig. 5.2).

Accuracy refers to how close a *sample* estimate is to the *population* value (on *average*). **Precision** refers to how close all the possible sample estimates are likely to be (that is, how much variation is likely in the sample estimates).

Definition 5.1 (Accuracy)**.** *Accuracy* refers to how close a *sample* estimate is to the *population* value, on average.

Definition 5.2 (Precision)**.** *Precision* refers to how close the sample estimates from different samples are likely to be to each other.

Using this language:

- The *type* of sampling (i.e., the way in which the samples in selected) impacts the *accuracy* of the sample estimate. In other words, the type of sampling impacts the *external validity* of the study.
- The *size* of the sample impacts the *precision* of the sample estimate.

For example, large samples are more likely to be *precise* estimates because each possible sample value will produced similar estimates, but they may or may not be accurate estimates. Similarly, random samples are likely to produce *accurate* estimates (and hence the study is more likely to be externally valid), but they may not be *precise* unless the sample is also large.

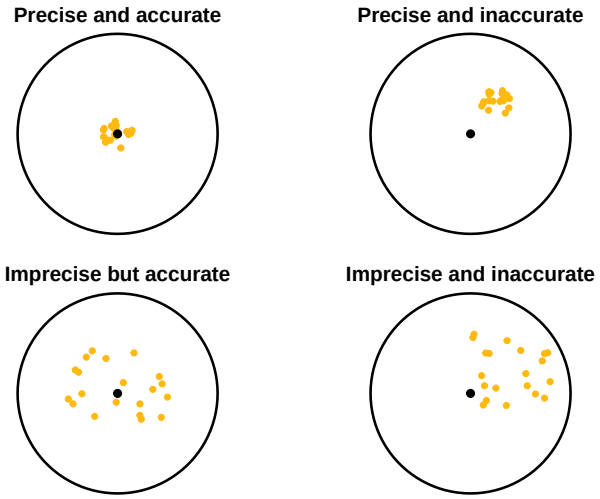


FIGURE 5.2: *Precision and accuracy: Each coloured dot is like a sample estimate of the population value (shown by the black central dot)*

Example 5.2 (Precision and accuracy)**.** To estimate the average age of *all Queenslanders*, we could ask 9000 Queensland school children (a large sample indeed!).

This will give a *precise* answer because the sample is large, but *inaccurate* answer because the sample is not representative of *all* Queenslanders. In fact, the sample may give a precise answer to a *different* question: 'What is the average age of Queensland school children?'

5.3 Types of sampling

One key to obtaining accurate estimates about the population is to ensure that the sample studied is representative of the population of interest (that is, to ensure the study is externally valid).

So, *how* can a representative sample of the population be found? Whenever a sample is taken, only *some* of the population is selected. The selected individuals can be chosen using either **random sampling** or **non-random sampling**.

The word *random* here has a specific meaning that is different than how it is often used in everyday use.

Definition 5.3 (Random)**.** In research and statistics, *random* means "determined completely by chance."

5.3.1 Random sampling methods

In a **random sample**, each individual in the population can be selected on the basis of impersonal chance. (Remember than *random* means that the sample is determined completly by chance!) Some examples of random sampling methods appear in the following sections (Table 5.2).

⚠ The results obtained from a random sample probably generalise to the population from which the sample is drawn; that is, *random samples* are likely to produce *externally valid* studies.

TABLE 5.2: *Comparing four types of random sampling*

Type	Stage 1	Stage 2	Reference
Systematic	Start at a *random* location	Take every nth element thereafter	Sect. 5.5
Stratified	Split into a few large groups ('strata')	Select **simple random sample* from every stratum	Sect. 5.6
Cluster	Split into many small groups ('clusters'); select *simple random sample* of clusters	Select all in the chosen clusters	Sect. 5.7
Multi-stage	Select *simple random sample* from the larger stage	Select *simple random sample* from those chosen in Stage 1; etc.	Sect. 5.8

Consider testing a pot of soup by 'sampling.' If the soup is stirred, we don't need to taste the whole pot of soup to see how the soup tastes.

The same principle applies in research: If we use a random sample (analagous to the stirring the soup), we don't need to study *every* member of the population. If we don't use a random sample (that is, we don't stir the soup), we do not get an *overall* impression of the population (or the soup).

5.3.2 Non-random sampling methods

A **non-random** sample requires some kind personal input. Examples of non-random samples include:

- *Judgement sample*: Individuals are selected, based on the researchers' judgement, depending on whether the researcher thinks they are likely to be agreeable or helpful. For example, researchers may decided to survey people who are not in a hurry.
- *Convenience sample*: Individuals are selected because they are convenient for the researcher. For example, researchers may gather data from their family and friends.
- *Voluntary* response (self-selecting) sample: Individuals participate if they wish to. For example, a voluntary response survey, or a TV station call-in survey.

In non-random sampling, those who *are* in the study may be different than those who *are not* in the study. That is, *non-random samples are not likely to be externally valid*.

⚠ Using a non-random sample means that the results may not generalise to the intended population: they probably do not produce externally valid studies.

Example 5.3 (Different ways to sample)**.** During the COVID-19 (coronavirus) pandemic in 2020, a Facebook poll[1] asked the question:

> Do you think a Coronavirus vaccine should be compulsory?

The result was reported as '79 per cent of Australians oppose a compulsory vaccination,' from a sample of over 53,000 responses.

However, this sample was a *voluntary response sample*, not a random sample, so the results may not be *accurate*. For example, many anti-vaccination groups instructed their members to flood the poll with 'No' responses (including celebrity chef Pete Evans), and the poll could have been completed by non-Australians as well as Australians.

A different study (Smith et al. 2020) asked Australians:

> The Federal Government's 'No Jab, No Pay' policy withholds certain benefits and payments from families who don't fully vaccinate their children. Do you agree with this policy?

In the sample of 1809 respondents, 83.7% either agreed or strongly agreed with this statement.

While this study did not use a *random* sample, the researchers made efforts to sample a *representative* cross-section of Australians:

> Researchers recruited Australian adults aged 18-years and older to participate in the study through a large, well-established online panel provider. While not a random sample of the Australian population, researchers made efforts to ensure the sample included individuals representing a wide range of demographics (e.g., age, gender, location, income, political preferences, religiosity).
> — Smith et al. (2020), p. 194

Further more, 'respondents were paid small token sum for their participation in the study' to encourage all selected respondents to provide an answer.

In Sect. 5.11, *random* and *non-random samples* are compared using an example.

5.4 Simple random sampling

Definition 5.4. In a *simple random sample*, *every* possible sample of the same size has *same* chance of being selected.

A simple random sample is chosen from a list of all members of the population (the *sampling frame*) using tables of *random numbers* (Appendix B.1) or even websites like `https://www.random.org`.

Definition 5.5 (Sampling frame)**.** The *sampling frame* is a list of *all* the members of the population (the *individuals*, or *cases*, or *subjects*).

Often, establishing the sampling frame is difficult or impossible, and so finding a random sample is also difficult.

For example, to study a simple random sample of wombats (Yang et al. 2018) would require having a list of all wombats, so some could be selected using random number tables. This is clearly absurd, and other random sampling methods, such as special ecological sampling methods[2], would be used instead (Manly and Alberto 2014).

Other good (random) sampling methods use a system to select randomly, rather than by human choice (discussed in the following sections).



This book always assumes simple random samples, for simplicity, unless otherwise noted.

Example 5.4 (Simple random sampling)**.** Suppose we are interested in this RQ:

> For students at a large course at a particular university, is the average number of letters typed on a keyboard in 10 seconds the same for females and males?

Suppose a sample of 40 students is needed. The sampling frame is the list of all students enrolled. Obtaining the *sampling frame* is feasible here (lecturers have access to this information for grading).

Think 5.2 (Simple random sampling)**.** *Suppose budget and time constraints mean only 40 students can be selected for the study above.*

Describe how *to use the course enrolment list to find a simple random sample of 40 students to study.*

[2] `http://www.countrysideinfo.co.uk/what_method.htm`

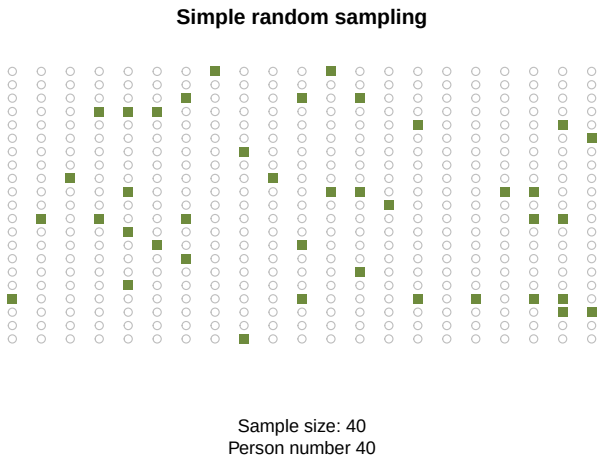


FIGURE 5.3: *Taking a simple random sample of size 40*

5.5 Systematic sampling

In **systematic sampling**, the first case is *randomly* selected; then, every (say) fifth element is selected thereafter.

In general, we say that every nth individual is selected.

⚠ There is no advantage in using a systematic sample to take every nth individual if it is just as easy and efficient to take every individual.

Example 5.5 (Systematic sampling)**.** Suppose for a particular study, a sample of 40 students in a particular course is needed.

If 441 students are enrolled. 40 students could be randomly selected, by choosing a number at random between 1 and $441/40$ (approximately 11) as a starting point; suppose the random number selected is 9. The first student selected is the 9th person in the list (which may be ordered alphabetically, by student ID, or any other means).

Thereafter, every $441/40$th person, or 11th person, in the list is selected: people numbered 9, 20, 31, 42…

Figure 5.4 shows a diagram of selecting 40 students from a class of 441 students using a systematic random sample, by starting at student number 9 and then talking every 11th person. (The online version has an animation.)

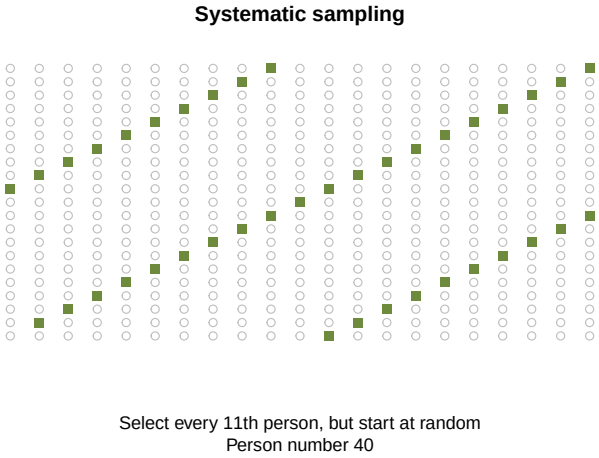


FIGURE 5.4: *Taking a systematic random sample of size 40, by starting at student number 9 and then talking every 11th person*

⚠ Care needs to be taken when using systematic samples to ensure a pattern is not hidden.

For example, consider a study where residents in a large, eight-level residential accommodation complex are to be visited and a survey administered. Each floor of the building has a similar layout (Fig. 5.5), with nine apartments per level.

If the researchers decide to systematically sample every *tenth* apartment, the very same apartment on each floor woud be chosen.

For example, suppose Apartments 1-10, 2-10, 3-10,... 8-10 were chosen. These apartments are all larger than all the other apartments. The residents of these apartments may be wealthier than the other residents, so the systematic sample will not be a representative sample of residents.

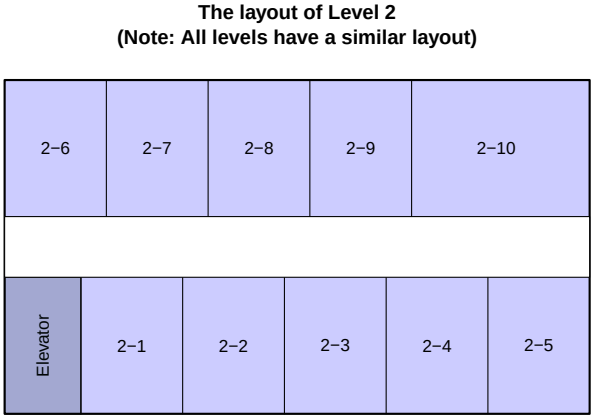


FIGURE 5.5: *The layout of each level in an eight-storey apartment building; Level 2 is shown*

5.6 Stratified sampling

In **stratified sampling**, the population is split into a small number of large (usually homogeneous) groups called *strata*, then cases are selected using a *simple random sample* from *each* stratum.

⚠ The strata must be unrelated to the variables.

For example, if the RQ is about comparing the percentage of females and males who wear hats at midday, a *stratified sample* of size 100 is **not** obtained by selecting 50 females and 50 males, for example. This is merely selecting people from each level of the explanatory variable.

The sex of the person is the explanatory variable; it does not define the strata.

Example 5.6 (Stratified sampling)**.** To select students in a large course at a particular university, 20 of the females and 20 of the males could be selected. The sample is stratified by *sex* of the person.

At the university where I work, about 67% of the students are females. So, I could ensure that two-thirds of the sample was females (around 26.7, say 27) and about one-third males (about 13.3, say 13).

Figure 5.6 shows a diagram of selecting a stratified random sample of size 40, by randomly selecting 20 female and 20 male students (The online version has an animation.) Similarly, Fig. 5.7 shows a diagram of selecting a stratified random sample of size 40, by randomly selecting 27 female and 13 male students. (Again, the online version has an animation.)

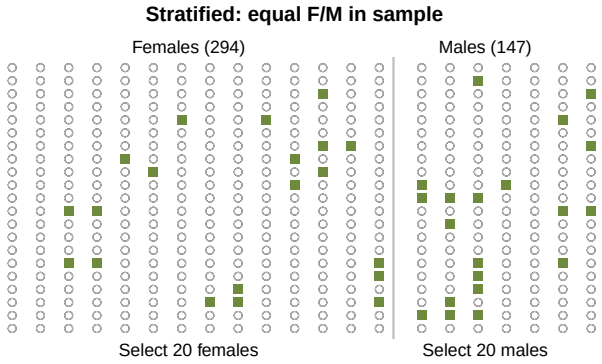


FIGURE 5.6: *Selecting a stratified random sample of size 40, by randomly selecting 20 female and 20 male students.*

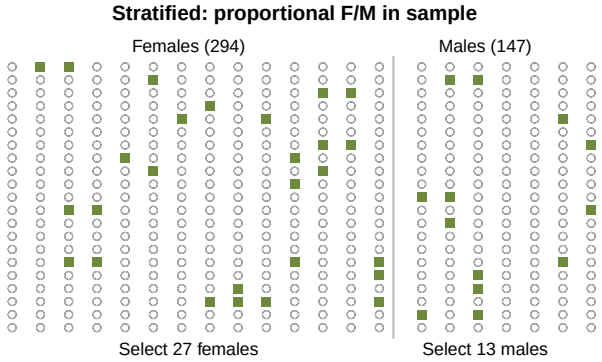


FIGURE 5.7: *Selecting a stratified random sample of size 40, by randomly selecting 27 female and 13 male students.*

5.7 Cluster sampling

In **cluster sampling**, the population is split into a large number of small groups called *clusters*, then a *simple random sample* of clusters is selected and *every* member of the chosen small groups is part of the sample.

> **Example 5.7** (Cluster sampling)**.** To select students in a large course at a particular university again, a simple random sample of (say) three of the many tutorials could be selected, and *every* student enrolled in those selected tutorials constitute the sample.

Figure 5.8 shows a diagram of selecting approximately 40 students using cluster sampling, using the tutorials as clusters. (The online version has an animation.)

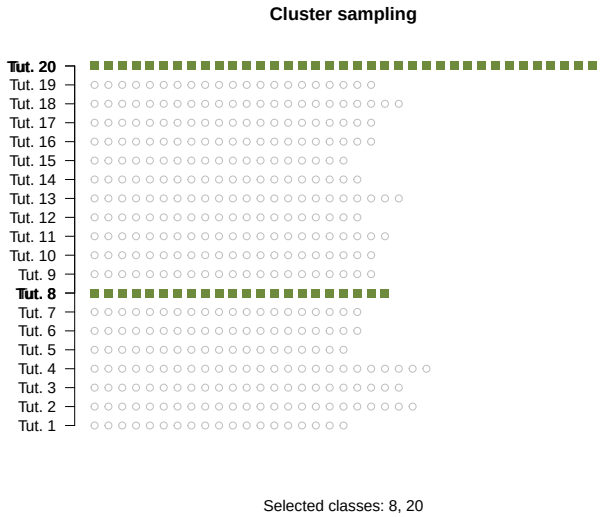


FIGURE 5.8: *Taking a cluster random sample of size 40, using the tutorials as clusters*

5.8 Multistage sampling

In **multistage sampling**, large groups are selected using a *simple random sample*, then smaller groups within those large groups are selected using a *simple random sample*. The simple randomly sampling can continue for as many levels as necessary.

Example 5.8 (Multistage sampling)**.** To select students in a large coirse at a particular university again, a *simple random* sample of (say) ten of the many tutorials could be selected (Stage 1), and then 4 people *randomly* selected from each of these 10 selected tutorials (Stage 2).

Figure 5.9 shows a diagram of selecting approximately 40 students using multistage sampling, by randomly selecting 10 classes at random in Stage 1, then randomly selecting students from each class in Stage 2. (The online version has an animation.)

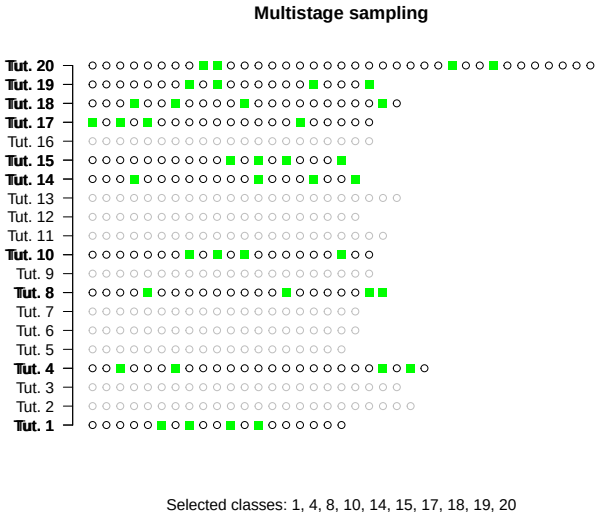


FIGURE 5.9: *Taking a multistage random sample of size 40, by randomly selecting 10 classes at random in Stage 1, then randomly selecting students from each class in Stage 2*

Example 5.9 (Multistage sampling)**.** Multistage sampling is often used by the Australian Bureau of Statistics (ABS). For example, to obtain a random sample of Queenslanders, one procedure is:

- Stage 1: Randomly select some cities in Queensland;
- Stage 2: Randomly select some suburbs in these chosen cities;
- Stage 3: Randomly select some streets in these chosen suburbs;
- Stage 4: Randomly select some houses in these chosen streets.

This is cheaper than simple random sampling, as data collectors can be employed in a smaller number of Queensland cities (only those chosen in Stage 1).

5.9 Representative sampling

Obtaining a truly random sample is usually hard or impossible, and the best we can do is to select a sample that we hope is *representative* of the population. Even so, the results from any non-random sample *may not generalise* to the intended population. The results will generalise to the population which the sample *does* represents.

Ideally, even if obtaining a random sample is impossible, prefer a sample where those *in* the sample are not likely to be different than those *not in* the sample, at least for the variables of interest.

Example 5.10 (Representative sample)**.** A randomly-chosen group of Queensland and Northern Territory residents is asked to evaluate two types of hand prosthetics.

It is probable (but not certain) that their views would be similar to those all of Australians. There is no obvious reason why residents of Queensland and the Northern Territory would be very different from residents in the rest of Australia, regarding their view of hand prosthetics.

Even though the sample is not a random sample of all Australians, the results *may* generalise to all Australians (though we cannot be sure).

Example 5.11 (Non-representative samples)**.** Suppose we wish to determine the average time per day that Australia households use their air-conditioners for *cooling* in summer.

If a group of Queensland and Northern Territory residents is asked, this sample would not be expected to represent all Australians: it would *over*-represent the average number of hours air-conditioners are used for *cooling* in summer.

In this case, those *in* the sample are likely to be very different to those *not in* the sample, regarding their air-conditioners usage for *cooling* in winter.

In contrast, suppose a group of Tasmanians was asked the same question. This second sample would not be expected to represent all Australians either (it would *under*-represent).

Again, those *in* the sample are likely to be very different to those *not in* the sample, regarding their air-conditioners usage for *cooling* in winter.

Sometimes, a *combination* of different sampling methods is used.

Example 5.12 (A combination of sampling methods)**.** In a study of pathogens present on magazines in doctors' surgeries in Melbourne, some suburbs can be selected at *random*, and then (within each suburb) surgeries are used which *volunteer* to be part of the study.

Extra example: In a study of diets of children at child-care centres, researchers used samples in 2010 and 2016, described as follows:

> In 2010, a stratified random sampling procedure was used to select representative cross-sections of providers working in licensed center-based programs and licensed providers of family home-based care from publically available lists. [...] Additional participants were also recruited in 2016 using a combination of stratified random and open, convenience-based sampling.
> — Larson et al. (2019), p. 336

Sometimes, practicalities override how the sample can be obtained, which may not result in a random sample. Even so, the impact of this on the conclusions shoud be noted (that is, in discussing the limitations of the study). Sometimes, ways exist to obtain a sample that is *more likely* to be representative.

⚠ Random samples are often difficult to obtain, and sometimes *representative* samples are the best that can be done.

In a good representative sample, those *in* the sample are not obviously different than those *not in* the sample. Try to ensure a broad cross-section of the target population appears in the sample.

Example 5.13 (Attempts to increase representativeness)**.** To find a sample of university students, students at Cafe A could be approached every Monday morning at 8am, for four consecutive weeks.

This is a *convenience* sample, and not a random sample. However, the sample would be *more likely* to be representative if a broader cross-section of students was approached:

- Students at Cafe A on Monday at 8am;
- Students at the Cafe B on Tuesday at 11:30am; and
- Students entering the Library on Thursdays at 2pm.

This is still not a *random* sample, but the sample now comprises more than just students who attend university on Mondays at 8am, at Cafe A.

Ideally, **student would not be included more than once in our sample**, though this is difficult to ensure.

Think 5.3 (Sampling)**.** *To assess the quality of bearings from a manufacturer, a researcher takes a random sample of 25 bearings from* each *of the three cases delivered.*

What type of sampling scheme is being used?

Answer: Stratified sampling.

Sometimes, information may be recorded from those in the sample, and this information used this to make some comment about whether our sample seems reasonably representative.

For example, the sex and age of a sample of university students may be recorded; if the proportion of females in the sample, and the average age of students in the sample, are similar

to those of the whole university population, then the sample may be somewhat representative of the population. (though we cannot be sure).

Example 5.14 (Comparing samples and populations)**.** A study of the adoption of electric vehicles (EVs) by Americans (Egbue et al. 2017) used a sample of $n = 121$ found through social media (such as Facebook) and professional engineering channels. This is not a random sample.

The authors compared some characteristics of the sample with the American population from the 2010 census (Table 5.3), stating:

> The sample has a higher representation of males and individuals in the 18–44 age group [...] compared to the US population. In addition, the sample has a higher representation of [...] wealthier individuals.
> — Egbue et al. (2017), p. 1931

In interpreting the results of this study, the authors say:

> ...the results of this study are more applicable to people with an engineering or technical background...
> — Egbue et al. (2017), p. 1931

TABLE 5.3: *Comparing the sample and the population (in percentages), for the EV study*

	Sample	Population
Gender		
Male	77.68	49.20
Female	22.32	50.80
Age		
Under 18	0.00	24.00
18–44	55.36	36.50
45–64	31.25	26.40
65 and older	13.39	13.00
Annual income		
Under $75,000	28.56	67.49
$75,000 and over	51.78	22.51
Prefer not to say	19.64	0.00

5.10 Bias in selecting samples

The sample may not be representative of the population for many reasons, all of which compromise how well the sample represents the population (i.e., compromises *external* validity). This is called *bias*. Biased samples are less likely to produce externally valid studies.

Definition 5.6 (Bias)**.** *Bias* is the tendency of a sample to over- or under-estimate a population quantity.

More formally:

> ... bias is the introduction of systematic error, subconsciously or otherwise, in the design, data collection, data analysis, or publication of a study.
> — Sedgwick (2014)

In *selection bias*, the wrong sampling frame may be used, or non-random sampling is used. The sample is biased because those *in* the sample may be different than those *not in* the sample.

Example 5.15 (Selection bias)**.** Consider Example 5.11, about estimating the average time per day that air conditioners are used for cooling in summer.

Using people only from Queensland and the Northern Territory in the sample is using the wrong sampling frame: the sampling frame does not represent the target population ('Australians'). This is *selection bias*.

Non-response bias occurs when chosen participants do not respond for some reason. The problem is that the responses from those who *do not* respond may be different than the responses who *do* respond. Non-response bias can occur because of a poorly-designed survey, using voluntary-response sampling, chosen participants refusing to participate, participants forgetting to return completed surveys, etc.

Example 5.16 (Non-response bias)**.** Consider a study to determine the average number of hours of overtime worked by various professions. People who work a large amount of overtime are likely to be too busy to answer the survey.

Those who answer the survey may be likely to work less overtime than those who do not answer the survey. This is an example of *non-response bias*.

Response bias occurs when participants provide *incorrect information*: the answers provided by the participants may not reflect the truth. This may be intentional (for example, if the survey questions are very personal or controversial in nature) or non-intentional (for example, if the question is poorly written or is misunderstood).

Think 5.4 (Sampling)**.** *One (true) survey concluded (Hieger (2001), cited in Bock et al. (2010), p. 283):*

> *All but 2% of the home buyers have at least one computer at home, and 62% have two or more. Of those with a computer, 99% are connected to the internet.*

The article later reveals the survey was conducted on-line *(and recall the survey was done in 2001...). What type of bias is apparent?*

Answer: Non-response bias.

Think 5.5 (Bias)**.** *For these samples, to what populations will results generalise?*

- *Obtaining data using a telephone survey.*
- *Obtaining data using a TV stations call-in.*
- *Asking your friends to participate because it is easier than finding a random sample.*

For each of the above samples, give an example of an outcome which would be likely to over-*estimate the true (population) value.*

5.11 Final example

As a demonstration sampling schemes (Marshman and Dunn Submitted), consider taking a *non-random* sample of 10% of the pixels of an image (Fig. 5.10). What is the image? Seeing the **big picture** is hard using these non-random samples.



FIGURE 5.10: *Non-random samples from an image: 5 percent of pixels (top left); 10 percent of pixels (top right); 25 percent of pixels (bottom left); 50 percent of pixels (bottom right)*

In contrast, taking *simple random sample* makes the **big picture** much clearer (Fig. 5.11).

Indeed, *any* type of random sample makes seeing the **big picture** easier.



FIGURE 5.11: *Random samples from an image: 5 percent of pixels (top left); 10 percentof pixels (top right); 25 percent of pixels (bottom left); 50 percent of pixels (bottom right)*

For example, for a *cluster sample* we treat each *column* as a cluster, and select some *columns* at random. Then, the entire chosen columns are selected.

For a *systematic sample*, we take:

- every 20th pixel for a 5% sample;
- every 10th pixel for a 10% sample;
- every 4th pixel for a 25% sample; and
- every second pixel for a 50% sample.

For a *multi-stage sample* we select some columns at random, then select some pixels in those columns at random.

For a *stratified sample*, we select :

- a simple random sample from the background greenery, and then
- a simple random sample from the person.

These two are then combined to get an overall random sample.

5.12 Summary

Almost always, the entire population of interest cannot be studied, so a **sample** (a subset of the population) must be studied. Samples can be **random samples** or **non-random samples**. Conclusions made from random samples can usually be generalized to the population (that is, they are externally valid).

Random sampling methods include **simple random samples**, **systematic samples**, **stratified samples**, **cluster samples** and **multi-stage samples**. Random samples are likely to be *externally valid*. Non-random sampling methods include **convenience samples**, **judgement samples** and **self-selecting samples**.

Random samples are often very difficult to obtain, so the best we can do is to aim for **reasonably representative** samples, where those who *are* in the sample are unlikely to be different than those who *are not* in the sample. Non-random samples *may not be externally valid*.

5.13 Quick review questions

1. Suppose we randomly select a student and send them a postal survey, but the student has moved address and so never receives the survey. What type of bias will this result in?
2. What is the main advantage of using a *random* sample?
3. What is the main advantage of using a *large* sample?
4. A *large* sample is always better than a *random* sample: True or false?
5. Suppose I classify a natural forest region into two zones, which are quite different: Region A is mostly dunes and lightly vegetated, and is on the coastal side of a ridge; Region B is more densely vegetated and on the inland side of the ridge. I then take samples of sugar ants (*Camponotus app*) from each zone to study their size. What is the best description of the *type* of sampling method being used?

5.14 Exercises

Selected answers are available in Sect. D.5.

Exercise 5.1. Suppose we needed to estimate the average number of pages in a book in a university library (including all five campuses), using a sample of 200 books.

1. Describe how you might select a *simple random sample* of books.
2. Describe how you might select a *stratified sample* of books.

3. Describe how you might select a *cluster sample* of books.
4. Describe how you might select a *convenience sample* of books.
5. Describe how you might select a *multi-stage sample* of books.
6. Which would be most *practical*?

Exercise 5.2. Suppose we need a sample of 20 residents from apartments in a large residential apartment complex, comprising 20 floors with 30 apartments in each floor. We plan to interview the residents of these apartments.

1. One approach to obtaining a sample is to randomly select five floors, then randomly select four apartments from each of those five floors, and interview the oldest resident of that apartment. What type of sampling scheme is this?
2. Another approach is to select one floor at random, and select the first 20 apartments on that floor then interview the oldest resident of that apartment. What type of sampling scheme is this?
3. Another approach is to wait at the ground-floor elevator, and ask people who emerge to participate in our interview. What type of sampling scheme is this?
4. Another approach is to select five floors at random, then wait by the elevator and interview residents as they arrive at the elevator. What type of sampling scheme is this?
5. Which of the above sampling methods are good, and which are poor? Explain your answers.

Exercise 5.3. Suppose a researcher needs a sample of customers who shop at a large, local shopping centre to complete a survey.

1. The researcher stations themselves outside the supermarket at the shopping centre one morning, and approaches every 10th person who walks past. What is the sampling method?
2. The researcher waits at the main entrance for 30 minutes at 8am every morning for a week, and approaches every 5th person. What is the sampling method?
3. The researcher leaves a pile of survey forms at an unattended booth in the shopping centre, and a locked barrell in which to place completed surveys. What is the sampling method?
4. The researcher goes to the shopping centre every day for two weeks, at a different time and location each day, and approaches someone every 15 minutes. What is the sampling method?
5. Which would the best sampling method?
6. Which (if any) of the methods produce a random sample?

Exercise 5.4. A study (Ridgewell et al. 2009) investigated how children in Brisbane travel to state schools. Suppose researchers randomly sampled four schools from a list of Brisbane state schools, and invited every family at each of those four schools to complete a survey.

What type of sampling method is this?

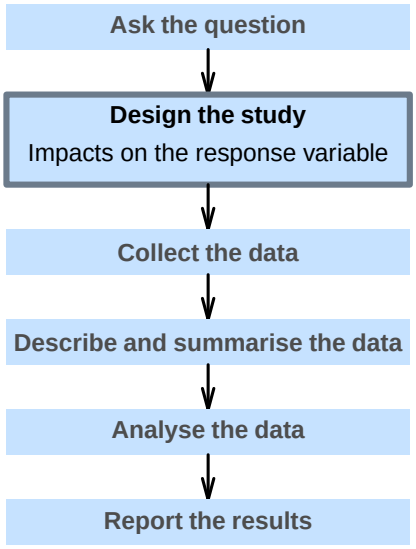
6

Overview of internal validity

So far, you have learnt to ask a RQ, identify different ways of obtaining data, and design the study.

In this chapter, you will learn how to ensure that the conclusions we can make are logical and sound. You will learn to:

- identify issues that might impact internally validity.
- identify issues that might impact the values of the response variables.
- identify extraneous, confounding and lurking variables.



6.1 Introduction

Consider the letter-typing RQ again (from Example 5.4), where this RQ was posed:

> For students in this course this semester, is the average number of letters typed on a keyboard in 10 seconds the same for females and males?

For this study:

- **P**: Students in this course this semester.
- **O**: Average number of letters typed in 10 seconds (say, 'typing speed').
- **C**: Between females and males.

- **I**: None. (The values of C cannot be allocated to students).

After measuring the typing speed (the *response variable*) of many individuals, a lot of variation will be observed in the values collected: Every student in the study is likely to have a different typing speed.

The measured typing speeds can be influenced by many issues (Fig. 6.1):

- *The explanatory variable* (Sect. 6.2): The values of the explanatory variable may influence the values of the response variable; of course, they may not either. The purpose of the study is to find out... In this example, the explanatory variable is the sex of the student.
- *Other variables* (Sect. 6.3): *Other* variables that aren't the focus of the study may influence the response variable, such as 'age' or 'whether or not the person wears glasses.' We can work with these other variables if we are careful.
- *Design issues* (Sect. 6.4): The way in which the study is *designed* can also influence the values of the response variable. These can mean *disaster* if not handled properly.
- *Chance*, or *randomness* (Sect. 6.5): Even the same person doing the same thing repeatedly will not record exactly the same reaction time every attempt. This influence is unavoidable, but we can live with it if we have some idea of the how large this variation is.

An internally-valid study is one where the association between the outcome and the comparison/connection can be attributed to that comparison/connection, rather than to other factors.

That is, an internally-valid study is one where the impacts of other possible explanations for that association (such as *extraneous variables*, *design issues*, and *chance*) have been accounted for, minimised, or are well-managed.

This is hugely important, and is the main focus of Chaps. 7 (experimental studies) and 8 (observational studies).

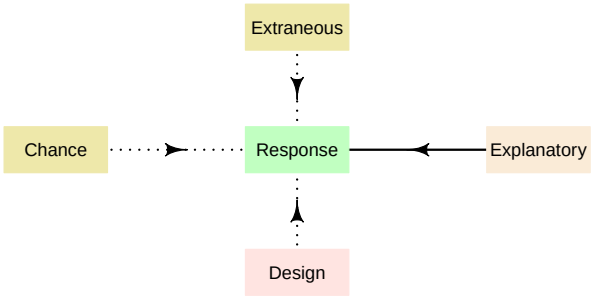


FIGURE 6.1: *What may influence the values of the response variable*

6.2 The explanatory variable and variation in the response

The explanatory variable may be associated with changes in the values of the response variable. However, it may not; after all, determining this is the purpose of the study.

If nothing else influenced the values of the response variable, life would be easy: Any change of a given size in the value of the explanatory variable would *always* results in change of the same size in the value of the response variable.

Example 6.1 (Explanatory variable)**.** In the typing-speed study (Example 5.4), the explanatory variable is the sex of the person. If nothing else influenced typing speed, all females would record the same typing speed every time, and all males would record the same typing speed every time. This is clearly unreasonable.

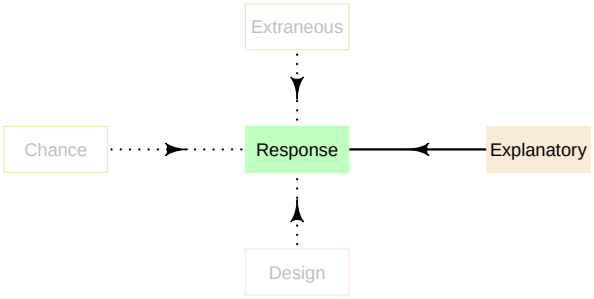


FIGURE 6.2: *Explanatory variables influence the values of the response variable*

6.3 Extraneous variables and variation in the response

Other variables probably exist which are associated with changes in the value of the response variable; these are called *extraneous variables*.

Definition 6.1 (Extranaeous variable)**.** An *extraneous variable* is any variable that is (potentially) associated with the response variable, but is not the explanatory variable.

Example 6.2. In the typing-speed study (Example 5.4), potential extraneous variables may include age, the presence or absence of certain medical conditions, the level of familiarity with computers, etc.

All extraneous variables are, by definition, related to the response variable. An extraneous variable may or may not be associated with the explanatory variable as well. Extraneous variables may have other names too (Table 6.1), though these names are used inconsistently by researchers (Dunn et al. 2016).

Definition 6.2 (Confounding variable)**.** A *confounding variable* (or a *confounder*) is an extraneous variable associated with the response *and* explanatory variables (Fig. 6.3).

Definition 6.3 (Confounding)**.** *Confounding* is when a third variable influences the relationship between the response and explanatory variable.

The problem with confounding is a relationship between the response and explanatory variables may be evident, but only because *both* of these variables are related to the confounding variable (Fig. 6.3).

Example 6.3 (Confounding variables)**.** A relationship exists between carrying cigarette lighters, and lung cancer: people who carry cigarette lighters are more likely to get lung cancer.

The only reason that this relationship exists is because of a *confounding variable*: whether or not the person is a smoker. A smoker is more likely to carry a cigarette lighter, and is also more likely to develop lung cancer.

Managing confounding is *very* important, as confounding can completely change the relationship between the response and explanatory variables (see the example in Sect. 14.1) and hence can compromise internal validity.

Ways of managing confounding are discussed in Sects. 7.2 and 8.2.

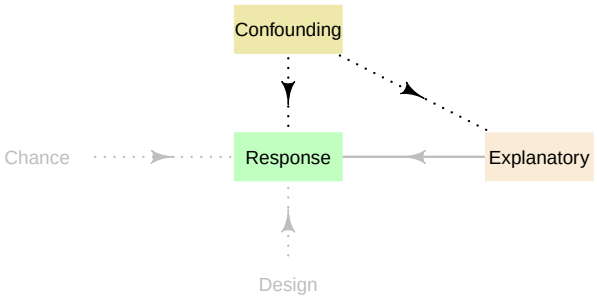


FIGURE 6.3: *Confounding variables are extraneous variables associated with the response and explanatory variables*

Sometimes confounding variables are not measured, assessed, described or recorded; these confounding variables are then called *lurking variables* (Fig. 6.4). Failure to acknowledge lurking variables can lead to wrong conclusions (for example, see Sect. 14.1).

Definition 6.4 (Lurking variable)**.** A *lurking variable* is an extraneous variable associated with the response *and* explanatory variables (that is, is a *confounding* variable), but whose values *are not* measured, assessed, described or recorded in the study.

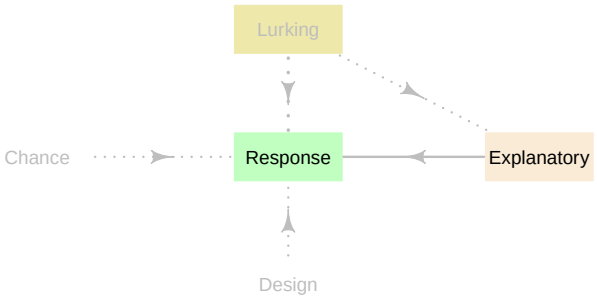


FIGURE 6.4: *Lurking variables are associated with the response and explanatory variables, but are not recorded*

Example 6.4 (Lurking variables)**.** Consider the relationship between carrying cigarette lighters, and developing lung cancer (Example 6.3).

In this study, we could define:

- the **response** variable as "whether or not a person gets lung cancer"; and
- the **explanatory** variable as "whether or not a person carries a cigarette lighter."

Now consider the variable "whether or not a person is a smoker." This variable is associated with the response variable (people who smoke are more likely to get lung cancer than those who do not smoke) and with the explanatory variable (people who smoke are more likely to carry a cigarete lighter than those who do not smoke).

Hence, if that information *was* recorded by the researchers, it would be called a **confounding** variable.

In contrast, if it was *not recorded* by the researchers, it would be called a **lurking variable** (Fig. 6.5).

Now consider the variable "whether or not the person worked closely with someone who smoked." This variable is possibly associated with the response variable (someone who works closely with a smoker would be slightly more likely to get lung cancer ('passive smoking') than someone who does not (Taylor et al. 2001)), but is very unlikely to be associated with owning a cigarette lighter (whether or not someone owns a cigarette lighter probably doesn't depend on whether or not they work closely with a smoker).

Hence, if that information *was* recorded, it would be an **extraneous** variable (but not a confounding variable).

If that variable *was not* recorded, the variation it produces in the response variable would just end up as part of the chance variation.

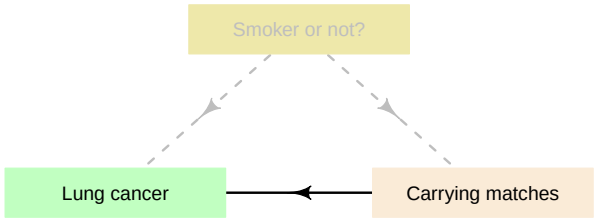


FIGURE 6.5: *An example of a lurking variable*

To clarify (Table 6.1):

- Extraneous variables are all related to the response variable, by definition.
- Some extraneous variables are also called *confounding variables*: if they are also related to the explanatory variable.
- Some confounding variables are also called *lurking variables*: if they are not measured, assessed, described or recorded.

Some unknown extraneous variables will be associated with the response variable only, and so become part of variation due to chance (i.e., unexplained). These terms are not always used consistently by all researchers (Flanagan-Hyde 2005).

To avoid lurking variables, researcher generally collect lots of information about the *individuals in the study* (such as age and sex if the study involves people) and *circumstances of the study* (such as the temperature) that may be relevant, in case they are confounding variables.

TABLE 6.1: *The relationship between the population and the individuals*

Type	Associated with response	Associated with response and explanatory
Measured or observed	No special name: extraneous	Confounding (not lurking)
Not measured or observed	Becomes part of 'chance'	Lurking

Example 6.5 (Lurking variables)**.** Consider the relationship between the number of fatalities in an incident, and number of paramedics sent to the incident. 'Severity of the incident' is the lurking variable, since more severe accidents would have more paramedics attending (in general), and also have more fatalities (in general).

Think 6.1 (Extraneous variables)**.** *Can you think of any other possible extraneous variables in the letter-typing study (Example 5.4)?*

6.4 Study design and variation in the response

Many aspects of the study design can influence the value of the response variable. Good design principles can be used to minimise the impact of these as much as possible, so the focus is on the influence of the explanatory variable on the response.

The study design principles are discussed at length soon (Chaps. 7 and 8).

Example 6.6 (Design)**.** The typing-speed study (Example 5.4) could be poorly designed.

For example, if females were *always* asked to use their dominant hand, and males *always* asked to use their non-dominant hand, the comparison would not be equivalent for females and males.

Females would probably have a faster average time, simply because they are using their dominant hands.

6.5 Chance and variation in the response

Natural (chance) variation refers to variation that cannot otherwise be explained: even repeating a study exactly the same way every time will not always produce the same values of the response variable. This is called *natural variation*, *chance variation*, or just *chance*.

Natural variation makes the influence of the explanatory variable (which we are wanting to study) hard to detect, so minimizing chance variation is important. *Minimise the amount of*

the chance variation, requires using good design principles, and measuring as many other extraneous variables that may explain variation in the response variable as is reasonable.

Chance can impact the values of the response variable in different ways: each *individual* can produce different values of the response variable each time the individual repeats the study (*within*-individuals variation); each individual in the study can produce different values of the response variable compared to *other* individuals (*between*-individuals variation):

- To estimate the amount of variation *within* individuals: Many observations are needed from each unit of analysis (individual).
- To estimate amount of variation *between* individuals Many units of analysis (individuals) are needed.

Since *between*-individual variation is usually more variable than the *within*-individual variation, using *many* individuals is usually more important than using a smaller number of individuals many times.

Think 6.2 (Chance)**.** *Consider the letter-typing study (Example 5.4) again. What are the advantages and disadvantage of:*

- *measuring one female 30 times?*
- *measuring 30 different females once each?*
- *measuring 10 different females three times each?*

6.6 Summary

In a research study, we are usually exploring relationships between a response variable and explanatory variable. However the values of the response variable can be influenced by things other than the explanatory variable, such as other variables that aren't really of interest (**extraneous variable**), the **study design** and by **chance**. Some extraneous variables are also related to the explanatory variable, and are called **confounding variables** (and are **lurking variables** if they cannot be measured, assessed, described or recorded).

6.7 Quick review questions

The Giant Mine in Yellowknife, Canada, ceased operation in 1999 after operating for 50 years, during which 237,000 tonnes of arsenic trioxide was released.

One study (Houben et al. 2016) examined the arsenic concentration in lake water from 25 lakes within a 25km radius of the mine (11 years after the mine closed), to determine if the arsenic concentration was related to the distance of the lake from the mine.

They also recorded the type of bedrock (volcanic; sedimentary; grandiorite), the ecology type

(lowland; upland), the elevation of the lake (in metres), the lake area (in hectares), and the catchment area (in hectares).

1. What is the *response* variable?
2. What is the *explanatory* variable?
3. Is the variable "Catchment area" likely to be a *lurking* variable?
4. Is the variable "Type of bedrock" likely to be a *confounding* variable?
5. What is the *best* description of the variable "Ecology type?"
6. What *type* of study is this?

6.8 Exercises

Selected answers are available in Sect. D.6.

Exercise 6.1. A study examined the relationship between diet quality and depression in Australian adolescents (Jacka et al. 2010). The researchers used a sample of 7114 adolescents aged 10–14 years old in their study, and also measured information about:

> ...age, gender, socioeconomic status, parental education, parental work status, family conflict, poor family management, dieting behaviours, body mass index, physical activity, and smoking...
> — Jacka et al. (2010), p. 435

After identifying the response and explanatory variables, which of these listed variable reasonably could be considered *extraneous variables*, *confounding variables* and *lurking variables*?

Exercise 6.2. A newspaper article (Anonymous 2012) reported on a study that found

> Women who drank green tea at least three times a week were 14 per cent less likely to develop a cancer of the digestive system.

However, the final paragraph of the article notes that:

> Nobody can say whether green tea itself is the reason, since green tea lovers are often more health-conscious in general.

Identify the explanatory and response variables, and explain that final sentence using terms introduced in this chapter.

Exercise 6.3. A study recorded the lung capacity (measured as Forced Expiratory Volume, or FEV, in litres) of children aged 3 to 19 (Tager et al. 1979; Kahn 2005), and also recorded whether not the children were smokers. One finding from the data (Dunn and Smyth 2018) is

that children who smoke have a *larger* average FEV (i.e. larger average lung capacity) than children who do *not* smoke.

Name a confounding variable that may explain this surprising finding. Would it be likely that this variable is a lurking variable?

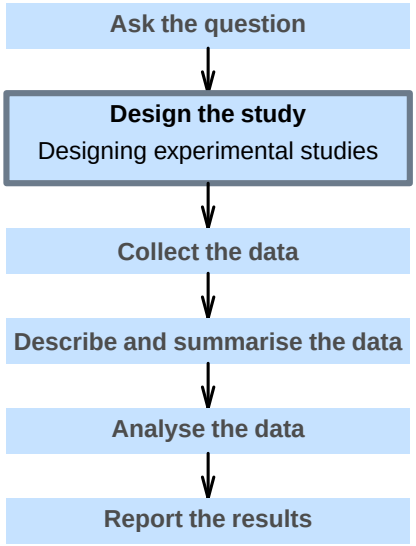
7

Internal validity and experimental studies

So far, you have learnt to ask a RQ, identify different ways of obtaining data, and design the study.

In this chapter, you will learn how to ensure that the conclusions we can make are logical and sound in *experimental* studies. You will learn to:

- maximise internal validity in experimental studies.
- manage confounding in experimental studies.
- explain, identify and manage the carry-over effect in experimental studies.
- explain, identify and manage the Hawthorne effect in experimental studies.
- explain, identify and manage the observer effect in experimental studies.
- explain, identify and manage the placebo effect in experimental studies.
- explain different descriptions of blinding.



7.1 Introduction

The conclusions drawn from a study are only as good as the data that the conclusions are based on, and the data are only as good as the study design that the data emerge from. The design of a study needs to be carefully considered.

A good study requires high *internal validity*: When studying the relationship between the response and explanatory variable, other possible issues that might influences the value of the

response variable should be eliminated. Many aspects of the design must be considered to achieve this goal, some of which are discussed in this chapter.



Data collection is often tedious, time consuming and expensive.

You usually get one chance to collect your data, but you can analyse your data as many times as you like. Since you usually get one chance to collect your data, design the study properly the first time!

Example 7.1 (Importance of internal validity)**.** A group of researchers (Beaman et al. 2013) describe an experiment where free fertilizer was provided to a sample of female farmers in Mali (at the recommended amount per hectare; or at half the recommended amount per hectare).

Since all the farmers knew they were being provided with fertilizer (that is, they were not *blinded*), the farmers changed their farm management: they employed more hired labour and used more herbicide. Consequently, all the farmers' yields changed.

However, it is difficult to know whether this change in yield was was due to the amount of fertilizer applied, the change in labour, the change in herbicides, or a combination of these. That is, the study had poor *internal validity*.

Specific design strategies that we consider for maximising internally validity are:

- Managing confounding;
- Managing the carry-over effect using washout periods;
- Managing the Hawthorne effect by blinding individuals;
- Managing the observer effect by blinding the researchers; and
- Managing the placebo effect using controls.

Not every design consideration will be relevant to every study.

In this chapter, we will work with this RQ (based on Bird et al. (2008)):

Among Australians, does eating provided food made from wholegrain *Himalaya 292* increase average faecal weight compared to eating provided food made from refined cereal?

Think 7.1 (Revision)**.** *For the* Himalaya 292 *study:*

1. *Determine P, O, C and I.*
2. *What are the variables?*
3. *What type of study is this?*

To answer this RQ, a study must be designed to collect the data. However, carefully thought must be given to *how* the study is designed. Some relevant design issues are discussed in this chapter for experimental studies. The next chapter considers design issues for observational studies.

Example 7.2 (Exclusion criteria)**.** In the *Himalaya* study (Bird et al. 2008), the exclusion criteria were:

> [. . .] a history of diabetes, gastrointestinal, renal, hepatic and cardiovascular disease, an intolerance to cereal-based foods, fasting plasma glucose concentrations > 6.1 mmol/l and medications or supplements likely to affect experimental endpoints
> — Bird et al. (2008), p. 1033

7.2 Managing confounding

Confounding has the potential to compromise the internal validity of the study and hence the interpretation of the results, so managing the impact of confounding is important. Suppose, for example, that the researchers created two groups:

- **Group A**: Women recruited at a female-only gym.
- **Group B**: Men recruited at a local nursing home.

The researchers then gave *Himalaya 292* to Group A, and the refined cereal to Group B. If a difference in faecal weight was found between the two groups, the difference may because:

- The *diet* (the explanatory variable) was different in each group;
- The *sex* of the participants was different in both groups, since Group A was all women and Group B was all men;
- The *age* of the participants in each group, since Group A is likely to be younger on average, and Group B is likely to be older on average;
- The health and fitness levels in each group: those in Group A would generally be far healthier than those in Group B.

If a difference is found between the *Himalaya 292* and refined cereal groups, it may not be because of the cereal (Table 7.1). For example, the age of the subject may be related to faecal weight (as older people tend to eat less), and the study design means that older people are more likely to consume the refined cereal. This study has extremely poor internal validity. This is an extreme case of *confounding*; usually confounding is more subtle than in this example.

TABLE 7.1: *Comparing Groups A and B: An extreme example of confounding*

Group A	Variable	Group B
Women	**Sex**	Men
Younger (in general)	**Age**	Older (in general)
Himalaya 292	**Diet**	Refined cereal
Very fit	**Fitness**	Less fit

⚠ The key point is that **the groups being compared should be as similar as possible**, apart from the difference being studied (in this example, the diet that they are given).

Example 7.3 (Comparing groups)**.** An experiment to study the effect of using ginko to enhance memory (Solomon et al. 2002) compared two groups: one using ginko ($n = 111$) and a pretend, non-active supplement ($n = 108$).

The authors randomly allocated participants to each group, but also compared the two groups to ensure that no obvious differences initially existed between the two groups that might explain any differences in the response variable (Table 7.2).

The table shows that the two groups are very similar on these variables, so any difference between the groups cannot be attributed to existing difference in the age, the percentage of men, or the years of education in the two groups.

TABLE 7.2: *Comparing the two groups in the ginko-memory study*

Characteristic	Group A (Ginko)	Group B (Pretend)
Average age (in years):	68.7	69.9
Men (number; percentage)	46 (41)	45 (42)
Average years of education	14.4	14.0

Extra example: Researchers explored the use of dominant and non-dominant hands for chest compression in student paramedics in an experimental study (Cross et al. 2019).

Students were randomly divided into two groups: DHOS (dominant hand on chest) and NDHOC (non-dominant hand on chest).

The two groups were then compared:

Demographic	All participants ($n = 75$)	DHOC ($n = 37$)	NDHOC ($n = 38$)
Average age (years)	23.4	22.5	24.3
Gender: percentage Female	51%	53%	47%

The two groups appear to be very similar in terms of average age of participants, and the percentage of female participants.

This means that, if any difference are observed in the study between DHOC and NDHOC groups, it is unlikely to be because the groups themselves are different in terms of age and sex of particpants.

Potentially, many extraneous variables exist. To demonstrate, we will consider just one: age. How can we make sure that the age of the participants does not cause confounding?

Confounding can be *managed* by:

- **Restricting** the study to a certain group (for example, only people under 30).
- **Blocking**. Analyse the data separately for different groups (for example, analyse the data separately for people under 30, and 30 and over).
- **Analysing** using special methods (after measuring the age of each subject).
- **Randomly allocating** people to groups: Older and younger people would be spread approximately evenly between groups.

The first two approaches (*restricting*; *blocking*) are useful if one or two variables are known, or thought likely, to cause confounding.

The third approach (*analysing*) requires recording all the variables suspected of being confounders.

The fourth approach (*randomly allocating*) is superior if it is possible, because it reduces the chance of confounding even for variables not even suspected as being confounding variables.

Notice that a common theme is to measuring, observing, assessing or recording any variables of potential concern, to ensure no *lurking variables* exist to compromise the results.

Of course, more than one of these approaches can be used, such as randomly allocating individuals to groups, but also measuring, observing, assessing or recording many other variables that can be managed through analysis (Example 7.3).

7.2.1 Restrictions

Sometimes the impact of confounding is managed by *restricting the study to some groups, based on potential confounding variables*, or keeping some variables constant. These variables are called *control variables*. If possible, a reason for this restriction should be given.

Example 7.4 (Restricting)**.** In the *Himalaya* study (Bird et al. 2008), the study might be restricted to subjects aged under 30. The control variable is 'age.'

7.2.2 Blocking

Sometimes *blocking* is used to minimise the impacts of confounding. Blocking refers to separating the units of analysis into a small number of groups that are similar to one another, then studying those groups separately. The *Himalaya* study, might be blocked on age (Fig. 7.1).

Definition 7.1 (Blocking)**.** *Blocking* is when units of analysis are arranged in groups (called *blocks*) that are similar to one another.

7.2.3 Analysis

Confounding variables can be accommodated in the analysis (using analysis methodology beyond what is in this book), *provided those variables have been measured, observed, assessed or recorded*. Because of this, *measuring, observing, assessing or recording all the information likely to be important for understanding the data* is important.

⚠ *Measure, observe, assess or record* all all the information that is likely to be important for understanding the data. This may include information about

- the individuals in the study; and
- the circumstances of the study.

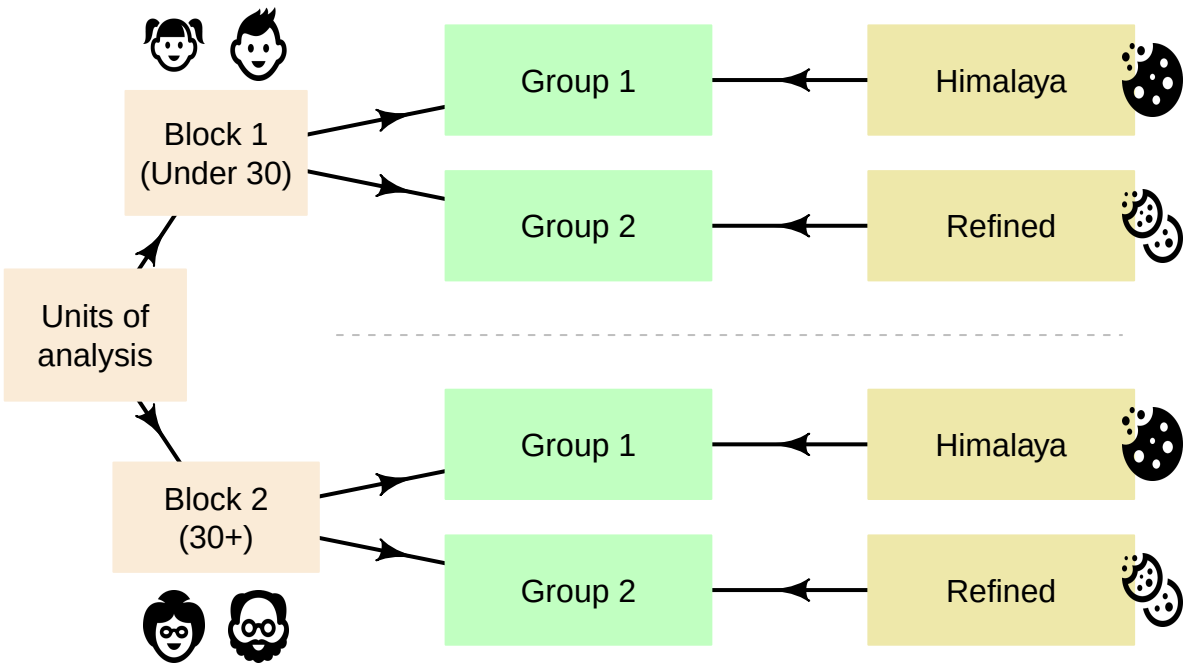


FIGURE 7.1: *Blocking in the *Himalaya* study, based on age*

For this reason, most studies involving people record the participants' age and sex, as these two variables are common confounders. Once a sample is obtained, recording this extra information usually requires little extra effort.

Example 7.5 (Analysis)**.** In the *Himalaya 292* study, the sex, age, pre-study weight and pre-study BMI were also recorded for each individual.

Example 7.6 (Analysis)**.** An experimental study (Schröder et al. 2015) compared nitrogen (N) and phosphorus (P) concentrations in maize, for evenly-injected liquid manure and band-injected liquid manure.

As potential confounding variables, the researchers also recorded the average temperature and the precipitation (between May 1 and September 30) at each site.

7.2.4 Random allocation

One way to minimise confounding is to random allocate individuals in the study to the treatment groups. (Remember that the word "random" has a special meaning.) The advantage of random allocation is that it should approximately evenly distribute potential confounding variables that have been identified (such as age) but also those variables that may *not* have even been considered as confounders, or are hard to measure or observe (such as genetic conditions).

In the *Himalaya* study, the units of analysis (the people in the sample) could be allocated to a group at random, and then the groups allocated a diet through a toss of a coin (Fig. 7.2).

Example 7.7 (Random allocation)**.** In the *Himalaya 292* study, the article reports that 'Subjects were allocated randomly to [. . .] dietary treatments. . . ' (Bird et al. (2008), p. 1033).

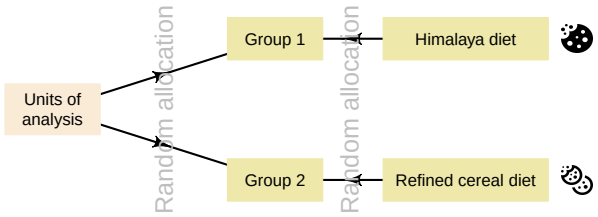


FIGURE 7.2: *Random allocation can occur in two places for the Himalaya study*

Random allocation may occur when randomly allocating individuals to groups (true experiment), and/or when randomly allocating treatments to groups (true or quasi-experiment). Random allocation can be shown, in general, as in Fig. 7.3.

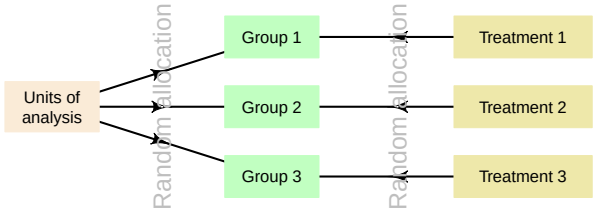


FIGURE 7.3: *Random allocation in general*

7.3 Random allocation vs random sampling

Random *sampling* and random *allocation* are two different concepts (Fig. 7.4), that serve two different purposes, but are often confused:

- **Random sampling** allows results to be generalised to a larger population, and impacts *external* validity. It concerns *how the sample is found* to study.
- **Random allocation** tries to eliminate confounding issues, by evening-out possible confounders across treatment groups. *Random allocation* of treatments helps establish cause-and-effect, and impacts *internal* validity. It concerns *how the members of the chosen sample get the treatments*.

7.4 Carry-over effect and washout periods

In the *Himalaya* study, what if patients spent two weeks on the *Himalaya 292* diet, then the next two weeks on the refined cereal diet?

Potentially, the influence of the first diet could still be impacting the subjects' faecal weight

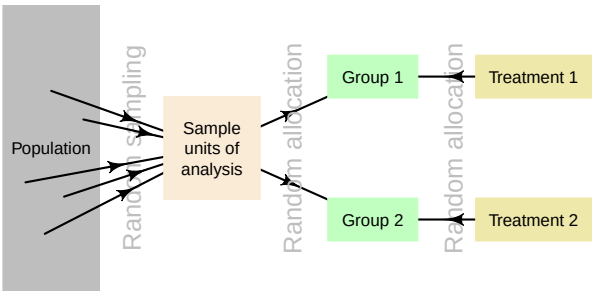


FIGURE 7.4: *Comparing random allocation and random sampling*

for a little while after stopping the first diet. This could compromise the internally validity of the study.

This is called the *carryover effect*.

Definition 7.2 (Carryover effect)**.** The carry-over effect is when the influence of past experience(s) of the individuals carry over to influence future experience(s) of the individuals.

In the context of experiments, this may means that the influence of one treatment carries over into the influence of the next treatment.

The impact of the carryover effect can be minimized by using a *washout period* or similar; for example, after finishing one diet, the participant spend four weeks on their usual (before study) diet, and then revert to the second diet being used.

Sometimes, researchers can *randomly allocation* the *order* in which the treatments (i.e., the diets) are used. That is, some participants start by spending four weeks on the *Himalaya 292* diet, then (after a washout period) four weeks on the refined cereal diet; meanwhile, other participants start by spending four weeks on the refined cereal diet, then (after a washout period) four weeks on the *Himalaya 292* diet.

Example 7.8 (Carry-over effect)**.** In the *Himalaya 292* study, the authors report:

> Subjects were allocated randomly to [. . .] dietary treatments according to a cross-over study design with each intervention phase lasting 4 weeks. There was no washout period between phases.
> — Bird et al. (2008), p. 1033

That is, subjects were randomly allocated to a diet: some subjects began the study on the *Himalaya 292* diet while others started on the refined cereal diet, No washout period was used; however, since the response variable was recorded after four weeks on the diets, no washout period was necessary.

Example 7.9 (Washout periods)**.** An engineering study (Miller and Boyle 2019) examined drivers' exposure to lane-keeping system on their driving performance. Subjects were exposed to a driving simulation that used a lane-keeping system, and then to a driving simulation without using a lane-keeping system.

The researchers found that there was a carryover effect when drivers moved from a simulation with a lane-keeping system to one without a lane-keeping system.

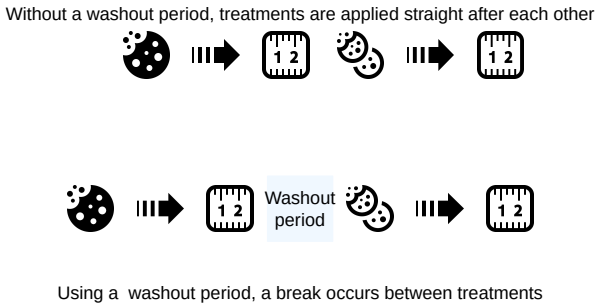


FIGURE 7.5: *Using a 'washout' period to minimize the carry-over effect*

7.5 Hawthorne effect and blinding individuals

What if the patients in the *Himalaya 292* study were being watched (or waited for) while defecating?

People often behave differently (either positively or negatively) if they know (or think) they are in a study or are being watched. This is called the *Hawthorne effect* (McCarney et al. 2007). This could also compromise the internally validity of the study.

Definition 7.3 (Hawthorne effect)**.** The Hawthorne effect is the tendency of people (or animals, or. . .) to behave differently if they know (or think) they are being observed.

Example 7.10 (Hawthorne effect)**.** People are more health-conscious if they know they will be followed up on a regular basis. For example, a study aiming to increase fruit and vegetable intake in young adults (Clark et al. 2019) noted that

> The changes that did occur could be explained by the Hawthorne effect [. . .] the intervention [. . .] can inherently cause participants to change behavior because they know they are being observed. . .
> — Clark et al. (2019)

The impact of the Hawthorne effect can be minimized by blinding the individuals in the experiment so that they do not know:

- that they are in a study;
- the aims of the study, and/or
- which treatment they are receiving.

Blinding *people* to knowing they are involved in a study is often difficult, as ethics usually requires individuals' informed consent.

For example, if the individuals do not know which treatment they are receiving, they cannot behave differently according to the treatment they know they are receiving.

Example 7.11 (Hawthorne effect)**.** In the *Himalaya 292* study, the authors report:

> The study was explained fully to the subjects, both verbally and in writing, and each gave their written, informed consent before participating.
> — Bird et al. (2008), p. 1033

That is, the subjects knew they were in a study. As is usual, this was an ethics requirement (in this case, from the Ethics Committee of the CSIRO). The Hawthorne effect may influence the results.

However, the subjects did to know which diet they were on:

> Volunteers were not told the identity of the test cereal in the foods provided to them.
> — (Bird et al. (2008), p. 1033)

Example 7.12 (Hawthorne effect)**.** In an experimental study (Lorenz et al. 2019) to compare the efficacy of a new type of toothpaste, participants were given two types of toothpaste to use (a new type, and an exisiting type), and evaluations of plaque remaining on the teeth were taken. The authors state that:

> ... a plaque-reducing effect was seen not only in the test group but also in the control group. This phenomenon is due to the so-called Hawthorne effect that can lead to an overestimation of the effect and false positive results.
> — Lorenz et al. (2019), p. 5

That is, since all participants knew they were being assessed after brushing their teeth, there may have been a tendency to brush their teeth better than usual. The authors then state:

> To minimize the Hawthorne effect, longer study durations of more than 6 months were suggested.
> — Lorenz et al. (2019), p. 6

7.6 Observer effect and blinding researchers

What if the *researchers* assessing the outcomes *knew the diet* allocated to each patient? Perhaps surprisingly, this can have an (unconscious) impact on the values of the response variable. This is called the *observer effect* (or, in experiments, sometimes called the *experimenter effect*). This could also compromise the internally validity of the study.

Definition 7.4 (Observer effect)**.** The observer effect is when the researchers *unintentionally* influence the behaviour of subjects.

The impact of the observer effect can be minimized by blinding the researchers so that they do not know: which treatments the individuals are receiving. That is, the people *giving* the treatment and the people *evaluating* the treatment do not know what treatment has been given. Instead, a third party can be used.

For example, the researchers may give an assistant two drugs labelled A and B. The assistant then administers the drug and evaluates the participants' response to the treatments. Later, the assistant tells the researchers whether Drug A or Drug B performed better, but only the researchers know what drugs the labels A and B refer to.

The observer effect does not just apply to situations where *people* are used as participants.

Example 7.13 (Observer effect)**.** 'Clever Hans' (https://en.wikipedia.org/wiki/Clever_Hans) was a horse that seemed to be able to perform simple mental arithmetic. After much study, Carl Stumpf realised that the horse was responding to involuntary (and unconscious) cues from the trainer. This was discovered, in part, by using an experiment where the people interacting with the horse were blinded.

The same effect has been observed in narcotic sniffer dogs (Bambauer 2012), who may respond to their handlers' unconscious cues.

⚠ The *observer effect* is about *unconsciously* influencing the outcome; that is, the researchers are not aware that it is occurring. When researchers *intentionally* influence the outcomes, this is called *fraud* (Sect. 4.3.2).

7.7 Placebo effect and using controls

What if people *thought* they were on the wholegrain diet, but they weren't? Perhaps surprisingly, individuals in a study may report effects of a treatment (either positive or negative), even if they have not received an active treatment. This could also compromise the internally validity of the study.

This is called the *placebo effect*.

Definition 7.5 (Placebo effect)**.** The placebo effect is when individuals report perceived or actual effects without having received the treatment.

Managing the placebo effect is difficult! However, impact of the placebo effect can be minimized using a *control group*: units of analysis without the treatment applied, but *as similar as possible* in every other way to those units of analysis receiving the treatment. This allows the effect of the treatment to be ssessed, over and above the placebo effect.

Definition 7.6 (Control)**.** A *control* is a unit of analysis without the treatment applied (but as similar as possible in every other way to other units of analysis).

Sometimes the control group receives a *placebo*. A *placebo* is a non-effective treatment. Those who receive the placebo should be selected through random allocation when possible. Sometimes, using a placebo is unethical. The Wikipedia entry about placebos[1] is intriguing.

Definition 7.7 (Placebo)**.** A *placebo* is a treatment with no intended effect or active ingredient.

Example 7.14 (Placebo effect)**.** In the *Himalaya 292* study, the authors report

> On each day of the intervention periods, volunteers were asked to consume a combination of bread, breakfast cereal, muffins and crackers that would supply in total 103g of the test cereal. The aim was for each volunteer to consume 60g cereal flakes (or puffed rice for the refined cereal diet), two slices of bread, one muffin and six savoury crackers each day. Volunteers were not told the identity of the test cereal in the foods provided to them
> — (Bird et al. (2008), p. 1033)

That is, the subjects were blinded to the diet they were exposed to. However, some may *think* they are on the refined cereal or *Himalaya* diet, and respond accordingly (perhaps unconsciously).

Think 7.2 (Controls)**.** *To test the effectiveness of a new drug, patients are to report to a GP to receive injections of a new drug. We wish to compare to people who do not get the injection. What is the control?*

Answer: The answer is given in the online book.

Example 7.15 (Placebo effect)**.** Three active analgesics (pain relievers) were compared to a placebo (Huskisson 1974). Four different coloured placebos were used. The most pain relief was experienced by those taking *red* placebos (Fig. 7.6), who experienced even more pain relief than those given true pain relievers.

Example 7.16 (Placebo effect)**.** A study of placebos (Waber et al. 2008) gave half the subjects a placebo, but told them that the pill was an expensive (impling 'very effective') pain killer ($2.50 per tablet). The other half were also given a placebo, but were told that the pill was a discount (impling 'less effective') pain killer ($0.10 per tablet).

About 85% of participants in the first group reported a pain reduction, yet only 61% in the second group reported a pain reduction. Remember that *both* groups actually received a placebo!

[1] http://en.wikipedia.org/wiki/Placebo

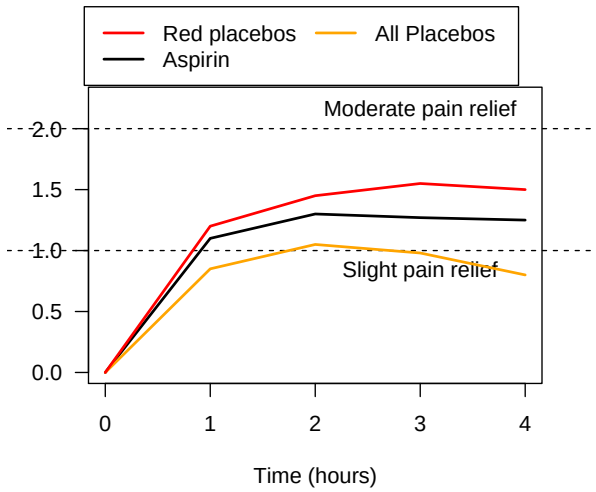


FIGURE 7.6: *Pain relief, for various pain relief medicine*

7.8 Comments on describing blinding

Blinding is when those involved in the study do not know information about the study.

Those involved in the study may not know:

- that they are in a study at all;
- the purpose of the study; and/or
- which comparison group they are in.

When participants are blinded in as many ways as possible, the internal validity of the study is increased. However, when people are the individuals, ethics requirements often mean that they need to know they are in a study and the purpose of the study.

Different individuals involved in the study can be blinded:

- A study can blind the **researcher** to knowing what comparison group the study individuals are in.
- A study can blind the **participants** to knowing what comparison group they are in.
- A study can blind the **analysts** to knowing what comparison group the individuals are in during analysis.

When as many participants are blinded as possible, the internal validity of the study is increased.

If *only* the participants are blinded, the study is called *single blind*. If *both* the researchers and participants are blinded, the study is called *double blind*. If the researchers, participants *and* the analyst are blinded, the study is called *triple blind*. However, for clarity, explicitly stating who or what is blinded is best. Blinding should be considered in all studies, when possible (and it is not always possible).

Blinding of participants does not just apply to people; it is also relevant with animals (Example 7.13 about Clever Hans).

Think 7.3 (Blinding)**.** *Why might it be necessary to blind the* analyst *to the treatments being used?*

Example 7.17 (Blinding)**.** In a study comparing chest compressions with dominant and non-dominant hands of student paramedics (Cross et al. 2019), the article states that:

> Participants were asked to participate in a study exploring general CPR performance but were blinded to the specific research question at any stage to reduce the chance of performance bias...
> — Cross et al. (2019), p. 2

Participants could not, however, be blinded to which group they were in (dominant hand on chest; non-dominant hand on chest). In this case, participants were only partically blinded.

Later, the article reports that:

> Data were analysed by a biostatistician blinded to group allocation.
> — Cross et al. (2019), p. 3

This means that the analyst was blinded to the treatments.

Example 7.18 (Double-blinding)**.** In a cropping study comparing yields from modern and traditional cowpea crops in Tanzania, the researchers wanted to use a double-blind study.

To do so:

> ...it was important that the traditional and modern seed looked exactly the same—the seed types must be indistinguishable in terms of size and color.
> While information about seed type may be gradually revealed as the crop matures in the field, this does not invalidate our design because key inputs were already provided.
> Since the modern seed was treated with purple powder, we also dusted the traditional type, and clearly communicated this to the farmers—they knew that seed type could not be inferred from the color.
> Bulte et al. (2014), p. 817–818; line breaks added

7.9 Design issues: Overview

In summary, issues to consider when designing a study, when possible, include:

- Minimising confounding (and *lurking variables*);
- Minimising the *carryover* effect;
- Minimising the *Hawthorne* effect;

- Minimising the *observer* effect;
- Minimising the *placebo* effect.

Ways to minimize the impact of these have been discussed (Fig. 7.7), but is not always possible. These effects are important to understand, so studies can be designed to manage or minimise their influence (to maximise internal validity). This ensures that the results and conclusions from our studies are correctly interpreted (that is, noting, for example, how the Hawthorne effect may have influenced the conclusions).

Often, however, some (or all) of these issues cannot be well managed. For instance, individuals often know they are involved in an experimental study (Hawthorne effect). In these cases, the impacts should be minimized as far as possible, and then the likely impact that these issues have on our conclusions discussed. The impact of these issues are often reported as *limitations* in a journal article (Chap. 9), perhaps part of the Discussion section.

Example 7.19 (Study limitations)**.** A study of alcohol use in college females reported these limitations of their study:

> The present study has several limitations. First, data were collected over 15 years ago [. . .] Second, only college females were assessed and findings may not generalize to college males or to broader groups of young adults [. . .] Third, alcohol and caffeine consumption variables were all self-reported. . .
> — Kelpin et al. (2018), p. 3

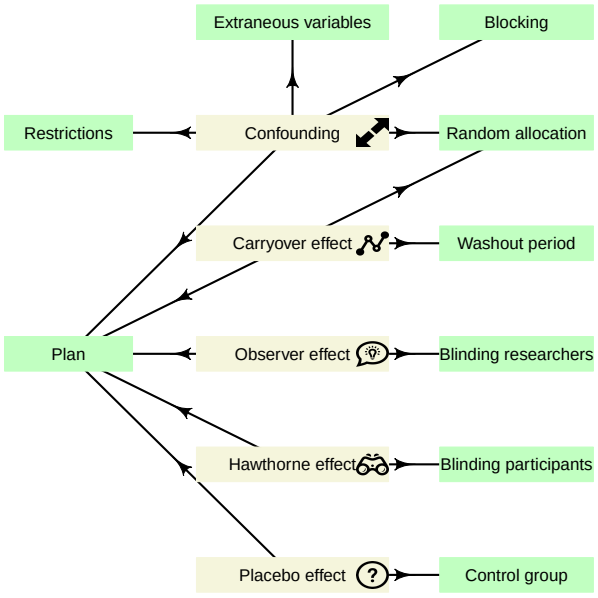


FIGURE 7.7: *Design considerations. Note: Lurking variables become confounding variables when measured, observed, assessed or recorded in the study, and then they can be managed. The arrows mean that the design issue can be* ***partially*** *managed by the indicated means*

Example 7.20 (Study design)**.** In a study of student paramedics comparing chest compressions with dominant and non-dominant hands (Cross et al. 2019), as discussed in Example 7.17, the participants were partially blinded: they were blinded to the *purpose* of the study, but not to which group they were allocated.

The analyst was also blinded to the group allocations.

Later, the article reports that:

> . . . participants were allocated randomly to one of two groups: 'dominant hand on chest' or 'non-dominant hand on chest.' Group allocation was determined by a computer-generated randomisation schedule. . .
> Cross et al. (2019), p. 3

This study used a number of good design features.

7.10 Summary

Designing effective *experimental* studies requires researchers to **manage or minimise confounding** where possible, by *restricing* the study to certain groups, by *blocking*, through special *analysis* methods, and/or through *random allocation*.

Well-designed experimental studies also try to manage the effects of the **carry-over effect** (for example, using a washout period, or randomly allocating treatments), the **Hawthorne effect** (for example, by blinding participants to the treatment), the **observer effect** (for example, by blinding the researchers to the treatments being applied), and the **placebo effect** (for example, by blinding participants to the treatments and by using *controls*).

7.11 Quick review questions

A study on the bruising of apples (Doosti-Irani et al. 2016) aimed to determine the relationship between the recorded surface temperature of apple, the depth of bruising.

The researchers purposefully hit apples with three different **forces** (200, 700 and 1200 mJ) to inflict bruises.

The researchers then recorded the **depth** of the bruising, and recorded the **surface temperature** at each bruise location.

The study was conducted separately for three different **regions** of the apple (lower; middle; upper), and each apple was only used once.

1. The *response variable* is
2. The *explanatory variable* is
3. What is the *best* description for the variable 'The location of the bruising?'

4. True or false: The researchers could minimise the effects of confounding by incorporating potential confounding variables in the anaysis.
5. True or false: The researchers could use random allocation of the treatments to the applied to the apples to minimise confounding.
6. True or false: The *carry-over* effect is likely to be a big problem in this study.
7. True or false: The *Hawthorne* effect is likely to be a big problem in this study.
8. True or false: The *placebo* effect is likely to be a big problem in this study.
9. True or false: The *observer* effect is likely to be a big problem in this study.

7.12 Exercises

Selected answers are available in Sect. D.7.

Exercise 7.1. A scientist is comparing the effects of two types of fertiliser on the yield of tomatoes (based on Klanian et al. (2018)). He plants tomato seedlings, and fertilises with Fertiliser I, and later measures the yield of tomatoes. He then immediately plants more tomato seedlings in the same field, and fertilises with Fertilizer II, and measures the yield of tomatoes.

What potential problems can you identify with the study design?

Exercise 7.2. A scientist is expecting that tap water will taste the same as bottled water in a taste test (based on Teillet et al. (2010)). The scientist provides people with a plastic cup of either bottled or tap water, and she asks them to give a rating of the taste on a scale of 1 (terrible) to 5 (fantastic).

What potential problems can you identify with the study design?

Exercise 7.3. Consider this RQ (based on Teillet et al. (2010))):

> Among university students, is the taste of tap water different than the taste of bottled water?

This RQ needs some clarification, but you decide to answer this question using an *experiment*. How would you manage:

1. Random allocation?
2. Blinding?
3. Double blinding?
4. Finding a control?
5. Finding a random sample?

Exercise 7.4. In a study of time spent applying sunscreen (Heerfordt et al. 2018) the Aim was to 'determine whether time spent on sunscreen application is related to the amount of sunscreen used' (Heerfordt et al. (2018), p. 117). The authors state this about the study design:

> The volunteers were asked to apply the provided sunscreen [...] the way they would normally do on a sunny day at the beach in Denmark [...] The volunteers wore swimwear during the whole session. No other information was given. Participants applied sunscreen behind a curtain and were not observed during application. Measurements of time and sunscreen weight were made without the subjects' being aware of this.
> — Heerfordt et al. (2018), p. 118

1. What are the response and explanatory variables?
2. The researchers also recorded age, height, weight and body surface area of each participant. Why would they have done this?
3. The researchers also compared the mean values of the response variable for males and females, and the mean values of the explanatory variable for males and females. Why would they have done this?
4. What design features are being used in the second quote?

8

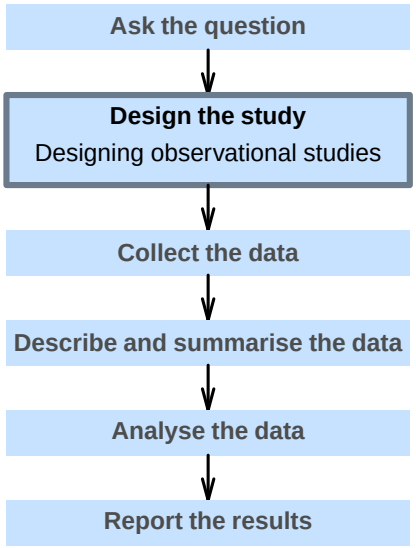
Internal validity and observational studies



So far, you have learnt to ask a RQ, identify different ways of obtaining data, and design the study.

In this chapter, you will learn how to ensure that the conclusions we can make are logical and sound in *observational* studies. You will learn to:

- maximise the internal validity of observational studies.
- manage confounding in observational studies.
- explain, identify and manage the carry-over effect in observational studies.
- explain, identify and manage the Hawthorne effect in observational studies.
- explain, identify and manage the observer effect in observational studies.
- explain, identify and manage the placebo effect in observational studies.



8.1 Introduction

In experimental studies, many aspects of the study design typically can be controlled by the researcher, so experimental studies are often easier to design to maximise [internal validity](#def:InternalValidity]. In contrast, *observational studies* have fewer design features that can be controlled by the researchers.

For example, treatments are not imposed in observational studies, so random allocation of

treatments is impossible, and hence *confounding* is always a potential threat to [internal validity](#def:InternalValidity] in observational studies.

Nonetheless, researchers should still consider aspects of research design when designing observational studies, and manage those aspects when possible to maximise the internal validity. Specific design strategies that we consider for maximising internally validity are:

- Managing confounding;
- Managing the carry-over effect;
- Managing the Hawthorne effect;
- Managing the observer effect; and
- Managing the placebo effect.

Not every design consideration will be relevant to every study.

Think 8.1 (Observational studies)**.** *Doll and Hill (1954) wrote to a large number of British doctors, and asked how much they smoked. Then they observed smokers and non-smokers for many years, and recorded who died of lung cancer.*

Why is this is an observational *study?*

Answer: The answer is given in the online book.

8.2 Managing confounding

In Sect. 7.2 different methods were listed for managing confounding in experimental studies. Some, but not all, of these are still possible in observational studies:

- **Restricting** the study to a certain group (for example, only people under 30).
- **Blocking**. Analyse the data separately for different groups (for example, analyse the data separately for people under 30, and 30 and over).
- **Analysing** using special methods (after measuring the age of each subject).
- **Randomly allocating** people to groups: Not possible in observational studies.

8.2.1 Restrictions

As with experimental studies, observational studies can be restricted to certain parts of the population. For example, in the smoking study of Doll and Hill (1954), participants were restricted to males aged under 35 years since, at the time of the study:

> ... lung cancer is relatively uncommon in women and rare in men under 35 [and] useful figures are unlikely to be obtained in these groups for some years to come.
> — Doll and Hill (1954), p. 1452.

The reason for the restriction should be justified if possible (as shown in the quote above).

8.2.2 Blocking

Blocking can be used with observational studies; for example, those aged under 30 and those aged 30 or over could be analysed separately.

This, of course, requires the age of the participants to be available to the researchers.

8.2.3 Analysis

The best advice for observational studies is to **measure, observe, assess or record all the information that is likely to be important for understanding the data**. While this strategy is also useful for experimental studies, it is particularly important for observational studies, as managing confounding through analysis (Sect. 7.2.3) is often one of the few practical means available.

⚠ *Measure, observe, assess or record* all all the information that is likely to be important for understanding the data. This may include information about

- the individuals in the study; and
- the circumstances of the study.

Example 8.1 (Analysis)**.** In a different smoking study, Doll and Hill (1950) recorded the the social class and place of residence of each participant, as potential confounding variables.

Example 8.2 (Confounding)**.** An observational study of 2599 kiwifruit orchards (Froud et al. 2018) explored the relationship between the time since a bacterial canker was first detected (in weeks) and the orchard productivity (in tray equivalents per hectare).

The researchers also recorded information such as 'whether or not the farm was organic,' 'elevation of the orchard' and 'whether or not general fungicides were used.'

They used these variables in their analysis to manage the potential effects of confounding. Their analysis showed that 'elevation of the orchard' and 'whether or not general fungicides were used' were important confounding variables, but 'whether or not the farm was organic' was not.

8.2.4 Random allocation

In observational studies, the study conditions are not allocated by the researchers (at random or otherwise), so random allocation of treatments is not possible

For this reason, confounding is often a major threat to internal validity in observational studies, as individuals who are in one group may be different to those who are in another group (Table 8.1). As a result, researchers often summarise the groups being compared on various potential confounding variables.

Example 8.3 (Comparing study groups)**.** In an observational study comparing the iron levels

of active and sedentary women aged 18 to 35 (Woolf et al. 2009), the authors compared the active women ($n = 28$) to the sedentary women ($n = 28$) on a variety of characteristics.

However, maybe the intrinsic physical differences between the women in the the two groups might explain any differences found between iron levels in two two groups.

To determine this, the researchers examined many characteristics of the women; some are shown in Table 8.1. d They conclude that active women in their sample tended to be (in general) slightly younger, slightly heavier and taller, and slightly more likely to use hormonal contraceptives. Hence, any difference in iron levels between the two groups may be because of the active/sedentary nature of the groups, or because the active group was (in general) younger, for example.

TABLE 8.1: *The demographic information for those in the study of iron levels in women*

Characteristic	Active women	Sedentary women
Average age (in years)	20	24
Average height (in cm)	169	166
Average weight (in kg)	68	62
Percentage using hormonal contraceptives	13	11

Extra example: In the smoking study of Doll and Hill (1954), doctors who chose to smoke may be inclined to undertake other risky behaviours, whereas those doctors who choose *not* to smoke may also be inclined to *not* undertake other risky behaviours. It may be those *other* risky behaviours that lead to lung cancer, and not the smoking itself.

In a different smoking study, Doll and Hill (1950) used a control group. The control group was chosen to be very similar to those in the lung-cancer group, in terms of age and sex. (That is, the numbers of females and males within each age group was very similar for those with lung cancer, and those without lung cancer.)

⚠ Observational studies *can* (and often do) have control groups. Indeed, one specific type of observational study is called a *case-control study*[1].

However, individuals are not *allocated* to the control group by the researchers in observational studies.

Extra example: A study (Gunnarsson et al. 2017) examined the difference between two types of helicopter transfer (physician-staffed; non-physician-staffed) of patients with a specific type of myocardial infarction (STEMI). The purpose of the study was:

> ...to evaluate the characteristics and outcomes of physician-staffed HEMS (Physician-HEMS) versus non-physician-staffed (Standard-HEMS) in patients with STEMI.
> — Gunnarsson et al. (2017), p. 1

The researchers

> ... studied 398 STEMI patients transferred by either Physician-HEMS ($n = 327$) or Standard-HEMS ($n = 71$) for [...] intervention at 2 hospitals between 2006 and 2014.
> — Gunnarsson et al. (2017), p. 1

Since the study is an observational study (patients were not allocated by the researchers to the type of helicopter transport), the researchers recorded information about the patients being transported. They compared the patients in both groups, and found (for example) that both groups had similar average ages, a similar percentages of females, a similar percentage of smokers, and so on. They also compared information about the transportation, and found (for example) that both groups had similar average flight times and flight distances.

One conclusion from the study was that 'Patients with STEMI transported by Standard-HEMS had longer transport times' (p. 1), but one limitation of the study was that:

> The patient cohorts received treatment by 2 different care teams at two hospitals, which is a potential confounder despite similar baseline characteristics
> — Gunnarsson et al. (2017), p. 5

In other words, the difference between hospitals and the staff may have been a confounding variable.

8.3 Carry-over effect and washout periods

The carry-over effect is a possible compromise to internal validity in observational studies. However, since treatments are not *allocated* in observational studies, carry-over effects may be difficult to prevent.

It may be possible, however, to *observe* individuals who are exposed to Condition A then Condition B, and other individuals who are exposed to Condition B and then Condition A.

Example 8.4 (Carry-over effects)**.** A study of the carry-over effect in ecological observational studies gave many examples, including:

> individuals occupying poor quality winter habitat may experience reduced reproductive success the following breeding season when compared to individuals occupying high quality winter habitat.
> — Norris (2005), p. 181

8.4 Hawthorne effect and blinding individuals

In observational studies, individuals *may* or *may not* know they are being observed. For example, in an observational study where subjects' blood pressure is measured (Verdecchia et al. 1995), subjects clearly will know that they are being observed. This has the potential to alter the data being recorded, and hence compromise internal validity.

As with experimental studies, efforts should be made to ensure that individuals do not know that they are being observed (that is, that the participants are *blinded*).

Example 8.5 (Hawthorne effect)**.** One study (Wu et al. 2018) examined hand hygiene (HH) in a tertiary teaching hospital, using *covert* observers (that is, the observers were not obviously watching the hand hygiene practices of staff) and *overt* observers (that is, the observers were obvious about watching the hand hygiene practices of staff). One conclusions was that

> The overall HH compliance was higher with overt observation than with covert observation (78% vs. 55%)...
> — Wu et al. (2018), p. 369

In other words, people's behaviour changed markedly when people knew they were being observed. This could easily change the observed relationship between the response and explanatory variables, and hence compromise internal validity.

8.5 Observer effect and blinding researchers

The observer effect can be an issue in observational as well as experimental studies. For example, consider a study where the blood pressure of smokers and non-smokers is recorded (Verdecchia et al. 1995).

This is an observational study (individuals cannot be allocated to be a smoker or non-smoker), but if the researchers *know* whether or not the individual is a smoker when they record the blood pressure, then the observer effect could still come into play (recalling that the observer effect is an *unconscious* effect).

In this example, the observer effect could be managed if the researchers *first* measured the blood pressure, and *then* asked if the individual was a smoker or not. That is, the researchers may be able to be *blinded* to whether or not the subject is a smoker.

This may only be partially successful; the researcher may see the subject carrying a packet of cigarettes, or can smell smoke on their breath, for example; nonetheless, it may prove at least partially successful, and is easy to implement.

Example 8.6 (Blinding in ecology)**.** A study of research articles in ecology found:

> Across all 492 EEB articles surveyed, we judged 50.4% ($n = 248$) to have potential for observer bias, but only 13.3% ($n = 33$ of 248) of these articles stated use of blind observation.
> Some articles explicitly stated the use of blind observation in the methods ($n = 24$), while others indicated indirectly that experiments had been done blind ($n = 9$; e.g., use of a naïve experimenter...).
> — Kardish et al. (2015); line breaks added

Blinding the observer is not always possible, but should be used when possible to improve the internal validity of the study.

Extra example: A study (Gamble and Walker 2016) found that bicycle riders who *wear* helmets are more likely to take risks compared to bicycle riders who *do not wear* helmets.

The paper states that the bicycle riders were blinded to the *purpose* of the study (reducing the impact of the Hawthorne effect), though clearly the participants knew they were involved in a study (so the impact was not completely eliminated).

However, the study was criticised (Radun and Lajunen 2018), since it was possible that

> ... the experimenters unconsciously conveyed their expectations to participants and thereby affected their responses [...] it is clear that the double-blind procedure has been developed for a reason and should have been used in this study.
> — Radun and Lajunen (2018), p. 1020

The lack of blinding compromised the study's internal validity.

8.6 Placebo effect and using controls

The placebo effect is concerned with *treatments*, so are not directly relevant to observational studies.

However, observational studies can still have a control group, but the individuals are not randomly *allocated* to the control group.

For example, in the Doll & Hill smoking study (Doll and Hill 1950), two groups were being compared: non-smokers (the control group) and smokers.

Subjects were *not allocated* to the groups, however, so confounding remains a possibility. Again, the groups in the study can be compared (Example 8.3) to see if the groups are different in other ways.

8.7 Summary

Designing effective *observational* studies requires researchers to maximise internal validity. This can be achieved by **managing confounding** where possible, as confounding is often a major threat to the internal validity of observational studies. This can be managed by *restricting* the study to certain groups, by *blocking*, or through special *analysis* methods. Random allocation is not possible in observational studies. For this reason, observing, measuring, assessing or recording all the information that is likely to be important for understanding the data is important, usually to be used in analysis.

Well-designed observational studies also try to manage the effects of the **carry-over effect**, the **Hawthorne effect**, the **observer effect**, and the **placebo effect**, though the means of doing so are often not under the control of the researchers.

8.8 Quick review questions

Formwork is used in construction using reinforced concrete. It is complicated and labour intensive. An observational study (Mine et al. 2015) examined the relationship between the floor area of the building (in m^2 per storey) and the number of hours of labour needed for the construction (in person-minutes per storey).

The researchers also recorded, among other things:

- the average age of the workers (in years);
- the average years of experience of the workers (in years); and
- the storey height (in meters)

for each of $n = 15$ multi-storey buildings in the study. Some data was obtained from

> ... interviews [..] conducted with the relevant person-in-charge of scheduling for each project.
> — Mine et al. (2015), p. 2

To record the number of person-hours of labour:

> Two observers made the site rounds observation of 2 to 40 workers. Observations were carried out continuously from the start to the end of work, excluding lunch time.
> — Mine et al. (2015), p. 2

1. The *explanatory variable* is
2. The *response variable* is
3. What is the *best* description for the variable 'Average age of the workers?'
4. What is the most likely way in which confounding would be managed in this study?

5. True or false: The *carry-over* effect is likely to be a big problem in this study.
6. True or false: The *Hawthorne* effect is likely to be a big problem in this study.
7. True or false: The *placebo* effect is likely to be a big problem in this study.
8. True or false: The *observer* effect is likely to be a big problem in this study.

8.9 Exercises

Selected answers are available in Sect. D.8.

Exercise 8.1. Consider this RQ (based on Teillet et al. (2010)):

> Among university students, is the taste of tap water different than the taste of bottled water?

You want to answer this question using an *observational study*. Describe what these might look like for this study:

1. Random allocation.
2. Blinding.
3. Double blinding.
4. Control.
5. Finding a random sample.

Exercise 8.2. Is it possible to have a control group in an observational study? Explain.

Exercise 8.3. Is the Hawthorne effect only a (potential) issue in experimental studies? Explain.

Exercise 8.4. A study of how well hospital patients sleep at night (Delaney et al. 2018) had the stated aim 'to investigate the perceived duration and quality of patient sleep [...] in hospital.' In discussing the limitations of the study, the researchers state:

> The researchers made no attempt to deceive clinical staff regarding the nature of the study so the influence of the Hawthorne Effect should be considered. The presence of the observer and environmental monitoring equipment in the clinical environment could have altered behaviour among patients and nursing staff seeking to conform to the presumed research objectives. As a result, the findings reported may be an underestimation of the magnitude of the issues that affect sleep.
> — (Delaney et al. (2018) p. 7)

Discuss these limitations in terms of the language used in this chapter.

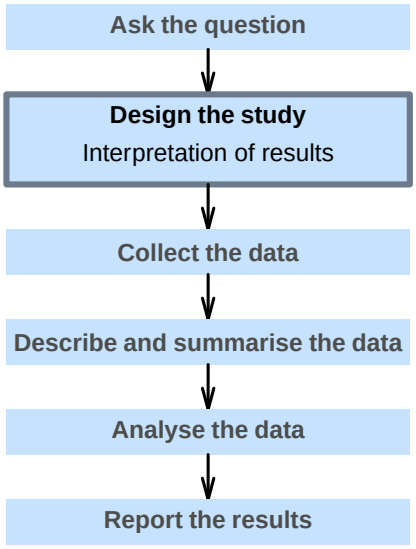
9

Identifying study limitations

So far, you have learnt to ask a RQ, identify different ways of obtaining data, and design the study.

In this chapter, you will learn to identify and describe the limitations of a study. You will learn to:

- identify when and how study results are internally valid.
- identify when and how study results are externally valid.
- identify when and how study results are ecologically valid.



9.1 Introduction

The *type* of study and *how* that study is designed can determine how the results of the study should be interpreted. Ideally, a study would be perfectly externally and internally valid, but in practice this is very difficult to achieve. Practically *every* study has limitations.

The results of a study should be interpreted in light of these limitations: The results should be discussed in light of what the study actually tells, and what it doesn't actually tell us.

Limitations can often be discussed through three components:

- **External validity** (the *applicability* of the study results outside the sample): The *generalisability* of the results (Sect. 9.2) to the intended population.

- **Internal validity** (the *effectiveness* of the study in the sample): The steps taken to maximise the internal validity of the study, and the impacts of these on the interpretation of results (Sect. 9.3).
- **Ecological validity** (the *practicality* of the results to real life): The *practicality* the results in the real world (Sect. 9.4); how the study methods, materials and context approximate the real situation being studied.

All these issues should be considered when considering the study limitations.

⚠ Almost every study has limitations. *Identifying* them, and *discussing the impact* that they have on the interpretation of the study results, is important and ethical.

Example 9.1 (Interpretation)**.** Smoking was once considered healthy and beneficial: some advertisements used doctors to promote cigarette smoking, and others promoted the benefits of smoking to athletes (Ingalls 1936).

An advertisement for Camel cigarettes[1] once claimed that 'More doctors smoke Camels[2] than any other cigarette' based on a survey of 113 597 doctors. That certainly is a *large* sample… but understanding how the data were collected is important.

In fact, the company that owned Camel cigarettes (RJ Reynolds Tobacco Company[3]) conducted the survey (which raises suspicions immediately). Even worse, RJ Reynolds[4] staff interviewed doctors and asked about their smoking habits *after RJ Reynolds provided free cartons of Camel cigarettes* to the doctors. No wonder more doctors smoked Camel brand!

Concluding that 'More doctors smoke Camels than any other cigarette' in light of this is highly unethical.

9.2 Limitations: External validity

Externally validity refers to the ability to *generalise* the results to other groups in the population apart from the sample studied (Sect. 3.6). For a study to be externally valid, it must first be internally valid.

⚠ Importantly, external validity refers to how well the sample is likely to represent the *target population* as given in the RQ.

For example, suppose the RQ is 'Among Queenslanders, what proportion own a smart speaker?'

The study is externally valid if the sample is representative of Queenslanders, and hence the results from the sample are likely to apply to Queenslanders as a whole.

The results **do not** have to apply to people in the rest of Australia. The intended population, as given in the RQ, is *Queenslanders*.

External validity refers to the *applicability* or the *generalisability* of the results to the *target (or*

intended*) population (Example 5.11), which depends on how the sample was obtained:* results from random samples are likely to generalise to the population and be externally valid*. when appropriately analysed. (The analyses in this book assume all samples are* simple random samples*.) Furthermore,* results from approximately representative samples may generalise to the population and be externally valid *if those* in* the study are not obviously different than those *not in* the study.

Example 9.2 (External validity)**.** A New Zealand study (Gammon et al. 2012) identified (for well-documented reasons) a particular group to study: 'women of South Asian origin living in New Zealand' (p. 21).

The women in the sample came from a study

> ... which investigated the health and lifestyle of women of South Asian origin living in New Zealand. Subjects [...] were recruited using a convenience sample method throughout Auckland, which is New Zealand's largest city, and the city in which most South Asian immigrants settle...
> — (Gammon et al. (2012), p. 21)

The results may not generalise to the *intended* population of 'women of South Asian origin living in New Zealand' because all the women in the sample came from only one city in New Zealand (Auckland), and the sample was not a random sample.

The results will not generalise to *all* New Zealand women, but this is *not* a limitation: the target population was only 'women of South Asian origin living in New Zealand.' The researchers did not intend the results to apply to *all* New Zealnd women.

Example 9.3 (Using biochar)**.** A study of using biochar (Farrar et al. 2018) to grow ginger only used one farm at Mt Mellum, Australia.

While the results may only apply to growing ginger at Mt Mellum, the encouraging results suggest that a wider, more general, study of the impact of using biochar to grow ginger would be worthwhile.

In addition, ginger is usually grown is similar types of climates and soils, so the results *may* apply to other ginger farms also.

9.3 Limitations: Internal validity

Internal validity refers to how reasonable and logical the results from the study are: the strength of the inferences that can be made from the sample (Sect. 3.6). That is, an internally valid study is *effective* in demonstrating that the conclusions made from the sample cannot be explained any other way.

Internal validity can be compromised by confounding, the carryover effect, the Hawthorne effect, the observer effect, and/or the placebo effect. Consequently, if any of these issues are

likely to compromise internal validity, the implications on the interpretation of the results should be discussed.

For example, if the study design implies that the Hawthorne effect is likely to be an issue (since the participants were not blinded), this should be clearly stated, and the conclusion should indicate that the individuals in the study may have behaved differently than usual because (for example) they knew the were in a study.

The internal validity of observational studies is often compromised because confounding can be less effectively managed than for experimental studies.

Example 9.4 (Internal validity)**.** In a study of the hand-hygiene practices of paramedics (Barr et al. 2017), *self*-reported hand-hygiene practices were very different than what was reported by *peers*:

> ...social desirability and identity may have led to the intentional misreporting of IPC [infection prevention and control] behaviors in favor of better compliance by the participants. Evidence for this is that participants reported much higher levels of compliance for themselves than their colleagues
> — Barr et al. (2017), p. 777.

That is, when participants knew they were being studied, they may have given responses that made their own behaviours appear better than their colleagues. This is a study limitation that was necessary to discuss.

Extra example: A study (Botelho et al. 2019) examined the food choices made when subjects were asked to shop for ingredients to make a last-minute meal.

Half were told to prepare a 'healthy meal,' and the other half told just to prepare a 'meal.' Part of the Discission stated:

> Another limitation is that results report findings from a simulated purchase. As participants did not have to pay for their selection, actual choices could be different. Participants may also have not behaved in their usual manner since they were taking part in a research study, a situation known as the Hawthorne effect.
> — Botelho et al. (2019), p. 436

9.4 Limitations: Ecological validity

The *practical* of the study results in the real worldshould also be discussed. This is called *ecological validity*.

Definition 9.1 (Ecological validity)**.** A study is *ecologically valid* if the study methods, materials and context approximate the real situation being studied.

Studies don't *need* to be ecologically valid to be useful; much can be learnt under special

conditions, as long as the potential limitations are understood when applying the results to the real world. Although ecological validity is not essential for a good study, ecological validity is useful if it is possible to achieve.

The ecological validity of experimental studies may be compromised because the experimental conditions are sometimes contrived.

Example 9.5 (Ecological validity)**.** Consider a study to determine how likely it is that people will buy a coffee in a reuseable cup.

We could *ask* people about their *intentions*. This study may not be ecologically valid, as how people *act* in the real world may not align with what the *say*, especially when social pressures exist to use reusable cups.

An alternative study involves *watching* people buy coffees at various coffee shops, and record what people actually *do* in practice.

This second study is more likely to be e*cologically valid*, as we are watching actual behaviour in the real world.

Extra example: A study was completed to observe the effect of using high-mounted rear brake lights (Kahane and Hertz 1998), which are now commonplace. The American study showed that such lights reduced rear-end collisions by about 50%. However, after making these lights mandatory, rear-end collisions reduced by only 5%. Why?

9.5 Summary

The limitations in a study need to be identified. Limitations may be related to **internal validity** (effectiveness), **external validity** (generalisability), or **ecological validity** (practicality).

9.6 Quick review questions

A study (Bingham et al. 2016) examined the effect of peer pressure from passengers among teenage male drivers; the aim was to

> ...experimentally test the effects of passenger presence and social influence [...] of male adolescent novices in a simulated driving task.
> Bingham et al. (2016), p. 126

The use of a driving simulator was justified as:

> ... driving simulation has been shown to be an externally valid predictor of real-world driving
> Bingham et al. (2016), p. 125

The *Discussion* section of the article includes a subsection called 'Strengths and limitations.' Part of that sub-section reads:

> participants were closely clustered around average rates of resistance to peer influence for this age group [...] so it is unclear to what extent these findings would generalize to participants with weaker resistance to peer influences.
> Bingham et al. (2016), p. 135

Later, the paper reports:

> ... the use of an age-peer [passenger assigned by the researchers] allowed substantial experimental control, it may have provided participants with an artificial experience compared to the influence of actual friends.
> Bingham et al. (2016), p. 135

1. The study used $n = 52$ 16- and 17-year-old males in the study. Should the external validity of the study be criticised for only using teenage *males* in the study, and not teenage females?
2. What does the *second last* quotation above mean?
3. True or false: The Hawthorne effect is likely to be an issue in this study.

9.7 Exercises

Selected answers are available in Sect. D.9.

Exercise 9.1. A student project at the university where I work had the RQ:

> Among USC students on-campus, is the percentage of word retention higher in male students than female students?

When they were discussing *external validity*, they said:

> We cannot say whether or not that the general public have better or worse word retention compared to the students that we will be studying.

Why is the statement not relevant?

Exercise 9.2. Despite their common use, no experimental scientific evidence shows that parachutes are effective (Smith and Pell 2003). To obtain evidence, researchers studied this scenario (Yeh et al. 2018). Part of the Abstract for the paper (slightly edited for clarity) says:

> **Objective** To determine if using a parachute prevents death or major traumatic injury when jumping from an aircraft.
> **Design** Randomized controlled trial.

> **Setting** Private or commercial aircraft between September 2017 and August 2018.
> **Participants** 92 aircraft passengers aged 18 and over were screened for participation. 23 agreed to be enrolled and were randomized.
> **Intervention** Jumping from an aircraft (airplane or helicopter) with a parachute versus an empty backpack (unblinded).
> **Main outcome measures** Composite of death or major traumatic injury (defined by an Injury Severity Score over 15) upon impact with the ground measured immediately after landing.
> **Results** Parachute use did not significantly reduce death or major injury (0% for parachute v 0% for control; $P > 0.9$).
> **Conclusions** Parachute use did not reduce death or major traumatic injury when jumping from aircraft in the first randomized evaluation of this intervention. However, the trial was only able to enroll participants on small stationary aircraft on the ground, suggesting cautious extrapolation to high altitude jumps [...]
> — Yeh et al. (2018)

Based on this information:

1. Carefully define POCI.
2. What *type* of study is this: observational or experimental?
3. What are the variables?
4. Comment on the ecological validity of this study.
5. Comment on the limitations of the study.
6. What are the conclusions?

Exercise 9.3. A study of how well hospital patients sleep at night (Delaney et al. 2018) set out to

> ...to investigate the perceived duration and quality of patient sleep and identify any environmental factors associated with patient-reported poor sleep in hospital.
> — Delaney et al. (2018) p. 1

In discussing the study, the researchers state:

> Patients and nursing staff were recruited for this study. Non-probability convenience sampling was used to recruit patients to participate...
> — Delaney et al. (2018) p. 2

Later, while discussing the limitations, the researchers state:

> While the multiple methods of data collection and inclusion of 15 clinical areas are strengths of this study, the results may not be generalisable to all hospitals or all ward areas [...] while most healthy individuals sleep primarily or exclusively at night, it is important to consider that patients requiring hospitalization will likely require some daytime nap periods. This study looks at sleep only in the night-time period 22:00–07:00h, without the context of daytime sleep considered.
> – Delaney et al. (2018) p. 7

Discuss these issues using the *language* introduced in this chapter.

Part III

Collecting data

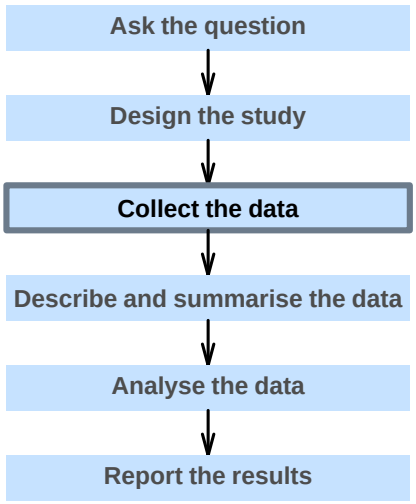
10

Procedures for collecting data

So far, you have learnt to ask a RQ, identify different ways of obtaining data, and design the study.

In this chapter, you will learn how to collect the data needed to answer the RQ. You will learn to:

- record the important steps in data collection.
- describe study protocols.
- ask survey questions.
- describe the basic differences between online and paper surveys.



10.1 Protocols

If the RQ is well-constructed, all terms are clearly defined, and the research design is clear and well explained, then collecting the data should be reasonably easy to implement. However, data collection may be time-consuming, tedious and expensive so getting the data collection correct first time is important.

Before collecting the data, a plan should be established and documented that explains exactly how the data will be obtained; Unforeseen complications are not unusual, so often a *pilot study* (or a *practice run*) is conducted before the real data collection takes place, to see if the planned procedure is practical and optimal. This plan is a draft **protocol**.

A pilot study allows the researcher to:

- Determine the feasibility of the data collection protocol.
- Identify unforeseen challenges.
- Obtain data that might help with sample size calculations.
- Potentially save time and money.

Definition 10.1 (Protocol)**.** A *protocol* is a procedure documenting the details of the design and implementation of studies, and for data collection.

Definition 10.2 (Pilot study)**.** A *pilot study* is a small test run of the study protocol, used to check that the protocol seems appropriate and practical, and to identify possible problems with the design or protocol.

After the pilot study, the planned protocol may need to be refined. Once the protocol has been finalised, then the data can be collected.

Protocols ensure repeatability of the study (to ensure successful replication of results by others) to enable others to confirm or compare results, and others can understand exactly what was done, and how:

> . . . scientific reports must describe experiments in sufficient detail to allow other researchers to reproduce them.
> Mendes (2018) (p. 3081)

Protocols should clearly indicate how design aspects (such as blinding the individuals, random allocation of treatments, etc.) will happen. This final record is the final **protocol**. Someone else should be able to read the protocol and approximately repeat the study (this is part of ethical research practice: Sect. 4). Diagrams can be useful to aid explanations. All studies should have a well-established protocol for describing how the study was done.

Example 10.1 (Protocol)**.** A study (Wojcik et al. 1999) examined the forward-leaning angle from which people could recover and not fall, to determine if this angle was different (on average) for younger and older people. The paper goes into great detail to explain the protocol (almost 1.5 pages, plus a diagram).

Example 10.2 (Protocol)**.** Consider this partial protocol, which shows ethics and honesty in describing a protocol:

> Fresh cow dung was obtained from free-ranging, grass fed, and antibiotic-free Milking Shorthorn cows (*Bos taurus*) in the Tilden Regional Park in Berkeley, CA. Resting cows were approached with caution and startled by loud shouting, whereupon the cows rapidly stood up, defecated, and moved away from the source of the annoyance. Dung was collected in ZipLoc bags (1 gallon), snap-frozen and stored at -80 C.
> — Hare et al. (2008), p. 10

Extra example: A study (Stensballe et al. 2005) examined three different types of male catheters, to compare 'withdrawal friction force.' The paper goes into great detail to explain the protocols (almost a whole page, plus a (painful-looking) diagram).

In addition, the exclusion criterion are given as:

> Subjects with experience of recurrent urinary tract infections, known congenital urogenital abnormalities or known urethral strictures were excluded from participation.
> — Stensballe et al. (2005), p. 979

One approach to documenting the data collection process is to use the STAR method[1]:

- S (**Structured**): The protocol is organised logically, with attention to detail.
- T (**Transparent**): All the necessary information is provided: the protocol is transparent, comprehensive and accurate.
- A (**Accessible**): The protocol is easy to access, easy to follow, and easy to comprehend.
- R (**Reporting**):
  The protocol is reported in whole and in detail, so it could be replicated.

10.2 Collecting data using surveys

Data may be collected in many ways (laboratory experiments, taking measurements, field observations, etc.). For both observational and experimental studies, though, collecting data using surveys is common. Surveys are very difficult to do well: question wording is crucial, and surprisingly difficult to get right (Fink 1995).

Questions in a survey may be *open-ended* (respondents can write their own answers) or *closed* (respondents select from a small number of possible answers, as in multiple-choice questions). Both open and closed questions have advantages and disadvantages. Answers to open questions usually lend themselves to qualitative analysis.

This section only briefly examines surveys:

- Asking survey questions; and
- Comparing online and paper surveys.

10.2.1 Asking survey questions

Many issues must be kept in mind when framing survey questions; here are some.

- **Avoid leading questions** which may indicate how respondents are expected to answer. Question wording is the usual reason for leading questions.
- **Avoid ambiguity**: Avoid terms that may be unfamiliar, and questions that are unclear.

[1] https://www.cell.com/star-methods

- **Avoid asking the uninformed**, and avoid asking respondents about issues they don't know about (for example, people will even give directions to places that do not even exist (Collett and O'Shea 1976)). Many people will tend to give a response even if they do not understand, but such responses are worthless.
- **Avoid complex and double-barrelled questions**; these are often hard to understand.
- **Avoid problems with confidentiality**, which would be considered unethical. Ethics committees usually look very carefully for question that are unethical. In special cases and with justification, ethics committees may allow such questions.
- **Ensure** that questions and clearly and precisely worded.
- **Ensure** that options for multiple-choice questions are *mutually exclusive* (all answers fit into only one category) and *exhaustive* (the categories cover *all* possible options).

Example 10.3 (Leading question)**.** This survey question is a *leading question*, because the expected response is obvious:

> Because bottles from bottled water create enormous amounts of non-biodegradable landfill and hence pose a threat to sensitive native wildlife, do you support a ban on bottled water in Australia?

Example 10.4 (Question wording)**.** Question wording can be important. These two questions would produce different percentages of respondents agreeing:

- Which is easier to *buy*: cigarettes, beer or marijuana?
- Which is easier to *obtain*: cigarettes, beer or marijuana?

Example 10.5 (Leading question)**.** Consider this survey question:

> Do you like this new orthotic?

Although not obvious, this question may incite respondents to please, since *liking* is the only option presented. Better would be to ask:

> Do you like or dislike this new orthotic?

Even better (but more difficult to implement) is to ask the second question above, but randomly chose the order of the 'like' and 'dislike'; that is, ask some respondents if they 'like or dislike' the new orthotic, and others if they 'dislike or like' the new orthotic.

Example 10.6 (Ambiguous question)**.** Consider this survey question:

> Do children run faster now?

This question is ambiguous: Faster now compared to *what* or *when*?

Example 10.7 (Asking the uninformed)**.** Consider this survey question:

> Is the use of fibre composites for waterside recreational purposes likely to cause the material to swell?

Only people involved in the industry are likely to be able to properly answer this question. Nonethless, many people will still give an opinion, even if they are uninformed. This data will be effectively useless (response bias), but the researcher may not realise this.

Example 10.8 (Unclear wording)**.** Consider this survey question:

> I don't go out of my way to purchase low-fat food unless they are also low in calories but not necessarily salt. Do you agree?

It is not clear what a 'yes' answer means.

Example 10.9 (Double-barrelled question)**.** Consider this survey question:

> Do you jog and swim for exercise?

This question would be better asked as two separate questions: one asking about jogging, and one about swimming.

Example 10.10 (Confidentiality)**.** Consider this survey question:

> Do you have a water tank that has been installed illegally, without council permission?

Respondents are unlikely to admit to breaking rules.

Think 10.1 (Survey questions)**.** *Consider this survey question:*

> *Consider this book that you are currently reading. How useful do you think this book would be for students and young professionals in the field?*

What is the biggest problem with this survey question?

Example 10.11 (Miutually exclusive options)**.** (#exm:MutuallyExclusive Qns) Consider this survey question (from Chan et al. (2008)):

> Approximately how much time do you spend on attending to patient's medications at the event of a non-critical case? (Includes writing down a medication list, searching for medications)

The options are:

- 0–5 minutes
- 5–10 minutes
- More than 10 minutes

This is a poor question, because a respondent does not know which option to select for an answer of "5 minutes."

Answer: The answer is given in the online book.

10.2.2 Online and paper surveys

Surveys may be conducted using paper-based surveys, or online surveys; both have advantages and disadvantages (Porter 2004).

Paper-based surveys require the survey information to be manually entered into the computer for later analysis, which is time consuming and expensive, and prone to data-entry errors. Paper-based surveys can also be costly to prepare, especially if physical mailing and photocopying is necessary. However, people may be more likely to complete paper-based surveys if they are presented with a survey face-to-face and someone waits to collect the completed survey.

Online surveys make data collection easier and data entry easier: data are entered directly onto a computer. This means less manual handling and less chance of data entry errors. Online surveys are also easier to share with a geographically-diverse group of people (for example, through email or social media), but only if the relevant contact details are available. However, online surveys may have a lower response rate, as respondents may be reluctant to click on links in emails (especially from unknown sources), may ignore emails, or the emails may be flagged a spam.

10.3 Summary

Having a detailed procedure for collecting the data (the **protocol**) is important. Using a **pilot study** to trial the protocol is often a good idea. Sometimes, data can be collected using surveys, either on **paper** or **online**. However, creating good survey questions is far more difficult than it looks...

10.4 Quick review questions

1. What is the biggest problem with this survey question: 'Do have bromodosis?' (Possible answers are: Yes/No)

2. What is the biggest problem with this survey question: Do you spend too much time connected to the internet? (Possible answers are: Yes/No)
3. What is the biggest problem with this survey question: 'Do you eat fruits and vegetables?' (Possible answers are: Yes/No)

10.5 Exercises

Selected answers are available in Sect. D.10.

Exercise 10.1. What is the problem with this survey question:

> What is your age? (Select one option)
> - Under 18
> - Over 18

Exercise 10.2. Which of these survey questions[2] is better, and why?

1. Should concerned dog owners vaccinate their pets?
2. Should dogs be required to be vaccinated or not?

Exercise 10.3. Before the 2019 *State of the Union* address, American president Donald Trump sent out an online survey[3] to gather information. Some of the questions are given below. Critique these questions:

1. Do you agree that President Trump is taking our country in the RIGHT DIRECTION?
   - Yes
   - No
   - No Opinion
   - Other, please specify:
2. Do you agree with President Trump's unwavering commitment to, and respect for, our incredible veterans and TROOPS?
   - Yes
   - No
   - No Opinion
   - Other, please specify:
3. Are you satisfied with President Trump's efforts to revitalize American manufacturing?
   - Yes
   - No
   - No Opinion

[2] https://surveytown.com/10-examples-of-biased-survey-questions/

[3] https://action.donaldjtrump.com/sotu-prep-survey?DCPe=a2VpdGhAd3Jrc2NzLmNvbQ&utm_medium=email&utm_source=ET_16&utm_campaign=20190201_8898_sotu-2019-prep-survey_donaldjtrump_tmagac&utm_content=gop_surveys_text_take_top_other_all

- Other, please specify:

Part IV

Describing and summarising data

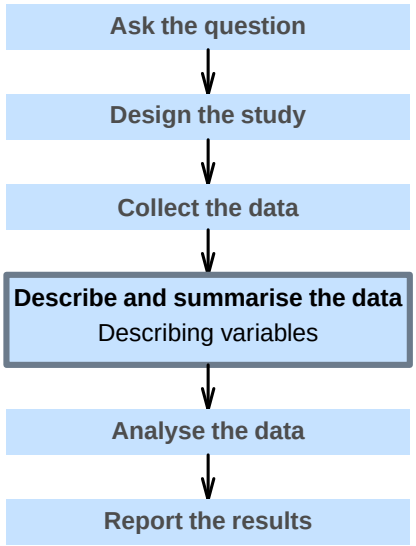
11

Describing data

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study and collect the data.

In this chapter, you will learn how to describe the data, because this determines how to proceed with the analysis. You will learn to:

- identify qualitative and quantitative variables.
- identify nominal and ordinal qualitative variables.
- identify continuous and discrete quantitative variables.
- describe data in ways suitable for use in software.



11.1 Quantitative and qualitative data

Understanding the *type* of data collected is essential before starting any analysis, because the *type* of data determines how to proceed with summaries and analyses. Broadly, data may be described as either:

- **Quantitative** data; or
- **Qualitative** data.

We can also talk about quantitative and qualitative *variables*. (Remember that variables are measured on the *individuals* in the study.) The *variable* is the description of what varies,

and the *data* are the values of the variables that are recorded. Quantitative variables record quantitative data, and qualitative variables record qualitative data.

⚠ *Quantitative research* summarises and analyses data using numerical methods (Sect. 1.7).

Quantitative research can include both *quantitative* and *qualitative* variables, because both *quantitative* and *qualitative* data can be summarised numerically (Chaps. 13 and 14 respectively) and analysed numerically.

Example 11.1 (Variables and data)**.** 'Age' is a *variable* because age can vary from individual to individual. The *data* would be values such as 13 years, 21 years and 76 years.

11.1.1 Quantitative data: Discrete and continuous data

Quantitative data are *mathematically* numerical. Most data that are counted or measured will be quantitative. Quantitative data is often (but not always) measured with measurement units (such as *kg* or *cm*).

Definition 11.1 (Quantitative data)**.** *Quantitative data* is *mathematically* numerical data: the numbers themselves have numerical meaning, and it makes sense to be able to perform mathematical operations on them. Most data that are counted or measured will be quantitative.

Be careful: Just because the data are numbers, it does not necessarily mean that the data are quantitative. *Mathematically numerical data are quantitative*; that is, numbers with numerical *meanings*.

Example 11.2 (Quantitative data)**.** Australian postcodes are numbers, but are *not* quantitative. The numbers are just labels. A postcode of 4556 isn't one 'better' or one 'more' than a postcode of 4555.

The values do not have numerical meanings. Indeed, rather than numerical postcodes, alphabetic postcodes could have been chosen. For example, the post code of Caboolture is 4510, but it could have been QCAB.

Quantitative data may be further defined as *discrete* or *continuous*. *Discrete* quantitative data has possible values that can be counted, at least in theory. Sometimes, the possible values may not have a theoretical upper limit, yet can be still considered 'countable.'

Definition 11.2 (Discrete data)**.** *Discrete* quantitative data has a countable number of possible values between any two given values of the variable.

Example 11.3 (Discrete quantitative data)**.** These (quantitative) variables are *discrete* (and so record *discrete* quantitative data):

- The *number* of heart attacks in the previous year experienced by women over 40. Possible values are 0, 1, 2, ...

- The *number* of cracked eggs in a carton of 12. Possible values are: 0, 1, 2, ... 12.
- The *number* of orthotic devices a person has ever used. Possible values are 0, 1, 2, ...
- The *number* of fissures in turbines after 5000 hours of use. Possible values are 0, 1, 2, ...

Continuous quantitative data has values that cannot, at least in theory, be recorded exactly. In other words, another value can always be found between any two given values of the variable, if we measure to a greater number of decimal places. In practice, though, the values need to be rounded to a reasonable number of decimal places.

Definition 11.3 (Continuous data)**.** *Continuous* quantitative data have (at least in theory) an infinite number of possible values between any two given values.

Height is continuous: between the heights of 179cm and 180cm, many heights exist, depending on how many decimal places are used to record height. In practice, however, heights are usually rounded to the nearest centimetre for convenience. All continuous data are rounded.

Example 11.4 (Continuous quantitative data)**.** These (quantitative) variables are *continuous* (that is, they record continuous quantitative data):

- The *weight* of 6-year-old Australian children. Values exist between any two given values of weight, by measuring to more decimal places of a kilogram; we would usually quote weight to the nearest kilogram
- The *energy consumption* of houses in a given city. Values exist between any two given values of energy consumption, by measuring to more and more decimal places of a kiloWatt-hour (kWh); we would usually quote to the nearest kWh.
- The *time* spent in front of a computer each day for employees in a given industry. Values exist between any two given times, by measuring to more decimal places of a second; we would usually quote the times to (say) the nearest minute, or the nearest 15 minutes.

11.1.2 Qualitative data: Nominal and ordinal data

Qualitative data has distinct labels or categories that are not mathematically numerical. These categories are called the *levels* or the *values* of the variable.

Definition 11.4 (Qualitative data)**.** *Qualitative data* is not *mathematically* numerical data: it consists of categories or labels.

Definition 11.5 (Levels)**.** The *levels* (or the *values*) of a qualitative variable refer to the names of the distinct categories.

Example 11.5 (Qualitative data)**.** 'Brand of mobile phone' is a qualitative variable. Many levels are possible (that is, many possible brands), but these could be simplified by defining the levels as 'Huawei,' 'Apple,' 'Samsung,' 'Google' and 'Other.'

Be careful: *numerical* data may be qualitative. Qualitative data are not *mathematically* numerical; that is, the numbers don't have numerical *meanings*.

Example 11.6 (Qualitative data)**.** Australian postcodes are numbers, but are *qualitative* (Example 11.2).

Think 11.1 (Types of qualitative data)**.** *Here are two survey questions that produce qualitative data.*

1. *What is your blood type?*

- *Type A.*
- *Type B.*
- *Type AB.*
- *Type O.*

2. *What is your age group?*

- *Under 20.*
- *20 to under 30.*
- *30 to under 50.*
- *50 or over.*

What features of the data collected from the questions are similar? What features are different?

Qualitative data can be further classified as *nominal* or *ordinal*. *Nominal* variables are qualitative variables where the levels have *no natural order*. *Ordinal* variables are qualitative variables where the levels do have a *natural order*. So in Extra Example 11.1, 'Blood type' is qualitative *nominal*, while 'Age group' is qualitative *ordinal*.

Definition 11.6 (*Nominal* qualitative variables)**.** A *nominal* qualitative variable is a qualitative variable where the levels *do not* have a natural order.

Definition 11.7 (*Ordinal* qualitative variables)**.** An *ordinal* qualitative variable is a qualitative variable where the levels *do* have a natural order.

Example 11.7 (Nominal data)**.** This survey question will produce *nominal* data:

How do you usually get to university?
- Car (as driver or passenger).
- Bus.
- Ride bicycle or walk.
- Other.

The data will be nominal with four levels. The levels can appear in any order: from largest group to smallest, or in alphabetical order. Because there is no *natural* order, the order used should be carefully considered: what is the most useful order when summarising the data?

Example 11.8 (Ordinal data)**.** This survey question will produce *ordinal* data:

Please indicate the extent to which you agree or disagree with this statement: 'Permeable pavements technology has the potential to revolutionise green building practices.'

- Strongly disagree.
- Disagree.
- Neither agree or disagree.
- Agree.
- Strongly agree.

The data will be ordinal with five levels. Treat the levels in the given order (or the reverse order) makes sense; It would not make sense, for example, to give the levels in alphabetical order.

Example 11.9 (Clarity in definitions)**.** Consider the variable 'Age.' Age is *continuous quantitative*, since we age continuously (on our birthday, we don't suddenly get one year greyer with one extra year's worth of wrinkles...).

Age is usually rounded to the number of completed years, for convenience. However, the age of young children may be given as '3 days' or '10 months,' instead of the nearest year.

Sometimes *Age group* is used instead (such as Under 20; 20 to under 30; 30 to under 50; 50 or over). 'Age group' is *qualitative ordinal*.

Ensure your RQ is clear about which is used!

Example 11.10 (Types of variables)**.** Consider a study to determine if the weight of 500g bags of pasta really is at least 500g. One approach is to record the weight of pasta in each bag (a *quantitative* variable), and compare the *average* weight to the target weight of 500g.

Another approach is to record whether or not each bag of pasta weighed at least 500g (bags are not underweight). This would be a *qualitative* variable, with two *levels* (underweight; not underweight). We could then report the *percentage* of bags that are underweight.

11.2 Describing data in jamovi and SPSS

In practice, quantitative research requires the use of a computer for producing graphs and completing calculations. In this book, two statistical software packages are described for analysis of data:

- **jamovi** and
- **SPSS**.

(For reason to avoid Excel and other spreadsheets, read this information from earlier in this book.)

This section makes only brief notes about setting up data in these software packages; consult a

comprehensive reference for more (and better) details. For both packages, however, declaring the variables correctly is very important (Table 11.1).

Practically all software, including jamovi and SPSS record data in a spreadsheet-like grid, with the *variables in the columns*, and the *units of analysis in the rows*.

TABLE 11.1: *Different types of variables, and their descriptions in jamovi and SPSS*

Type of variable	Further classification	In jamovi	In SPSS
Qualitative	Nominal	Nominal	Nominal
	Ordinal	Ordinal	Ordinal
Quantative	Discrete	Continuous, Integer	Scale
	Continuous	Continuous, Decimal	Scale

11.2.1 Using jamovi

In jamovi, nominal variables are called *Nominal*, and ordinal variables are called *Ordinal* (Table 11.1). In jamovi, *continuous* quantitative variables are called *continuous decimal*, and *discrete* quantitative variables are called (confusingly) *continuous integer*.

To add this information to jamovi, double-click on the variable name at the top of the data worksheet (Fig. 11.1), which produces Fig. 11.2. This opens an area where the data can be described:

- Nominal qualitative variables are set as `Nominal`, and the levels described in the `Levels` area to the right
- Ordinal qualitative variables are set as `Ordinal`, and the levels described in the `Levels` area to the right.
- Quantitative *continuous* variables are set as `Continuous` with the *Data type* as `Decimal`.
- Quantitative *discrete* variables are set as `Continuous` with the *Data type* as `Integer`.

When the information has been entered, clicking the up-arrow on the top right (Fig. 11.2) closes this window.

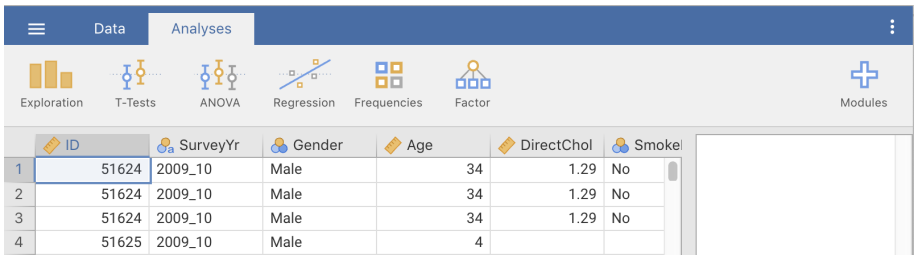


FIGURE 11.1: *jamovi: The variable names at the top of the columns of data*

11.2.2 Using SPSS

In SPSS, variables are described in the *Variable View* window (*not* the *Data View* window). Each variable is then described in the *Measure* column (Fig. 11.3):

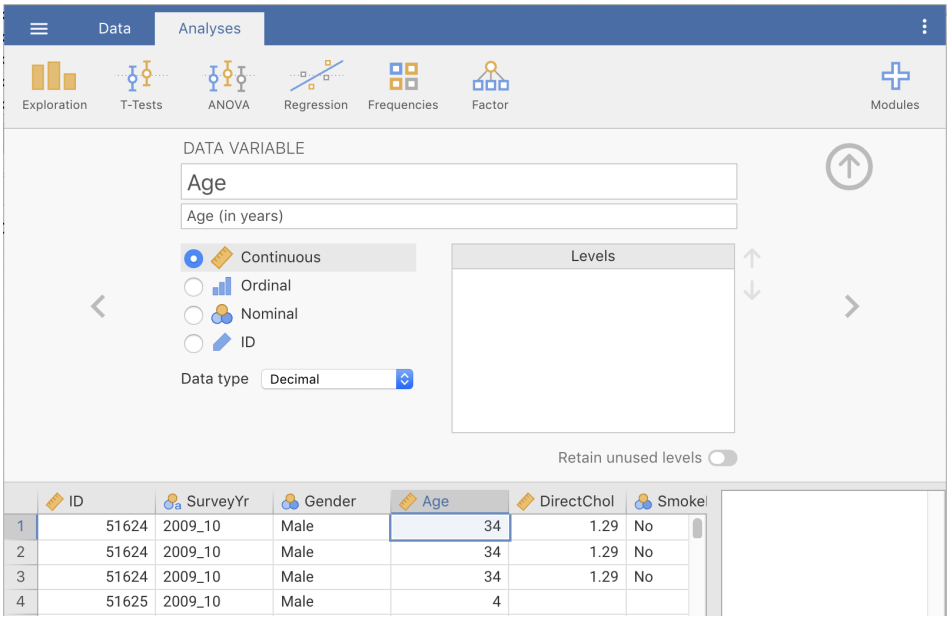


FIGURE 11.2: *jamovi: Setting the variable type*

- Nominal qualitative variables are called `Nominal`.
- Ordinal qualitative variables are called `Ordinal`.
- Quantitative variables are called `Scale`, regardless of whether they are discrete or continuous.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	ID	Numeric	5	0		None	None	8	Right	Scale	Input
2	SurveyYr	String	7	0		None	None	7	Left	Nominal	Input
3	Gender	Numeric	6	0		{1, Female}...	None	6	Right	Nominal	Input
4	Age	Numeric	2	0	Age (in years)	None	None	8	Right	Scale	Input
5	DirectChol	Numeric	4	2	Direct HDL chol...	None	None	8	Right	Scale	Input
6	SmokeNow	Numeric	3	0	Does responde...	{0, No}...	99	3	Right	Nominal	Input
7	Weight	Numeric	5	1	Weight (in kg)	None	None	8	Right	Scale	Input
8	Diabetes	Numeric	3	0		{0, No}...	99	10	Right	Nominal	Input
9	AgeDecade	String	6	0		None	None	6	Left	Nominal	Input
10	AgeMonths	Numeric	3	0		None	None	8	Right	Scale	Input
11	Race1	String	8	0		None	None	8	Left	Nominal	Input

FIGURE 11.3: *SPSS: Setting the variable type*

11.3 Summary

Data and variables can be described as either **quantitative** (either **discrete** or **continuous**) or **qualitative** (either **nominal** or **ordinal**). Variables should be correctly defined in jamovi and SPSS.

11.4 Quick revision questions

A study on the bruising of apples (Doosti-Irani et al. 2016) aimed to determine the relationship between the recorded surface temperature of apple, the depth of bruising.

The researchers purposefully hit apples with three different **forces** (200, 700 and 1200 mJ) to inflict bruises.

This was repeated at three different **locations** of the apple (lower; middle; upper).

The researchers then recorded the **depth** of the bruising, and recorded the **surface temperature** at each bruise location.

1. How would the variable 'region of the apple' be best described?
2. How would the variable 'depth of bruising?' be best described
3. How would the variable 'temperature of the bruise location' be best described?
4. The variable 'force of hit' could be considered as quantitative continuous variable. However, since only a small number of forces are used, it could also be considered as qualitative ordinal.
   If it was considered as *qualitative ordinal*, how many *levels* would the variable have?

11.5 Exercises

Selected answers are available in Sect. D.11.

Exercise 11.1. A study of lime trees (*Tilia cordata*) recorded these variables for 385 lime trees in Russia (Schepaschenko et al. 2017; Dunn and Smyth 2018):

- the foliage biomass, in kg;
- the tree diameter (in cm);
- the age of the tree (in years); and
- the origin of the tree (one of Coppice, Natural, or Planted).

Describe the variables in the study using the language of this chapter.

Exercise 11.2. Are these variables quantitative (discrete or continuous, and with what units of measurement?), or qualitative (nominal or ordinal, and with what levels?)?

1. Systolic blood pressure.
2. Program of enrolment.
3. Academic grade (HD; DN; CR; PS; FL).
4. Number of times a person visited the doctor last year.

Exercise 11.3. A study of body mass index and its relationship with use of social media (Alley et al. 2017) recorded these variables (among others) from a group of 1140 participants:

1. Age (under 45; 45 to 64; 65 or over)
2. Gender (male; female)
3. Location (urban; rural)
4. Social media use (none; low; high)
5. BMI (body mass index; the body mass in kg, divided by the square of height in cm)
6. Total sitting time, in minutes per day

For each variable, determine the *type* of variable: quantitative (discrete or continuous, and with what units of measurement?), or qualitative (nominal or ordinal, and with what levels)?

Exercise 11.4. In a study of the influence of using ankle-foot orthoses in children with cerebral palsy (Swinnen et al. 2017), the data in Table 11.2 describe the 15 subjects. (GMFCS is used to describe the impact of cerebral palsy on their motor function, where *lower* levels means *better* functionality: the Gross Motor Function Classification System[1].) Describe the variables in the study.

TABLE 11.2: *Describing the sample in the orthoses data set*

Gender	Age (years)	Height (cm)	Weight (kg)	GMFCS
M	9	136	34.5	1
M	7	106	16.2	2
M	7	129	21.1	1
M	12	152	40.4	1
M	11	146	39.3	2
M	5	113	18.1	1
M	6	112	16.7	2
M	8	112	19.1	1
M	8	138	28.6	1
M	6	116	19.3	1
F	7	113	17.6	1
M	11	141	34.9	1
M	7	136	34.5	1
F	9	128	21.9	1
F	8	133	23.0	1

Exercise 11.5. A study of fertilizer use (Lane 2002; Dunn and Smyth 2018) recorded the soil nitrogen after applying different fertilizer doses. These variables were recorded:

- the *fertilizer dose*, in kilograms of nitrogen per hectare;
- the *soil nitrogen*, in kilograms of nitrogen per hectare; and
- the *fertilizer source*; one of 'inorganic' or 'organic.'

Describe the variables in the study.

Exercise 11.6. A study (Brunton et al. 2019) recorded the response of kangaroos to drones (one of 'Vigilance,' 'No vigilance,' 'Flee $<$ 10m,' or 'Flee $>$ 10m') and the altitude of the drone (30m, 60m, 100m or 120m). The mob size and sex of the kangaroo was also recorded. Describe the variables in the study.

[1] https://en.wikipedia.org/wiki/Gross_Motor_Function_Classification_System

Exercise 11.7. A study of people who died while taking selfies (Dokur et al. 2018) recorded the location (Table 11.3). Which of the following are the *variables* in the table? For each that is a variable, describe the variable.

1. The location.
2. The number of people who died at each location.
3. The percentage of people who died at each location.

TABLE 11.3: *Locations of people dying while taking selfies*

	Number	**Percentage**
Nature, associated environments	48	43.2
Train, railway, associated structures	22	19.9
Buildings, associated structures	17	15.3
Road, bridge, associated structures	12	10.8
Dams, associated structures	7	6.3
Fields, farms, associated structures	4	3.6
Others	1	0.9

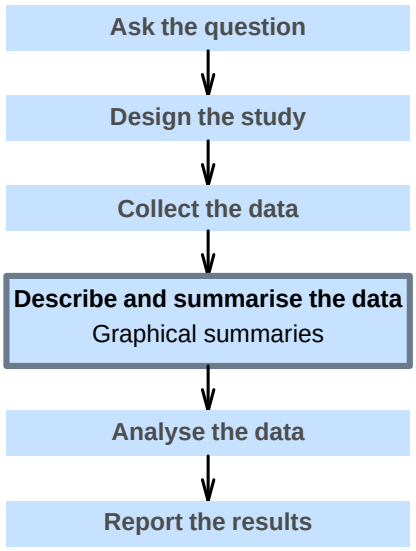
12

Graphical summaries of data

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data and describe the data.

In this chapter, you will learn to graph the data, so we can understand the data used to answer the RQ. You will learn to:

- select the appropriate graphic to graphically summarise data.
- graphically summarise data using quality, appropriate graphs.
- interpret graphs.
- identify badly prepared graphics, giving reasons.



12.1 Introduction

To answer a RQ, a study is designed to collect data, because *data are the means by which the research question is answered*. But before analysis, understanding, describing and summarising the collected data is important. This chapter discusses the use of *graphs* to summarise data.

The purpose of a graph is to display the information in the clearest, simplest possible way, to help the reader understand the message(s) in the data.

Graphs are not produced just for the sake of it; what the graph tells us should always be explained, especially regarding the RQ.

12.2 One quantitative variable

For quantitative data, a graph shows the *distribution* of the data: what values are present in the data, and how often those values appear.

The graphs discussed in this section usually work best for *continuous* quantitative data, but may also be useful for discrete quantitative data if many possible values are present. Sometimes, *discrete* quantitative data with very few values are best graphed using the graphs discussed in Sect. 12.3.

Three different types of graphs can be used to show how the values of one quantitative variable are *distributed*:

- **Stemplot** or **stem-and-leaf plot**: Best for small amounts of data; useful only in some cases.
- **Dot chart**: Best for small amounts of data; good for moderate amounts of data.
- **Histogram**: Best for moderate to large amounts of data.

Whatever graph is used, what the graph shows should be described.

12.2.1 Stem-and-leaf plots

Stem-and-leaf plots (or *stemplots*) are best described and explained using an example, so consider the data in Table 12.1 (which shows just the first 10 of the 44 observations).

The data give the weights (in kg) of babies born in a Brisbane hospital on one day (Steele 1997; Dunn 1999).

The data set also includes the gender of each baby, and the number of minutes after midnight that each birth occurred.

The data are given in the order in which the births occurred.

TABLE 12.1: *The first ten observations (of 44) of the baby-births data*

Gender	Weight (in kg)	Minutes since midnight
Female	3.8	5
Female	3.3	64
Male	3.6	78
Male	3.8	115
Male	3.6	177
Female	2.2	245
Female	1.7	247
Male	2.8	262
Male	3.2	271
Male	3.5	428

For these data, the weights (quantitative) are to one decimal place of a kilogram. In a stemplot, part of each number is placed to the left (the *stem*) of a vertical line, and the rest of each

number to the right (the *leaf*). Here, the whole number of kilograms is placed to the left (as a *stem*), and the first decimal place is placed on the right (as a *leaf*). Figure 12.1 shows the stemplot starting to be built, and Fig. 12.2 shows the final stemplot. (The online version has an animation.) From this plot, most birthweights are seen to be 3-point-something kilograms.

1.7	2.1	2.2	2.2	2.4	2.6	2.6	2.8	2.9	3.0	3.1
3.2	3.2	3.2	3.3	3.3	3.3	3.3	3.3	3.4	3.4	3.4
3.4	3.4	3.4	3.4	3.5	3.5	3.5	3.5	3.5	3.5	3.6
3.6	3.6	3.7	3.7	3.7	3.8	3.8	3.8	3.9	3.9	4.2

Stemplot of birthweights
(2|6 means 2.6 kg)

1	7
2	122
3	
4	

FIGURE 12.1: *Starting to make the stemplot for the baby-weight data: the first 4 observations added*

1.7	2.1	2.2	2.2	2.4	2.6	2.6	2.8	2.9	3.0	3.1
3.2	3.2	3.2	3.3	3.3	3.3	3.3	3.3	3.4	3.4	3.4
3.4	3.4	3.4	3.4	3.5	3.5	3.5	3.5	3.5	3.5	3.6
3.6	3.6	3.7	3.7	3.7	3.8	3.8	3.8	3.9	3.9	4.2

Stemplot of birthweights
(2|6 means 2.6 kg)

1	7
2	12246689
3	012223333444444455555566677788899
4	2

FIGURE 12.2: *The final stemplot for the baby-weight data*

For stem-and-leaf plots:

- Place the larger unit (e.g., kilograms) on the left (stems).
- Place the next smallest unit (e.g., first decimal place of a kilogram) on the right (leaves).
- Some data do not work well with stem-and-leaf plots.
- Sometimes, data might need to be suitably rounded before creating the stem-and-leaf plot.
- The numbers in each row should be evenly spaced, so that the numbers in the columns are under each other. This allows patterns to be seen.
- Within each row, the observations are *ordered* on each stem so patterns can be seen.
- Add an explanation for reading the stem-and-leaf plot. For example, the caption for the stem-and-leaf plot for the baby-birth data in Sect. 12.2.1 says '2 | 6 means 2.6kg,' which explains what the stem plot means. For instance, '2 | 6' could mean 26kg, or 0.26kg.

Example 12.1 (Stem-and-leaf plots)**.** A study of krill (Greenacre 2016) produced 15 measurements of the number of eggs. The stemplot shows that the number of eggs is usually under 10, but occasionally a large number of eggs are seen.

Figure shows the stemplot. (The online version has an animation.)

0	0	0	0	1
1	1	2	2	3
8	16	20	26	31

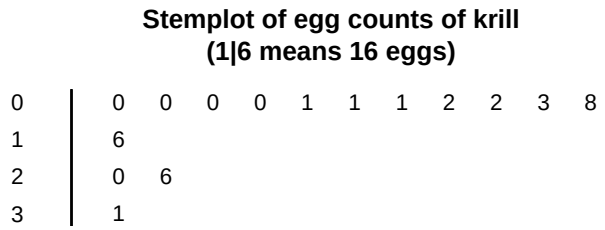


FIGURE 12.3: *The final stemplot for the krill data*

12.2.2 Dot charts (quantitative data)

Dot charts show the original data on a single axis, with each observation represented by a dot.

Example 12.2 (Dot charts)**.** A study examined the serving size of fries at McDonald's (Wetzel 2005), and produced the dot chart in Fig 12.4 (based on Wetzel (2005), Fig. 2).

The mass of fries is almost always *under* the target, and often substantially so. An alternative way to look at these data is to measure the percentage that each serving is in relation to the target serving (Fig. 12.5), where 100% means the serving size was exactly the target weight.

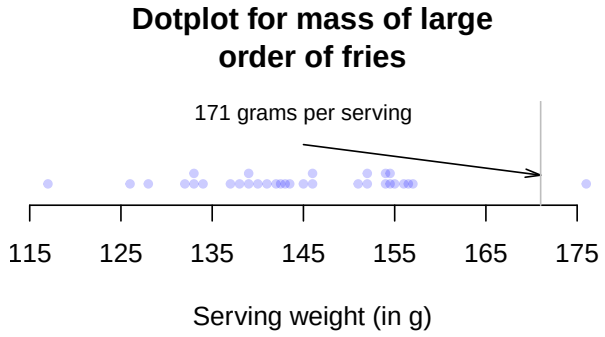


FIGURE 12.4: *Mass measurements for large orders of french fries*

Example 12.3 (Dot charts)**.** Consider again the weights (in kg) of babies born in a Brisbane hospital in one day. Again, a dot chart (Fig. 12.6) shows that most babies are between 3 and 4kg.

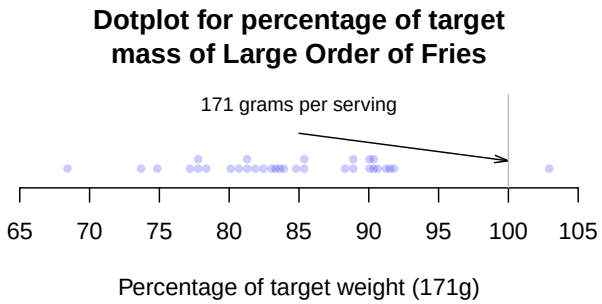


FIGURE 12.5: *Percentage variation from target mass, for large orders of french fries*

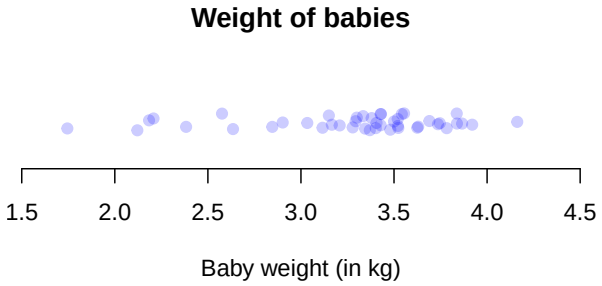


FIGURE 12.6: *A dot chart of the baby-weight data*

12.2.3 Histograms

Histograms are a series of boxes, where the width of the box represents a range of *values* of the variable being graphed, and the height of the box[1] represents the *number* (or *percentage*) of observations within that range of values.

Example 12.4 (Histograms)**.** Consider again the weights (in kg) of babies born in a Brisbane hospital in one day (Dunn 1999). A histogram (below) can be constructed for these data. When an observation occurs on a boundary between the boxes, software usually (but not universally) places it in the *higher* box (so 2.5kg would be counted in the '2.5 to 3.0kg' box, not the '2.0 to 2.5kg' box). The histogram shows, for example, that 17 babies weighed 3.0kg or more, but under 3.5kg.

Figure 12.7 shows the histogram starting to be built, and Fig. 12.8 shows the final histogram. (The online version has an animation.)

Example 12.5 (Histograms)**.** A study of 'headache attributed to ingestion or inhalation of a cold stimulus' (HICS), commonly known as a brain freeze from eating cold food (e.g., ice cream) or drinking a cold drink, measured the duration of the brain freeze (Mages et al. 2017).

A histogram of the data (Fig 12.9, based on Mages et al. (2017), Figure 2b), shows that 11 people experience HICS symptoms less than 5 seconds in length.

In addition, 9 people experienced symptoms for at least 5 but less than 10 seconds, and 1 person experienced symptoms for at least 35 seconds but under 40 seconds.

[1] Technically, the *area* of the box is proportional to the number of observations. Since we only consider histograms where the bars are all the same width, this is equivalent.

1.7	2.1	2.2	2.2	2.4	2.6	2.6	2.8	2.9	3.0	3.1
3.2	3.2	3.2	3.3	3.3	3.3	3.3	3.3	3.4	3.4	3.4
3.4	3.4	3.4	3.4	3.5	3.5	3.5	3.5	3.5	3.5	3.6
3.6	3.6	3.7	3.7	3.7	3.8	3.8	3.8	3.9	3.9	4.2

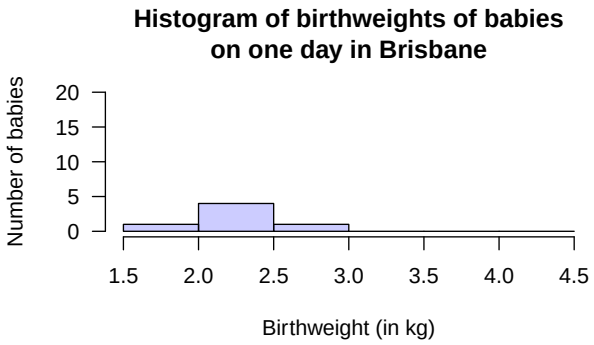


FIGURE 12.7: *Starting to make the histogram for the baby-birth data: the first 6 observations added*

1.7	2.1	2.2	2.2	2.4	2.6	2.6	2.8	2.9	3.0	3.1
3.2	3.2	3.2	3.3	3.3	3.3	3.3	3.3	3.4	3.4	3.4
3.4	3.4	3.4	3.4	3.5	3.5	3.5	3.5	3.5	3.5	3.6
3.6	3.6	3.7	3.7	3.7	3.8	3.8	3.8	3.9	3.9	4.2

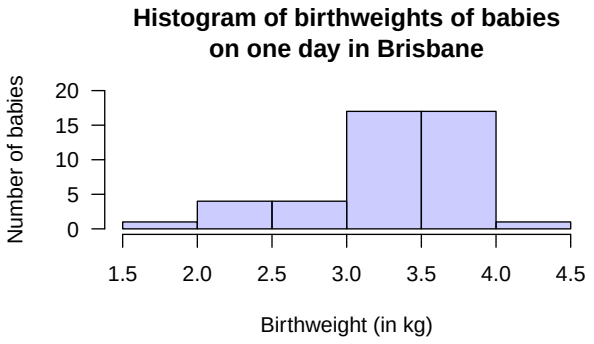


FIGURE 12.8: *The final histogram for the baby-birth data*

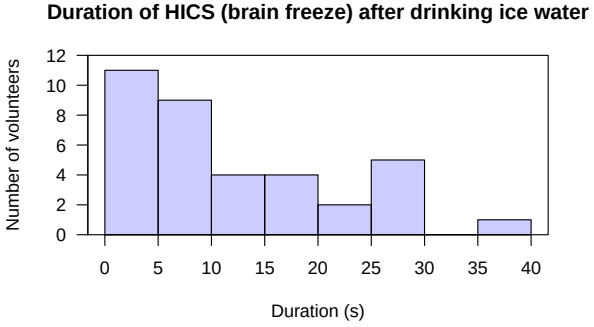


FIGURE 12.9: *Duration of HICS (brain freeze) after drinking ice water*

12.2.4 Describing the distribution

Graphs are constructed to help us understand the data. After producing a graph for one quantitative variable, then, we need to summarise what we learn. For one *quantitative variable*, describe:

1. *Average*: What is an "average" or typical value?
2. *Variation*: How much variation is present in the bulk of the data?
3. *Shape*: How are the values distributed? That is, are most of the values smaller values, or larger values, or about even distributed between smaller and larger values?
4. *Outliers* (observations unusually large or small) or unusual features: Are there any unusual observations, or anything else of interest?

Describing the *shape* can be tricky, but terminology may help:

- Skewed *right*: the bulk of the data is smaller, but there are some larger values (to the *right*).
- Skewed *left*: the bulk of the data is larger, but there are some smaller values (to the *left*).
- Symmetric data (and perhaps bell-shaped): There are approximately equal numbers of values that are smaller and larger.
- Bimodal data: There are two peaks in the distribution.

Typical shapes are shown in Fig. 12.10. Sometimes, no suitable short descriptions is suitable.

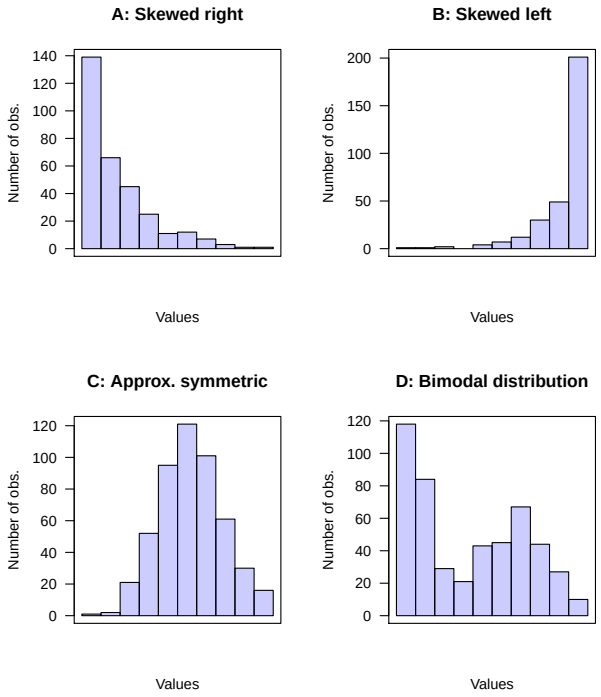


FIGURE 12.10: *Some common shapes of the distribution of qualitative data*

Example 12.6 (Bimodal data)**.** The *Old Faithful* geyser in Yellowstone National Park (USA) erupts regularly (Härdle and others 1991).

The time between eruptions (Fig. 12.11) is clearly bimodal, with a peak near 55 minutes and another near 80 minutes.

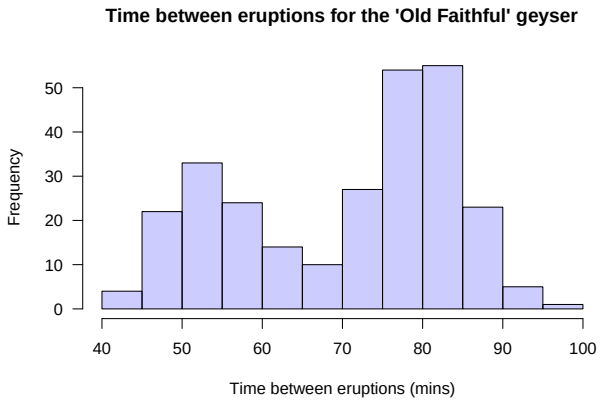


FIGURE 12.11: *Histogram of the times between eruptions for the Old Faithful geyser*

Example 12.7 (Describing quantitative data)**.** For the baby-weight data displayed in, for example, Fig. 12.6:

- The *average* weight is somewhere between 2.5 to 3 kilograms.
- The *variation* in weights is between 1.5 and 4.5 kilograms approximately.
- The *shape* is slightly skewed to the left. That is, occasional small birth weights appear (probably premature babies).
- There doesn't appear to be any outliers or anything unusual.

Think 12.1 (Histograms)**.** *Describe the histogram in Example 12.9, the brain freeze durations.*

Answer: The answer is given in the online book.

12.3 One qualitative variable

For qualitative data, graphs show how often each level of the variable occurs in the data. The three options for graphing qualitative data are:

- *Dot chart*: Usually a good choice.
- *Bar chart*: Usually a good choice.
- *Pie chart*: Only useful in special circumstances, and can be harder to interpret.

Comparing these graphs is useful too; indeed, sometimes a graph may not even be needed.

For *nominal* data, the order in which the levels of the variables appear is unimportant, so categories could be ordered alphabetically, by size, by personal preference, or any other way. Since you have a choice, think about the order that is most useful to readers. For *ordinal* data, the natural order of the levels should almost always be used.

Sometimes these graphs are also used for discrete quantitative data with a small number of possible options.

12.3.1 Dot charts (qualitative data)

Dot charts indicate the counts (or the corresponding percentages) in each level, using dots on a line starting at zero. The levels can be on the horizontal or vertical axis; placing the level names on the vertical axis often makes for easier reading, and room for long labels.

Example 12.8 (Dot plots)**.** A study of spider monkeys (Chapman 1990) examined the social groups present in a sample. A dot chart (Fig. 12.12) show the most common social group has many females plus offspring (with almost 50 social groups).

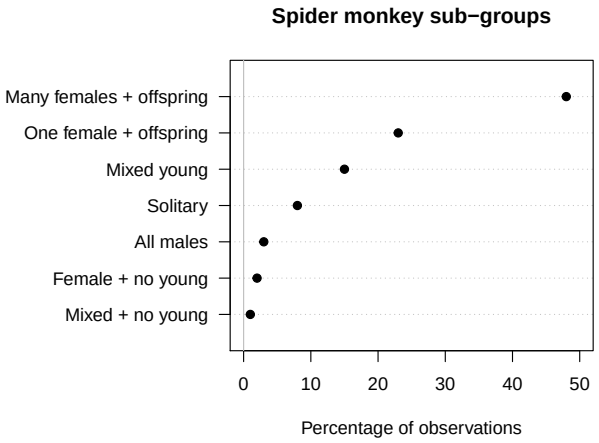


FIGURE 12.12: *Dot chart of spider monkey family groups*

12.3.2 Bar charts

Bar charts indicate the counts in each category using a bar starting from zero. As with dot charts, the levels can be on the horizontal or vertical axis, but placing the level names on the vertical axis often makes for easier reading, and room for long labels.

Example 12.9 (Bar charts)**.** In a study of functional independence (Ocepek et al. 2013), the type of diagnoses were graphed using a bar chart (Fig. 12.13). For example, two people in the sample have cerebral palsy.

The reason for the different coloured bars becomes apparent in Sect. 12.3.3.

For bar charts and dot charts:

- Place the qualitative variable on the horizontal or vertical axis (and label with the levels of the variable).
- Use counts or percentages on the other axis.
- For nominal data, dots and bars can be ordered any way: *Think about the most helpful order*.
- Bars have gaps between bars, as the bars represent distinct categories. In contrast, the bars in histograms are butted together (except when an interval has a count of zero), as the bars represent a numerical scale.

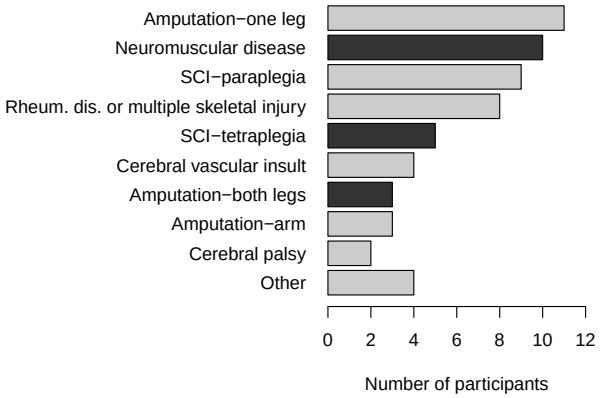


FIGURE 12.13: *Diagnoses of participants*

12.3.3 Pie charts

In pie charts, a circle is divided into segments proportional to the number in each level of the qualitative variable.

Example 12.10 (Pie charts)**.** In a study of functional independence (Ocepek et al. 2013), the severity of the diagnoses were graphed using a pie chart (Fig. 12.14). This picture actually conveys one thing only ("69% of patients had a less severe injury"), so a graph of any kind is probably unnecessary.

The pie chart colours explain the colours used in the bar chart in Example 12.9. This is called *encoding* extra information into the bar chart.

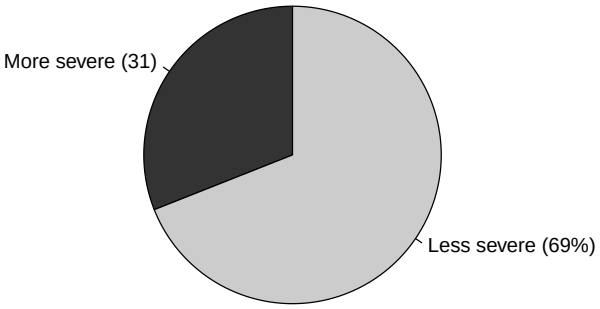


FIGURE 12.14: *Severity of diagnoses of participants*

Pie charts presents challenges:

- Pie charts only work when graphing parts of a whole.
- Pie charts only work when *all* options are present ('exhaustive').
- Pie charts only work when each unit can appear in just one group ('mutually exclusive').
- Pie charts are difficult to use when levels with zero counts, or small counts, are present.
- Pie charts are difficult to read when many categories are present.
- Pie charts are hard to read: In general, human brains are better at comparing *lengths* (as used in bar and dot charts) than comparing *angles* (as used in pie charts) (Friel et al. 2001).

Think 12.2 (Pie charts)**.** *In which of these situations is a pie chart* appropriate*?*

1. *The percentage of people who use these web browsers: Firefox, Chrome, and Safari.*
2. *For each state of Australia, the percentage of people who own an iPhone.*
3. *The percentage of students awarded different grades in this course last semester.*

Answer:

1. A pie chart is **not suitable**.

   Each individual (person) has information recorded on *two* qualitative variables:

   (a) which browser is being asked about (three levels); and
   (b) whether or not they use that browser ('yes' or 'no').

   The three browsers are not *mutually exclusive* (people can use more than one of these browsers) nor *exhaustive* (some people may use browsers not listed, such as Edge, Brave, Vivaldi, etc.). For example, the percentages *could* be that 65% use Firefox, 84% use Chrome, and 20% use Safari. These add to more than 100%.

2. A pie chart is **not suitable**, as the percentages are not parts of a whole.

   Again, each individual (person) has information recorded on *two* qualitative variables:

   (a) which state the person lives in (many levels); and
   (b) whether or not they own an iPhone ('yes' or 'no').

   For example, the percentages *could* be 53% in Queensland, 61% in NSW, 41% in Victoria, and so on. They could possibly add to more than 100%.

3. A pie chart **is suitable**.
   Only one qualitative variable is recorded for each individual (person): their grade.

A bar chart or dot chart could be used for all three situations.

12.3.4 Comparing pie charts and bar charts

Consider the pie chart (using data in Andersen (1977)) in the top panel of Fig. 12.15.

The pie chart displays the number of lung cancer deaths in Fredericia between 1968 and 1971 inclusive, for various age groups (qualitative).

A pie chart *is* appropriate: only one variable is recorded on each individual (the age of each individual person), and the counts are parts of a whole. However, notice that determining which age groups have the most lung cancer deaths is hard.

The equivalent bar chart (lower panel) makes the comparison easy: clearly the age groups '65 to 69' and 'Over 74' have slightly fewer deaths than the other age groups.

Recall that the *purpose of a graph is to is to display information in the clearest, simplest possible way, to help the reaader understand the message(s) in the data*. Adding an artificial third dimension usually makes the message hard to see (Siegrist 1996); see Example 12.11.

Example 12.11 (Comparing graphs)**.** In the NHANES study (Center for Disease Control and Prevention (CDC) 1988--1994), the age of each participant was recorded.

Rank the age groups from largest group to smallest group using each graph in Fig. 12.16, all constructed from the same data.

Which graph makes it easiest to compare the sizes of the categories?

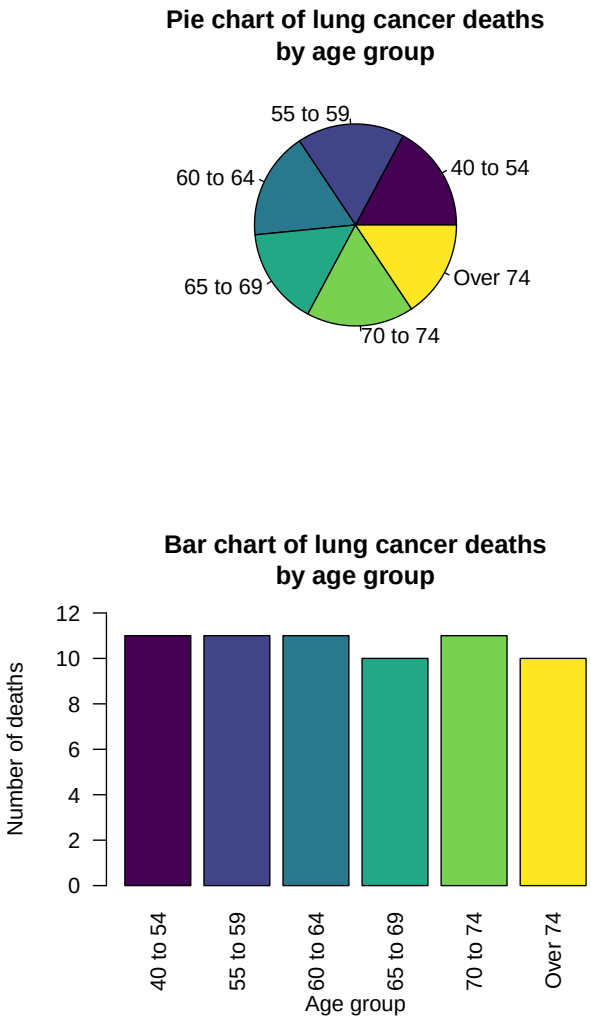


FIGURE 12.15: *Graphs from a study of hospital admission of children with asthma*

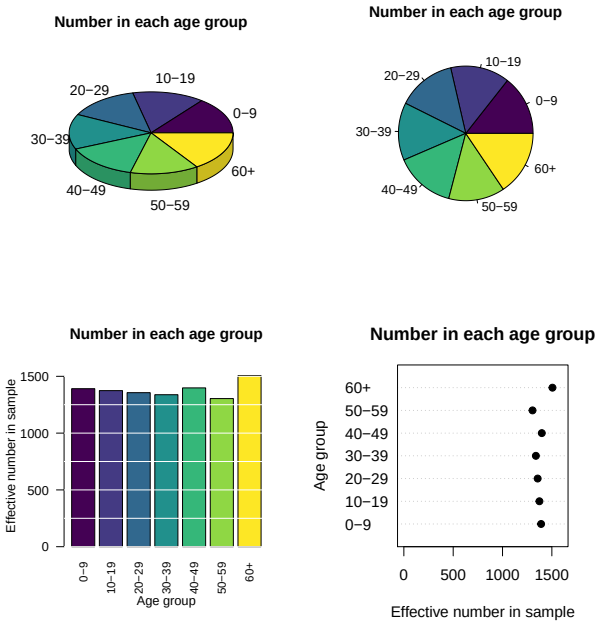


FIGURE 12.16: *Four different graphs for the same data*

12.3.5 Is a graph needed?

Although graphs are excellent for summarising data, sometimes a graphic is not the best way to display information, especially for qualitative data. Sometimes just writing the information is better ('69% of diagnoses were less severe'; Fig. 12.14).

Sometimes a table may be better, such as when a small number of levels is present, or if the details are important. Compare different ways of presenting the NHANES age data in Example 12.11: Fig. 12.16 and Table 12.2 display the same data. Which do you think is 'best,' and why?

TABLE 12.2: *The NHANES age distribution, displayed as a table*

Age group	Number	Percentage
0-9	1391	14.4
10-19	1374	14.2
20-29	1356	14
30-39	1338	13.8
40-49	1398	14.5
50-59	1304	13.5
60+	1506	15.6

12.4 One qualitative variable and one quantitative variable

Relationships between *one qualitative variable* and *one quantitative variable* can be displayed using:

- **Back-to-back stem-and-leaf plot**: Best for small amounts of data when the qualitative variable only has *two levels*;
- **2-D dot chart**: Best choice for small to moderate amounts of data;
- **Boxplot**: Best choice, except for small amounts of data.

12.4.1 Back-to-back stem-and-leaf

Back-to-back stem-and-leaf plots are essentially two stem-and-leaf plots (Sect. 12.2.1) sharing the same stems; one group has the leaves going left-to-right from the stem, and the second group has the leaves going right-to-left from the stem. Back-to-back stem-and-leaf plots can only be used when *two* groups are being compared.

Example 12.12 (Back-to-back stem-and-leaf plots)**.** A study of krill (Greenacre 2016) produced the observations shown in Table 12.3. A back-to-back stem-and-leaf plot of these data makes it easy to compare the two groups visually (Fig. 12.17).

The plot for the *Treatment* data goes from right-to-left, and the data for the *Control* group goes from left-to-right, sharing the same stems. The control group tends to produce more eggs, in general.

TABLE 12.3: *The number of eggs laid by krill, for those in a treatment group and for those in a control group*

Treatment group		Control group	
0	18	0	2
0	21	0	3
1	26	0	8
1	30	0	16
3	35	1	20
8	48	1	26
8	50	1	31
12		2	

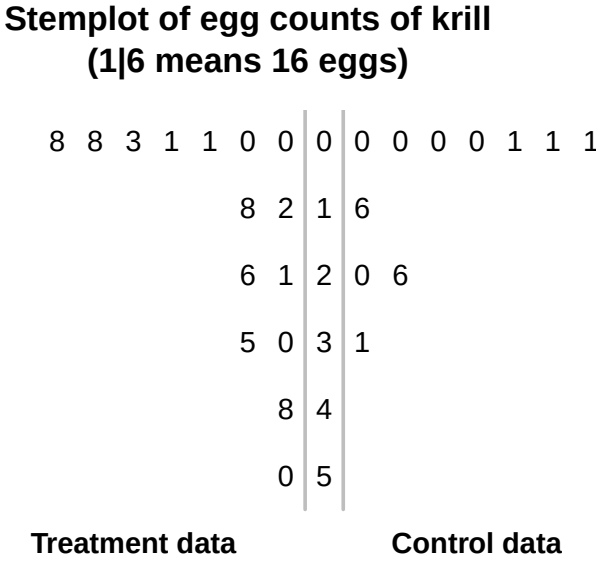


FIGURE 12.17: *The number of eggs from krill, for control and treatment groups*

12.4.2 2-D dot charts

A 2-D dot chart places a dot for each observation, but separated for each level of the qualitative variable (also see Sect. 12.3.1). For the same krill data used in Example 12.12, a dot chart is shown in Fig. 12.18.

Many observations are the same, so some points would be *overplotted* if points were not *stacked* (top panel). Another way to avoid overplotting is to add a bit of randomness (called a 'jitter') in the vertical direction to the points before plotting (bottom panel).

12.4.3 Boxplots

Understanding boxplots takes some explanation, and so boxplots will be discussed again later (Sect. 13.3.3). For the same krill data used in Example 12.12, a boxplot is shown in Fig. 12.19.

To explain boxplots, first focus on just one boxplot from Fig. 12.19: the boxplot for the *Treatment* group. Boxplots have five horizontal lines; from the top to the bottom of the plot (Fig. 12.20):

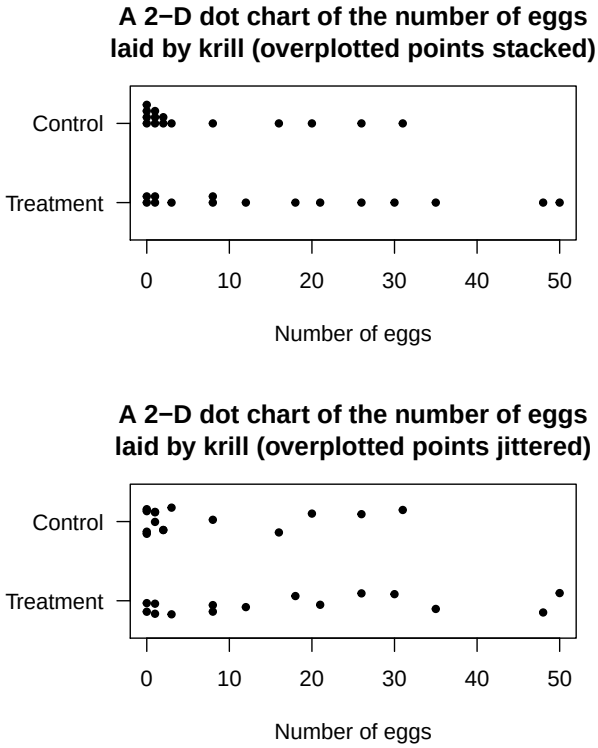


FIGURE 12.18: *Two variations of a 2-D dot chart for the krill-egg data: stacking and jittering*

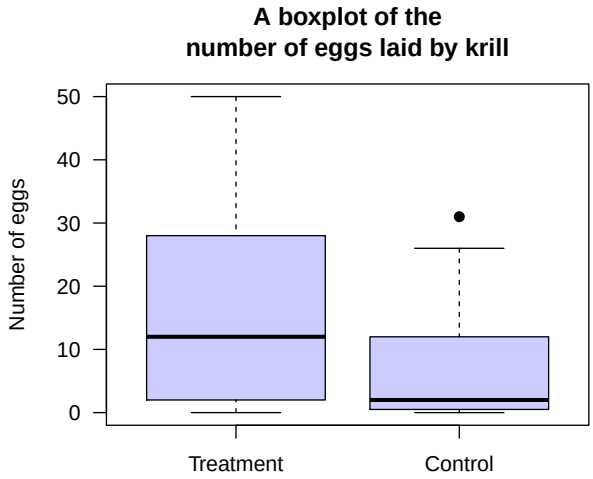


FIGURE 12.19: *A boxplot for the krill-egg data*

- **Top line**: The *largest* number of eggs is 50: This is the line at the top of the boxplot.
- **Second line from the top**: About 75% of the observations are smaller than about 28, and this is represented by the line at the top of the central box. This is called the *third quartile*, or Q_3.
- **Middle line**: About 50% of the observations are smaller than about 12, and this is represented by the line in the centre of the central box. This is an 'average' value for the data, or the *second quartile*, or Q_2.
- **Second line from the bottom**: About 25% of the observations are smaller than about 2, and this is represented by the line at the bottom of the central box. This is called the *first quartile*, or Q_1.
- **Bottom line**: The smallest number of eggs is 0. This is the line at the bottom of the boxplot.

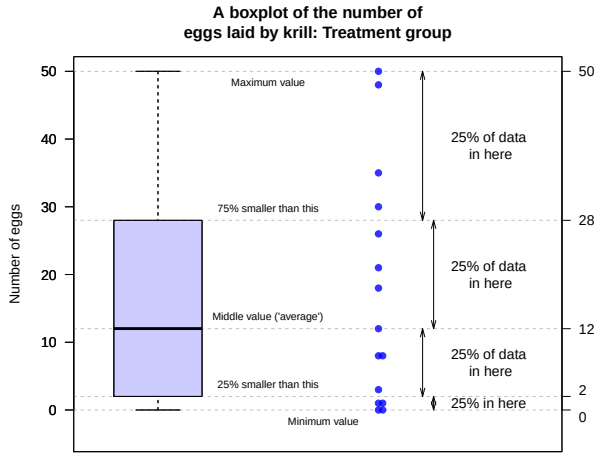


FIGURE 12.20: *A boxplot for the krill-egg data; the boxplot and dotplot just for the treatment group*

However, the box for the krill in the *Control* group is slightly different (Fig. 12.19): One observation is identified with a point, *above* the top line. Computer software has identified this observation as potentially unusual (in this case, unusually *large*), and so has plotted this point separately. (Unusually large or small observations are called *outliers*.)

The values of the quantiles (Q_1, Q_2 and Q_3) are computed as usual.

So, for the *Control* data, the largest observation (31 eggs) is deemed unusually large (using arbitrary rules explained in Sect. 13.5.3). Then the boxplot is constructed like this:

- The *largest* number of eggs (*excluding* the outlier of 31 eggs) is about 26: This is the line at the top of the boxplot.
- 75% of the observations (*including* the 31 eggs) are smaller than about 12, and this is represented by the line at the top of the central box. This is called the *third quartile*, or Q_3.
- 50% of the observations (*including* the 31 eggs) are smaller than about 2, and this is represented by the line in the centre of the central box. This is an 'average' value for the data, the *second quartile*, or Q_2.
- 25% of the observations (*including* the 31 eggs) are smaller than about 0.5, and this is represented by the line at the bottom of the central box. This is called the *first quartile*, or Q_1.

  Clearly we cannot have 0.5 eggs, but with 15 observations it is not possible to exactly determine the value for which 25% of observations are smaller. Software uses approximations to compute these values. (Different software may use different rules.)
- The smallest number of eggs is 0. This is the line at the bottom of the boxplot.

Example 12.13 (Boxplots explained)**.** The NHANES study collects large amounts of information from about 10,000 Americans each year (Sect. 12.10). Consider the boxplot of the age of these Americans.

The boxplot is shown in Fig. 12.22. (The online version has an animation.) The "average" age of the subjects is about 38 years, and the ages range from almost zero to about 80 years of age.

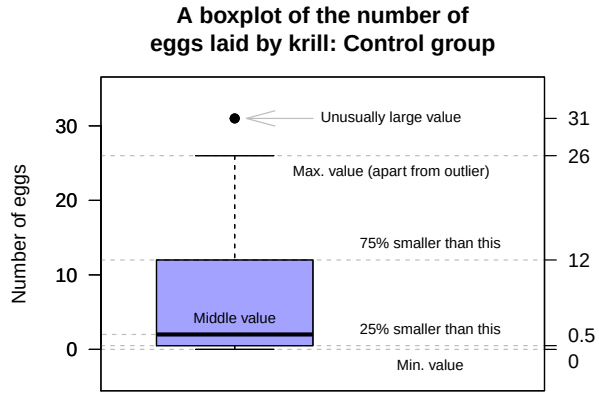


FIGURE 12.21: *A boxplot for the krill-egg data; the boxplot just for the control group*

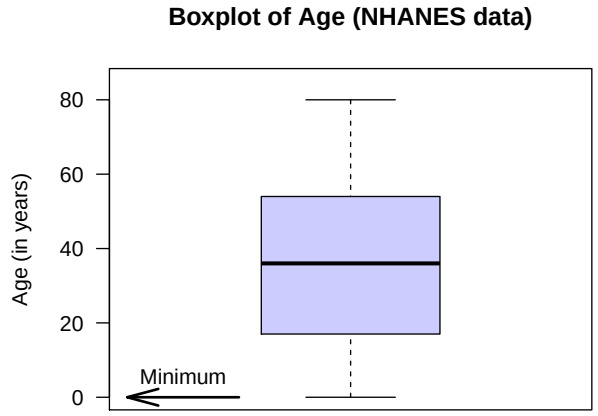


FIGURE 12.22: *The boxplot for the age of people in the NHANES data*

Example 12.14 (Boxplots)**.** Boxplots can be plotted horizontally too, which leaves room for long labels. In Fig. 12.23 (based on Silva et al. (2016)), the three cements are quite different regarding their push-out forces.

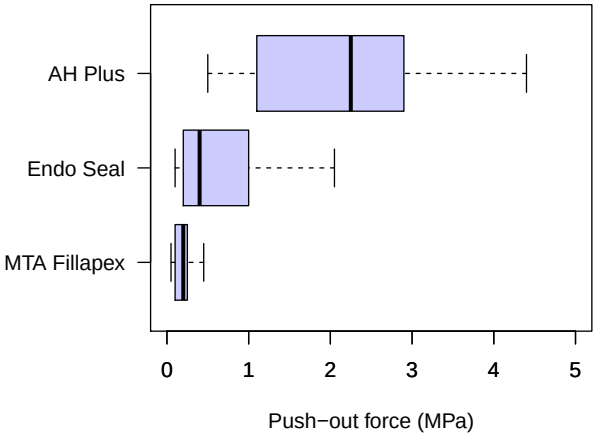


FIGURE 12.23: *Comparing three push-out values for three cements*

Example 12.15 (Boxplots)**.** A study of different engineering project delivery methods (Hale

et al. 2009 12.24

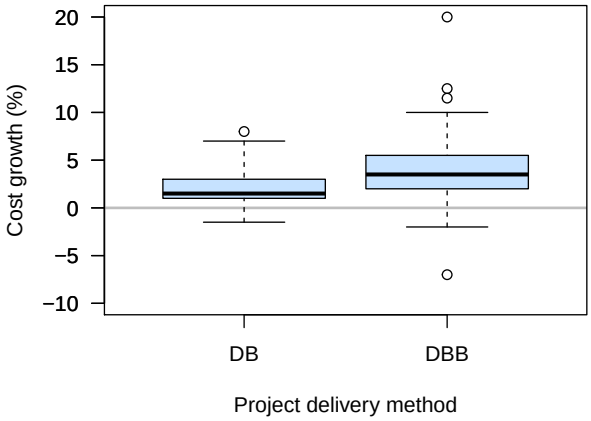


FIGURE 12.24: *Comparing two engineering project delivery methods*

12.5 Two quantitative variables

Scatterplots display the relationship between *two quantitative variables*. Conventionally, the "response" variable is on the vertical axis, and the the "explanatory" variable is on the horizontal axis.

As with any graph, explaining what the graph tells us is important because the purpose of a graph is to display the information in the clearest, simplest possible way, to help the reader understand the message(s) in the data. *Scatterplot* can be described by briefly explaining how the variables are related to each other. Scatterplots are studied again later (Sect. 12.5).

Example 12.16 (Scatterplots)**.** A study of *mdx* mice (mice with muscular dystrophy) (Laws 2005) recorded the lung weight of mice at various ages. The scatterplot in Fig 12.25 shows that the average lung weight increases, then declines after about 50 weeks of age; a lot of variation exists within mice of similar ages.

Example 12.17 (Scatterplots)**.** A study of athletes at the Australian Institute of Sport (AIS) measured numerous physical and blood measurements from high performance athletes (Telford and Cunningham 1991).

Many relationships were of interest. Fig. 12.26 shows the relationship between the sum of skin folds (SSF) and percentage body fat.

Each point represents the percentage body fat and the SSF for one athlete. Clearly, the greater

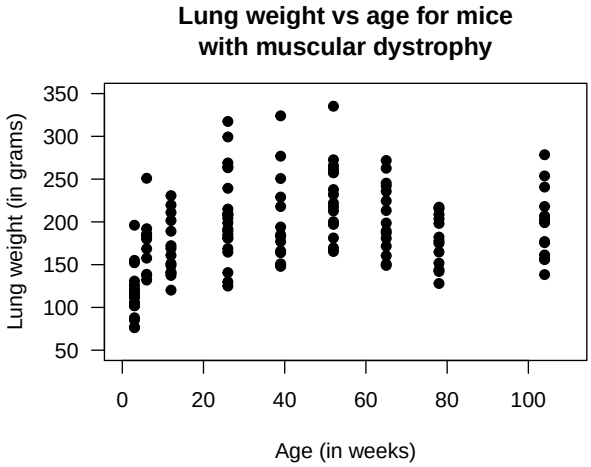


FIGURE 12.25: *Scatterplot of lung weight vs age for mice with muscular dystrophy*

the percentage body fat, the greater the sum of skin folds, in general.

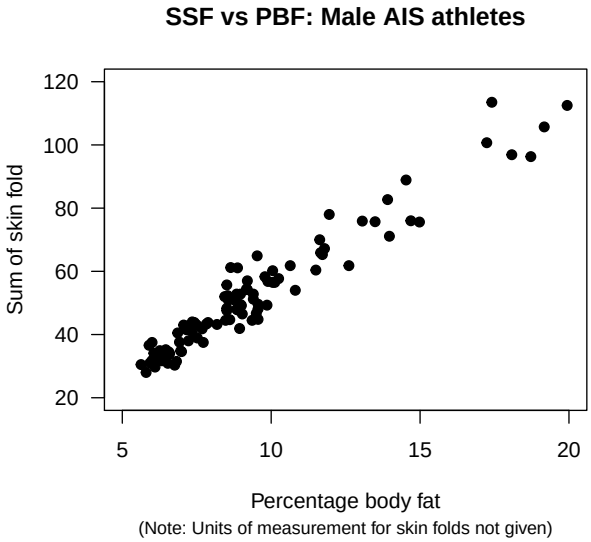


FIGURE 12.26: *Scatterplot of SSF against percentage body fat for male AIS athletes*

12.6 Two qualitative variables

The relationship between two qualitative variables can be explored using:

- **Stacked bar charts**;
- **Side-by-side bar charts**; or
- **Dot charts**.

Many variations of these graphs are possible.

As an example, a study of road kill (Russell et al. 2009) produced the data in Table 12.4. There are two qualitative variables: the season (ordinal, with four levels) and the sex (nominal, with three levels including 'Unknown').

TABLE 12.4: *The number of possums found as road kill, by sex and season*

	Unknown	M	F
Autumn	75	25	21
Winter	74	27	22
Spring	71	10	18
Summer	58	10	12

12.6.1 Stacked bar charts

The data can be graphed by using a bar for each season, *stacking* the bars by sex on top of each other, within each season (Fig. 12.27).

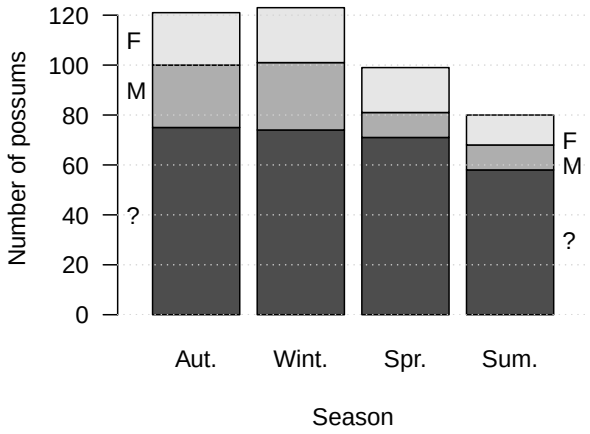


FIGURE 12.27: *The number of possums found as road kill, by sex and season*

12.6.2 Side-by-side bar charts

Instead of stacking the bars within each season on top of each other, the bars can be placed *side-by-side* within each season (Fig. 12.28).

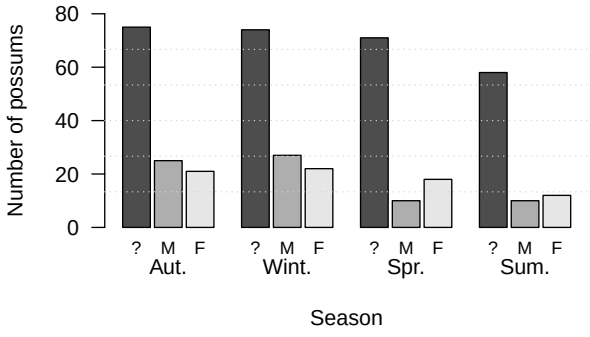


FIGURE 12.28: *The number of possums found as road kill, by sex and season*

12.6.3 Dot charts

Instead of bars, dots (or other symbols) can be used in place of a side-by-side bar chart (Fig. 12.29).

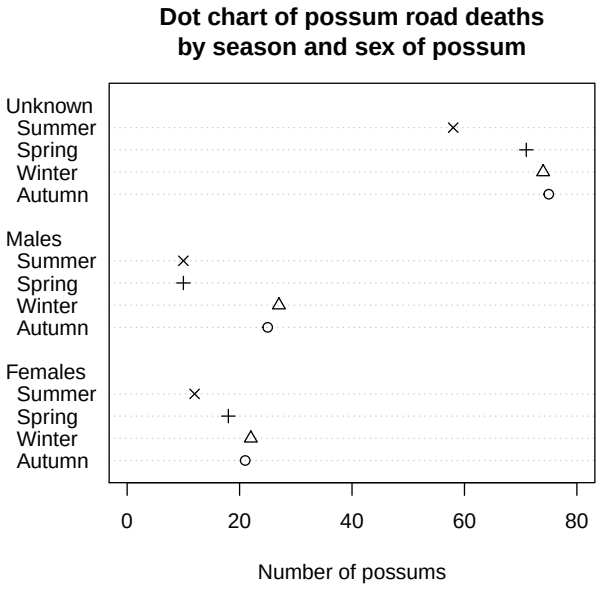


FIGURE 12.29: *A dot chart of the number of possums found as road kill, by sex and season*

12.6.4 Other variations

Many variations of these bar charts are possible. We can choose:

- horizontal or vertical bars;
- percentages or counts;
- stacked bar charts, side-by-side bar charts, or dot charts;
- either the sex of the possum or the season as the first division of the data.

Many variations exist; some are shown in Fig. 12.30. Another display is to construct a two-way table, of either counts (Table 12.4) or percentages (Table 12.5).

TABLE 12.5: *The percentages of possums found as road kill by sex, within each season (rows sum to 100%)*

	Unknown	M	F
Aut.	62.0	20.7	17.4
Wint.	60.2	22.0	17.9
Spr.	71.7	10.1	18.2
Sum.	72.5	12.5	15.0

Think 12.3 (The "best" graph)**.** *Of all these displays, which one do you think best communicates the message in the data? (Indeed, what* is *the main message that you would like to get across?)*

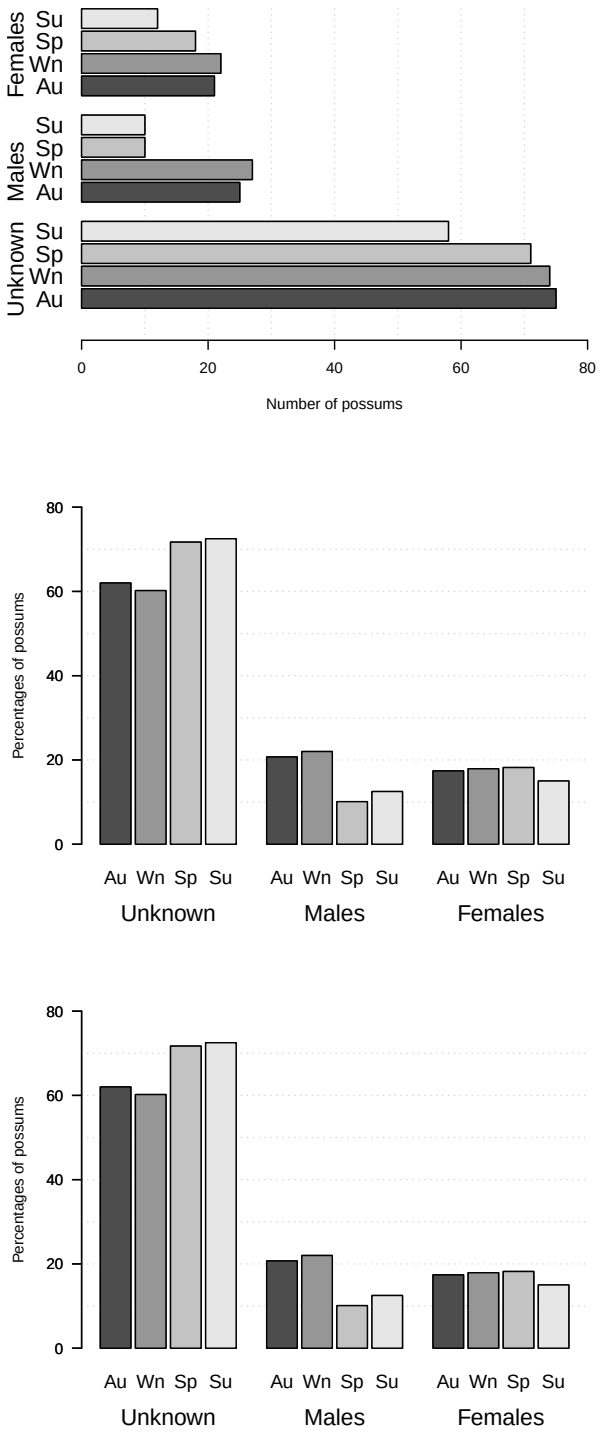


FIGURE 12.30: *Three graphs: The number or percentage of possums found as road kill, by sex and season*

12.7 Comparing 2-D and 3-D graphs

Always remember the purpose of a graph: to display the information in the *simplest* and *clearest* possible way, to help the reader understand the message(s) in the data.

Instead of aiming to communicate information, sometimes graphs are prepared to look fancy or clever, or to show off the researchers computing skills.

One way that people try to be fancy is to use a third dimension when it is not needed. This is a bad idea: the resulting graphs can be misleading and hard to read (Siegrist 1996).

Think 12.4 (Graphs and tables)**.** *In the NHANES study (Center for Disease Control and Prevention (CDC) 1988--1994), the age and sex of each participant were recorded.*

Using Fig. 12.31, can you easily determine if more females or more males are in each age group?

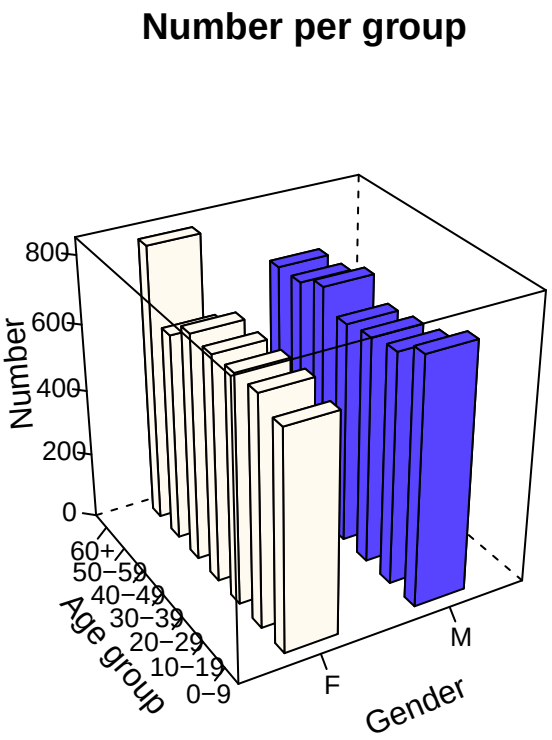


FIGURE 12.31: *A three-dimensional bar chart of the NHANES data*

The artificial third dimension makes quickly determining the heights of the bars hard. In contrast, a 2-D graph (a side-by-side bar graph; Fig. 12.31) makes it clear whether each age group has more females or more males.

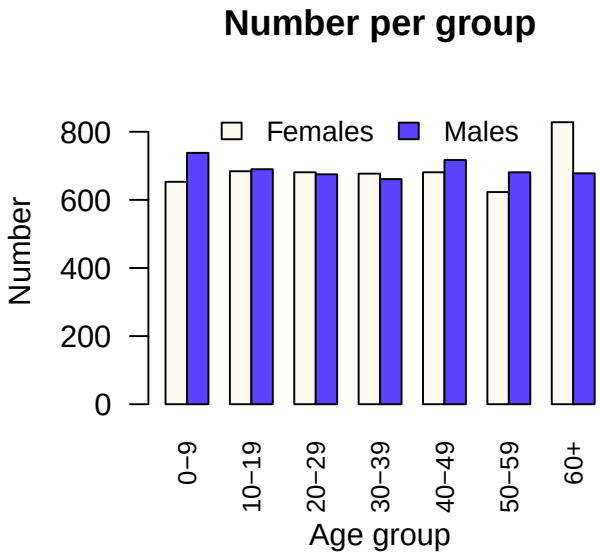


FIGURE 12.32: *A side-by-side bar chart of treatment data*

12.8 Other types of graphs

Many types of graphs have been studied, for specific types of data. But sometimes, other plots are useful. Usually the best plot is one of those just described, but sometimes the best plot is something different, perhaps even unique. Always remember the driving principle:

> The purpose of a graph is to display the information in the clearest, simplest possible way, to help the reader understand the message(s) in the data.

Importantly, a graphs needed that helps answer the research question. In this section, graphs for some other situations are discussed:

- **Geographic plots**: Useful when the RQ is about comparing geographical regions.
- **Case-profile plots**: Useful when the same units are measured over a small number of time points, or are otherwise connected.
- **Histogram of differences**: Useful when the same units are measured over *two* time points, or are otherwise connected.
- **Time plots**: Useful when the units are measured over a large number of time points.

12.8.1 Geographic plots

When data are summarised over a geographic area, plotting accordingly can be useful.

> **Example 12.18** (Geographics plots)**.** A study examining lower-limb amputation incidence in Australia (based on Dillon et al. (2017)) produced the graphs in Figs. 12.33 and 12.34.
>
> Clearly, the incidence of amputation is higher in the Northern Territory than other parts of

Australia for both females and males; furthermore, males have higher incidence of amputation than females in every state/territory.

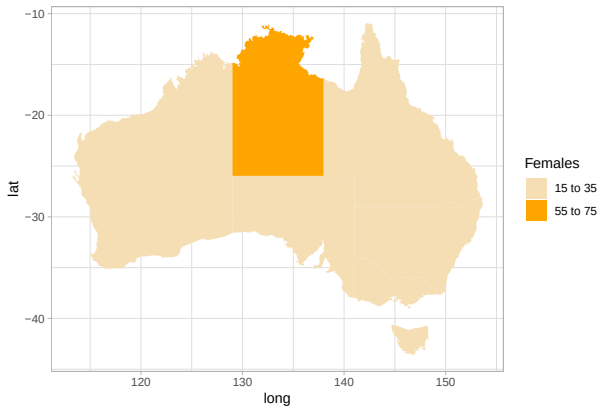


FIGURE 12.33: *Age-adjusted incidence of lower limb amputations in Australia, from August 2007 to December 2011: females. Numbers are incidents per 100 000.*

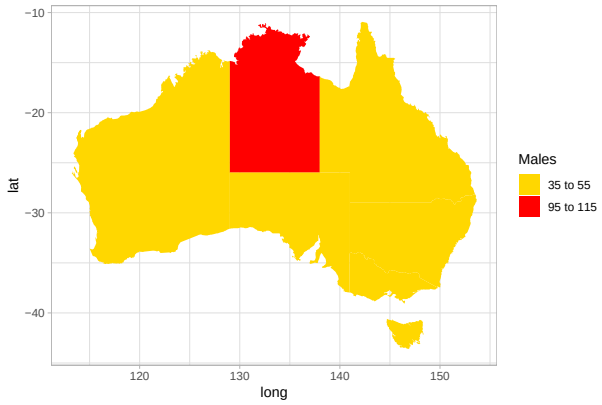


FIGURE 12.34: *Age-adjusted incidence of lower limb amputations in Australia, from August 2007 to December 2011: males. Numbers are incidents per 100 000.*

12.8.2 Case-profile plots

Sometimes the same variable is measured on each unit of analysis more than once (i.e. many observations per unit of analysis) but only a small number of times. Then a *case-profile plots* can be used: plots that show how the response variable changes for each unit of analysis. Examples of this type of data include:

- Measurements of household water consumption before and after installing water-saving devices, for many households.
- Blood pressure measurements taken from left arms and right arms, for many people.

In both cases, the data are *paired* (*two* observations per unit of analysis) as each unit of analysis gets a pair of similar measurements.

Example 12.19 (Case-profile plots)**.** A study of children with atopic asthma (Lothian et al. 2006) included the graph in Fig. 12.35. Each line on the graph represents one person.

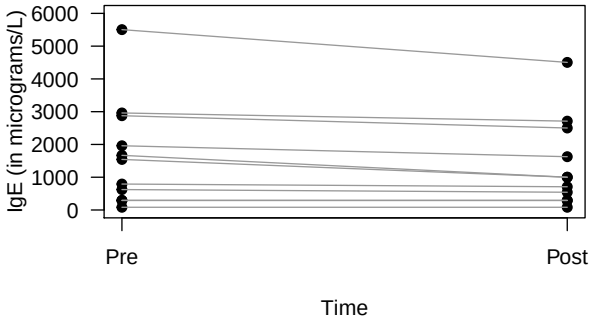


FIGURE 12.35: *A case-profile plot. Each line represents one subject, joining that person's pre-intervention score to their post-intervention score*

12.8.3 Histogram of differences

An alternative way to present paired data (two observations made for each unit of analysis) is to produce a histogram of the *changes* for each individual. This may be easier to produce in software than a case-profile plot.

Consider the person in the case-profile plot whose line is at the top of the plot in Fig. 12.35. Their 'pre' IgE level is about 5500 micrograms/L, and their 'post' IgE level is about 4500 micrograms/L, which is a *change* (or more descriptively, a *reduction*) of about 1000 micrograms/L. These reductions could be computed for each person (Table 12.6).

TABLE 12.6: *The IgE before and after an intervention, and the change in IgE (in micrograms/L)*

IgE (before)	IgE (after)	IgE reduction
83	83	0
292	292	0
293	292	1
623	542	81
792	709	83
1543	1000	543
1668	1000	668
1960	1626	334
2877	2502	375
2961	2711	250
5504	4504	1000

Then a histogram can be constructed based on these *reductions* in IgE, with one reduction for each person (Fig. 12.36).

Example 12.20 (Graphing paired data)**.** The Electricity Council in Bristol wanted to determine

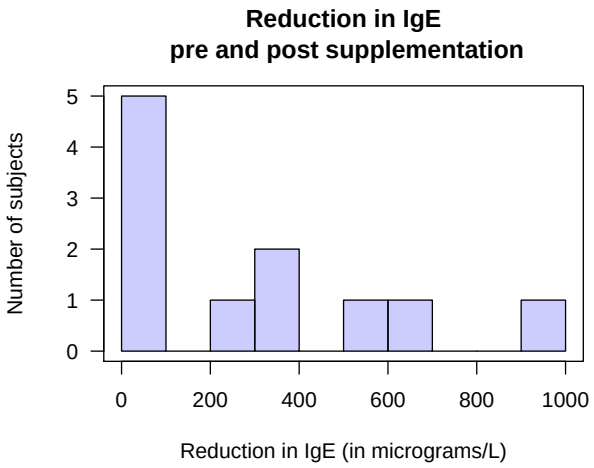


FIGURE 12.36: *An alternative to a case-profile plot: A histogram of the differences*

if a certain type of wall-cavity insulation was effective in reducing energy consumption in winter (The Open University 1983). Their RQ was:

> Is there an *average reduction* in energy consumption due to adding insulation?

The data collected, shown below, can be graphed using a case-profile plot (Fig. 12.37, top panel): A dashed line has been used to show an *increase* in energy usage, and a solid line for houses where energy was *saved* after installing insulation. (Again, this is *encoding* extra information.)

TABLE 12.7: *The house insulation data: Energy consumption before and after adding insulation, and the energy saving (all in MWh)*

Before	After	Energy savings
12.1	12.0	0.1
11.0	10.6	0.4
14.1	13.4	0.7
13.8	11.2	2.6
15.5	15.3	0.2
12.2	13.6	-1.4
12.8	12.6	0.2
9.9	8.8	1.1
10.8	9.6	1.2
12.7	12.4	0.3

For these data, finding the difference in energy consumption for each house seems sensible. The same unit of analysis is measured twice on the same variable: energy consumption *before* adding insulation and then *after* adding insulation. The difference in energy consumption (or the energy *saving* more specifically) for each house can be computed, then graphed using a histogram, bar chart, stemplot, or dot chart (Fig. 12.37, bottom panel).

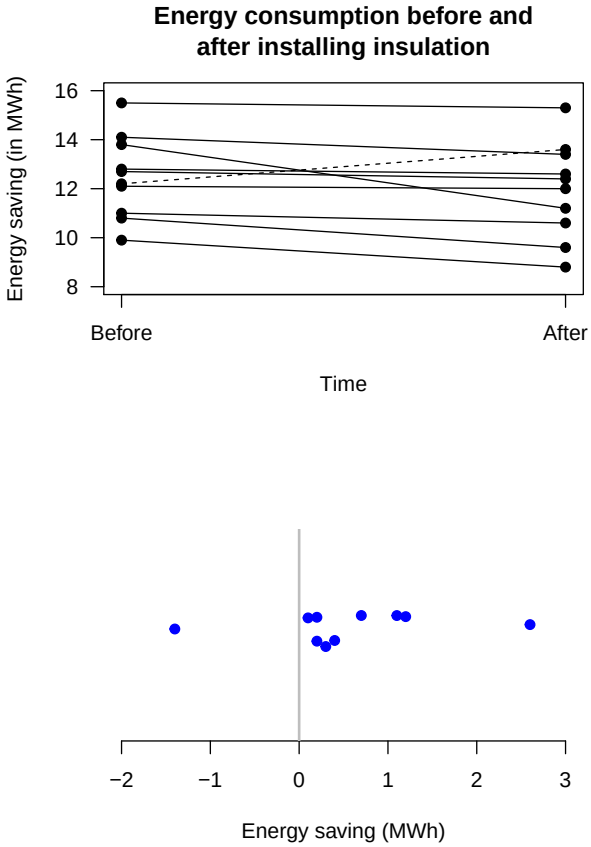


FIGURE 12.37: *Two plots of the energy consumption data. The dotted line in the top panel shows the one home where energy consumption increased.*

Example 12.21 (Graphing paired data)**.** A study (Dawson et al. 2017) examined the average number of bacteria on birthday cakes *before* and *after* blowing out the candles.

This question could be studied by taking two measurements from each cake: before and after blowing out candles. The *change* in the number of bacteria could be computed for each cake, and a histogram of the differences plotted.

12.8.4 Time plots

Sometimes, a variable is measured over many time points. A *time plot* can be used, which show how the variable changes over time.

Example 12.22 (Time plots)**.** The baby-birth data (in Sect. 12.2.1) recorded the time of each birth. A time plot shows how the weights varied over time (Fig. 12.38).

Example 12.23 (Time plots)**.** A study of the number of lynx trapped in the Mackenzie River area of Canada (Elton and Nicholson 1942) each year from 1821 to 1934 produced the data shown in Fig. 12.39. A regular cycle is apparent, where the number trapped is very large.

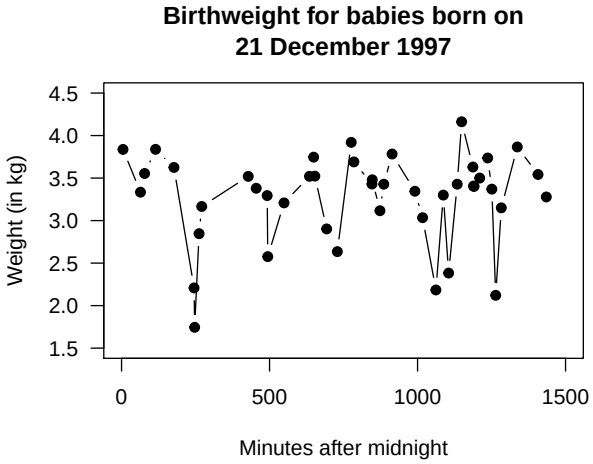


FIGURE 12.38: *The weight of babies born on 21 December 1997 at a Brisbane hospital, plotted over time*

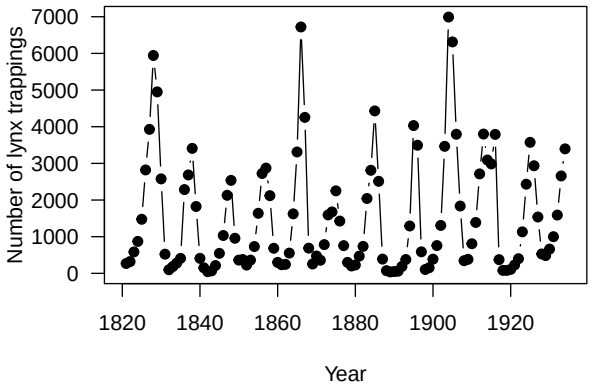


FIGURE 12.39: *The number of lynx trapped in the Mackenzie River district of north-west Canada, from 1821 to 1934*

12.9 Notes on constructing graphs

Always remember:

> ⚠ The purpose of a graph is to display the information in the clearest, simplest possible way, to help the reader understand the message(s) in the data.

Helping readers to understand the data is the essense of producing a good graph:

> Data graphics should draw the viewer's attention to the sense and substance of the data, not to something else [. . .] essentially statistical graphics are instruments to help people reason about quantitative information.
> — Tufte et al. (1998), p. 91

You should be able to construct the graphics by hand, but we will generally use software (e.g.,

jamovi or SPSS). Importantly, given the purpose of a graph, what the graph communicates should be explained: Graphs need to be clear and well-labelled with captions.

Ensure that you:

- **Do not** add artificial third dimensions, or other 'chart junk' (Su 2008).
- **Do not** add optical illusions.
- **Do not** make any errors.

Ensure that you:

- **Do** add units of measurement or value labels where appropriate.
- **Do** add titles and axis labels.
- **Do** ensure the axis scales are appropriate.
- **Do** add any necessary explanations.
- **Do** make it easy for the reader to be able to consider the RQ (for example, to easily make the comparison of greatest interest).

Example 12.24 (Truncating bar charts)**.** One optical illusion often appearing in graphs is when the frequency (or percentage) axis on a bar chart is truncated so that it doesn't start at zero. For example, consider data recording the number of lung cancer cases in various Danish cities (Andersen 1977).

Figure 12.40 shows the original bar chart with the count (vertical) axis starting at zero; the counts in each age group look very similar. In contrast, if the vertical axis starts at 9.75, the counts look very different (Fig. 12.41) for two age categories, suggesting large difference between the number of lung cancer cases. (The online version has an animation.) The graph is misleading when the graph does not start at a count of zero.

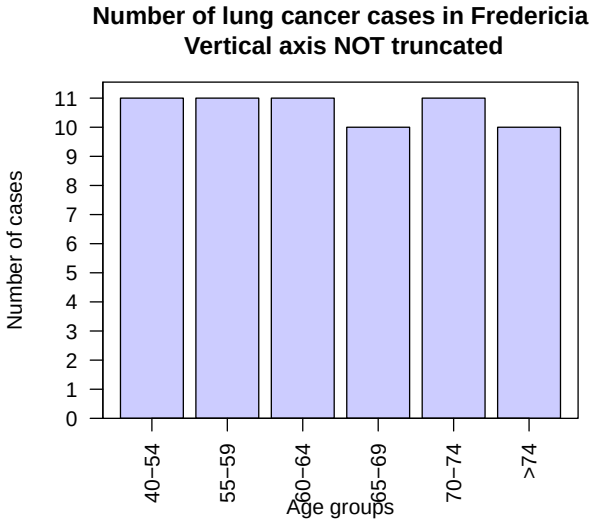


FIGURE 12.40: *The bar chart, without truncating the vertical axis*

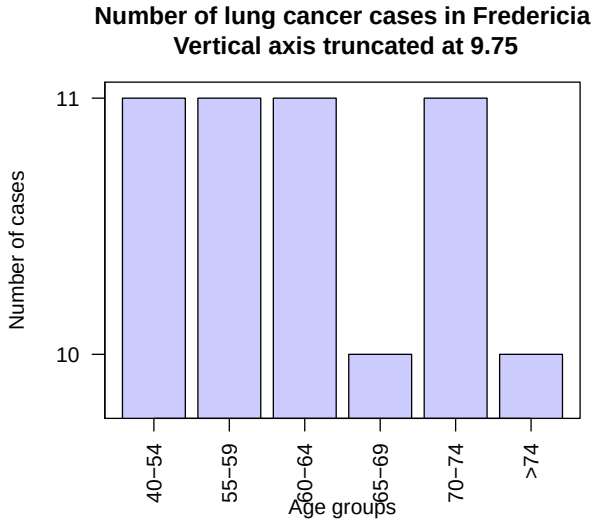


FIGURE 12.41: *The bar chart, after truncating the vertical axis*

12.10 Case Study: The NHANES data

To demonstrate how graphs can help answer RQs, consider the following RQ:

> Among Americans, is the average direct HDL cholesterol different for those who current smokers and non-smokers?

From the RQ, the Population is 'Americans,' the Outcome is the 'average direct HDL cholesterol levels,' and the Comparison is 'Between those who currently do and do not smoke.' There is no intervention; this is a *relational RQ*, that can be answered using an *observational study*.

> **Think 12.5** (Confounding variables)**.** *As with any study, managing confounding should be considered, by thinking about possible extraneous variables that could be measured or observed.*
>
> *Can you think of any possible extraneous variables that have the potential to be confounding variables?*

In this study, clearly we cannot collect primary data. However, data to answer this (and many other) RQs are obtained from the American *National Health and Nutrition Examination Survey* (NHANES) (Center for Disease Control and Prevention (CDC) 1988--1994; Center for Disease Control and Prevention 1996; Pruim 2015). From the NHANES webpage[2], this survey:

> . . . examines a nationally *representative* sample of about 5,000 persons each year. . . located in counties across the country, 15 of which are visited each year.
> — NHANES webpage (emphasis added)

[2] http://www.cdc.gov/nchs/nhanes/about_nhanes.htm

The data available are equivalent to a "a simple random sample from the American population" (Pruim 2015). In total, 10,000 observations on scores of variables are available (from the 2009/2010 and the 2011/2012 surveys).

For any RQ, exploring and understanding the data is important, and using graphs is a great way to do so (especially with large data sets). Begin by graphing both the response and explanatory variables involved in the RQ.

For the NHANES data, the *response variable* is direct HDL cholesterol (quantitative continuous). The histogram (Fig. 12.42) shows that about 2200 people had a direct HDL cholesterol concentration between 1.25 and 1.50 mmol/L; and about 1200 had a concentration between 0.75 and 1,00 mmol/L. In general, the direct HDL cholesterol is usually around 1.5mmol/L, and varies between about 0.5 and 3 mmol/L. The data are slightly skewed right.

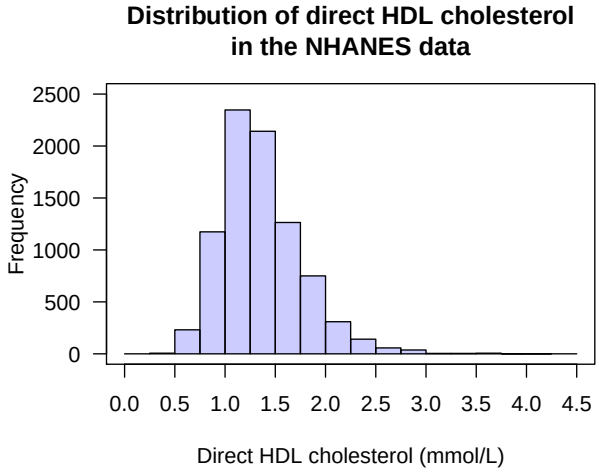
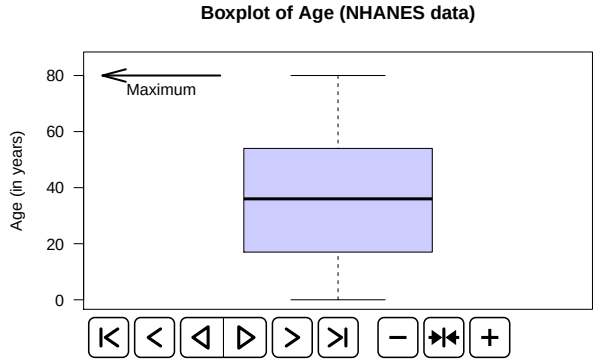


FIGURE 12.42: *A histogram of the direct HDL cholesterol in the NHANES study*

The *explanatory variable* is 'current smoking status,' which could be graphed using a bar chart (Fig. 12.43) or pie chart, for example. But only *one* piece of information is present, so a graph probably isn't necessary): That 45.7% (just under half) of the respondents currently smoke.

The NHANES data contains 10,000 respondents, yet the bar chart clearly does not have 10,000 responses. Many respondents did not answer this question. A bar could be added to the bar chart to show the number of non-responses (though it probably isn't necessary).

The main RQ involves the *relationship* between average direct HDL cholesterol and current smoking status, so a graph displaying this relationship is needed, such as a boxplot (Fig. 12.44). From the plot, is there a difference in the average HDL cholesterol concentrations between the two smoking groups?



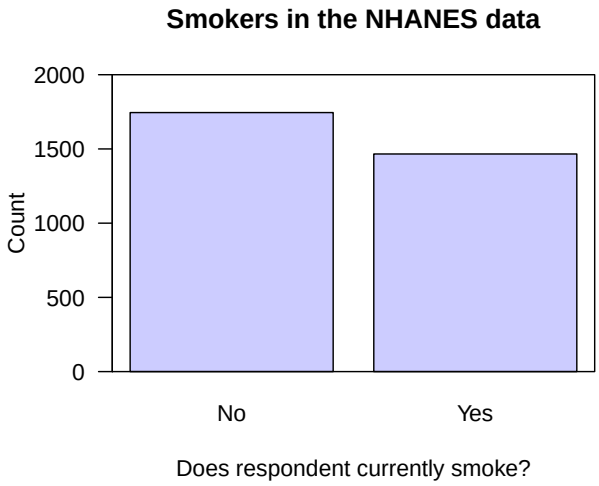


FIGURE 12.43: *A bar chart of current smoking status for the NHANES data. No response was recorded for many subjects*

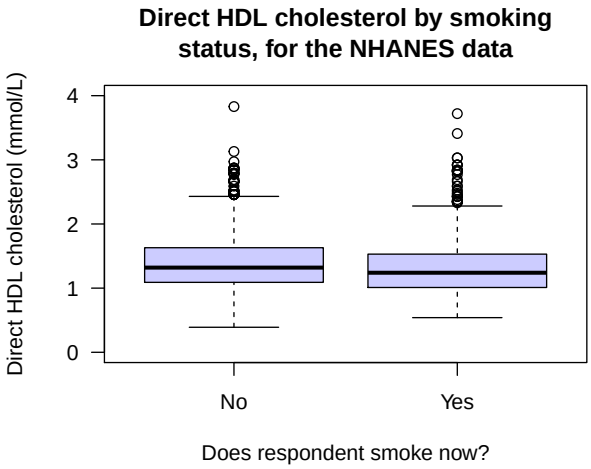


FIGURE 12.44: *Boxplot of direct HDL cholesterol against current smoking status, for the NHANES data*

The NHANES study is an observational study, where confounding is a potential problem (Sect. 6.3). To examine potential confounding, exploring the relationships between the response and extraneous variables, and between the explanatory and extraneous variables, is useful. Some useful plots are shown in Fig. 12.45; what do they suggest?

Think 12.6 (Confounding)**.** *What do these graphs suggest about possible confounding relationships?*

12.11 Summary

To summarise one variable, these graphs can be used:

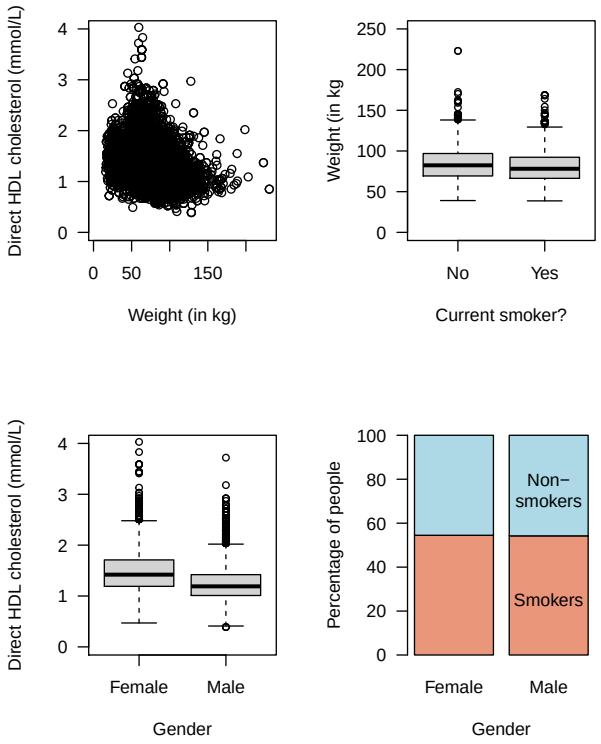


FIGURE 12.45: *Some plots from the NHANES data*

- For *one quantitative variable*: Histogram; stem-and-leaf; dot chart.
- For *one qualitative variable*: Bar chart; pie chart; dot chart.

For two variables, the graphs shown in Table 12.8 can be used. In general, the 'response' variable is on the vertical axis.

TABLE 12.8: *Graphs to use for most situations*

	Explanatory: Qualitative	**Explanatory: Quantitative**
Response: Qualitative	Stacked bar chart; side-by-side bar chart	Boxplot; back-to-back stem; 2-D dot chart
Response: Quantitative	Boxplot; back-to-back stem; s2-D dot chart	Scatterplot

12.12 Quick review questions

A study on the bruising of apples (Doosti-Irani et al. 2016) purposefully hit apples with three different forces (200, 700 and 1200 mJ), in three different regions of the apple (lower; middle; upper).

The researchers then recorded the depth of the bruising (the *response* variable), to determine if it could be estimated from the recorded the surface temperature (the *explanatory* variable) at the bruise location (an *extraneous* variable).

1. Which one of the following could be used to graph the *response* variable?

2. Which one of the following could be used to graph the *explanatory* variable?
3. Which one of the following could be used to graph the *relationship* between the response and explanatory variables?
4. Which one of the following could be used to graph the *relationship* between the 'surface temperature' and 'bruise location?'

12.13 Exercises

Selected answers are available in Sect. D.12.

Exercise 12.1. A study (Henderson and Velleman 1981) recorded the number of cylinders in many models of cars (Table 12.9). The *number* of cylinders is quantitative discrete, but with so few different values, this variable could be plotted with some of the graphs used for graphing qualitative data.

For these data:

1. Produce a dot chart.
2. Produce a histogram.
3. Produce a bar chart.
4. Produce a pie chart.

What graph do you think is best? Why?

TABLE 12.9: *The number of cylinders in cars in a study*

Number of cylinders	Number of cars
4	11
6	7
8	14

Exercise 12.2. A study of lime trees (*Tilia cordata*) recorded these variables for 385 lime trees in Russia (Schepaschenko et al. 2017; Dunn and Smyth 2018):

- the foliage biomass, in kg;
- the tree diameter (in cm);
- the age of the tree (in years); and
- the origin of the tree (one of Coppice, Natural, or Planted).

The purpose of the study is to estimate the foliage biomass from the other variables. What graphs would be useful?

Exercise 12.3. In a study of the influence of using ankle-foot orthoses in children with cerebral palsy (Swinnen et al. 2017), the data in Table 11.2 describe the 15 subjects. (GMFCS is an *ordinal* variable used to describe the impact of cerebral palsy on their motor function: the

Gross Motor Function Classification System[3].) Sketch some graphs to explore the *relationships* between these variables.

Exercise 12.4. A study of fertilizer use (Lane 2002; Dunn and Smyth 2018) recorded the soil nitrogen after applying different fertilizer doses. These variables were recorded:

- the fertilizer dose, in kilograms of nitrogen per hectare;
- the soil nitrogen, in kilograms of nitrogen per hectare; and
- the fertilizer source; one of 'inorganic' or 'organic.'

What graphs would be useful for understanding the data?

Exercise 12.5. A survey of voice assistants[4] (e.g., Amazon Echo; Google Home; etc.) conducted by Nielsen[5] asked respondents to indicate how they used their voice assistant; options given were:

- Listening to music;
- Search for real-time info (e.g., traffic; weather);
- Search for factual info (e.g., trivia; history);
- Listen to news;
- Chat with voice assistant for fun;
- Use alarms, timer.

What would be the best graph for displaying respondents answers? Would a pie chart be suitable? Explain your answer.

Exercise 12.6. A study of athletes at the Australian Institute of Sport (AIS) measured numerous physical and blood measurements from high performance athletes (Telford and Cunningham 1991). The graph in Fig. 12.46 compares the heights of females in two similar sports[6]: basketball and netball. How would you describe the heights of the athletes in the two sports?

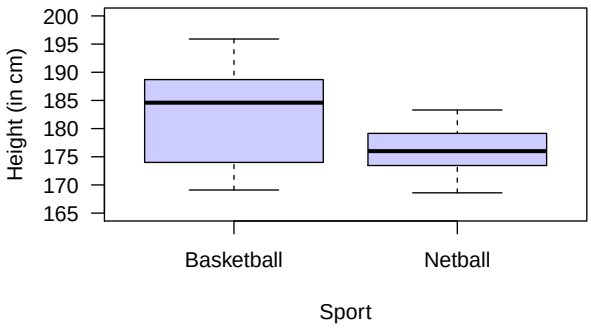


FIGURE 12.46: *The heights of female basketball and netball players attending the AIS*

Exercise 12.7. A study of noisy miners (a small Australian bird) counted the number of noisy miners and the number of eucalyptus trees in random quadrats (Maron 2007; Dunn and Smyth 2018). Critique the graph of the data (Fig. 12.47).

[3] https://en.wikipedia.org/wiki/Gross_Motor_Function_Classification_System

[4] https://www.nielsen.com/us/en/insights/news/2018/smart-speaking-my-language-despite-their-vast-capabilities-smart-speakers-all-about-the-music.print.html

[5] https://www.nielsen.com/au/en.html

[6] Netball was derived from basketball: https://en.wikipedia.org/wiki/Netball#History

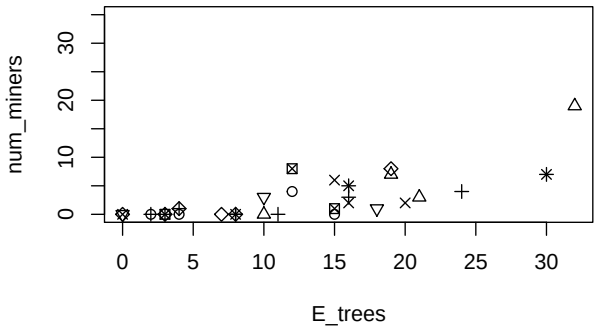


FIGURE 12.47: *The number of noisy miners and the number of eucalyptus trees*

Exercise 12.8. A study of 173 female horseshoe crabs (Brockmann 1996; Dunn and Smyth 2018) recorded, among other things, the colour of the carapace (one of 'Light medium,' 'Medium,' 'Dark medium' or 'Dark') and the condition of the carapace (one of 'Both OK,' 'One OK,' 'None OK'). Critique the scatterplot (Fig. 12.48) used to explore the data.

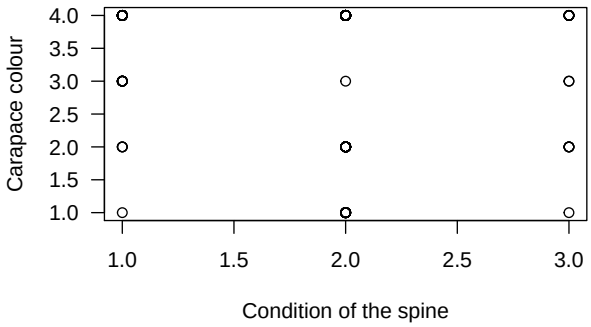


FIGURE 12.48: *A scatterplot of the colour of female horseshoe crabs and the condition of their spines. There are no missing values.*

Exercise 12.9. A study (Danielsson et al. 2014) examined the change in MADRS (a *quantitative* scale measuring level of depression) and treatment group (whether each person was treated using: exercise; body awareness; or advice).

1. What is the response variable?
2. What is the explanatory variable?
3. What graphs would be useful for exploring the data and the relationships of interest?

Exercise 12.10. In a study of the temperature in offices, Paul and Taylor (2008) compared the temperature in three offices (during working hours) at Charles Sturt University (Australia); the data are summarised in Table 12.10. Using this information, draw the boxplot comparing the three offices. What do we learn from this graph?

Exercise 12.11. A study of high-performance athletes at the Australian Institute of Sport (AIS) (Telford and Cunningham 1991) recorded numerous variables about athletes. A plot for the sports played by the athletes is shown in Fig. 12.49. How would you describe the data: Left skewed, right skewed, approximately symmetrical? Or something else?

TABLE 12.10: *A summary of the temperature (in degrees C) in three offices at CSU during working hours according to current smoking status*

	Office A	**Office B**	**Office C**
Mean	24.1	25.3	25.7
Minimum	16.4	15.9	20.1
Q_1	22.8	23.8	24.6
Median	24.4	25.5	26.1
Q_3	25.5	26.9	27.2
Maximum	27.4	31.0	30.3

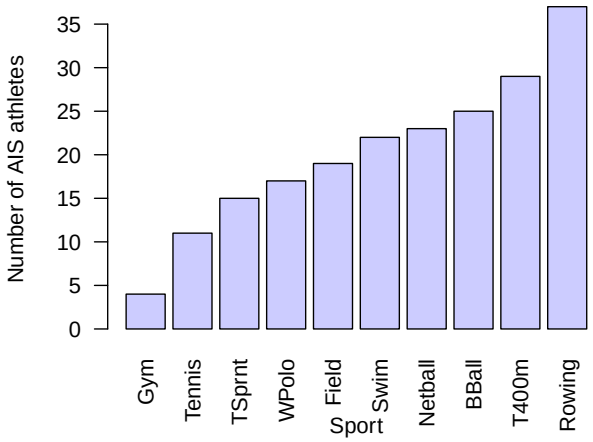


FIGURE 12.49: *Sports played by athletes in the AIS study*

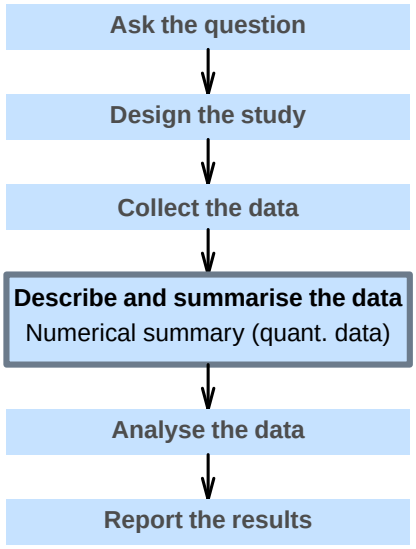
13

Numerical summaries: quantitative data

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, and graphically summarise data.

In this chapter, you will learn to numerically describe *quantitative* data. Both quantitative and qualitative *data* are described numerically in *quantitative research*. You will learn to:

- numerically summarise quantitative data using the appropriate statistics.
- describe quantitative data by average, variation, shape and unusual features.



13.1 Introduction

In the last chapter (Sect. 12.10), this RQ was posed:

> Among Americans, is the average direct HDL cholesterol different for current smokers and non-smokers?

Graphs were used to understand the data in Sect. 12.10, where information contained in the graphs was given. In some cases, the features of the data displayed in the graph can be described *numerically*. That is the purpose of this chapter: to learn how to summarise *quantitative* data numerically.

Example 13.1 (Describing quantitative data)**.** For the RQ above, understanding the response variable (direct HDL cholesterol values) is important; a histogram is useful (Fig. 13.1).

What does the histogram tell us?

- **Average**: The average value is about 1.5 mmol/L.
- **Variation**: The values range from about 0.5 to 3 mmol/L, but with some larger values (that are hard to see on the histogram).
- **Shape**: The distribution is slightly skewed right.
- **Outliers**: Some large outliers are present (that are hard to see on the histogram).

Describing some of these features more precisely, with *numbers*, can be helpful.

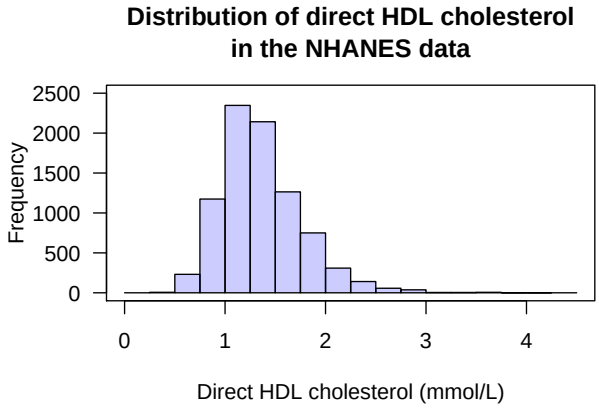


FIGURE 13.1: *The histogram of the direct HDL cholesterol from the NHANES study*

A number that describes a feature of a *population* is called a *parameter*. The values of parameters are usually unknown.

In contrast, a number that describes a feature of a *sample* is called a *statistic*. That is:

- **S**amples are numerically described by **s**tatistics;
- **P**opulations are numerically described by **p**arameters.

Definition 13.1 (Parameter)**.** A **parameter** is a number describing some feature of a population.

Definition 13.2 (Statistic)**.** A **statistic** is a number describing some feature of a sample (to estimate a population *parameter*).

⚠ The RQ identifies the population, but in practice a sample is studied. *Statistics* are estimates of *parameters*, and the value of the *statistic* is not the same for every possible *sample*.

13.2 Computing the average value

The average (or *location*, or *centre*, or *typical value*) for *quantitative sample data* can be described in many ways; the two most common ways are:

- the **sample mean** (or *sample arithmetic mean*), which estimate the population mean; and
- the **sample median**, which estimates the population median.

In both cases, the population parameter is *estimated* by a sample statistic. Understanding whether to use the mean or median is important.



The word 'average' can refer to either mean or median (or other measures of centre too). Use the precise terms 'mean' or 'median,' rather than 'average,' when necessary!

Think 13.1 (Difference between averages)**.** *Consider the daily river flow volume (called 'streamflow') at the Mary River from 01 October 1959 to 17 January 2019, summarised by month in Table 13.1 (from Queensland DNRM[1]).*

The 'average' daily streamflow in February could be quoted using either the mean or the median; but the two give very *different values for the 'average':*

- *the* mean *daily flow is 1123.2ML.*
- *the* median *daily flow is 146.1ML.*

These two common ways of measuring the same thing (the 'average' daily streamflow in February) give very different answers. Why? Which is the best 'average' to use? To decide, both measures of average will need to be studied.

TABLE 13.1: *The daily streamflow at Mary River (Bellbird Creek), in ML, from 01 October 1959 to 17 January 2019; average for each month*

Month	Mean	Median
Jan	849.3	71.3
Feb	1123.2	146.1
Mar	793.9	194.9
Apr	622.5	141.7
May	348.4	118.4
Jun	378.7	83.6
Jul	259.3	68.8
Aug	108.6	55.5
Sep	100.9	48.0
Oct	151.2	37.6
Nov	186.6	45.3
Dec	330.8	64.1

13.2.1 Computing the average: The mean

The mean of the population is denoted by μ, and its value is almost always unknown.

Instead, the mean of the population is *estimated* by the mean of the sample, which is denoted by $\bar{x}$ (an x with a line above it). In this context, the unknown *parameter* is μ, and the *statistic* is $\bar{x}$. The sample mean is used to *estimate* the population mean.



The Greek letter μ is pronounced 'myoo,' as in **mu**sic.

The symbol $\bar{x}$ is pronounced 'ex-bar.'

Example 13.2 (A small data set to work with)**.** To demonstrate ideas, consider a small data set for answering this descriptive RQ:

> For mature Jersey cows, what is the average percentage butterfat in their milk?

The *population* is 'milk from Jersey cows,' and an estimate of the population mean percentage butterfat is sought. The population mean is denoted by μ.

Clearly, milk from every Jersey cow cannot be studied; a *sample* is studied (Sokal and Rohlf 1995; Hand et al. 1996): The unknown population mean is estimated using the sample mean ($\bar{x}$). Measurements were taken from milk from 10 cows, in percentages (Table 13.2).

TABLE 13.2: *The butterfat percentage from a sample of milk from 10 Jersey cows*

Butterfat percentages				
4.8	5.2	5.2	5.4	5.2
6.5	4.5	5.7	4.8	5.2

The *sample mean* is what people usually think of as the 'average.' The sample mean is actually the 'balance point' of the observations (Figure 13.2). (The online version has an animation.) Alternatively, the mean is the value such that the positive and negative distances of the observations from the mean add to zero (Fig. 13.3; again, the online version has an animation.) Both of these explanations seem reasonable for identifying the 'average' of the data.

Definition 13.3 (Mean)**.** The **mean** is one way to measure the 'average' value of quantitative data. The *arithmetic mean* can be considered as the 'balance point' of the data, or the value such that the positive and negative distances from the mean add to zero.

To find the *value* of the sample mean:

- *Add* (shown using the symbol $\sum$) all the observations (denoted by x); then
- *Divide* by the number of observations (denoted by n).

In symbols:

$$\bar{x} = \frac{\sum x}{n}.$$

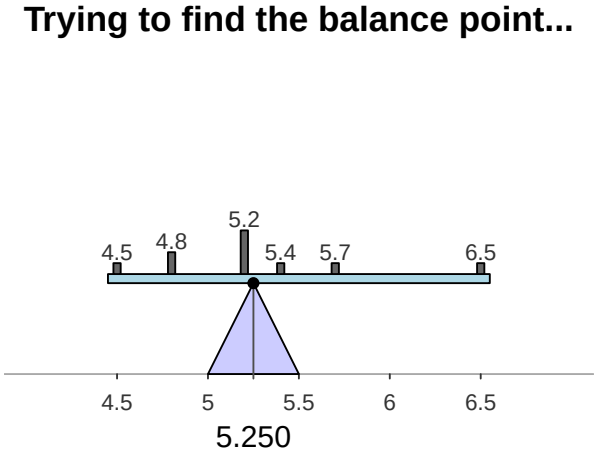


FIGURE 13.2: *The mean is the balance point of the data*

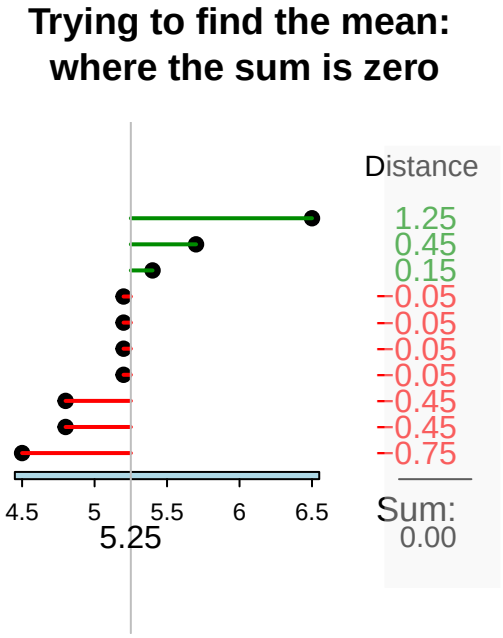


FIGURE 13.3: *The mean is the value such that the positive and negative distances sum to zero*

This means to add up (indicated by $\sum$) the observations (denoted by x), then divide by the size of the sample (denoted by n).

Example 13.3 (Computing a sample mean)**.** For data for the Jersey cow data (Example 13.2), an estimate of the population mean percentage butterfat is found using the sample information: sum all $n = 10$ observations and divide by n:

$$\begin{aligned}\overline{x} &= \frac{\sum x}{n} = \frac{4.8 + 6.5 + \cdots + 5.2}{10} \\ &= \frac{52.5}{10} = 5.25.\end{aligned}$$

The sample mean, the best estimate of the population mean, is 5.25 *percent*.

Usually, software (such as jamovi or SPSS) or a calculator (in *Statistics Mode*) will be used to compute the sample mean. However, knowing *how* these quantities are computed is important.

Think 13.2 (Mean)**.** *For the butterfat data (Table 13.2), what is the value of μ, the* population *mean?*

Answer: We do not know! We know the value of the *sample* mean, but not the *population* mean. We only have an *estimate* of the value of the population mean.

Think 13.3 (Estimating a mean)**.** *A study of eyes (Ehlers 1970) aimed to estimate the average thickness of eyes affected by glaucoma. The collected data (in microns) are shown in Table 13.3. Estimate the population mean corneal thickness.*

TABLE 13.3: *The thickness of the cornea (in microns) in eyes affected by glaucoma*

Corneal thickness			
484	492	436	464
478	444	398	476

Software and calculators often produce numerical answers to many decimal places, some of which may not be meaningful or useful. A useful rule-of-thumb is to round to one or two more significant figures than the original data.

For example, the butterfat data are given to one decimal place. The *sample mean* weight can be given to two decimal places: $\bar{x} = 5.25\%$.

13.2.2 Computing the average: The median

The median is a value separating the larger half of the data from the smaller half of the data. In a data set with n values, the median is *ordered* observation number $\frac{n+1}{2}$. The median is:

- **not** equal to $\frac{n+1}{2}$.
- **not** halfway between the minimum and maximum values in the data.

Most calculators cannot find the median.

The median has no commonly-used symbol.

Definition 13.4 (Median)**.** The **median** is one way to measure the 'average' value of some data. The *median* is a value such that half the values are larger than the median, and half the values are smaller than the median.

Example 13.4 (Find a sample median)**.** To find the sample median for the Jersey cow data (Example 13.2), first arrange the data *in numerical order* (Table 13.4). The median separates the larger 5 numbers from the smaller 5 numbers. With $n = 10$ observations, the median is the ordered observation located between the fifth and sixth observations (i.e., at position $(10 + 1)/2 = 5.5$; the *median itself is not 5.5*). So the sample median is between 5.2 (ordered observation five) and 5.2 (ordered observation six): the sample median is 5.20 percent.

TABLE 13.4: *The butterfat percentage from a sample of milk from 10 Jersey cows, in increasing order*

Butterfat percentages				
4.5	4.8	4.8	5.2	5.2
5.2	5.2	5.4	5.7	6.5

Think 13.4 (Median)**.** *For the butterfat data (Table 13.2), what is the* population *median?*

Answer: We do not know! We know the value of the *sample* median, but not the *population* median. We only have an *estimate* of the value of the population median.

Think 13.5 (Medians)**.** *A study of eyes (Ehlers 1970) aimed to estimate the average thickness of eyes affected by glaucoma.*

Using the collected data (Table 13.3), estimate the population median *corneal thickness. What is the population median?*

Answer: With $n = 8$ observations, the median is ordered observation number $(8 + 1)/2 = 4.5$, halfway between ordered observation numbers 4 and 5. After sorting into increasing order, the two middle numbers (the 4th and 5th) are 464 and 476. The median could be *any* number between 464 and 476, but the usual answer would be that the median is $(464 + 476)/2 = 470$.

The *sample* median is 470 microns; the value of the *population* median remains unknown.

To clarify:

- If the sample size n is *odd*, the median is the middle number when the observations are ordered.
- If the sample size n is *even* (such as in Think 13.5), the median is halfway between the two middle numbers, when the observations are ordered.

Some software uses different rules when n is even.

13.2.3 Which average to use?

Consider again estimating the average daily streamflow at the Mary River (Bellbird Creek) during February (Table 13.1): The *mean* daily streamflow is 1123.2ML, and the *median* daily streamflow is 146.1ML. Which is the 'best' average to use?

A dot chart of the daily stream flow (Fig. 13.4) shows that the data are *very* highly right-skewed, with many *very* large outliers: the maximum value is 156586.4ML, more than one hundred

times larger than the mean of 1123.2ML). In fact, about 86% of the observations are *less* than the mean. In contrast, about 50% the values are less than the median (by definition). For these data, the mean is hardly a *central* value…

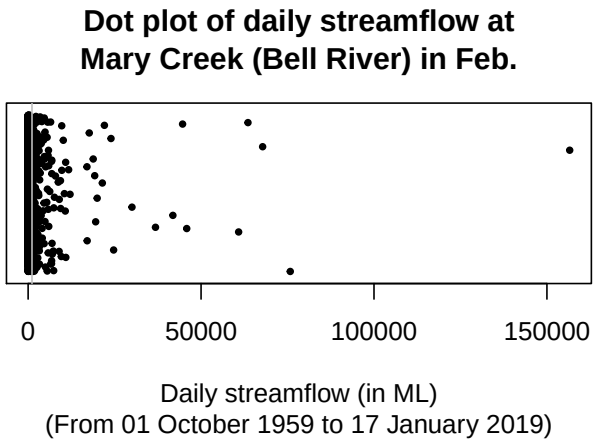


FIGURE 13.4: *A dot plot of the daily streamflow at Mary River from 1960 to 2017, for February. The vertical grey line is the mean value. Many large outliers exist, so the data near zero are all squashed together*

The streamflow data are very highly skewed (to the right), which is important and relevant:

- *Means* are best used for approximately symmetric data: the mean is influenced by outliers and skewness.
- *Medians* are best used for data that are skewed or contain outliers: the median is not influenced by outliers and skewness.

Means tend to be too large if the data contains large outliers or severe right skewness, and too small if the data contains small outliers or severe left skewness.

For the Mary River data, the large outliers—and the fact that they are so *extreme* and abundant—result in the mean being substantially influenced by the outliers, which explains why the mean is much larger than the median. *The median is the better measure of average for these data.*

The mean is generally used if possible (for practical and mathematical reasons), and is the most commonly-used measure of location. However, the mean *is* influenced by outliers and skewness; the median *is not* influenced by outliers and skewness. The mean and median are similar in approximately symmetric distributions. Sometimes, quoting *both* the mean and the median may be appropriate.

Think 13.6 (Which average to use)**.** *An engineering study (Hald 1952) was studying a new building material to determine the average permeability time.*

The time (in seconds) taken for water to permeate $n = 81$ pieces of material. Using a histogram of the data (Fig. 13.5), estimate the value of the population mean and median. Which would be best to use (for example, to quote an average permeability time on a specification sheet)?

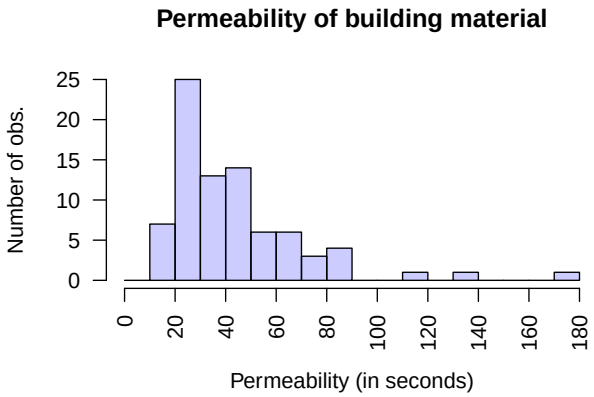


FIGURE 13.5: *A histogram of the permeability of a type of building material*

13.3 Computing the variation

For quantitative data, the amount of *variation* in the bulk of the data should be described. Many ways exist to measure the variation in a data set, including:

- The **range**: very simple and simplistic, so not often used.
- The **standard deviation**: commonly used.
- The **interquartile range (or IQR)**: commonly used.
- **Percentiles**: sometimes used.

As always, a value computed from the *sample* (the statistic) estimates the unknown value in the *population* (the parameter). Knowing which measure of variation to use is important.

13.3.1 Computing the variation: Range

The range is the simplest measure of variation.

Definition 13.5 (Range)**.** The range is the maximum value *minus* the minimum value.

The range is not often used, because only the two extreme observations are used, so it is highly influenced by outliers. Sometimes, the *range* may be given by stating both the maximum and the minimum value in the data instead of giving the *difference* between the maximum and the minimum values. The range is measured in the same measurement units as the data.

Example 13.5 (The range)**.** For Jersey cow data (Example 13.2), the range is:

$$\text{Range} = \overbrace{6.5}^{\text{largest}} - \overbrace{4.5}^{\text{smallest}} = 2.0 \text{ percent.}$$

So the sample median percentage butterfat is 5.20 percent, with a *range* of 2.00 percent.

13.3.2 Computing the variation: Standard deviation

The population standard deviation is denoted by σ ('sigma,' the parameter) and is estimated by the sample standard deviation s (the statistic). The standard deviation is the most commonly-used measure of variation, but is complicated to compute manually (but you don't need to do it manually!). The *standard deviation* is (roughly) the mean distance that the observations are away from the mean. This seems like a reasonable way to measure the amount of variation in some data.



The Greek letter σ ('sigma') is pronounced as expected: 'sigma.'

The sample standard deviation s is mostly found using computer software (e.g., jamovi or SPSS) or a calculator (in *Statistics Mode*).

Definition 13.6 (Standard deviation)**.** The *standard deviation* is, approximately, the average distance that observations are away from the mean.

You do not have to use the formula to calculate s, but we will demonstrate for those who might find it useful to understand exactly what s calculates. The formula is:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}},$$

where $\bar{x}$ is the sample mean, x represents the data values, and n is the sample size. To use the formula, follow these steps:

- Calculate the sample mean: $\overline{x}$;
- Calculate the *deviations* of each observation x from the mean: $x - \bar{x}$;
- Square these deviations (to make them all *positive* values): $(x - \bar{x})^2$;
- Add these values: $\sum(x - \bar{x})^2$;
- Divide the answer by $n - 1$;
- Take the (positive) square root of the answer.



You do not need to use the formula! You **must** know how to use software or a calculator to find the standard deviation.

Example 13.6 (Standard deviation)**.** For the Jersey cow data (Example 13.2), the *deviations* of each observation from the mean of 5.25 can be found (Fig. 13.6). Then follow the steps outlined. **You don't have to do this manually!** From Fig. 13.6, the sum of the squared distances is 2.7650. Then, the sample standard deviation is:

$$s = \sqrt{\frac{2.765}{10 - 1}} = \sqrt{0.3072222} = 0.5542763.$$

The sample mean percentage butterfat is 5.25 percent, with a sample *standard deviation* of 0.554 percent.

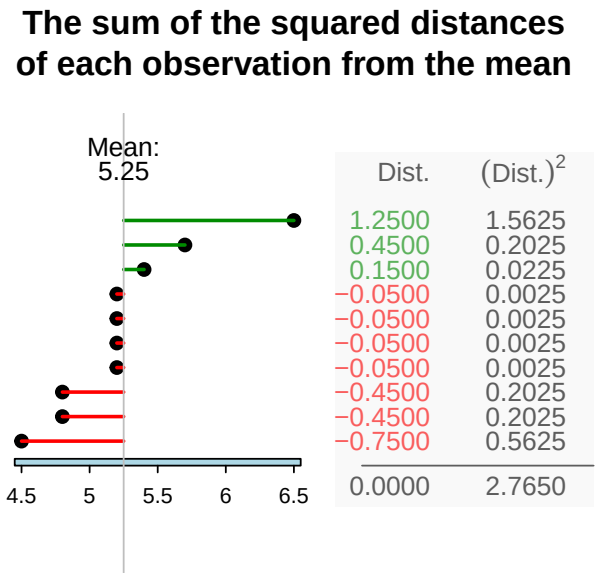


FIGURE 13.6: *The standard deviation is related to the sum of the squared-distances from the mean*

Think 13.7 (Standard deviations)**.** *The standard deviation for Dataset A in Fig. 13.7, is 2.00. What do you estimate the standard deviation of Dataset B will be:* smaller *than 2.00 or* greater *than 2.00? Why?*

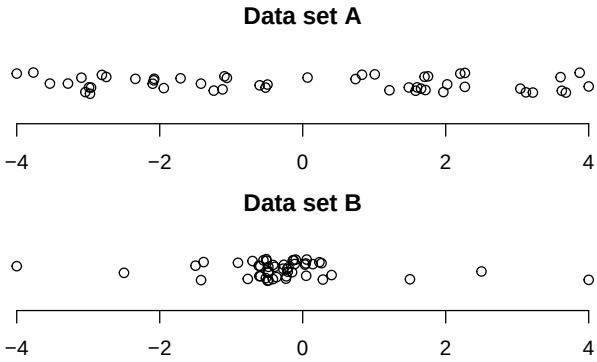


FIGURE 13.7: *Dotplots of two sets of data*

The sample standard deviation is:

- Positive (unless all observations are the same, when it is zero: there is *no* variation);
- Best used for (approximately) symmetric data;
- Usually quoted with the mean;
- The most commonly-used measure of variation;
- Measured in the same units as the data;
- Influenced by *skewness* and outliers, like the mean.

Think 13.8 (Standard deviation)**.** *Consider again the Jersey cow data (Example 13.2). Using your calculator's Statistics Mode, find the* population *standard deviation and the* sample *standard deviation.*

Answer: The *population* standard deviation is unknown. The best estimate is the *sample* standard deviation: $s = 0.554\%$.

13.3.3 Computing the variation: IQR

The standard deviation uses the value of $\bar{x}$, so is affected by skewness like the sample mean. Another measure of variation that is *not* affected by skewness is the inter-quartile range, or IQR. To understand the IQR, understanding *quartiles* first is important.

Definition 13.7 (Quartiles)**.** *Quartiles* to describe the variation and shape of data:

- The first quartile Q_1 is a value that separates the smallest 25% of observations from the largest 75%. The Q_1 is like the median of the *smaller* half of the data, halfway between the minimum value and the median.
- The second quartile Q_2 is a value that separates the smallest 50% of observations from the largest 50%. (This is the *median*.)
- The third quartile Q_3 is a value that separates the smallest 75% of observations from the largest 25%. The Q_3 is like the median of the *larger* half of the data, halfway between the median and the maximum value.

Quartiles divide the data into four parts of approximately equal numbers of observations, and *a boxplot is a picture of the quartiles*. The **inter-quartile range**, or the **IQR** is the difference between Q_3 and Q_1. The IQR measures the range of the middle 50% of the data, and is a measure of variation not influenced by outliers. The IQR is measured in the same measurements units as the data.

Definition 13.8 (IQR)**.** The *IQR* is the range in which the middle 50% of the data lie; the difference between the third and the first quartiles.

Quartiles were previously discussed in the context of boxplots (Sect. 12.4.3). For example, a boxplot of the egg-krill data (Greenacre 2016) was shown in Example 12.12; the data are repeated in Table 13.5, and the boxplot in Fig. 13.8.

For the **Treatment** group:

- 75% of the observations are smaller than about 28, and this is represented by the line at the top of the central box. This is Q_3, or the **third quartile**.
- 50% of the observations are smaller than about 12, and this is represented by the line in the centre of the central box. This is Q_2, the **second quartile** or the **median**.
- 25% of the observations are smaller than about 2, and this is represented by the line at the bottom of the central box. This is Q_1, the **first quartile**.

The IQR is $Q_3 - Q_1 = 28 - 2$, so that IQR $= 26$. Figure 13.9 shows how the IQR is found. (The online version uses an animation.)

TABLE 13.5: *The number of eggs laid by krill, for those in a treatment group and for those in a control group*

Treatment group		Control group	
0	18	0	18
0	21	0	21
1	26	0	26
1	30	0	30
3	35	1	35
8	48	1	48
8	50	1	50
12		2	

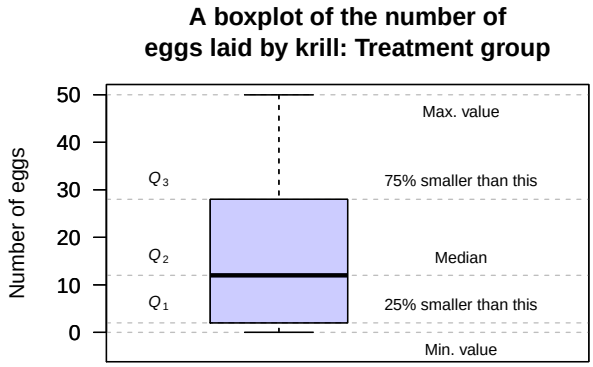


FIGURE 13.8: *A boxplot for the krill-egg data; the boxplot just for the treatment group*

Example 13.7 (Boxplots)**.** Consider the NHANES data (Center for Disease Control and Prevention (CDC) 1988--1994; Center for Disease Control and Prevention 1996; Pruim 2015).

The boxplot for the age of respondents in the NHANES data set is shown in Fig. 13.10. (The online version has an animation.) For these data:

- No outliers are identified.
- The oldest person is 80.
- About 75% of the subjects are aged less than about 54 (Q_3): the third quartile $Q_3 = 54$, the median of the *largest half* of the data.

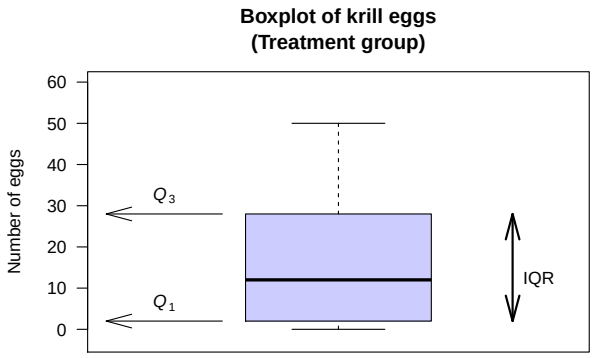


FIGURE 13.9: *Computing the IQR for the krill Treatment-group data*

- About 50% of the subjects are aged less than about 36 (Q_2, the median): the second quartile $Q_2 = 36$, the median of the data set.
- About 25% of the subjects are aged less than about 17 (Q_1): the first quartile $Q_1 = 17$, the median of the *smallest half* of the data.
- The youngest subject is aged 0.

Then, $Q_3 = 54$ and $Q_1 = 17$, so the IQR $= Q_3 - Q_1 = 54 - 17 = 37$ years. The middle 50% of the participants have an age range of 37 years.

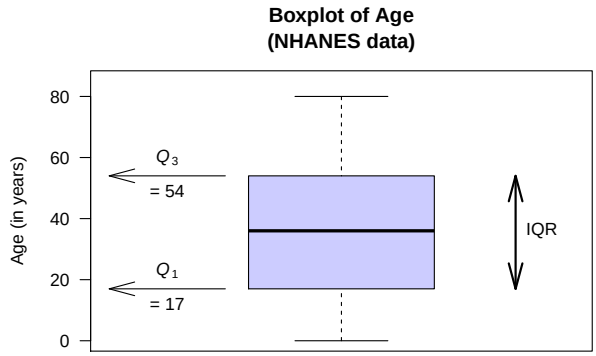


FIGURE 13.10: *Computing the IQR for the age in the NHANES data set*

13.3.4 Computing the variation: Percentiles

Percentiles can be computed, which are similar to quantiles; for example:

- The 12th percentile is a value separating the smallest 12% of the data from the rest.
- The 67th percentile is a value separating the smallest 67% of the data from the rest.
- The 94th percentile is a value separating the smallest 94% of the data from the rest.

Percentiles are measured in the same measurements units as the data.

Definition 13.9 (Percentiles)**.** The pth percentile of the data is a value separating the smallest p% of the data from the rest.

By this definition, the first quartile Q_1 is also the 25th percentile, the second quartile Q_2 is also the 50th percentile (and the median), and the third quartile Q_3 is also the 75th percentile.

Percentiles are especially useful for very skewed data and in certain applications. For instance, scientists who monitor rainfall and stream heights, and engineers who use this information, are more interested in extreme weather events rather than the 'average' event. Engineers, for example, may design structures to withstand 1-in-100 year events (the 99th percentile) or similar, which are unusual events.

Example 13.8 (Percentiles)**.** For the streamflow data at the Mary River (Table 13.1), the February data is highly right-skewed (Fig. 13.4). The median (50th percentile) is 146.1 ML. The 95th percentile is 3,480 ML, and the 99th percentile is 19,043 ML. Constructing infrastructure to cope with the *median* streamflow is clearly silly.

13.3.5 Which measure of variation to use?

Which is the 'best' measure of variation for quantitative data? As with measures of location, it depends on the data.

Since the standard deviation calculation uses the mean, it is impacted in the same way as the mean by outliers and skewness, so the standard deviation is best used with approximately symmetric data. The IQR is best used when data are skewed or asymmetric. Sometimes, both the standard deviation and the IQR can be quoted.

13.4 Describing shape

Describing the skewness numerically is possible; however, in this book the shape will be described just using words (skewed, approximately symmetric, bimodal, etc.) as before (Sect. 12.2.4).

Example 13.9 (Skewness)**.** The Australian Bureau of Statistics (ABS[2]) records the age at death of Australians[3]. The histograms of the age of death for females and males (Fig. 13.11) show that both distributions are *left* skewed: Few Australians die at a very young age, and most die at an older age.

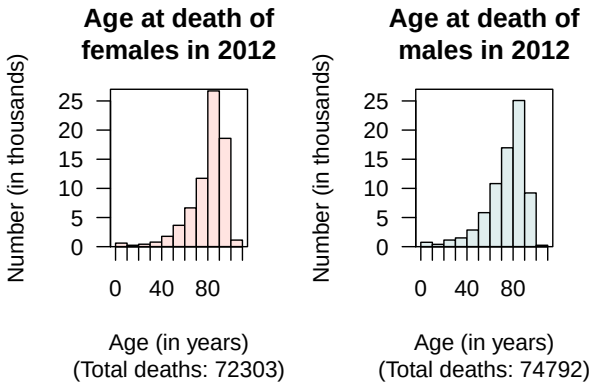


FIGURE 13.11: *Histograms of age at death for Australians in 2012*

13.5 Identifying outliers

Outliers are 'unusual' observations: observation quite different (larger or smaller) than the bulk of the data. Deciding whether or not an observation is 'unusual' is arbitrary, so 'rules' for identifying outliers are somewhat arbitrary too.

Definition 13.10 (Outliers)**.** An *outlier* is an observation that is 'unusual' compared to the bulk of the data (either larger or smaller). Rules for identifying outliers are arbitrary.

Two rules for identifying outliers are:

- The *standard deviation rule*, useful when the data have an approximately symmetric distribution.
- The *IQR rule*, useful in other situations.

Understanding the first rule requires studying bell-shaped distributions first. Knowing which rule to use is important.

13.5.1 Bell-shaped (normal) distributions and the 68–95–99.7 rule

To begin, identifying outliers will be studied for data approximately symmetrically distributed. More specifically, symmetric distributions with a bell shape will be studied. For example, the heights of husbands in the UK (Badiou et al. 1988; Hand et al. 1996) have an approximate bell shape (Fig. 13.13, left panel). Most men are between 160 and 185cm; a few are shorter than 160cm and a few taller than 185cm. More formally, *bell-shaped distributions are called normal distributions*.

These data are from a sample. Of course, every sample is likely to contain different men, and every sample of men will produce a slightly different histogram.

For convenience then, histograms may be *smoothed*, so that the smoothing produces a shape that represents an 'average' of all these possible sample histograms (in other words, an estimate of how the heights may be distributed in the *population*). For example, Fig. 13.12 shows a histogram from *one* sample of men (the online version has an animation), but every sample will be different. The solid line represents the average of many sample histograms.

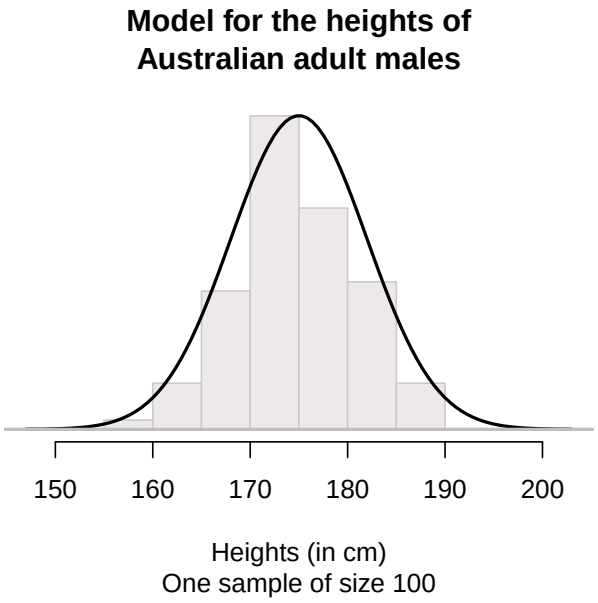


FIGURE 13.12: *The model for heights of Australian adult males, plus the histogram from one specific sample if size* $n = 100$ *of Australian adult males*

The smoothed histogram can be drawn can be considered as representing 100% of the observations; after all, *every* husband in the sample has a height, so is represented somewhere in the

histogram. When we do this, the *areas* under the normal curve are theoretical percentages of the total number.

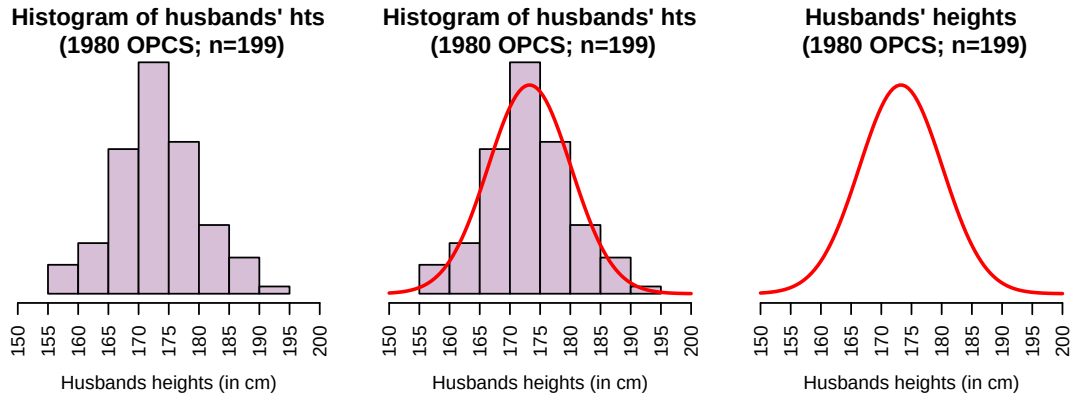


FIGURE 13.13: *The heights of husbands have an approximate normal distribution*

The smoothed histogram represents *all* of the husbands' heights (that is, 100%). Using this idea, areas of the histogram can be shaded (Fig. 13.14) to represent various percentages of the husbands' heights. For example:

- The middle 50% of husbands (Fig. 13.14, centre panel) are between about 168 and 178cm tall.
- The tallest 20% of husbands (Fig. 13.14, right panel) are taller than about 179cm.

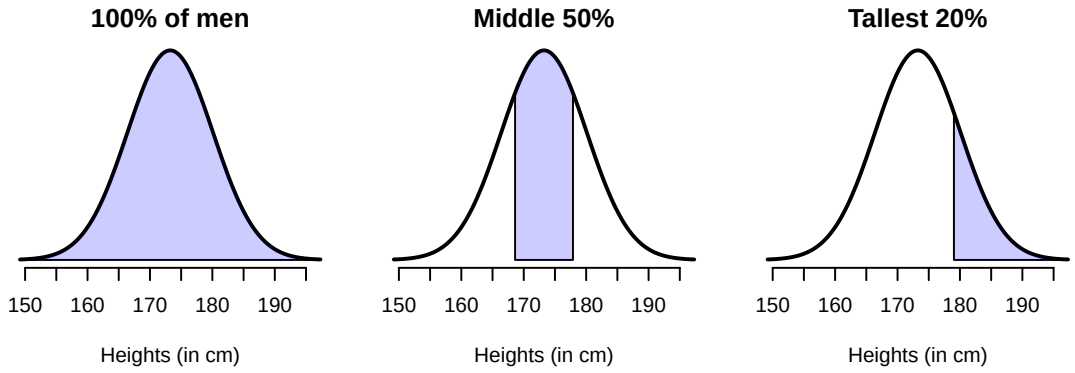


FIGURE 13.14: *The heights of husbands, with certain percentages shaded*

Importantly, for *any* normal distribution, whatever the mean or standard deviation, the areas under the smoothed curve approximately follow this important rule: *The 68–95–99.7 rule*.

Definition 13.11 (The 68–95–99.7 Rule (or the Empirical Rule))**.** For *any* bell-shaped distribution, *approximately*:

- 68% of observations lie within one standard deviation of the mean;
- 95% of observations lie within two standard deviations of the mean;
- 99.7% of observations lie within three standard deviations of the mean.

⚠ The **68–95–99.7 rule**, or the **empirical rule**, is one of the **most important rules we will see**.

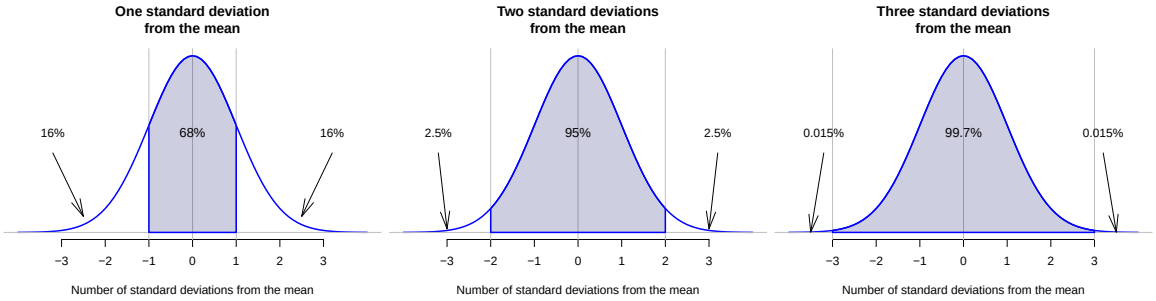


FIGURE 13.15: *The 68–95–99.7 rule: About 95% of observations are within two standard deviations of the mean*

The 68–95–99.7 rule is shown in Fig. 13.15 (the online version has an animation).

The percentages given in the 68–95–99.7 rule are *approximate*; the exact percentages are 68.27%, 95.45% and 99.73% respectively.

The 68–95–99.7 rule can be used to understand variables that have an approximate normal distribution. For example, consider the heights of husbands again (Fig. 13.16); the sample mean height is $\bar{x} = 173.2$cm; the sample standard deviation is $s = 6.88$cm. Using the 68–95–99.7 rule, approximately 68% of the husbands would have heights between

- $173.2 - 6.88 = 166.3$cm and
- $173.2 + 6.88 = 180.1$cm.

(In fact, 71% of husbands in the sample are between 166.3cm and 180.1cm tall, close to the expected 68%.) Similarly, approximately 95% of the husbands would have heights between

- $173.2 - (2 \times 6.88) = 159.4$cm and
- $173.2 + (2 \times 6.88) = 187.0$cm.

Think 13.9 (69–95–99.7 rule)**.** *For the husbands' heights, the sample mean height is* 173.2*cm; the sample standard deviation is* 6.88*cm. Using the 68–95–99.7 rule, about 99.7% of the husbands are between what heights?*

Answer: The answer is given in the online book.

The empirical rule indicates that 99.7% of observations are within 3 standard deviations of the mean. That is, *almost* all observations are within three standard deviations of the mean.

This suggests a rule for identifying outliers in approximately bell-shaped distributions: any observation more than 3 standard deviations away from the mean is unusual, so may be considered an *outlier*. More generally, this rule is often applied to approximately symmetric distributions.

Bell-shaped (normal) distributions are studied further later (for example, Chap. 17).

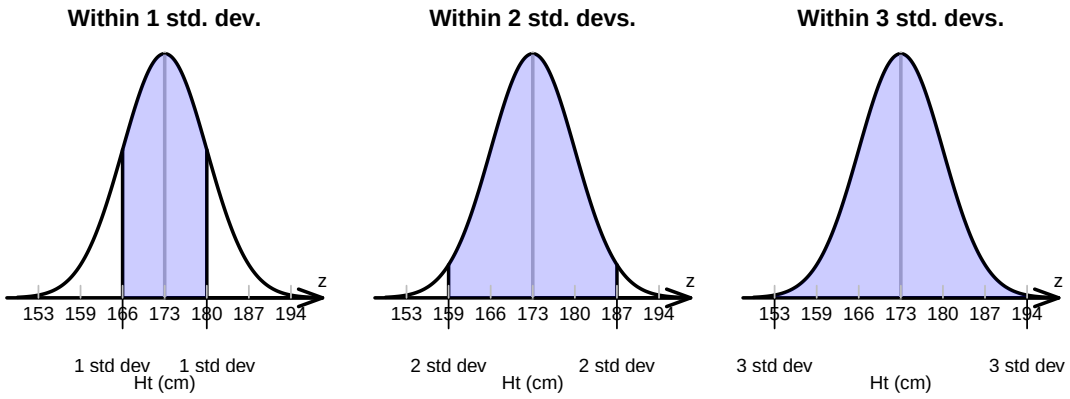


FIGURE 13.16: *The heights of husbands, showing the 68–95–99.7 rule in use*

13.5.2 The standard deviation rule for identifying outliers

One rule for identifying outliers is based on the 68–95–99.7 rule.

Definition 13.12 (Standard deviation rule for identifying outliers)**.** For approximately symmetric distributions, an observation more than three standard deviations from the mean may be considered an outlier.

This rule uses the mean and the standard deviation, so this rule is suitable for approximately symmetric distributions (when means and standard deviations are sensible numerical summaries to use). Although this rule is based on normal distributions, it has proved useful for many approximately-symmetric distributions.

All rules for identifying outliers are arbitrary. For example, the standard deviation rule is sometimes given slightly differently; for example, outliers identified as observations more than 2.5 standard deviations away from the mean. Since all rules for identifying outliers are arbitrary, both rules are acceptable.

13.5.3 The IQR rule for identifying outliers

Since the standard deviation rule for identifying outliers relies on the mean and standard deviation, it is not appropriate for non-symmetric distributions. Another rule is needed for identifying outliers in these situations: the IQR rule.

Definition 13.13 (IQR rule for identifying outliers)**.** The IQR rule identifies mild and extreme outliers as:

- *Extreme outliers*: observations $3 \times$ IQR more unusual than Q_1 or Q_3.
- *Mild outliers*: observations $1.5 \times$ IQR more unusual than Q_1 or Q_3 (that are not also extreme outliers).

This definition is *much* easier to understand using an example.

Example 13.10 (IQR rule for identifying outliers)**.** An engineering project (Hald 1952) studied a new building material, to estimate the average permeability.

Measurements of permeability time (the time for water to permeate the sheets) were taken from 81 pieces of material (in seconds). For these data $Q_1 = 24.7$ and $Q_3 = 50.6$, so we find that IQR $= 50.6 - 24.7 = 25.9$. Then, **extreme** outliers observations are $3 \times 25.9 = 77.7$ more unusual than Q_1 or Q_3. That is, *extreme* outliers are observations:

- more unusual than $24.7 - 77.7 = -53.0$ (that is, *less* than -53); or
- more unusual than $50.6 + 77.7 = 128.3$ (that is, *greater* than 128.3).

Mild outliers observations are $1.5 \times 25.9 = 38.9$ more unusual than Q_1 or Q_3 (that are not also extreme outliers). That is, *mild* outliers are

- more unusual than $24.7 - 38.9 = -14.2$ (that is, *less* than -14.2); or
- more unusual than $50.6 + 38.9 = 89.5$ (that is, *greater* than 89.5).

The outliers are identified when constructing a boxplot: the 'whiskers' extended to the most extreme observation remaining *after* excluding mild and extreme observations; then, *mild outliers* are shown using a $\circ$, and *extreme outliers* are shown using a $\star$.

You don't need to *do* this (that's what software is for), but you do need to *understand* what the software is doing. The final boxplot is shown in Fig. 13.17. (The online version has an animation.)

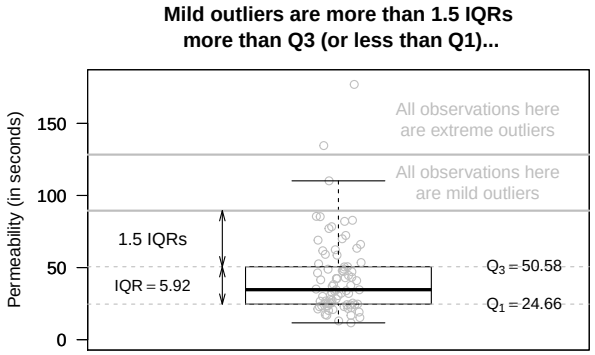


FIGURE 13.17: *Mild and extreme outliers, using the IQR rule, for the permeability data*

13.5.4 When to use which rule?

In summary, two common ways to identify outliers are:

- For *approximately symmetric distributions*: use the standard deviation rule.
- For *any distribution*, but primarily for those skewed or with outliers: use the IQR rule.

But remember: All rules for identifying outliers are arbitrary!

Example 13.11 (Boxplots and histograms)**.** For the permeability data (Hald 1952), compare the boxplot and histogram (Fig. 13.18). Can you see how the boxplot identifies the observations in the histogram that seem to be outliers?

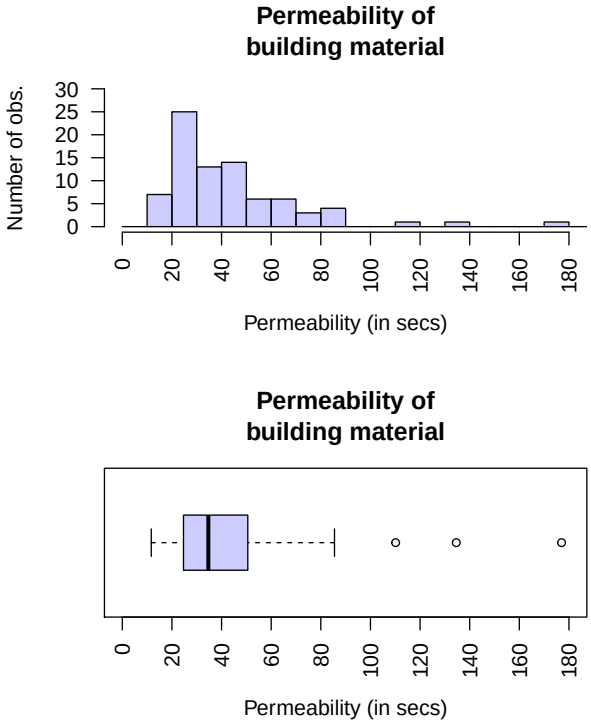


FIGURE 13.18: *A boxplot and histogram for the permeability data*

Think 13.10 (Understanding data summaries)**.** *In an American study (Tager et al. 1979), the lung capacity (FEV) of youth aged 3 to 19 was measured. The data are slightly skewed right, and the average FEV is about 2.6 litres, The FEV varies from about 0.8 to 5.8 litres, with no outliers.*

Using this information, sketch the boxplot and the histogram for the data.

13.6 Compiling tables of numerical summary information

Here are some tips for compiling tables of numerical summary information:

- Round numbers appropriately (don't necessarily use all decimals provided by software).
- Place captions *above* tables.
- In general, use **no** vertical lines and **very few** horizontal lines.
- Align numbers in the table by decimal point when possible, for easier reading.
- Ensure the table allows readers to easily make the important comparisons.

Example 13.12 (Tables for summarising data)**.** Consider a study (Ejtahed et al. 2012) assessing the effects of probiotic and conventional yoghurt on blood glucose and antioxidant status in Type 2 diabetic patients. A randomised controlled trial (i.e., an experiment) collected data from 60 patients.

Compare the two numerical summary tables in Tables 13.6 and 13.7: Table 13.6 makes comparing the two groups easier, but Table 13.7 is the more conventional orientation (for practical purposes: fewer columns).

TABLE 13.6: *Baseline characteristics of study participants. A superscript a indicates the data are summarised using means ± standard deviation; a superscript b indicates the data are summarised using medians ± IQR*

Yoghurt	Age[a]	Weight (kg)[a]	BMI[a] (kg/m^2)	Metformin/d[b]	Glibenclamide/d[b]
Conventional ($n = 30$)	51.0 ± 7.3	75.42 ± 11.28	29.14 ± 4.30	2 ± 1.25	1 ± 1
Probiotic ($n = 30$)	50.9 ± 7.7	76.18 ± 10.94	28.95 ± 3.65	2 ± 1.25	2 ± 2

TABLE 13.7: *Baseline characteristics of study participants. A superscript a indicates the data are summarised using means ± standard deviation; a superscript b indicates the data are summarised using medians ± IQR*

Variable	Conventional yoghurt ($n = 30$)	Probiotic yoghurt ($n = 30$)
Age[a]	51.00 ± 7.32	50.87 ± 7.68
Weight (kg)[a]	75.42 ± 11.28	76.18 ± 10.94
BMI (kg/m^2)[a]	29.14 ± 4.30	28.95 ± 3.65
Metformin/d[b]	2 ± 1.25	2 ± 1.25
Glibenclamide/d[b]	1 ± 1	2 ± 2

Think 13.11 (Sample differences)**.** *Do you think a difference exists between the mean BMI in the two groups* in the population*, based on Tables 13.6 and 13.7? Explain.*

13.7 Observing relationships: The NHANES study

In Sect. 12.10, the NHANES data were introduced [Center for Disease Control and Prevention (CDC) (1988--1994), Center for Disease Control and Prevention (1996), Pruim (2015)), and graphs were used to understand the data relevant to answering this RQ:

Among Americans, is the mean direct HDL cholesterol different for current smokers and non-smokers?

Using the software output (jamovi: Fig. 13.19; SPSS: Fig. 13.20), the direct HDL cholesterol can be summarised numerically:

- Average value:
  - Sample mean: $\bar{x} = 1.36$mmol/L.
  - Sample median: 1.29mmol/L.
- Variation:

  - Sample standard deviation: $s = 0.399$mmol/L.
  - Sample IQR: 0.49mmol/L.
- Shape: Slightly skewed right (from Fig. 13.1 or 12.44).
- Outliers: SPSS identified some outliers (Fig. 12.44), mostly unusually large values.

Descriptives

Descriptives

	DirectChol
N	8474
Missing	1526
Mean	1.36
Median	1.29
Standard deviation	0.399
Minimum	0.390
Maximum	4.03
25th percentile	1.09
50th percentile	1.29
75th percentile	1.58

FIGURE 13.19: *jamovi output for direct HDL cholesterol*

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Direct HDL cholesterol (in mmol/L)	8474	84.7%	1526	15.3%	10000	100.0%

Descriptives

			Statistic	Std. Error
Direct HDL cholesterol (in mmol/L)	Mean		1.3649	.00434
	95% Confidence Interval for Mean	Lower Bound	1.3564	
		Upper Bound	1.3734	
	5% Trimmed Mean		1.3425	
	Median		1.2900	
	Variance		.159	
	Std. Deviation		.39926	
	Minimum		.39	
	Maximum		4.03	
	Range		3.64	
	Interquartile Range		.49	
	Skewness		1.018	.027
	Kurtosis		2.069	.053

FIGURE 13.20: *SPSS output for direct HDL cholesterol*

The RQ is about *comparing* the mean direct HDL cholesterol in the two smoking groups, so compiling a table of summaries for each group is useful, using different output (jamovi: Fig. 13.21; SPSS: Fig. 13.22). Table 13.8 shows the numerical summaries of direct HDL cholesterol for each group.

Descriptives

Descriptives

	SmokeNow	DirectChol
N	No	1668
	Yes	1388
Missing	No	77
	Yes	78
Mean	No	1.39
	Yes	1.31
Median	No	1.32
	Yes	1.24
Standard deviation	No	0.428
	Yes	0.424
Minimum	No	0.390
	Yes	0.540
Maximum	No	3.83
	Yes	3.72
25th percentile	No	1.09
	Yes	1.01
50th percentile	No	1.32
	Yes	1.24
75th percentile	No	1.63
	Yes	1.53

FIGURE 13.21: *jamovi output for direct HDL cholesterol, by current smoking status*

TABLE 13.8: *Summarising quantitative data*

Group	Sample size	Mean	Median	Std. dev.	IQR
All participants:	8474	1.36	1.29	0.399	0.49
Smokers:	1388	1.31	1.24	0.424	0.52
Non-smokers:	1668	1.39	1.32	0.428	0.54



Notice that information about current smoking status is unavailable for all people in the study. This could impact the results, especially if those who provide data and those who do not are different regarding direct HDL.

The RQ, as usual, asks about the *population*. The RQ cannot be answered with certainty, only using a sample, since every sample is likely to be different.

Clearly, the *sample* means are different, but the RQ asks if the *population* means are different. Broadly, two possible reasons could explain why the *sample* mean direct HDL cholesterol is different for current smokers and non-smokers:

- **The *population* means are the same**, but the *sample* means are *different* simply because of the people who ended up in the sample. Another sample, with different people, might produce different sample means. *Sampling variation* explains the *difference in the sample percentages*.
- **The *population* means are different**, and the difference between the *sample* means simply reflects this difference between the *population* means.

Does respondent currently smoke?

Case Processing Summary

	Does respondent currently smoke?	Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
Direct HDL cholesterol (in mmol/L)	No	1668	95.6%	77	4.4%	1745	100.0%
	Yes	1388	94.7%	78	5.3%	1466	100.0%

Descriptives

	Does respondent currently smoke?			Statistic	Std. Error
Direct HDL cholesterol (in mmol/L)	No	Mean		1.3924	.01048
		95% Confidence Interval for Mean	Lower Bound	1.3718	
			Upper Bound	1.4129	
		5% Trimmed Mean		1.3652	
		Median		1.3200	
		Variance		.183	
		Std. Deviation		.42792	
		Minimum		.39	
		Maximum		3.83	
		Range		3.44	
		Interquartile Range		.54	
		Skewness		1.028	.060
		Kurtosis		1.474	.120
	Yes	Mean		1.3077	.01137
		95% Confidence Interval for Mean	Lower Bound	1.2854	
			Upper Bound	1.3300	
		5% Trimmed Mean		1.2812	
		Median		1.2400	
		Variance		.179	
		Std. Deviation		.42353	
		Minimum		.54	
		Maximum		3.72	
		Range		3.18	
		Interquartile Range		.52	
		Skewness		1.088	.066
		Kurtosis		2.117	.131

FIGURE 13.22: *SPSS output for direct HDL cholesterol, by current smoking status*

The difficulty, of course, is knowing which of these two reasons ('hypotheses') is the most likely reason for the difference between the sample means. This question is of prime importance (after all, it answers the RQ), and is addressed at length later in this book.

13.8 Summary

Quantitative data can be summarised numerically, and the most common techniques are indicated in Table 13.9. The **mean** and **standard deviation** are usually used whenever possible, for practical and mathematical reasons. Sometimes quoting both the mean and median (and the standard deviation and IQR) may be appropriate.

TABLE 13.9: *Summarising quantitative data*

	For distributions that are:	
Feature:	Approximately symmetric	Not symmetric, or outliers
Average:	Mean	Median
Variation:	Standard deviation	IQR
Shape:	Verbal description only	Verbal description only
Outliers:	Standard deviation rule	IQR rule

13.9 Quick review questions

A study of fulmars (a type of seabird) (Furness and Bryant 1996) explored the metabolic rate of the birds. The mass of the female birds were (in grams): 635; 635; 668; 640; 645; 635

1. From your calculator, the *sample* mean is
2. From your calculator, the *sample* standard deviation is
3. The *sample* median is
4. The *population* standard deviation is

13.10 Exercises

Selected answers are available in Sect. D.13.

Exercise 13.1. The histogram of the direct HDL cholesterol from the NHANES study is shown in Fig. 13.1. Should the mean or median be used to measure location?

Exercise 13.2. The average monthly SOI[4] values in August from 1995 to 2000 are shown in Table 13.10. Use your calculator (where possible) to calculate the:

1. sample mean
2. sample median.
3. range.
4. sample standard deviation.

Exercise 13.3. The activity below contains histogram and boxplots.

1. Match the histogram with the corresponding boxplot.
2. For which data sets would the mean and standard deviation be the appropriate numerical summary?

   For which data sets would the median and IQR be the appropriate numerical summary?

[4] http://www.bom.gov.au/climate/current/soihtml1.shtml

TABLE 13.10: *The average monthly SOI values in August from 1995 to 2000*

Year	Monthly average SOI
1995	0.8
1996	4.6
1997	-19.8
1998	9.8
1999	2.1
2000	5.3

Exercise 13.4. A study of the productivity of construction workers (Gatti et al. 2013) recorded, among other things, the rate at which concrete panels could be installed by workers. Data for three different female workers in the study are shown in Table 13.11. Construct the boxplot comparing the three workers. What does it tell you?

TABLE 13.11: *The productivity of three females workers installing concrete panels (in panels per minute)*

	Worker 1	Worker 2	Worker 3
Mean	1.24	1.73	1.36
Minimum	0.59	1.13	0.86
1st quartile	0.88	1.51	1.16
Median	1.35	1.70	1.38
3rd quartile	1.49	1.91	1.58
Maximum	1.88	3.00	2.17
Range	1.28	1.87	1.31

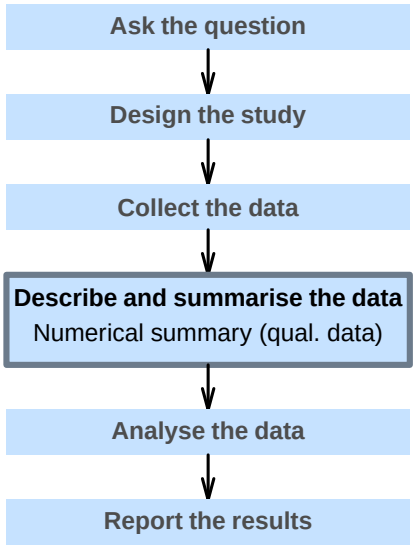
14

Numerical summaries: qualitative data

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, and graphically summarise data.

In this chapter, you will learn to numerically describe *qualitative* data. Both quantitative and qualitative *data* are described numerically in *quantitative research*. You will learn to:

- present and numerically summarise qualitative data.
- compute and understand row and column proportions (and percentages).
- compute and understand odds and odds ratios.
- describe relationships between qualitative variables.



14.1 Proportions and percentages

14.1.1 Introduction

In a study by Charig et al. (1986), the aim was to:

> ...compare (two) different methods of treating *renal calculi*... to establish which was the most [...] successful.
> — Charig et al. (1986), p. 879

(*Renal calculi* are better known as kidney stones.) Data were collected from 700 UK patients, on two qualitative variables:

- The treatment method used ('A' or 'B'): The explanatory variable. Each treatment was used on 350 patients.
- The result ('success' or 'failure' of the procedure): The response variable.

Both variables are *qualitative* with two *levels*. Treatment A was used from 1972–1980, and Treatment B from 1980–1985; that is, the treatments were not *randomly* allocated, and so *confounding* may be an issue. For this reason, the researchers also recorded the size of the kidney stone (also a qualitative variable) as a possible confounding variable, as 'small' or 'large.'

Firstly, consider just the **small stones** (Julious and Mullee 1994). The data can be compiled using a two-way table (Table 14.1), and graphed using a side-by-side or stacked bar chart, for example.

***TABLE 14.1: Numbers** for **small** kidney stones*

	Success	Failure	Total
Method A	81	6	**87**
Method B	234	36	**270**

Qualitative data can be numerically summarised by computing *proportions* or *percentages*. These can be computed:

- from the **table as a whole**;
- by **rows**; or
- by **columns**.

These are demonstrated within each section, and in a separate Example.

Definition 14.1 (Proportion)**.** A *proportion* is a fraction out of a total. Proportions are numbers between 0 and 1.

Definition 14.2 (Percentage)**.** A *percentage* is a proportion, multiplied by 100. Percentages are numbers between 0% and 100%.

14.1.2 Overall proportions and percentages

From Table 14.1, the overall *sample proportion* of successes (denoted $\hat{p}$) is:

$$\hat{p} = \frac{\text{Number of successes}}{\text{Number of procedures}} = \frac{81 + 234}{6 + 81 + 36 + 234} = 0.882.$$

The *sample* proportion of successful procedures for *small* kidney stones is 0.882. Sample proportions are denoted using $\hat{p}$. The *sample* proportion (a *statistic*) is an estimate of the unknown *population* proportion (a *parameter*), which is denoted p.



The symbol $\hat{p}$ is pronounced 'pee-hat,' and refers to the *sample* proportion.

The proportion could also be expressed as a *percentage*:

$$0.882 \times 100 = 88.2\%.$$

The sample *percentage* of successful procedures for *small* kidney stones is 88.2%. The sample *proportion* and sample *percentage* are both *statistics*

Notice that, when computing percentages and proportions, we divide the relevant number by the *total number* relevant to the context.

14.1.3 Row proportions and percentages

For the **small** kidney stones (Table 14.1), *row proportions* (or percentages), and *column proportions* (or percentages), can be computed

The *row proportions* (Table 14.2) give the proportion of successes for each *Method*, since the rows contain the counts for Method A and Method B. *Row* proportions allow the proportions within the *rows* to be compared: $81 \div 87 = 0.931$ (or 93.1%) of operations in the sample were successful for Method A, and 0.867 (or 86.7%) of operations were successful in the sample for Method B. This suggests that, for small kidney stones, Method A is more successful than Method B in the sample.

***TABLE 14.2: Row percentages** for **small** kidney stones (from Table 14.1)*

	Success	Failure	Total
Method A	93.1	6.9	**100**
Method B	86.7	13.3	**100**

14.1.4 Column proportions and percentages

For the **small** kidney stones (Table 14.1), *column proportions* can also be computed (Table 14.3). The *column proportions* give the proportion of successes within each method (since the columns contain the procedure results). *Column* proportions allow the proportions (or percentages) within *columns* to be compared: $81 \div (81 + 234) = 0.257$ (or 25.7%) of all *successful* operations came from using Method A, and 0.143 (or 14.3%) *failures* came from using Method A.

While both row and column proportions (or percentages) can be computed, row percentages seems more intuitive here: they compare the success percentage for each treatment method.

14.1.5 Example: Large kidney stones

The data in Table 14.1 are for **small** kidney stones. Data were also recorded for the **large** kidney stones (Table 14.4).

***TABLE 14.3: Column percentages** for **small** kidney stones (from Table 14.1)*

	Success	Failure
Method A	25.7	14.3
Method B	74.3	85.7
Total	**100.0**	**100.0**

For both small and large stones, the *success proportions* can be computed for Methods A and B (i.e., row percentages), and hence the better method (in the sample) can be identified.

***TABLE 14.4: Numbers** for **large** kidney stones*

	Success	Failure	Total
Method A	192	71	**263**
Method B	55	25	**80**

Think 14.1 (Percentages)**.** *The success proportion for Method A is greater than the success proportion for Method B for **small** stones (Table 14.1). Now, compute the success proportions for the **large** stones too (Table 14.4):*

- *For large stones, the success proportion with Method A is:*
- *For large stones, the success proportion with Method B is:*

*Which method has the higher success proportion for **large** stones?*

Answer: The answer is given in the online book.

Method A has a higher success proportion in the sample for both *small* (0.931 vs 0.867) and *large* kidney stones (0.730 vs 0.688). Perhaps the data for small (Table 14.1) and large kidney stones (Table 14.4) can therefore be combined, to produce a single two-way table of just Method and Result (Table 14.5), ignoring size.

***TABLE 14.5: Numbers** for **all** kidney stones combined, ignoring the size of the kidney stone*

	Success	Failure	Total
Method A	273	77	**350**
Method B	289	61	**350**

In summary, the sample shows that:

- For **small** stones (Table 14.1), **Method A has a higher success proportion**: Method A: 0.93; Method B: 0.87
- For **large** stones (Table 14.4), **Method A has a higher success proportion**: Method A: 0.73; Method B: 0.69
- Combining **all** stones together (Table 14.5), **Method B has a higher success proportion**: Method A: 0.78; Method B: 0.83

That seems strange... Method A performs better for small and for large kidney stones, but Method B performs better when combined (and size is ignored).

Think 14.2 (Explanation?)**.** *How can Method A be better when small and large stones are considered separately, but Method B be better when they are combined? Can you see* why*?*

Answer: The answer is given in the online book.

The *size of the stone* is a *confounding variable* (Fig. 14.1): The size of the stone is related to success proportion (small stones have a greater success proportion) *and* the size of the stone is related to the method used (small stones are treated more often with Method B).

This confounding could have been avoided by randomly allocating a treatment methods to patients. However, random allocation was not possible in this study, so the researchers used a different method to manage confounding: *recording* the size of the kidney stones (and other variables also: the age and sex of the patient); see Sect. 8.2.3.

In this example, acknowledging the size of the kidney stone is important, otherwise the wrong (opposite) conclusion is reached: one would think that Method B is better if the size of the stones was ignored, when the best method really is Method A.

This is called *Simpson's paradox*[1]. If the size of the kidney stone had not been recorded, size would have been a lurking variable, and the incorrect conclusion would have been reached.

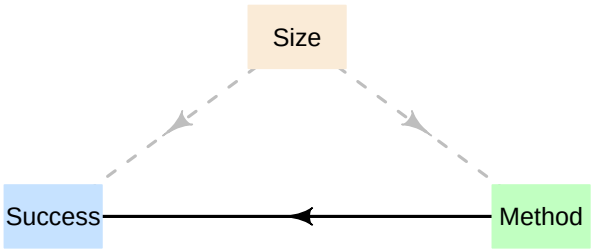


FIGURE 14.1: *The size of the stones is related to both the success percentage and the method*

14.2 Odds

Consider again the **small** kidney stone data (Table 14.1).

For *Method A*, the sample contains 81 successes and 6 failures. Apart from proportions and percentages, another way to numerically summarise this information is to see that there are $81 \div 6 = 13.5$ *times* as many successes than failures in the sample.

In other words, for small kidney stones, the **odds** of success for Method A is 13.5 (in the sample). The sample odds is a *statistic*, and the population odds is a *parameter*.

[1] https://en.wikipedia.org/wiki/Simpson%27s_paradox

Definition 14.3 (Odds)**.** The **odds** are the proportion (or percentage, or number) of times that an event *happens*, divided by the proportion (or percentage, or number) of times that the event does *not happen*:

$$\text{Odds} = \frac{\text{Proportion of times that something happens}}{\text{Proportion of times that something doesn't happen}},$$

or (equivalently)

$$\text{Odds} = \frac{\text{Number of times that something happens}}{\text{Number of times that something doesn't happen}}.$$

The *odds* show how many *times* an event *happens* compared to *not happening*. Alternatively, it is how many times the event *happens* for every 100 times that it does *not* happen.

Notice that, when computing odds, we divide the relevant number by the *remaining number*, which is different than how percentages are computed.

percentages and proportions, we divide the relevant number by the *total number* relevant to the context.

Software usually works with odds rather than percentages (for good reasons that we will not delve into). However, understanding *how* software computes the odds is important.



Software usually computes odds as comparing either

- Row 1 to Row 2; or
- Column 1 to Column 2.

Here then, based on Table 14.1, the odds for comparing the Methods would be computed as Method A compared to Method B (rather than Method B to Method A).

Example 14.1 (Interpreting odds)**.** For the *small* kidney stone data, the odds of a success for Method A is $81 \div 6 = 13.5$ (in the sample). This can be interpreted as:

- There are 13.5 *times* as many successes as failures (in the sample);
- There are $13.5 \times 100 = 1350$ successes for every 100 failures (in the sample).

Either way, successes are *far* more common than failures, for small kidney stones using Method A.

Think 14.3 (Odds)**.** *What are the odds of finding a* failure *for Method A? How is this value interpreted?*

Answer: $6 \div 81 = 0.0741$. For every 100 successes, about $0.0741 \times 100 = 7.4$ failures.

Example 14.2 (Odds)**.** Suppose that about 67% of students at a particular university were female. The *population* odds of finding a female is about $67/(100 - 67) = 2.03$: about twice as many females are students as non-females.

Suppose one tutorials had 18 females and 5 non-females. The *sample* odds of finding a female in this class is $18/5 = 3.60$. Another classes had 16 females and 9 non-females. The *sample* odds of finding a female in this class is $16/9 = 1.79$.

Example 14.3 (Computing odds)**.** Consider again the **small** kidney stone data (Table 14.1). The odds of a success using *Method B* can also be found (Table 14.1):

$$\begin{aligned}&\text{Odds(Success with Method B)}\\ =&\frac{\text{Number of successes for Method B}}{\text{Number of failures for Method B}} = \frac{234}{36} = 6.52.\end{aligned}$$

Working with the proportions (or percentages) (Table 14.2) rather than the numbers, the same value results:

$$\begin{aligned}&\text{Odds(Success with Method B)}\\ =&\frac{\text{Percentage of successes for Method B}}{\text{Percentage of failures for Method B}} = \frac{86.7}{13.3} = 6.52.\end{aligned}$$

⚠ When interpreting odds:

- When the odds are *greater* than one: the event is *more* likely to happen than to not happen.
- When the odds are *equal to* one: the event is just as likely to happen as it is to not happen.
- When the odds are *less* than one: the event is *less* likely to happen than to not happen.

14.3 Odds ratios

To summarise the **small** kidney stone data:

- For *Method A*, the odds of success are 13.5; there are 13.5 *times* as many successes as failures. (Alternatively, there are 1350 successes for every 100 failures.)
- For *Method B*, the odds of success are 6.5; there are 6.5 *times* as many successes as failures. (Alternatively, there are 650 successes for every 100 failures.)

The odds of success for Method A and Method B are very different: in the sample, the odds of success for Method A is many *times* greater than for Method B. In fact, in the sample, the odds of success for Method A is

$$\frac{13.5}{6.5} = 2.08$$

times the odds of a success for Method B. This value is called the **odds ratio** (OR); see Fig. 14.2. The sample odds ratio is a *statistic*, and the population odds ratio is a *parameter*.

Definition 14.4 (Odds Ratio (OR))**.** The **odds ratio** is how many *times* greater the odds of an event are in one group, compared to the odds of the same event in another group.



Understanding how software computes the odds ratio is important for understanding the output. jamovi and SPSS compute the odds ratio as *either*:

- The *odds* compare Row 1 to Row 2, then the odds ratio compares the Row 1 odds to the Row 2 odds.
- The *odds* compare Column 1 to Column 2, then the odds ratio compares the Column 1 odds to the Column 2 odds.

In other words, the odds and odds ratios are relative to the **first row** or **first column**.

The OR compares the odds of an event in two groups. This means that a 2×2 table can be summarised using one number: the odds ratio (OR).

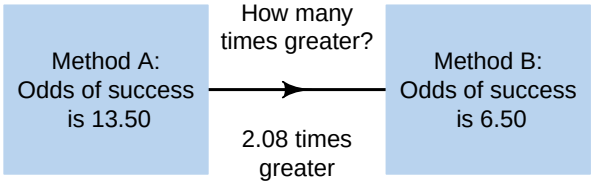


FIGURE 14.2: *The odds ratio for the small kidney stones data*



When using odds ratios (or ORs):

- When the odds ratio is *greater* than one: the odds of the event for the group in the top of the division is *greater* than the odds of the event for the group in the bottom of the division.
- When the odds ratio is *equal to* one: the odds of the event for the group in the top of the division is *equal to* the odds of the event for the group in the bottom of the division.
- When the odds ratio is *less* than one: the odds of the event for the group in the top of the division is *less* than the odds of the event for the group in the bottom of the division.

14.4 Observing relationships

For the **small** kidney stone data, the odds of a success for Method A is different than the odds of a successes for Method B, in the *sample*. Broadly, two possible reasons exist to explain the differences *in the sample*:

- The **odds** in the **population** are the same for Method A and Method B, but a difference is

observed in the *sample* odds simply because of who ended up in the sample. Every sample is likely to be different, and the *sample* we ended up with happened to show a difference. *Sampling variation* explains the *difference in the sample odds*.

- The **odds** in the **population** are different for Method A and Method B, and the difference in the sample odds simply reflects this difference between the *population* odds.

Similarly, the proportion (or percentage) of successes for Method A and B are quite different in the *sample*, and two possible reasons exist to explain the differences *in the sample*:

- **No difference exists between the proportion (or percentage)** in the **population**, but a difference is observed in the *sample* simply because of who ended up in the sample. *Sampling variation* explains the *difference in the sample proportion (or percentage)*.
- **A difference does exist between the proportion (or percentage)** in the **population**, and this difference in the sample simply reflects this difference between the *population* proportion (or percentage).

The difficulty, of course, is knowing which of these two reasons ('hypotheses') is the most likely reason for the difference between the sample odds. This question is of prime importance (after all, it answers the RQ), and is addressed at length later in this book.

14.5 Example: Skipping breakfast

The data in Table 14.6 come from a study of Iranian children aged 6–18 years old (Kelishadi et al. 2017). From this table:

- The *proportion* of females who skipped breakfast is $\hat{p}_F = 2\,383/6\,640 = 0.359$;
- The *proportion* of males who skipped breakfast is $\hat{p}_M = 1\,944/6\,846 = 0.284$.

Also,

- Odds(Skips breakfast, among F) $= 2\,383/4\,257 = 0.5598$;
- Odds(Skips breakfast, among M) $= 1\,944/4\,902 = 0.3966$.

For example, about 55.98 females *skip* breakfast for every 100 females who *eat* breakfast. The *odds ratio* (OR) comparing the odds of skipping breakfast, comparing females to males, is

$$\begin{aligned} \text{OR} &= \frac{\text{Odds(Skipping breakfast, for females)}}{\text{Odds(Skipping breakfast, for males)}} \\ &= \frac{0.5598}{0.3966} = 1.41; \end{aligned}$$

the odds of females skipping breakfast are 1.41 *times* the odds of males skipping breakfast. The data can then be summarised numerically (Table 14.7).

TABLE 14.6: *The number of Iranian children aged 6 to 18 who skip and do not skip breakfast*

	Skips breakfast	Doesn't skip breakfast	Total
Females	2383	4257	**6640**
Males	1944	4902	**6846**

TABLE 14.7: *Numerical summary of the Iranian-breakfast data: Odds and percentage of those who skip breakfast*

	Percentage	Odds	Sample size
Females	35.9	0.560	6640
Males	28.4	0.397	6846
Odds ratio		1.412	

14.6 Case Study: The NHANES data

In Sects. 12.10 and 13.7, the NHANES data were introduced (Center for Disease Control and Prevention (CDC) 1988--1994; Center for Disease Control and Prevention 1996; Pruim 2015), and graphs and numerical summaries used to understand the data relevant to answering this RQ:

> Among Americans, is the mean direct HDL cholesterol different for those who smoke now, and those who do not smoke now?

The data can be summarised numerically: the response variable (HDL cholesterol), the explanatory variable (current smoking status), and potential extraneous and confounding variables. Different summaries are needed for quantitative (means and standard deviations; medians and IQR) and qualitative (percentages; odds) variables (Table 14.8).

A number of interesting questions emerge from Table 14.8:

- How can the mean age of all respondents be 36.7 years, but the mean age for non-smokers and smokers *both* be much larger than this (54.3 and 42.7 years respectively)?
- Similarly, the percentage of females in the whole sample is 50.2%, but the percentage of females is less than this for both non-smokers and smokers (43.8% and 43.5% respectively)?

Table 14.9 summarises the relationship between current smoking status and having a diabetes diagnosis. Again, questions emerge:

- For current non-smokers, the *percentage* of diabetics is 15.32%.
- For current smokers, the *percentage* of diabetics is 7.23%.

The percentage of diabetics *in the sample* is different for non-smokers and smokers. Why? Similarly,

- For current non-smokers, the *odds* of finding a diabetic is 0.181.
- For current smokers, the *odds* of finding a diabetic is 0.078.

TABLE 14.8: *A summary of some variables in the NHANES data set, according to current smoking status (current smoking status was not reported for 6789 respondents, and some other variables were not reported for all respondents). Quantitative variables are summarised using either the mean and standard deviation, or median and IQR; qualitative variables using percentages. There are many missing values.*

Quantity	Statistic	Overall	Non-smokers	Smokers
Sample size		10000	1745	1466
Direct HDL (mmol/L)	*n*	8474	1668	1388
	Mean	1.36	1.39	1.31
	Std. dev.	0.4	0.43	0.42
Gender	*n*	10000	1745	1466
	% Female	50.2	43.8	43.5
Age (years)	*n*	10000	1745	1466
	Mean	36.74	54.28	42.68
	Std. dev.	22.4	16.64	14.79
Height (cm)	*n*	9647	1726	1459
	Mean	161.88	170.06	170.43
	Std. dev.	20.19	9.75	9.27
Weight (kg)	*n*	9922	1727	1458
	Mean	70.98	84.5	80.54
	Std. dev.	29.13	20.73	19.72
BMI (kg/m-sq)	*n*	9634	1726	1458
	Mean	26.66	29.09	27.7
	Std. dev.	7.38	6.19	6.42
Diabetes	*n*	9858	1743	1466
	% Yes	7.7	15.3	7.2
Urine volume (mL)	*n*	9013	1723	1447
	Median	94	97	102
	IQR	114	118.5	104

The odds of finding a diabetic *in the sample* is different for non-smokers and smokers. Why?

As noted before (Sect. 14.4), two possible reasons could explain this difference in percentages and odds *in the sample*:

- *Sampling variation*: The percentages (and odds) are the *same* in the population, but difference in the *sample* occur because of the people that happened to end up with the sample. *Sampling variation* explains the *difference in the sample percentages (and odds)*.
- The percentages (and odds) are *different* in population*: for non-smokers and smokers, and the difference in the sample percentages (and odds) simply reflects* a difference* between non-smokers and smokers in the population.

In the next chapters, tools for deciding which of these explanations is the most likely are discussed.

TABLE 14.9: *The two-way table of diabetes diagnosis against current smoking status*

	Doesn't smoke now	Smokes now
Not diabetic	1476	1360
Diabetic	267	106

14.7 Summary

One qualitative variable can be numerically summarized using **percentages** or **odds**. With two qualitative variables, data can be compiled into a two-way table of counts, and the data can be numerically summarised using **row percentages**, **column percentages**, **odds**, or **odds ratios**.

14.8 Quick revision questions

A study (Alley et al. 2017) examined social media (SM) use, using a

> ... sample of Australian adults [...] randomly selected from a database with Queensland landline telephone numbers. To be eligible, participants must be aged 18 or more and reside in Queensland.
> Alley et al. (2017), p. 92

Part of the data are summarised in Table 14.10.

1. Compute the *sample proportion* of *urban* residents who use social media, $\hat{p}_U$.
2. Compute the *sample proportion* of *rural* residents who use social media, $\hat{p}_R$.
3. Compute the *sample odds* of *urban* residents who use social media.
4. Compute the *sample odds* of *rural* residents who use social media.
5. Compute the *sample odds ratio* of using social media, comparing *urban* to *rural* residents.

TABLE 14.10: *The number of Queenslanders using and not using social media (SM) in rural and urban locations in a sample*

	Doesn't use SM	Uses SM	Total
Rural	78	89	**167**
Urban	416	568	**984**

14.9 Exercises

Selected answers are available in Sect. D.14.

Exercise 14.1. A study of hangovers (Köchling et al. 2019) recorded, among other information, when people vomited after consuming alcohol. Table 14.11 shows how many people vomited after consuming beer followed by wine, and how many people vomited after consuming just wine.

1. Compute the *row proportions*. What do these mean?
2. Compute the *column percentages*. What do these mean?
3. Compute the *overall percentage* of drinkers who vomited.
4. Compute the *odds* a wine-only drinker vomited.
5. Compute the *odds* that a beer-then-wine drinker vomited.
6. Compute the *odds ratio*, comparing the odds of vomiting for wine-only drinkers to beer-then-wine drinkers.
7. Compute the *odds ratio*, comparing the odds of vomiting for beer-then-wine drinkers to wine-only drinkers.

TABLE 14.11: *How many people vomited and did not vomit, by type of alcohol consumed*

	Beer then wine	Wine only
Vomited	6	6
Didn't vomit	62	22

Exercise 14.2. In a study of wallabies at the East Point Reserve (Darwin) (Stirrat 2008), the sex of adult and young wallabies was recorded. In December 1993, 91 males and 188 female *adult* wallabies were recorded. At the same time, 13 male and 22 female *young* wallabies were recorded.

1. For *adult* wallabies, what *proportion* of adult wallabies were males?
2. For *adult* wallabies, what are the *odds* that a female was observed?
3. For *young* wallabies, what are the *odds* that a female was observed?
4. For *young* wallabies, what *percentage* of wallabies were males?
5. What is the odds ratio of observing an adult wallaby to a young wallaby, for just the female wallabies?

Exercise 14.3. The *Southern Oscillation Index* (SOI) is a standardised measure of the pressure difference between Tahiti and Darwin, and has been shown to be related to rainfall in some parts of the world (Stone et al. 1996), and especially Queensland (Stone and Auliciems 1992; Dunn 2001).

As an example (Dunn and Smyth 2018), the rainfall at Emerald (Queensland) was recorded for Augusts between 1889 to 2002 inclusive, for months when the monthly average SOI was positive, and for months when the SOI was non-positive (that is, zero or negative), as shown in Table 25.12.

1. Compute the *percentage* of Augusts with no rainfall.

2. Compute the *percentage* of Augusts with no rainfall, in Augusts with a *non-positive SOI*.
3. Compute the *percentage* of Augusts with no rainfall, in Augusts with a *positive SOI*.
4. Compute the *odds* of no August rainfall.
5. Compute the *odds* of no August rainfall, in Augusts with a *non-positive SOI*.
6. Compute the *odds* of no August rainfall, in Augusts with a *positive SOI*.
7. Compute the *odds ratio* of no August rainfall, comparing Augusts with *non-positive SOI* to Augusts with a *positive SOI*.
8. Interpret this OR.

TABLE 14.12: *The SOI, and whether rainfall was recorded in Augusts between 1889 and 2002 inclusive*

	Non-positive SOI	**Positive SOI**
No rainfall recorded	14	7
Rainfall recorded	40	53

Exercise 14.4. A study (Haselgrove et al. 2008) asked boys and girls in Western Australia about back and pain from carrying school bags (Table 14.13).

1. Compute the *percentage* of boys reporting back pain from carrying school bags.
2. Compute the *percentage* of girls reporting back pain from carrying school bags.
3. Compute the *odds* of boys reporting back pain from carrying school bags.
4. Compute the *odds* of girls reporting back pain from carrying school bags.
5. Compute the *odds* of a child reporting back pain.
6. Compute the *odds ratio* of reporting back pain, comparing boys to girls.
7. Interpret this OR.

TABLE 14.13: *The number of boys and girls reporting back pain from carrying school bags*

	Males	**Females**
No	330	226
Yes	280	359

Part V

Tools for answering RQs

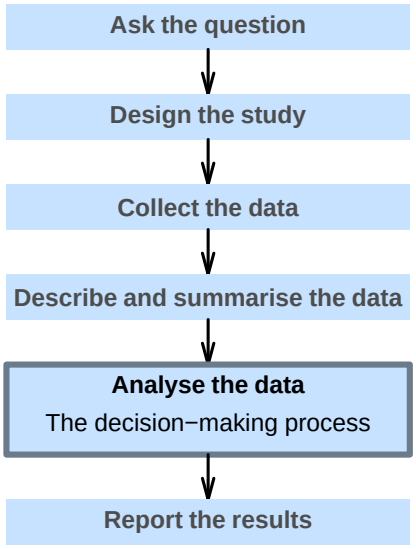
15

Making decisions: An introduction

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, and summarise data graphically and numerically.

In this chapter, you will learn how decisions are made in science, so we can answer RQs. You will learn to:

- explain the two broad reasons why differences are seen between sample statistics.
- explain how decisions are made in research.



15.1 Introduction

In Sect. 14.6, the NHANES data (Center for Disease Control and Prevention (CDC) 1988--1994) were numerically summarised. The *sample mean* direct HDL cholesterol concentration was different for smokers ($\bar{x} = 1.31$mmol/L) and for non-smokers ($\bar{x} = 1.39$mmol/L).

What does this difference between the **sample** means imply about the **population** means?

Two reasons could explain why the sample means are different:

1. The *population* means are the *same*. The *sample* means are different because every sample is likely to be different (each possible sample includes different people), so, sometimes the sample means are different by chance. This is called **sampling variation**.

2. Alternatively, the *population* means are *different*, and the sample means simply reflect this.

Similarly, in Sect. 14.6 the *odds* of being diabetic were different for smokers (0.181) and non-smokers (0.084). What does this difference between the **sample** odds imply about the **population** odds?

Again, two possible reasons could explain why the sample odds are different:

1. The *population* odds are the same. The *sample* odds are different because every sample is likely to be different (each possible sample includes different people), so sometimes, the sample odds are different by chance. This is called 'sampling variation.'
2. Alternatively, the odds are different in the *population*, and the sample odds simply reflect this.

In both situations (means; odds), the two possible explanations ('hypotheses') have special names:

1. There is *no difference* between the population parameters: this is the *null hypothesis*, or H_0.
2. There is *a difference* between the population parameters; this is the *alternative hypothesis*, or H_1.

(The word hypothesis just means 'a possible explanation.') A decision needs to be made about which of these two explanation is the most likely. However, because a sample is studied, conclusions about the *population* are never certain.

15.2 The need for making decisions

In research, decisions need to be made about *population parameters* based on *sample statistics*. The difficulty is that every sample is likely to be different (comprise different individuals from the population), and each sample will produce different summary *statistics*. This is called *sampling variation*.

⚠ Sampling variation refers to how much a sample estimate (a *statistic*) is likely to vary from sample to sample, because each sample is different.

However, sensible decisions *can* be (and *are*) made about population parameters based on sample statistics. For example, to determine if a pot soup is ready to serve, we don't have to consume the whole pot of soup (the 'population'); a sensible decision can be made from a small taste (the 'sample'). Likewise, in research sensible decisions about the population parameter can be made from the sample statistic.

To do this though, the process of *how* decisions are made needs to be articulated. In this chapter, the logic of making decisions is discussed.

To begin, consider the following scenario. Suppose I produce a standard pack of cards, and shuffle them well. The pack of cards can be considered a *population*.



A standard pack of cards has 52 cards, with four *suits*: spades and clubs (which are both black), and hearts and diamonds (which are both red). Each *suit* has 13 *denominations*: 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack (J), Queen (Q), King (K), Ace (A). Most packs also contain two jokers, but these special cards are not usually considered part of a *standard* pack.

Suppose I draw a *sample* of 15 cards from the pack, and notice that *all* are red cards. How likely is it that this would happen simply by chance? (See Fig. 15.1; the online version has an animation.) Is that evidence that the pack of cards is somehow unfair, or rigged?

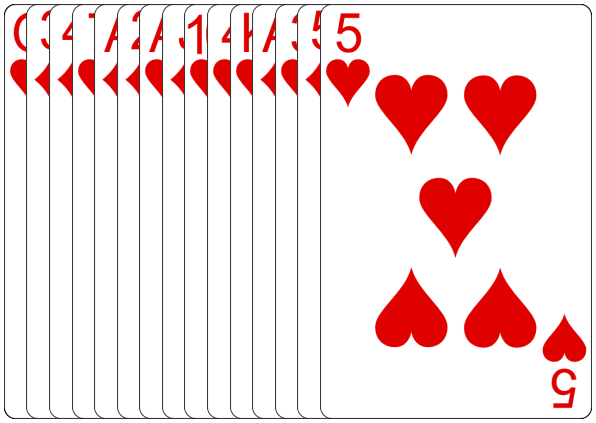


FIGURE 15.1: *How likely is it that you would get 15 red cards in a row from a fair pack?*

Getting 15 reds cards out of 15 seems very unlikely, so perhaps you may conclude that the pack is unfair in some way. But importantly, *how* did you reach that decision? Your unconscious decision-making process may have worked like this:

1. You **assumed**, quite reasonably, that this is a standard, well-shuffled pack of cards, so that half the cards are red and half the cards are black.
2. Based on that assumption then, you, quite reasonably, **expected** about half the cards in the sample of 15 to be red, and about half to be black. You wouldn't necessarily expect to see *exactly* half red and half black, but you'd probably expect something close to that.
3. But what you **observed** was nothing like that: *All* 15 cards were red. Since what you observed ('all red cards') was not like what you were expecting ('about half red cards'), the 15 cards in my hand *contradict* what you were expecting, based on your assumption of a fair pack... so your assumption of a fair pack is probably wrong.

Of course, getting 15 red cards in a row is *possible*... but very *unlikely*[1]. For this reason, we would probably conclude that the most likely explanation is that the pack is not a fair pack.

You probably didn't *consciously* go through this process, but it does seem reasonable. This process of decision making is similar to the process used in research.

[1] In fact, the probability of getting 15 cards *of the same colour* (either red *or* black) is about 0.0001025%.

15.3 How decisions are made

Based on the ideas in the last section, a formal process of decision making in research can be described as follows.

TABLE 15.1: *The decision-making process*

Assumption	Make a reasonable assumption about the value of a *population parameter*	
Expectation	Based on this assumption, describe what values of the *sample statistic* might reasonably be observed	
Observation	Observe the *sample statistic*. Then, if the observed *sample statistic* is:	...*unlikely* to happen by chance, it **contradicts** the assumption.
		...*likely* to happen by chance, it is **consistent with** the assumption.

To expand:

1. **Assumption**: Make a reasonable assumption about the population, such as the value of a *population parameter*, or state a value for the *population parameter* to be confirmed.
2. **Expectation**: Based on this assumption, describe what might be observed in the sample, such as values of the *sample statistic* that might reasonably be observed from all possible samples.
3. **Observation**: If the observed *sample statistic* is:
    - *unlikely* to happen by chance, it **contradicts** the assumption about the *population parameter*, and the assumption is probably **wrong**. The *evidence* suggests that the assumption is wrong (but it is not *certainly* wrong).
    - *likely* to happen by chance, it is **consistent with** the assumption about the *population parameter*, and the assumption may be **correct**. No *evidence* exists to suggest the assumption is wrong (though it may be wrong).

This is one way to describe the formal process of decision making in science (Fig. 15.2).

This approach is similar to what we use every day without really thinking about it For example, suppose I ask my son to brush his teeth (Budgett et al. 2013), and later I want to decide if he really did brush his teeth.

1. **Assumption**: I *assume* my son brushed his teeth (because I told him to).
2. **Expectation**: Based on that assumption, I *expect* to find a damp toothbrush when I check later.
3. **Observation**: When I check later, I observe a *dry* toothbrush. The evidence seems to contradict my assumption, as I did not find what I expected, so my assumption is probably *false*: He probably *didn't* brush his teeth.

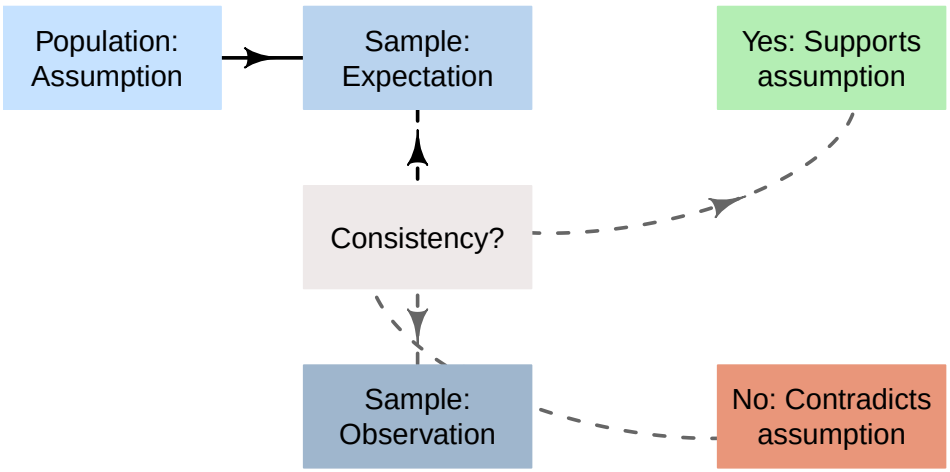


FIGURE 15.2: *A way to make decisions*

Of course, I may have made the wrong decision: He may have brushed his teeth, but his brush is now dry (he may have dried his brush with a hair dryer; he's that sort of kid). However, based on the evidence, quite probably he has not brushed his teeth.

The situation may have ended differently:

1. **Assumption**: I *assume* my son brushed his teeth (because I told him to).
2. **Expectation**: Based on that assumption, I *expect* to find a damp toothbrush when I check later.
3. **Observation**: When I check later, I observe a *damp* toothbrush. The evidence seems consistent with my assumption, as I found what I expected, so my assumption is probably *true*: He probably did brush his teeth.

Again, I may be wrong: He may have just ran his toothbrush under a tap (again, it wouldn't surprise me). I don't have any evidence that he didn't brush his teeth, though; I can hardly get him into trouble.

This logic underlies most decision making in science[2].

Example 15.1 (The decision-making process)**.** Consider the cards example from Sect. 15.2 again. The formal process might look like this:

1. **Assumption**: Initially *assume* the pack is fair and well-shuffled pack of cards (you have no evidence to doubt this).

   In other words, the proportion of red cards is 0.5 (the value of the *parameter*).

2. **Expectation**: Based on this assumption, roughly (but not necessarily *exactly*) equal numbers of red and black cards would be expected in a sample of 15 cards.

   In other words, the proportion of red cards in any *sample* is expected to be close to, but maybe not exactly, 0.5 (the value of the *statistic*).

[2]Other ways exist to make decisions too, such as incorporating prior knowledge. For example, if my son had a reputation for wetting his toothbrush under the tap instead of brushing his teeth, that information can be used to help with the decision making. This approach is called *Bayesian statistics*, but is too advanced for this book.

3. **Observation**: Suppose I then deal 15 cards, and *all* 15 are red cards.

This seems unlikely to occur if the pack is fair and well-shuffled; the data seem *inconsistent* with what I was expecting based on the assumption (Fig. 15.3). The evidence suggests that the assumption is probably false.

Of course, getting 15 red cards out of 15 is not *impossible*, so I may be wrong… but it is *very* unlikely. Based on the evidence, concluding that a problem exists with the pack of cards seems reasonable.

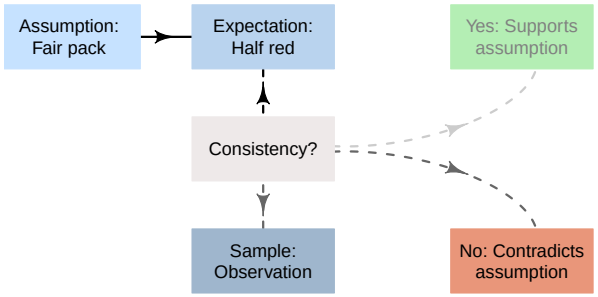


FIGURE 15.3: *A way to make decisions for the cards example*

15.4 Making decisions in research

Let's think about each step in the decision-making process (Fig. 15.2) individually.

- The **assumption** about the parameter;
- The **expectation** of the statistic; and
- The **observations**.

15.4.1 Assumption about the population parameter

Usually a reasonable assumption can be made about the *population* parameter. For example:

- We might **assume** that *no difference* exists between the parameter for two groups in the *population*, since we don't have any evidence yet to say there *is* a difference.
  For example, we might assume that the mean HDL cholesterol (or the odds of a diabetes diagnosis) is the same for current smokers and non-smokers in the *population*, for the NHANES data. If we already *knew* there was a difference, why would we be performing a study to see if there is a difference?
- We might be interested in testing a claim, or evaluating a benchmark, about a *population* parameter, to determine if the evidence supports this claim or benchmark,

These assumptions about the population parameter are called *null hypotheses*.

Example 15.2 (Assumptions about the population)**.** Most dental associations, such as the *American Dental Association* and the *Australian Dental Association*, recommend brushing teeth for two minutes. One study (Macgregor and Rugg-Gunn 1979) recorded the tooth-brushing time for 85 uninstructed schoolchildren (11 to 13 years old) from England.

We could *assume* the *population* mean tooth-brushing time in the population ('school children (11 to 13 years old) from England') is two minutes, as recommended. After all, we don't have evidence to suggest any other value for the mean. A sample can then be obtained to determine if the sample mean is consistent with, or contradicts, this assumption.

15.4.2 Expectations of sample statistics

Having made an assumption about the population parameter, the *second step* is to determine **what values to expect from the sample statistic**, based on this assumption.

Since every sample is likely to be different ('sampling variation', the value of the sample statistic depends on which of the possibe samples we end up with: the sample statistic is likely to be different for every sample.

Think about the cards in Sect. 15.2. Assuming a fair pack, then *half* the cards are red in the *population* (the pack of cards), so the *population* proportion is assumed to be $p = 0.5$.

In a *sample* of 15 cards, what values could be reasonably expected for the *sample* proportion $\hat{p}$ of red cards (the statistic)? If samples of size 15 were repeatedly taken, the sample proportion of red cards would vary from hand to hand, of course.

How much would $\hat{p}$ vary from sample to sample? Perhaps 15 red cards out of 15 cards happens reasonably frequently. Or perhaps it doesn't. How could we find out? We could:

- Use mathematical theory to determine how likely it is that we would get 15 red cards out of 15 cards.
- We could repeatedly shuffle a pack of cards, and repeatedly deal 15 cards many hundreds or thousands of times, then compute how often we get 15 red cards of out 15 cards.
- More reasonably, we could *simulate* (using a computer) dealing 15 cards many hundreds or thousands of times, and count how often we get 15 red cards of out 15 cards.

The third option is the most practical... To begin, suppose we simulated only ten hands of 15 cards each; Fig. 15.4 shows the sample proportion of red cards from ten repetitions. (The online version has an animation.) Not one of those ten hands produced 15 red cards in 15 cards.

Suppose we repeated this for *hundreds* of hands of 15 cards, and for each hand we recorded the sample proportion of cards that were red. The proportion of red cards would vary from sample to sample ('sampling variation'), and we could record the proportion of red cards from each of those hundreds of hands.

For these hundreds of sample proportions, we could draw a histogram; for example, Fig. 15.5 shows a histogram of the sample proportions from 1000 simulations of a hand of 15 cards. (The online version has an animation.)

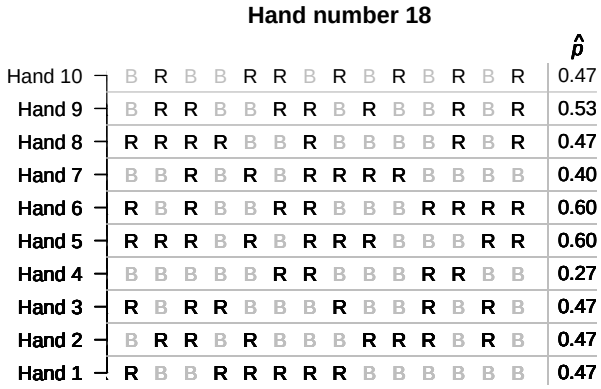


FIGURE 15.4: *Ten hands of 15 cards: The sample proportion that is red varies from hand to hand (as shown on the right-hand side)*

This histogram shows how we might expect the sample proportions $\hat{p}$ to vary from sample to sample, when the *population* proportion of red cards is $p = 0.5$.

We can see that observing 15 red cards out of 15 cards is quite rare: it never happened once in the 1000 simulations.

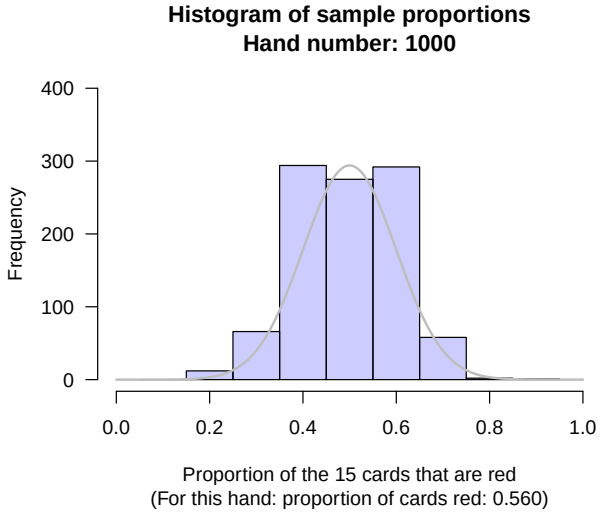


FIGURE 15.5: *A histogram of the sample proportion of red cards, in hands of 15 cards, for 1000 repetitions*

15.4.3 Observations about our sample

From this histogram, based on a simulation of one thousand hands, we could conclude that we would *almost never* find 15 red cards in 15 cards... *if the assumption of a fair pack was true*. But we *did* find 15 red cards in 15 cards... so the assumption ('a fair pack') is probably wrong.

What if we had observed 4 red cards in a hand of 15 cards (a sample proportion of $\hat{p} = 4/15 = 0.267$), rather than 15 red cards out of 15? The conclusion is not quite so obvious then: these values of $\hat{p}$ are uncommon, but they certainly do happen when $p = 0.5$. In these situations, a more sophisticated approach for making a decision is needed.

Special tools are needed to describe what to **expect** from the *sample* statistic after making **assumptions** about the *population* parameter. These special tools are discussed in the next chapter.

Example 15.3 (Sampling variation)**.** Most dental associations, such as the *American Dental Association* and the *Australian Dental Association*, recommend brushing teeth for two minutes. One study (Macgregor and Rugg-Gunn 1979) recorded the tooth-brushing time for 85 uninstructed schoolchildren from England (11 to 13 years old).

The *sample* mean tooth-brushing time may or may not be two minutes. Of course, every possible sample of 85 children will include different children, and so produce a different sample mean $\bar{x}$. Even if the *population* mean toothbrushing time really is two minutes ($\mu = 2$), the *sample* mean probably won't be exactly two minutes, because of sampling variation.

We could *assume* the population mean tooth-brushing time is two minutes ($\mu = 2$). *If* this assumption is true, we then could describe what values of the sample statistic $\bar{x}$ to *expect*. Then, after obtaining a sample and computing the sample mean, we could determine if the sample mean seems *consistent* with the assumption of two minutes, or whether it seems to *contradict* this assumption.

15.5 Tools for describing sampling variation

As we have observed previously, making decisions about population parameters[3] based on a sample statistic can be difficult: Every sample is likely to be different, and can produce a different value of the sample statistic[4].

In this chapter, though, a process for making decisions has been studied (Fig. 15.2). To apply this process to research, we need to describe *how* sample statistics vary from sample to sample (sampling variation). To do so, some of those tools are discussed in the following chapters:

- Tools to describe the population and the sample: Chap. 17.
- Tools to describe how sample statistics vary from sample to sample (sampling variation), and hence what to expect from the sample statistic: Chap. 18.
- Tools to describe the random nature of what happens with sample statistics, and so determine if the sample statistic is consistent with the assumption: Chap. 16.

15.6 Summary

Decisions are often made by first making an **assumption** about the population parameter, which leads to an **expectation** of what might occur in the sample statistics. We can then

[3] StatisticsAndParameters
[4] StatisticsAndParameters

make **observations** about our sample, to see if it seems to support or contradict the initial assumption.

15.7 Quick review questions

1. True or false: Parameters describe *populations*.
2. True or false: Both $\bar{x}$ and μ are *statistics*.
3. True or false: The value of a statistic is likely to be different in every sample.
4. True or false: *Sampling variation* refers to how the value of a *statistic* varies from sample to sample.
5. True or false: The initial assumption is made about the *sample statistic*.

15.8 Exercises

Selected answers are available in Sect. D.15.

Exercise 15.1. Suppose you are playing a die-based game, and your opponent rolls a **6** ten times in a row.

1. Do you think there is a problem with the die?
2. Explain how you came to this decision.

Exercise 15.2. In a 2012 advertisement, an Australian pizza company claimed that their 12-inch pizzas were 'real 12-inch pizzas,' unlike another brand (Dunn 2012).

1. What is a reasonable assumption to make to test this claim?
2. The claim is based on a sample of 125 pizzas, for which the sample mean pizza diameter was $\bar{x} = 11.48$ inches. What are the two reasons why the sample mean is not 12-inches?
3. Does the claim appear to be supported by, or contradicted by, the data? Why?
4. Would your conclusion change if the sample mean was $\bar{x} = 11.25$ inches, rather than 11.48 inches? Does the claim appear to be supported by, or contradicted by, the data? Why?
5. Does your answer depend on the sample size? For example, is observing a sample mean of 11.25 inches from a sample of size 10 equivalent to observing a sample mean of 11.25 inches from a sample of size 125?

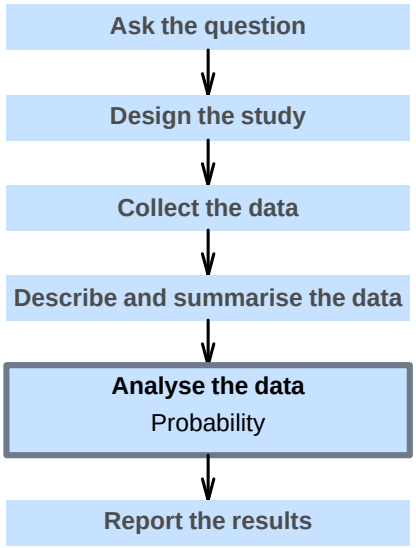
16

Probability

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, and understand the decision-making process.

In this chapter, you will learn about *probability* to describe the random nature of sample statistics. You will learn to:

- explain probabilities.
- apply the classical approach to probability in simple situations.
- apply the relative frequency approach to probability in simple situations.
- apply the subjective approach to probability in simple situations.
- identify events that are independent.



16.1 Introduction

In this chapter, *probability* is discussed. In short, *probability* quantifies the chance that something (an 'event') will happen in the future.

More formally, probability is discussed in the context of a procedure whose result is unknown beforehand. A list of all possible results from this procedure is called the *sample space*. An *event* is then defined as any combination of these elements of the sample space.

Definition 16.1 (Sample space)**.** The *sample space* is a list of all possible and distinct results after administering a procedure, whose result is unknown beforehand.

Definition 16.2 (Event)**.** An *event* is any combination of the elements in the sample space.

Example 16.1 (Sample spaces and events)**.** Consider rolling a fair, six-sided die. We do not know what face will be uppermost until we roll the die.

However, the *sample space* for this procedure can be listed: **1**, **2**, **3**, **4**, **5**, **6**. These are all distinct results (no overlap), and the sample space is *discrete*.

Many *events* could be defined using this sample space; for example:

- Rolling a **4**: This event includes one element of the sample space: **4**.
- Rolling a even number: This event includes three elements of the sample space: **2**, **4** and **6**.
- Rolling a number larger than **2**: This event includes four elements of the sample space: **3**, **4**, **5** and **6**.

Example 16.2 (Sample spaces and events)**.** Consider the distance which you can throw a cricket ball. We do not know what distance your throw will be until you throw, but we can describe the *sample space* for this procedure: the distance could be anywhere between (say) 0 and 200 metres (and of course, some of those distances are fairly unlikely to occur...).

This sample space is *continuous*.

Many *events* could be defined using this sample space; for example:

- Throwing more than 50 metres: This event includes elements of the sample space greater than 50m.
- Throwing between 10 and 40 metres: This event includes elements of the sample space between 10m and 40m.
- Throwing less than 20 metres: This event includes elements of the sample space less than 20m.

Because the sample space is continuous, throwing an *exact* distance (such as *exactly* 10 metres) is technically not possible.

Once a sample space is defined, a *probability* can be defined.

Definition 16.3 (Probability)**.** A *probability* is a number between 0 and 1 inclusive (or between 0% and 100% inclusive) that quantifies the likelihood that a certain *event* will occur.

A probability of zero (or 0%) means the event is 'impossible' (will never occur). At the other extreme, a probability of one (or 100%) means that the event is 'certain' to happen (will always occur). Most events have a probability between the extremes of 0% and 100%.

Example 16.3 (Probabilities)**.** Consider these cases:

- The probability of receiving negative rainfall is zero: It is impossible.
- The probability of receiving some rain in Buderim in the next decade is one. It is certain.
- The probability of receiving rain on any given day… is somewhere between 0 and 1.

Different ways exist to calculate probabilities of events, including:

- the *classical approach* (Sect. 16.2);
- the *relative frequency approach* (Sect. 16.3); and
- the *subjective approach* (Sect. 16.4).

16.2 Classical approach

What is the probability of rolling a  on a die? The sample space has six possible outcomes (listed in Example 16.1) that are *equally likely*, and the event ('rolling a ') comprises just *one* of those. Thus,

$$\begin{aligned}\text{Prob. of rolling a four} &= \frac{\text{The number of results that are a 4}}{\text{The number of possible results}} \\ &= \frac{1}{6}.\end{aligned}$$

We can say that 'the probability of rolling a **4** is 1/6,' or 'the probability of rolling a  is 0.1667.' The answer can also be expressed as a *percentage* ('the probability of rolling a  is 16.7%').

The answer could also be interpreted as 'the *expected* proportion of rolls that are a  is 0.167.' This approach to computing probabilities is called the *classical* approach to probability.

The chance of rolling a  in the future is 0.167, but a roll of the die will either produce a , or will not produce a  … and we don't know which will occur.

Example 16.4 (Describing probability outcomes)**.** Consider rolling a standard six-sided die again.

- The *probability* of rolling an even number is $3 \div 6 = 0.5$.
- The *percentage* of rolls that are expected to be even numbers is $3 \div 6 \times 100 = 50\%$.
- The *odds* of rolling an even number is $3 \div 3 = 1$.

Definition 16.4 (Classical approach to probability)**.** In the *classical approach to probability*, the probability of an event occurring is the number of elements of the sample space included in the event, divided by the total number of elements in the sample space, *when all outcomes are equally likely*.

By this definition:

$$\text{Prob. of an event} = \frac{\text{Number of equally-likely outcomes in the event of interest}}{\text{Total number of equally-likely outcomes}}$$

Example 16.5 (Simple events)**.** What is the *probability* of rolling a **2** on a die? What are the *odds* of rolling a **2** on a die?

Since the six possible outcomes in the sample space are *equally likely*:

$$\text{Prob. of rolling a two} = \frac{\text{One outcome is a 2}}{\text{Six equally-likely outcomes}}.$$

So the probability is $\frac{1}{6} = 0.1667$, or about 16.7%. Also, since the six possible outcomes are *equally likely*:

$$\text{Odds of rolling a two} = \frac{\text{One outcomes is a two}}{\text{Five of the possible outcomes are not a two}}.$$

So the odds of rolling a two is $\frac{1}{5} = 0.2$.

Example 16.6 (More complicated events)**.** Consider rolling a standard six-sided die. There are six equally likely outcomes (Example 16.1) each with probability $1/6$ (or 16.7%) of occurring. The probability of rolling a **1** *or* a **2** is $2/6$ (or 33.3%).

⚠ *Probabilities* describe the likelihood that an event will occur *before* the outcome is known.

Odds and *proportions* can be used either *before* or *after* the outcome is known, provided the wording is correct. For example:

- *Proportions* describe how often an event has occurred *after* the outcome is known.
- *Expected proportions* describe the likelihood that an event will occur *before* the outcome is known.

The following example may help also.

Example 16.7 (Probabilities, proportions, odds)**.** *Before* a fair coin is tossed:

- The *probability* of throwing a head is $1/2 = 0.5$.
- The *expected proportion* of heads in many coin tosses is 0.5.
- The *odds* of throwing a head is $1/1 = 1$.

If we have already tossed a coin 100 times and found 47 heads:

- The *proportion* of heads is $47/100 = 0.47$.
- The odds that we *threw* a head is $47/53 = 0.887$.

It makes no sense to talk about the 'probability that we just threw a head,' because the event has already occurred.

16.3 Relative frequency approach

What is the probability that a new baby will be a boy? The sample space could be listed as: *boy* and *non-boy*. The classical approach could be used, since the sample space has two elements: $1 \div 2 = 0.5$. This approach is fine *if* boys and non-boys are *equally likely* to be born. But are they?

In Australia in 2015[1], 305 377 live births occurred, with 157 088 male births and 148 289 non-male births. Then, the *proportion* of boys born in 2015 is

$$\frac{157\,088}{305\,377} = 0.514,$$

or about 51.4%. An *estimate* of the probability that the next birth will be a boy is is about 0.514 (or 51.4%). This is the *relative frequency* approach to calculating probabilities: based on past proportions.

The probability that the next birth will be a boy is *approximately* 0.514, but the next birth will either be a boy, or will be a not-boy… and we don't know which will occur.

Definition 16.5 (Relative frequency approach to probability)**.** In the *relative frequency approach to probability*, the probability of an event is (approximately) the number of times the outcomes of interest has appeared in the past, divided by the number of 'attempts' in the past.

Example 16.8 (Relative frequency probability)**.** Based on this information, the *odds* that a new baby will be a boy is *approximately* $0.514 \div (1 - 0.514) = 1.058$.

According to the ABS[2]:

> The sex ratio for all births registered in Australia generally fluctuates around 105.5 male births per 100 female births.

This is close to the odds of 1.058 found above.

Think 16.1 (Probability)**.** *The data in Table 16.1 concern students enrolling in a library introductory session in O-Week. (SSE is the School of Science and Engineering; SHSS is the School of Health and Sport Science.)*

Find the probability *that a randomly chosen student will be:*

1. *An SSE student.*
2. *An SHSS student aged Over 30.*
3. *Over 30,* if we already know the student is from SHSS.

[1] http:www.abs.gov.au/ausstats/abs@.nsf/0/B8865D71D84F5210CA2579330016754C?opendocument

Answer: 1: $159/255 = 62.4\%$; **2:** $40/255 = 15.7\%$; **3:** $40/96 = 41.7\%$.

TABLE 16.1: *Attendance at a library O-Week session*

	30 and under	Over 30	Total
SHSS	56	40	**96**
SSE	68	91	**159**
Total	**124**	**131**	**255**

16.4 Subjective approach

Many probabilities cannot be computed using the classical or relative frequency approach; for example:

What is the probability that Queensland will experience a Category 1 cyclone next year?

In this case, only a *subjective probability* can be given.

'Subjective' probabilities are not 'made up'; it means the probability can be estimated by considering all the relevant issues that may impact the probability (and may, for example, be based on mathematical models that incorporate information from numerous inputs). Depending on how these other issues are considered and combined, different individuals may give different subjective probabilities.

Weather forecasts are one example: weather forecasts incorporate data from sea surface temperatures, topography, air pressures, air temperatures and so on. Different models use different inputs, and then may coombine these inputs differently to produce different (subjective) forecast probabilities.

Definition 16.6 (Subjective approach to probability)**.** In the *subjective approach to probability*, various factors are incorporated, perhaps subjectively, to determine the probability of an event occurring.

Example 16.9 (Subjective probability)**.** Many farmers, based on many years of experience, can give a subjective probability of the chance of receiving rainfall in the coming month.

Think 16.2 (Probability approaches)**.** *Which* approach *is best used to estimate a probability in these situations?*

1. *The probability that the Reserve Bank will drop interest rates next month.*
2. *The probability that a randomly-chosen person writes left-handed.*
3. *The probability that a King will be randomly chosen from a pack of cards?*
4. *The probability that Buderim receives more than 100mm of rain next May.*

Answer: 1: Subjective; 2: Relative frequency; 3: Classical; 4: Probably subjective.

16.5 Independence

One important concept in probability is **independence**. Two events are *independent* if the probability of one event happening is the same, whether or not the other event has happened.

For example, if you toss a coin, the probability of getting a head is the same whether you are sitting or standing. That is, the result of a coin toss is *independent* of your position.

Definition 16.7 (Independence)**.** Two events are *independent* if the probability of one event is the same, whether or not the other event has happened.

Example 16.10 (Independence)**.** Consider drawing two cards from a fair pack (of 52 cards), without returning the first card.

For the *first* card, the sample space lists every card in the pack, and drawing any one card is as equally likely as drawing any other. Since four cards are **Aces**, the probability of drawing an **Ace** on the first draw is 4/52 (using the classical approach to probability).

If we drew an **Ace** for the first card, the probability of drawing an **Ace** for the *second* card is 3/51 (*three* **Aces** remain among the 51 remaining cards).

Alternatively, if we *don't* draw an **Ace** for the first card, the probability of drawing an **Ace** second time is 4/51 (*four* **Aces** remain among the 51 remaining cards).

In summary, the probability of drawing an **Ace** for the second card *depends* on whether or not an **Ace** was drawn for the first card. The two events 'Drawing an **Ace** for the first card' and 'Drawing an **Ace** for the second card' are *not independent* events.

16.6 Summary

At least three ways exist to compute simple probabilities: the **classical approach**, which requires all outcomes to be *equally likely*; the **relative frequency** approach; and the **subjective**

approach. Two events are **independent** if the probability of one event is the same, whether or not other event has happened.

16.7 Quick review questions

- Suppose **Event A** is defined as '*I will roll a* ⚀ *or a* ⚄ *on a fair die*.'

a. What is the best way to compute the probability of Event A occurring?

b. What is the *probability* of Event A occurring?

c. Suppose **Event B** is defined as '*A randomly-chosen university student will like pizza*.' What is the best way to compute the probability of Event B occurring?

d. True or false: **Events A** and **Event B** are *independent*.

- Consider these three events, then answer the questions that follow:
  - **Event 1** is 'The *first* card I pick from a standard 52-card pack will be an **Ace**';
  - **Event 2** is 'The *second* card I pick from a standard 52-card pack will be an **Ace**, if I *do not* return the first card'; and
  - **Event 3** is 'The *second* card I pick from a standard 52-card pack is an **Ace**, if I *do* return the first card to a random location.'

e. True or false: **Event 1** and **Event 2** are *independent*.

f. True or false: **Event 1** and **Event 3** are *independent*.

g. The probability of **Event 2** is:

16.8 Exercises

Selected answers are available in Sect. D.16.

Exercise 16.1. Suppose you have a well-shuffled, standard pack of 52 cards.

1. What is the *probability* that you will draw a **King**?
2. What are the *odds* that you will draw a **King**?
3. What is the *probability* that you will draw a picture card (**Ace**, **King**, **Queen** or **Jack**)?
4. What are the *odds* that you will draw a picture card (**Ace**, **King**, **Queen** or **Jack**)?
5. Suppose I draw two cards from the pack. Are the events 'Draw a **King** first' and 'Draw a **Queen** second' independent events?
6. Suppose I draw one card from the pack and roll a six-sided die. Are the events 'Draw a **Jack** from the pack of cards' and 'Roll a **5** on the die' independent events?

Exercise 16.2. On October 13, the American television programme *Nightline* interviewed Dr Richard Andrews, director of the California Office of Emergency Services. They discussed

various natural disasters that were being predicted as a result of an El Nino. In the interview, Dr Andrews said:

> ... we have to take these forecasts very seriously [...] I listen to earth scientists talk about earthquake probabilities a lot and in my mind every probability is 50–50, either it will happen or it won't happen...

Explain why Dr Andrews is incorrect when he says that "every probability is 50–50." Give an example to show why he must be incorrect. (Based on a report in Chance News 6.12.)

Exercise 16.3. The data in Table 16.2 were obtained from an investigation into aviation deaths of private pilots in Australia (Ruscoe and Dunn 2003).

1. What is the probability that a randomly chosen death in 1997 was of a pilot 50 or older?
2. What proportion of deaths from 1997 to 1999 were of pilots aged under 30?
3. What other information may be useful in studying the effect of age on pilot deaths?

TABLE 16.2: *Aviation deaths of private pilots in Australia from 1997 to 1999 according to the pilot's age*

	1997	**1998**	**1999**
Under 30	2	1	3
30 to 49	5	12	5
50 or over	9	11	9

Exercise 16.4. Are these pairs of two events likely to be *independent* or *not independent*? Explain.

1. 'Whether or not I walk to work tomorrow morning,' and 'Whether or not rain is expected tomorrow morning.'
2. 'Whether or not a person smokes more than 10 cigarettes per week on average' and 'Whether or not a person get lung cancer.'
3. 'Whether or not it rains today' and 'Whether or not my rubbish is collected today.'

Exercise 16.5. In disease testing, two keys aspects of the test are:

- Sensitivity*: the probability of getting a* positive* test result among people *who do* have the disease; and
- *Specificity*: the probability of getting a *negative* test result among people *who do not* have the disease.

Both are important for understanding how well a test works in practice. Consider a test with a *sensitivity* of 0.99 and a *specificity* of 0.98.

1. Suppose 100 people who *do* have a disease are tested. How many would be expected to return a positive test result?
2. Suppose 100 people who *do not* have a disease are tested. How many would be expected to return a positive test result?

Exercise 16.6. Consider the following argument:

> When I toss two coins, there are only three outcomes: a Head and a Head, a Tail and a Tail, or one of each. So the probability of obtaining two Tails must be one-third.

The reasoning is incorrect. Explain why.

Exercise 16.7. Since my wife and I have been married, I have been called to jury service three times. The latest notice reads:

> Your name has been selected at random from the electoral role…

In the same length of time, my wife has *never* been called to jury service.

Do you think the selection process really is 'at random?' Explain.

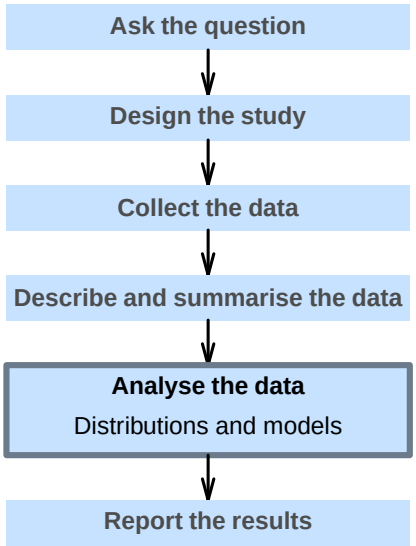
17

Distributions and models

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, and understand the decision-making process.

In this chapter, you will learn about *distributions and models* to describe the distribution of populations and samples. You will learn to:

- describe distributions.
- describe populations using normal distributions.
- use z-scores to compute probabilities related to normal distributions.
- use z-scores to 'work backwards' from probabilities for normal distributions.



17.1 Introduction

In the decision-making process used in statistics, an **assumption** is made about the population parameter, and then, based on this assumption, the values **expected** from the sample statistics can be described.

The **expectations** about the sample statistic are based around how the statistic (such as a sample mean, or a sample proportion, or a sample odds ratio) is *distributed*: what values it can take, and how often.

A *model* is used to describe this *sampling distribution*. For example, if I deal 15 cards, the

statistic could be 'the proportion of red cards in a hand of 15.' The *model* would describe how often we would see 0 red cards in 15, 1 red card in 15, 2 red cards in 15, ... up to 15 red cards in 15 (Sect. 15.4).

Under certain circumstances, many different statistics have a similarly-shaped distribution: a *bell-shaped (or normal) distribution*. We now study this distribution, as it often is the basis for describing what values the statistic can be **expected** to take, based on the **assumption** about the population that we begin with.

17.2 Distributions: An example

To begin, consider the heights of *all* Australian adult males. Clearly, the height of *all* Australian adult males is unknown: no-one has ever, or could ever realistically, measure the height of all Australian adult males. The Australian Bureau of Statistics (ABS)[1], however, takes samples of Australians to compute estimates of the heights and other measurements.

A model could be **assumed** for the heights of all Australian adult males. This is a *theoretical* idea that might be a useful description of the heights of Australian adult males in the *population*. Suppose a *model* for the heights of Australian adult males is adopted that has:

- a symmetric distribution,
- with a *mean height* of 175 cm, and
- a *standard deviation* of 7 cm.

Then, the *distribution* of the heights of Australian adult males may look like Fig. 17.1. That is, most Australian adult males are between about 168 and 182cm, and very few are taller than 196cm or shorter than 154cm.

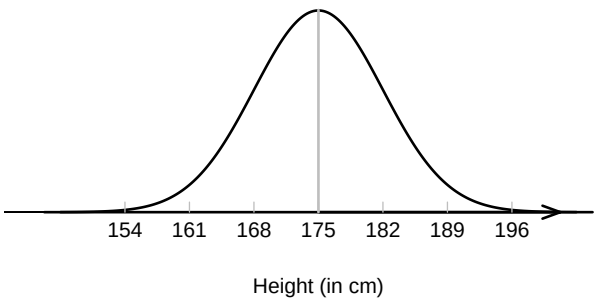


FIGURE 17.1: *A model for the heights of Australian adult males*

This model represents an idealised, or *assumed*, picture of the histogram of the heights of all Australian adult males in the *population*. If this model is a accurate, the distribution of heights in any *sample*, may be shaped a bit like this, but *sampling variation* will exist.

Any one sample will look a bit different than this model, but this model captures the general feel of the histogram from many of these samples. For example, Fig. 17.2 shows a histogram

[1] http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/4338.0main+features212011-13

from *one* sample of$n = 100$ men (the online version has an animation), but every sample will be different.

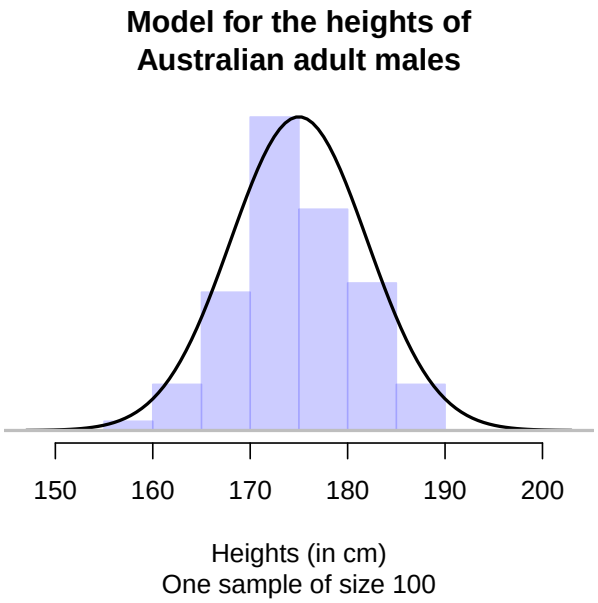


FIGURE 17.2: *The model for heights of Australian adult males, plus the histogram from one specific sample if size* $n = 100$ *of Australian adult males*

The model of heights has approximately a bell-shape: that is, most values are near the average height, but a small number of men are very tall or very short. A bell-shaped distribution is formally called a **normal distribution** or a **normal model**. A normal distribution is a way of *modelling* the population.

A *model* is a theoretical or ideal concept. In the same way that a model skeleton isn't 100% accurate (wire joins?) and certainly not exactly like *your* skeleton, it suitably approximates reality. None of us probably have a skeleton *exactly* like the model, but the model is still useful and helpful.

Likewise, no variable has *exactly* a normal distribution, but the model is still useful and helpful. The model is a *theoretical* way of describing the distribution in the population.

17.3 Normal distributions

A suitable model for the heights of *all* Australian adult males may be described (Fig. 17.1) as having:

- An approximately normal shape,
- With a mean height of $\mu = 175$ cm, and
- A standard deviation of $\sigma = 7$ cm.

This model for the heights of Australian adult males is a *theoretical idea* about the unknown *population*: it does not represent any particular sample of data. The model can be thought of as an 'average' of the histograms of the data from many samples.

Indeed, if this model turns out to be poor at describing what appears in these many samples, the *parameters* of the model (that is, the values of μ and σ) can be adjusted so the model *does* describe the sample data well.

In fact, sample evidence suggests that the average height of Australians has been increasing (Loesch et al. 2000) and so the mean of the model may need to be changed at various times to remain a good model for heights of Australian adult males.

17.4 Standardising (z-scores)

Since many statistics have a normal distribution (under certain circumstances), the 68–95–99.7 rule can be used to understand the distribution of sample statistics.

Recall that the 68–95–99.7 rule states that, for *any* normal distribution (Fig. 13.16):

- 68% of values lie within 1 standard deviation of the mean;
- 95% of values lie within 2 standard deviations of the mean; and
- 99.7% of values lie within 3 standard deviations of the mean.

These percentages only depend on how many standard deviations (σ) a value (x) is from the mean (μ). This information can be used to learn about how values are distributed.

Example 17.1 (The 68–95–99.7 rule)**.** Suppose heights of Australian adult males have a mean of $\mu = 175$cm, and a standard deviation of $\sigma = 7$cm, and (approximately) follow a normal distribution. Using this model, what proportion of Australian adult men are *taller* than 182cm?

Drawing the situation is helpful (Fig. 17.3). Notice that $175 + 7 = 182$cm is one standard deviation *above* the mean. We know that 68% of values are within one standard deviation of the mean, so that 32% are outside that range (smaller or larger) (Fig. 17.3). Hence, 16% are taller than one standard deviation above the mean, so the answer is about 16%. (Another 16% are less than one standard deviation *below* the mean, or less than $175 - 7 = 168$cm in height.)

Again, the percentages only depend on how many standard deviations (σ) the value (x) is from the mean (μ), and not the actual values of μ and σ.

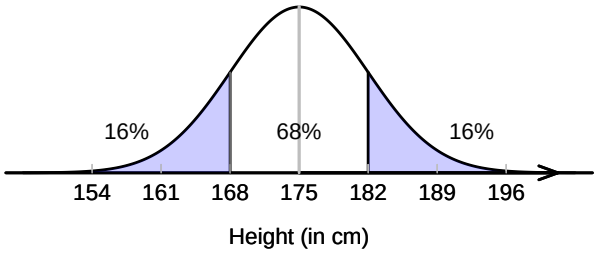


FIGURE 17.3: *What proportion of Australian adult males are taller than 182cm?*

Example 17.2 (The 68–95–99.7 rule)**.** Suppose heights of Australian adult males have a mean of $\mu = 175$cm, and a standard deviation of $\sigma = 7$cm, and (approximately) follow a normal

distribution. Using this model, what proportion are *shorter* than 161cm? Again, drawing the situation is helpful (Fig. 17.4).

Since $175 - (2 \times 7) = 161$, then 161cm is two standard deviation *below* the mean. Since 95% of values are within two standard deviation of the mean, 5% are outside that range (half smaller, half larger; see Fig. 17.4), so that 2.5% are *shorter* than 161cm. (Another 2.5% are *taller* than $175 + 14 = 189$cm.)

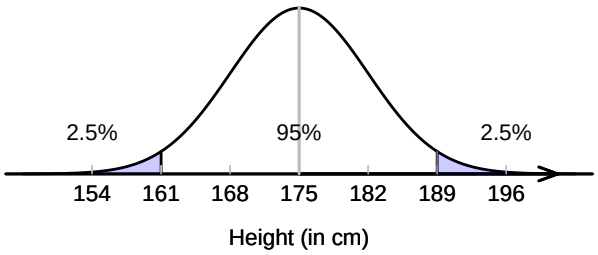


FIGURE 17.4: *What proportion of Australian adult males are shorter than 161cm?*

Again, the percentages only depend on how many standard deviations (σ) the value (x) is from the mean (μ). The number of standard deviations that an observation is from the mean is called a *z-score*. A z-score is computed using

$$z = \frac{x - \mu}{\sigma}.$$

Converting values to z-scores is called *standardising*.

Definition 17.1 (z-score)**.** A *z-score* measures how many standard deviations a value is from the mean. In symbols:

$$z = \frac{x - \mu}{\sigma}, \tag{17.1}$$

where x is the value, μ is the mean of the distribution, and σ is the standard deviation of the distribution.

Example 17.3 (z-scores)**.** In Example 17.1, the z-score for a height of 182cm is

$$z = \frac{x - \mu}{\sigma} = \frac{182 - 175}{7} = 1,$$

one standard deviation *above* the mean.

In Example 17.2, the z-score for a height of 161cm is

$$z = \frac{x - \mu}{\sigma} = \frac{161 - 175}{7} = -2,$$

two standard deviations *below* the mean (a *negative* z-score means the value is *below* the mean).

The z-score is the *number of standard deviations the observation is away from the mean*. The z-score is also called the *standardised value* or *standard score*, and is calculated using Equation (17.1). Note that:

- z-scores are negative for observations *below* the mean, and positive for observations *above* the mean.
- z-scores are numbers without units (that is, it is not in kg, or cm, etc.).

Example 17.4 (The 68–95–99.7 rule)**.** Consider the model for the heights of Australian adult males: a normal distribution, mean $\mu = 175$, standard deviation $\sigma = 7$ (Fig. 17.1).

Using this model:

- The mean is zero standard deviations from the mean: $z = 0$.
- 168cm and 182cm are one standard deviation from the mean: $z = -1$ and $z = 1$ respectively.
- 161cm and 189cm are two standard deviations from the mean: $z = -2$ and $z = 2$ respectively.
- 154cm and 196cm are three standard deviations from mean: $z = -3$ and $z = 3$ respectively.

17.5 Approximating areas using the 68–95–99.7 rule

Suppose again that heights of Australian adult males have a mean of $\mu = 175$cm, and a standard deviation of $\sigma = 7$cm, and (approximately) follow a normal distribution (Fig. 17.5).

Example 17.5 (Normal distribution areas)**.** Using this model, what proportion of men are *shorter* than 160cm?

Again, drawing the situation is helpful (Fig. 17.6).

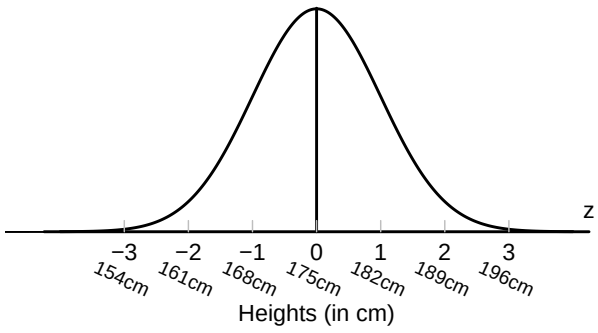


FIGURE 17.5: *The empirical rule and heights of Australian adult males*

Proceeding as before, we need to ask 'How many standard deviation below the mean is 160cm?' Using Equation (17.1) to compute the z-score, 160cm is

$$z = \frac{160 - 175}{7} = -2.14,$$

or 2.14 standard deviations, *below* the mean. What percentage of observations are less than this? This case is not covered by the 68–95–99.7 rule, though we can use the 68–95–99.7 rule to make some *rough estimates*.

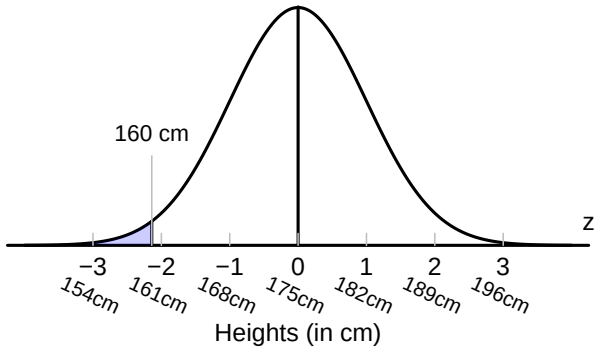


FIGURE 17.6: *What proportion of Australian adult males are shorter than 160cm?*

About 2.5% of observations are less than 2 standard deviations below the mean (Example 17.1); that is, about 2.5% of men are shorter than 161cm.

So the percentages males even shorter than 161cm (that is, further into the tail of the distribution), will be *less* than 2.5%. While we don't know the probability exactly, it will be smaller than 2.5%.

Estimates in this way are crude, but often serviceable. However, better estimates of 'areas under the normal curve' are found using tables compiled for this very purpose.

These tables are in Appendices B.2 and B.3. 'Percentages' under a normal curve are also called 'areas' under the normal curve. The *total area* under a normal curve is one (or 100%), since it represent all possible values that could be observed.

We now learn how to use these tables, then come back to Example 17.5.

17.6 Exact areas from normal distributions

Areas under normal distributions can be found using:

- **online** tables, or
- **hard copy** tables.

17.6.1 Using the online tables

The online tables work differently to the hard-copy tables. Consider the same example again: the percentage of observations *smaller than* $z = -2$.

Like the hard-copy tables, the online tables work with two decimal places, so consider the z-score as $z = -2.00$. In the tables, enter the value `-2` in the search region just under the column labelled `z.score` (see the animation below). After pressing `Enter`, the answer is shown in the column headed `Area.to.left`: the probability of finding a z-score less than -2 is 0.0228, or about 2.28%.

Using either the hard-copy or online tables gives an answer of about 2.28%. Using the

68–95–99.7 rule, the answer we obtained was 2.5%. Recall that the 68–95–99.7 rule is an *approximation* only.

17.6.2 Using the hard-copy tables

To demonstration the use of the normal distribution tables, consider the percentage of observations *smaller than* $z = -2$ (that is, two standard deviations *below* the mean) in a normal distribution.

The hard-copy tables (Appendices B.3 and B.3) work with z-scores to two decimal places, so consider the z-score as $z = -2.00$. On the tables, find -2.0 in the left margin of the table, and find the second decimal place (in this case, 0) in the top margin of the table (Fig. 17.7): where these intersect is the area (or probability) *less than* the z-score. So the probability of finding a z-score less than $z = -2$ is 0.0228, or about 2.28%. (The online tables work differently.)

	0.00	0.01
-4.0	0.0000	0.0000
-3.9	0.0000	0.0000
-3.8	0.0001	0.0001
-3.7	0.0001	0.0001
-3.6	0.0002	0.0002
-3.5	0.0002	0.0002
-3.4	0.0003	0.0003
-3.3	0.0005	0.0005
-3.2	0.0007	0.0007
-3.1	0.0010	0.0009
-3.0	0.0013	0.0013
-2.9	0.0019	0.0018
-2.8	0.0026	0.0025
-2.7	0.0035	0.0034
-2.6	0.0047	0.0045
-2.5	0.0062	0.0060
-2.4	0.0082	0.0080
-2.3	0.0107	0.0104
-2.2	0.0139	0.0136
-2.1	0.0179	0.0174
-2.0	0.0228	0.0222
-1.9	0.0287	0.0281
-1.8	0.0359	0.0351
-1.7	0.0446	0.0436
-1.6	0.0548	0.0537

FIGURE 17.7: *Using the hard-copy tables to compute the probability that z is less than -2*

⚠ The tables give the area to the *left* of the z-score that is looked up.

17.7 Comparing exact and approximate areas

Armed with knowledge of obtaining exact areas, let's return to Example 17.5:

Example 17.6 (Using normal distributions)**.** Suppose heights of Australian adult males have a mean of $\mu = 175$cm, and a standard deviation of $\sigma = 7$cm, and (approximately) follow a normal distribution. Using this model, what proportion are *shorter* than 160cm?

The general approach to computing probabilities from normal distributions is:

- **Draw a diagram**: Mark on 160 cm (Fig. 17.6).
- **Shade** the required region of interest: 'less than 160 cm tall' (Fig. 17.6).
- **Compute** the z-score using Equation (17.1).
- **Use** the z tables in Appendices B.2 and B.3.
- **Compute** the answer.

The number of standard deviations that 160cm is from the mean is using Equation (17.1):

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{160 - 175}{7} = \frac{-15}{7} = -2.14. \end{aligned}$$

That is, 160cm is 2.14 standard deviations *below* the mean, so use $z = -2.14$ in the tables. The diagram at the top of the tables reminds us that this is the probability (area) that the value of z is *less* than $z = -2.14$ (Fig. 17.6). The probability of finding an Australian man less than 160cm tall is about 1.6%.

More complicated questions can be asked too, as shown in the next section.

17.8 Examples using z-scores

Example 17.7 (Normal distributions)**.** Aedo-Ortiz et al. (1997) simulated mechanized forest harvesting systems (Devore and Berk 2007).

As part of their study, they assumed that the specific trees in their study would vary in diameter, with

- a normal distribution; with
- a mean of $\mu = 8.8$ inches; and
- a standard deviation of $\sigma = 2.7$ inches.

Using this model, what is the probability that a tree has a diameter *greater than* than 6 inches?

Follow the steps identified earlier:

- **Draw** a normal curve, and mark on 6 inches (Fig. 17.8, top panel).
- **Shade** the region corresponding to 'greater than 6 inches' (Fig. 17.8, bottom panel).
- **Compute** the z-score using Eq. (17.1). Here, $x = 6$, $\mu = 8.8$, $\sigma = 2.7$, so $z = (6 - 8.8)/2.7 = -2.8/2.7 = -1.04$ to two decimal places.

- **Use** tables: The probability of a tree diameter *shorter* than 6 inches is 0.1492. (The tables always give area *less* than the value of z that is looked up.)
- **Compute** the answer: Since the *total* area under the normal distribution is one, the probability of a tree diameter *greater* than 6 inches is $1 - 0.1492 = 0.8508$, or about 85%.

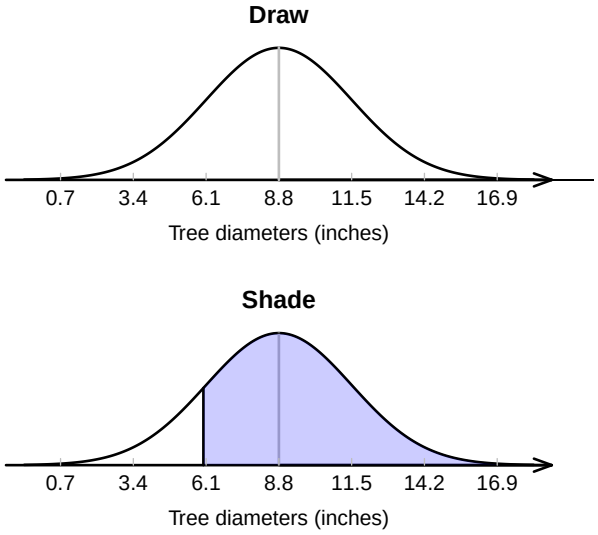


FIGURE 17.8: *What proportion of tree diameters are greater than 6 inches?*

⚠ The normal-distribution tables in the Appendix **always** provide area to the **left** of the z-scores that is looked up. Drawing a picture of the situation is important: it helps visualise how to get the answer from what the table give us.

Remember: The *total* area under the normal distribution is one.

Think 17.1 (Drawing diagrams)**.** *Match the diagram in Fig. 17.9 with the meaning for the tree-diameter model (recall: $\mu = 8.8$ inches):*

1. *Tree diameters greater than 11 inches.*
2. *Tree diameters* between *6 and 11 inches.*
3. *Tree diameters less than 11 inches.*
4. *Tree diameters between 3 and 6 inches.*

Answer: 1: B; 2: C; 3: D; 4: A.

Example 17.8 (Normal distributions)**.** Using the model for tree diameters in Example 17.7 (Aedo-Ortiz et al. 1997), what is the probability that a tree has a diameter *between* 6 and 11 inches?

First, **draw** the situation, and **shade** 'between 6 and 10 inches' (Fig. 17.10). Then, **compute** the z-scores for *both* tree diameters:

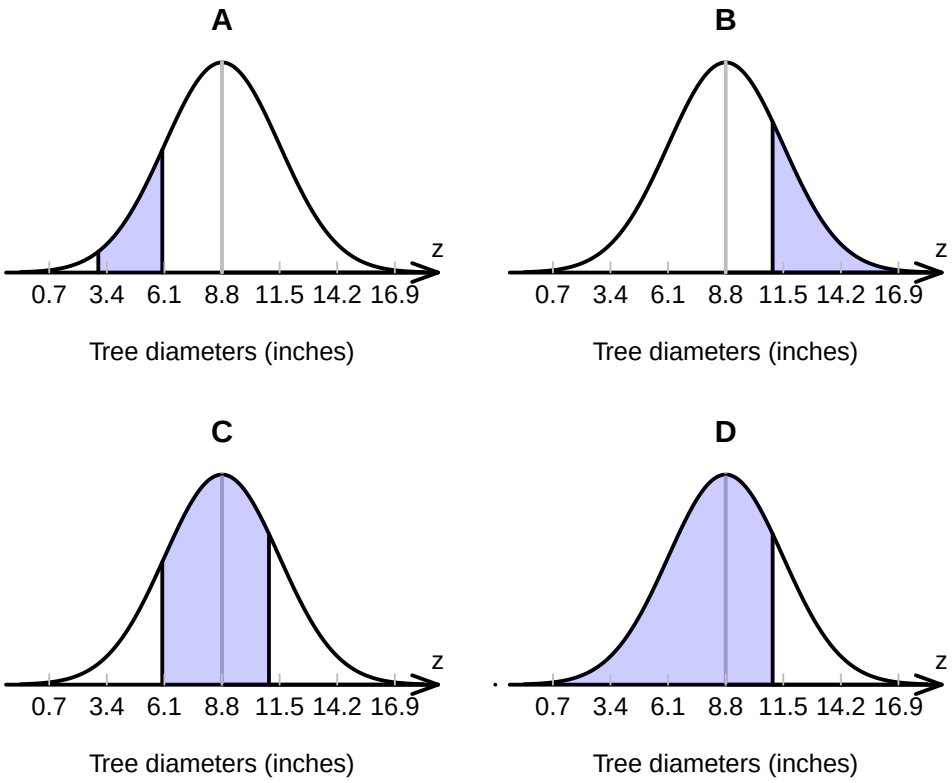


FIGURE 17.9: *Match the diagram with the description*

$$\text{6 inches: } z = \frac{6 - 8.8}{2.7} = -1.04;$$

$$\text{11 inches: } z = \frac{11 - 8.8}{2.7} = 0.81.$$

Table B can then be used to find the area to the *left* of $z = -1.04$, and also the area to the *left* of $z = 0.81$. However, neither of these provide the area *between* $z = -1.04$ and $z = 0.81$ (Fig. 17.11).

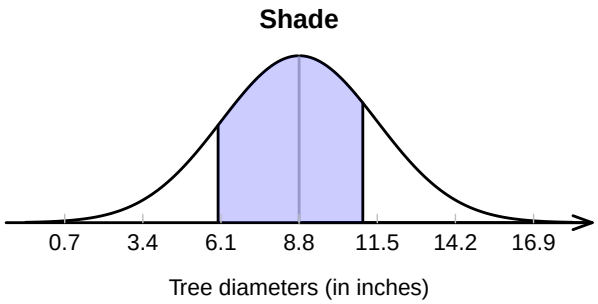


FIGURE 17.10: *What proportion of tree diameters are between 6 and 11 inches?*

Looking carefully at the areas from the tables and the area sought, that area between the two z-scores is

$$0.7910 - 0.1492 = 0.6418;$$

see Fig. 17.12 (the online version has an animation). The probability that a tree has a diameter between 6 and 11 inches is about 0.6418, or about 64%.

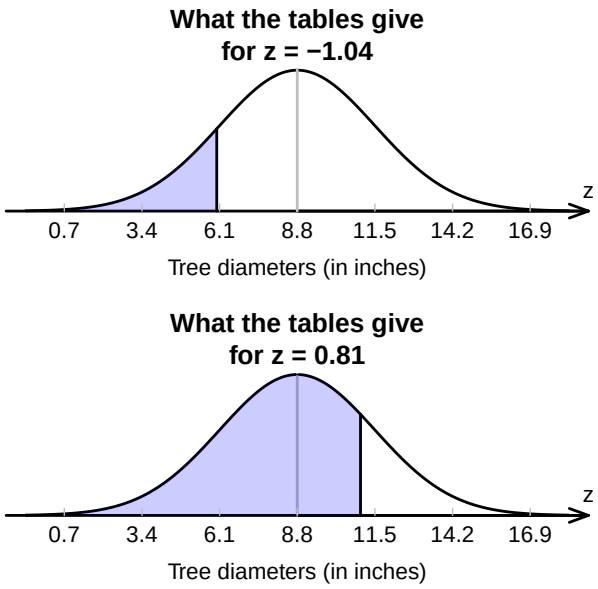


FIGURE 17.11: *What proportion of tree diameters are between 6 and 11 inches? The two shaded areas given are what we find by using the tables with $z = -1.04$ and $z = 0.81$, but neither give us the area we are seeking*

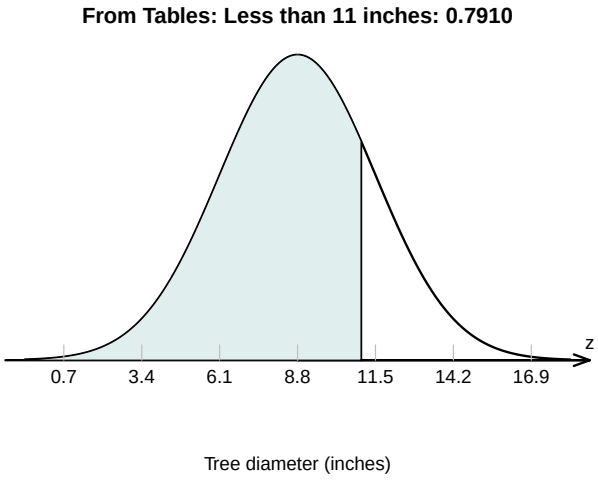


FIGURE 17.12: *Finding the area (probability) between two z-scores*

17.9 Unstandardising: Working backwards

Using the model for tree diameters in Example 17.7 (Aedo-Ortiz et al. 1997) again, suppose now the diameters of the *smallest* 10% of trees needs to be identified. What are these diameters?

Example 17.9 (Normal distributions backwards)**.** Consider again the trees study. The tree diameters can be modelled with

- a normal distribution; with
- a mean of $\mu = 8.8$ inches; and
- a standard deviation of $\sigma = 2.7$ inches.

Identify the diameters of the *smallest* 10% of trees,

This is a different problem than before; previously, the *tree diameter* was known, so a z-score could be computed, and hence a probability (Fig. 17.13, top panel). This time, the *probability* is known, and a tree diameter is sought. That is, working 'backwards' is needed (Fig. 17.13, bottom panel), so the z-tables need to be used 'backwards' too.

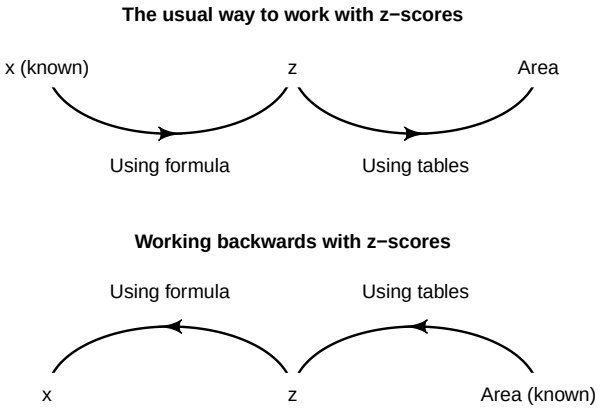


FIGURE 17.13: *Working with z-scores*

17.9.1 Using the hard-copy tables

When the z scores (in the *margins* of the tables (in Appendices B.3 and B.3) were known, the *areas* were found in the *body* of the table. If the area (or probability) is known (found in the body of the table), the corresponding z-score can be found (in the *margins* of the table), and hence the observation x; see the animation below. The closest area to 10% in the tables is 0.1003, or 10.03%.

To identify the diameters of the smallest 10% of trees, the z-score that has an area to the left of 10% (or 0.10) need to be found (at least, as close as possible to 0.10).

17.9.2 Using the online tables

When the area (or probability) is known, special online tables can be used . In these tables, enter the area to the left in search box under `Area.to.left`, and the corresponding z-scores appears under the `z.score` column .

Using either the hard-copy or online tables, the appropriate z-value is 1.28 standard deviations *below* the mean (Fig. 17.14). Then, the z-score can be converted to an observation value x using the *unstandardising* formula[2]:

$$x = \mu + z\sigma.$$

Using this unstandardising formula:

[2]This is found by re-arranging Equation (17.1).

$$x = \mu + (z \times \sigma)$$
$$= 8.8 + (-1.28 \times 2.7) = 5.344;$$

that is, about 10% of trees have diameters less than about 5.3 inches.

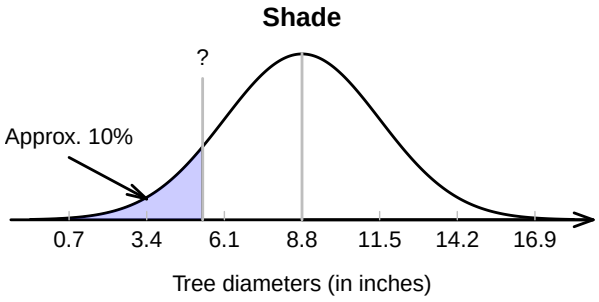


FIGURE 17.14: *Tree diameters: The smallest 10%*

Definition 17.2 (Unstandardizing formula)**.** When the z-score is known, the corresponding value of the observation x is

$$x = \mu + z\sigma. \tag{17.2}$$

This is called the *unstandardising formula*.

Think 17.2 (Normal tables backwards)**.** *Ball bearings labelled as "50mm bearings" actually have diameters that follow a normal distribution with mean 50mm and standard deviation 0.1mm. The* smallest *15% of bearings are too small for sale. What size bearings cannot be sold?*

Answer: The closest area from the tables is 0.1492, corresponding to $z = -1.04$. Using the unstandardising formula, $x = 50 + (-1.04 \times 0.1)$, or 49.896: Bearings less than about 49.9 mm in diameter cannot be sold.

Example 17.10 (Normal distributions backwards)**.** Using the model for tree diameters in Example 17.7 (Aedo-Ortiz et al. 1997) again, suppose now the diameters of the *largest* 25% of trees needs to be identified. What are these diameters?

The tree diameters can be modelled with

- a normal distribution; with
- a mean of $\mu = 8.8$ inches; and
- a standard deviation of $\sigma = 2.7$ inches.

Again, we need to work 'backwards' (Fig. 17.15, bottom panel), so the z-tables need to be used 'backwards' too. The *largest* 25% implies large trees, so we would expect a diameter larger than the mean.

Using a diagram is important (Fig. 17.15): the tables work with the area to the *left* of the value of interest, which is 75%.

Using either the hard-copy or online tables, the appropriate z-value is $z = 0.674$. Then, the z-score can be converted to an observation value x using the *unstandardising* formula:

$$x = \mu + (z \times \sigma)$$
$$= 8.8 + (0.674 \times 2.7) = 10.621;$$

that is, about 25% of trees have diameters larger than about 10.6 inches.

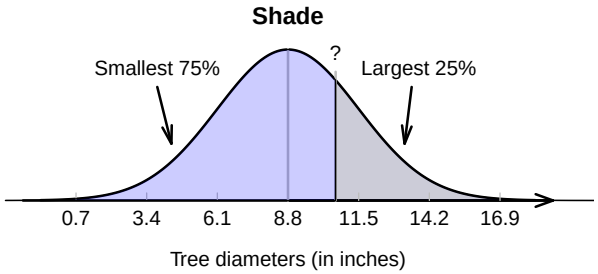


FIGURE 17.15: *Tree diameters: The largest 25% is the same as the smallest 75%*

17.10 Summary

A **model** is a way of theoretically describing the distribution of some quantitative variable in a population. One common model is a **normal model** or **normal distribution**, which is a bell-shaped distribution with a theoretical mean μ and a theoretical standard deviation σ. Probabilities can be computed from normal distributions using z**-scores**.

17.11 Quick revision questions

Consider again the model for tree diameters in Example 17.7 (Aedo-Ortiz et al. 1997): a normal distribution with $\mu = 8.8$ inches, and $\sigma = 2.7$ inches.

1. A tree diameter of 10.2 inches corresponds to a z-score (to two decimal places) of:
2. The probability that a tree has a diameter *less* than 10.2 inches is (as a *decimal value*):
3. The probability that a tree has a diameter *greater* than 10.2 inches is (as a *decimal value*):
4. A tree diameter of 6.9 inches corresponds to a z-score (to two decimal places) of (as a *decimal value*):
5. The probability that a tree has a diameter *less* than 6.9 inches is (as a *decimal value*):
6. The probability that a tree has a diameter *greater* than 6.9 inches is (as a *decimal value*):

17.12 Exercises

Selected answers are available in Sect. D.17.

Exercise 17.1. Consider again the study by Aedo-Ortiz et al. (1997), who studied the diameter of trees in certain forests. The tree diameters can be modelled with

- a normal distribution; with
- a mean of $\mu = 8.8$ inches; and
- a standard deviation of $\sigma = 2.7$ inches.

For these trees:

1. What is the probability that a tree will have a diameter *less than* 8 inches?
2. What is the probability that a tree will have a diameter *greater than* 9 inches?
3. What is the probability that a tree will have a diameter *between* 7 and 10 inches?
4. The largest 15% of trees have what diameters?
5. The smallest 25% of trees have what diameters?

Exercise 17.2. In a study (Snowden and Basso 2018) to help understand factors influencing preterm births, the researchers modelled the gestation length of healthy babies as having a normal distribution with a mean of 40 weeks, and a standard deviation of 1.64 weeks. Using this model:

1. What proportion of births are *longer* than 39 weeks (that is, nine months)?
2. In Australia, a premature birth is defined as a birth occuring before 37 weeks[3]. What proportion of births are expected to be premature?
3. According to Health Direct[4], 'Babies born between 32 and 37 weeks may need care in a special care nursery.' What proportion of healthy births would be expected to be born between 32 and 37 weeks gestation?
4. How long is the gestation length for the *longest* 5% of pregnancies?
5. How long is the gestation length for the *shortest* 5% of pregnancies?

Exercise 17.3. IQ scores are designed to have[5] a mean of 100 and a standard deviation of 15. Mensa[6] is a society for people with a high IQ:

> Membership of Mensa is open to persons who have attained a score within the upper two percent of the general population on an approved intelligence test that has been properly administered and supervised.
> — Mensa webpage[7]

What IQ score is needed to join Mensa?

Exercise 17.4. IQ scores are designed to have[8] a mean of 100 and a standard deviation of 15. Zagorsky (2016) reports that

[3] https://www.pregnancybirthbaby.org.au/premature-baby
[4] https://www.pregnancybirthbaby.org.au/premature-baby
[5] https://en.wikipedia.org/wiki/IQ_classification
[6] https://www.mensa.org/
[8] https://en.wikipedia.org/wiki/IQ_classification

> ...Congress requires the Pentagon to reject all military recruits whose IQ is in the bottom 10% of the population...
> — Zagorsky (2016), p. 403

What IQs scores lead to a rejection from the US military?

Exercise 17.5. IQ scores are designed to have[9] a mean of 100 and a standard deviation of 15. Match the diagram in Fig. 17.16 with the meaning.

1. IQs greater than 110.
2. IQs between 90 and 115.
3. IQs less than 110.
4. IQs greater than 85.

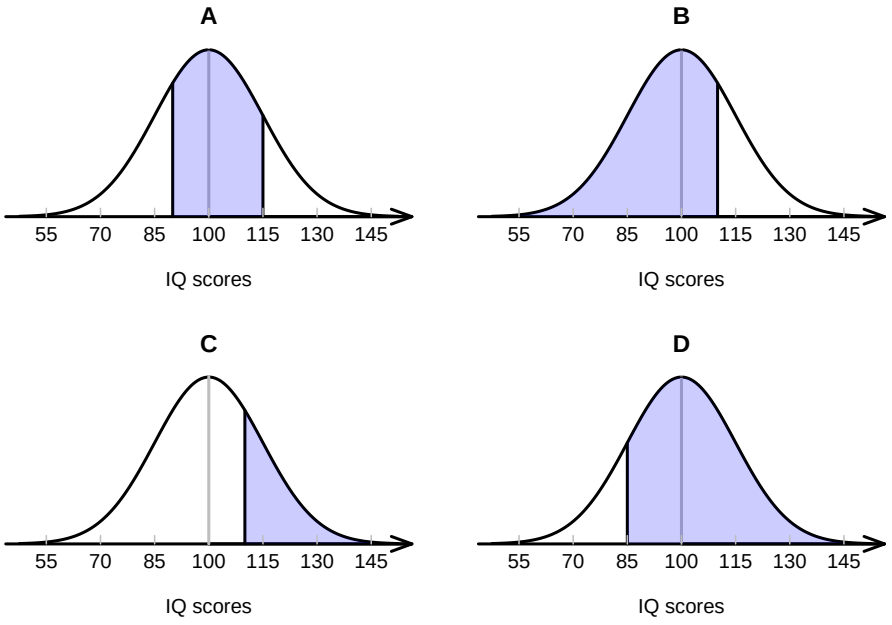


FIGURE 17.16: *Match the diagram with the description*

Exercise 17.6. IQ scores are designed to have[10] a mean of 100 and a standard deviation of 15. Match the diagram in Fig. 17.16 with the meaning.

1. The *largest* 25% of IQ scores.
2. The *smallest* 10% of IQ scores.
3. The *largest* 70% of IQ scores.
4. The *smallest* 60% of IQ scores.

Exercise 17.7. A study of the impact of charging electric vehicles (EVs) on electricity demands (Affonso and Kezunovic 2018) modelled the *time* at which people began charging their EVs at home. Based on a survey (US Department of Transportation 2011), they modelled the time at which EVs began charging as having a mean of 5:30pm, with a standard deviation of 2.28 hrs. For this model:

[9] https://en.wikipedia.org/wiki/IQ_classification
[10] https://en.wikipedia.org/wiki/IQ_classification

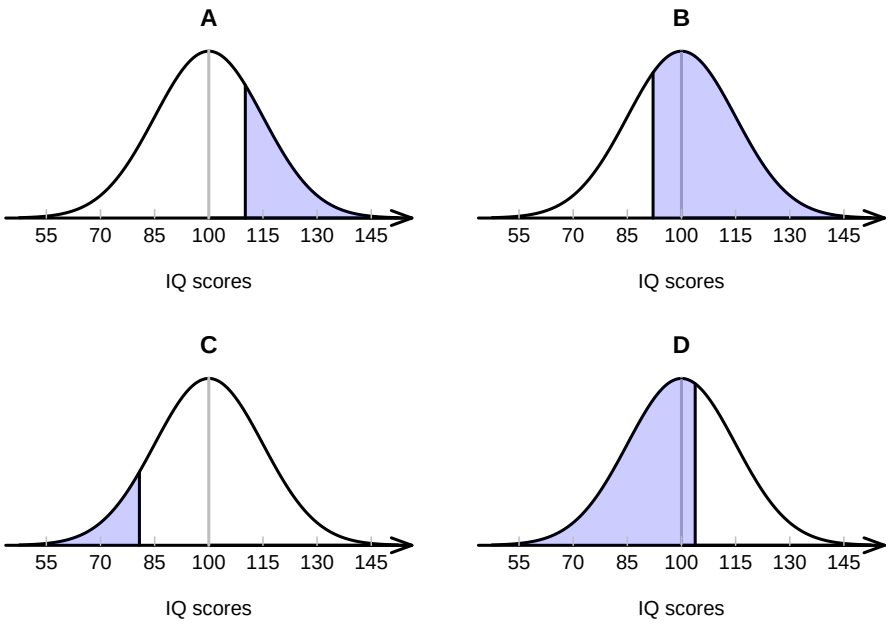


FIGURE 17.17: *Match the diagram with the description*

1. What is the probability that an EVs will begin charging after 9pm?
2. What is the probability that an EVs will begin charging before 5pm?
3. What is the probability that an EVs will begin charging between 5pm and 6pm?
4. 30% of the EVs begin charging after what time?
5. The earliest 15% of charging begins when?

Hint: This question is much easier if you convert times into ‘minutes after midnight.’

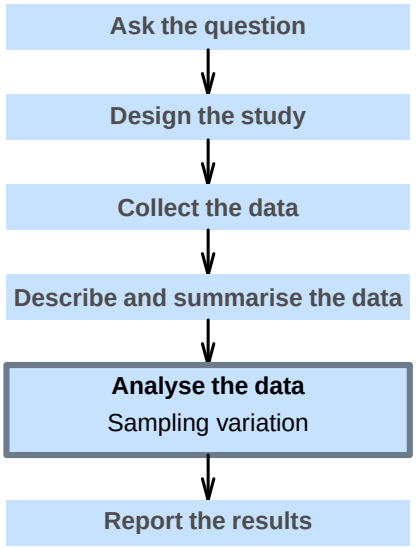
18

Sampling variation

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, and understand the decision-making process.

In this chapter, you will learn to describe *sampling variation*. You will learn to:

- explain what a sampling distribution describes.
- explain the difference between variation between individuals and variation in sample statistics.
- determine when a standard error is appropriate to use.
- explain the difference between standard errors and standard deviations.



18.1 Introduction

The last three chapters introduced tools to apply the decision-making process (Sect. 15.4) used in research:

1. Make an **assumption** about the population *parameter*.
2. Based on this assumption, describe what values the sample *statistic* might reasonably be **expected** from all possible samples.
3. **Observe** the sample data, and see if it seems *consistent* with the expectation, or if it *contradicts* the expectation.

One key observation is that, under certain conditions, the variation of many *sample statistics* (such as the sample mean, etc.) from sample to sample can be described *approximately by a normal distribution*. As a result, the expected behaviour of these statistics can be *described*, so we know what to **expect** from the sample *statistic*.

This has been alluded to before. In Sect. 15.4, the sample proportion of red cards in a sample of 15 varied from hand to hand, and was approximately distributed as a normal distribution. This is no accident: **Many sample statistics vary from sample to sample with an approximate normal distribution** if certain conditions are met. This is called the *Central Limit Theorem*.

A *sampling distribution* describes the distribution of the sample statistic: How the value of the sample statistic varies from sample to sample for many samples. The *sampling distribution* here is a normal distribution.

Definition 18.1 (Sampling distribution)**.** A **sampling distribution** is the distribution of some sample statistic, showing how its value varies from sample to sample.

18.2 Sample proportions have a distribution

As with any sample statistic, sample proportions vary from sample to sample (Sect. 15.4); that is, *sampling variation* exists, so the sample proportions have a *sampling distribution*.

Consider a European roulette wheel (Fig. 18.1; the online version has an animation): a ball is spun and can land on any number on the wheel from 0 to 36 (inclusive).

Using the classical approach to probability, the probability of the ball landing on an *odd* number (an '*odd-spin*') is $p = 18/37 = 0.486$. However, if the wheel is spun (say) 15 times, the *sample* proportion of odd-spins $\hat{p}$ can vary. Of course, the *sample* proportion $\hat{p}$ of odd-spins can vary after spinning the wheel 30, 50 or 100 times also. How does it vary from spin to spin?

Computer simulation can be used to demonstrate what happens if the wheel was spun $n = 15$ times, over and over and over again, and the proportion of odd-spins was recorded for each repetition. Clearly, the proportion of odd spins $\hat{p}$ can vary from sample to sample (sampling variation) for $n = 15$ spins, as shown by the histogram (Fig. 18.2, top left panel).

If the wheel was spun (say) $n = 40$ times, something similar occurs (Fig. 18.2, top right panel): the values of $\hat{p}$ vary from sample to sample.

The same process can be repeated for (say) $n = 70$ and $n = 100$ spins (Fig. 18.2, bottom panels). Notice that as the sample size n gets larger, the *distribution* of the values of $\hat{p}$ look more like an approximate normal distribution, and the variation gets smaller.

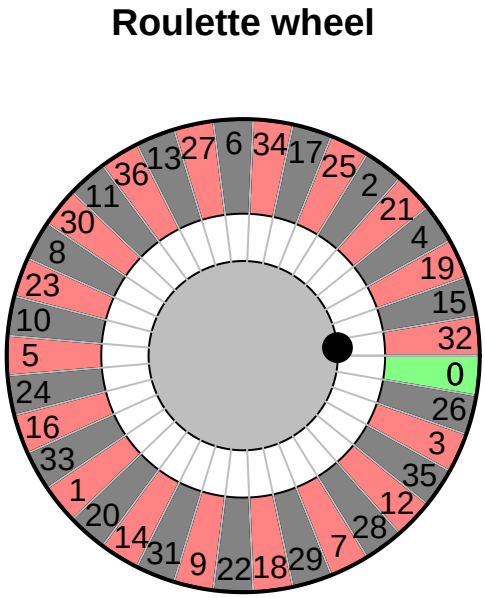


FIGURE 18.1: *A European roulette wheel*

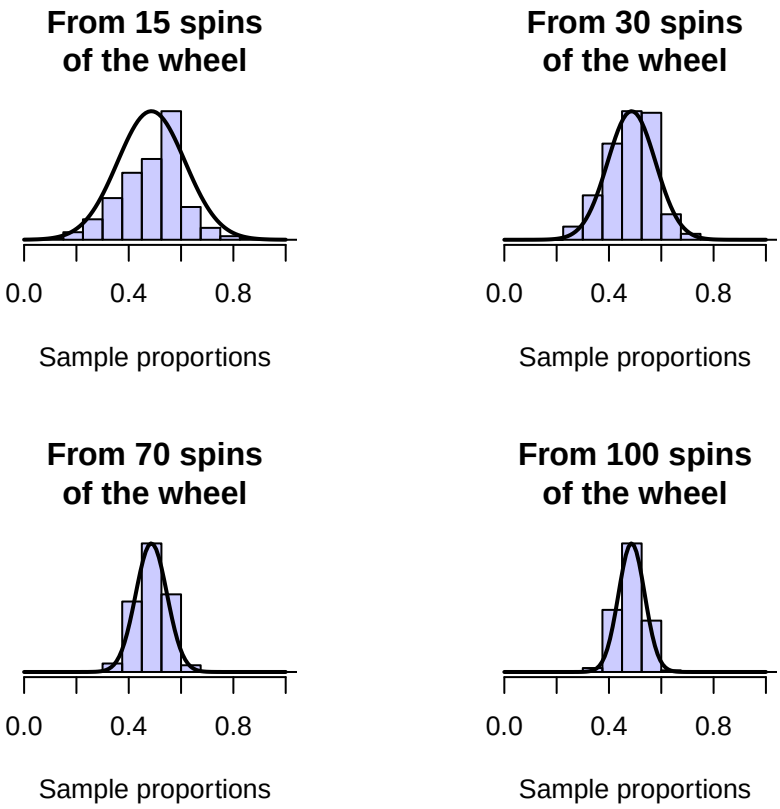


FIGURE 18.2: *Sampling distributions for the proportion of roulette wheel spins that show an odd number*

⚠ The values of the sample proportion vary from sample to sample. The distribution of the possible values of the sample statistic (in this case the *sample* proportion) from sample to sample is called a *sampling distribution*.

Under certain conditions, the sampling distribution of a sample proportion is described by an approximate a normal distribution. In general, the approximation gets better as the sample size gets larger.

18.3 Sample means have a distribution

As with any sample statistic, the sample mean varies from sample to sample (Sect. 15.4) just like sample proportions; that is, *sampling variation* exists, so the sample means have a *sampling distribution*.

Consider a European roulette wheel again (Fig. 18.1).

Rather than recording the sample proportion of odd-spins, the *sample mean* of the numbers spun can be recorded. So, for example, if the wheel is spun (say) 15 times, the *sample* mean of the spins $\bar{x}$ will vary.

Of course, spinning the wheel 30, 50 or 100 times also shows that the *sample* mean $\bar{x}$ can vary too. How much can it vary?

Again, computer simulation can be used to demonstrate what could happen if the wheel was spun 15 times, over and over and over again, and the mean of the spun numbers was recorded for each repetition.

Clearly, the sample mean spin $\bar{x}$ can vary from sample to sample (sampling variation) for $n = 15$ spins, as shown by a histogram (Fig. 18.3, top left panel).

When $n = 15$, the sample mean $\bar{x}$ indeed varies from sample to sample, and the *distribution* of the values of $\bar{x}$ have an approximate normal distribution. If the wheel was spun more than 15 times (say, $n = 50$ times) something similar occurs (Fig. 18.3, top right panel): the values of $\bar{x}$ vary from sample to sample, and the values have an approximate normal distribution. In fact, the values of $\bar{x}$ have a normal distribution for other numbers of spins also (Fig. 18.3, bottom panels).

⚠ The values of the sample mean vary from sample to sample. The distribution of the possible values of a sample statistic, in this case the *sample* mean, is called a *sampling distribution*.

Under certain conditions, the sampling distribution of a sample mean is a normal distribution.

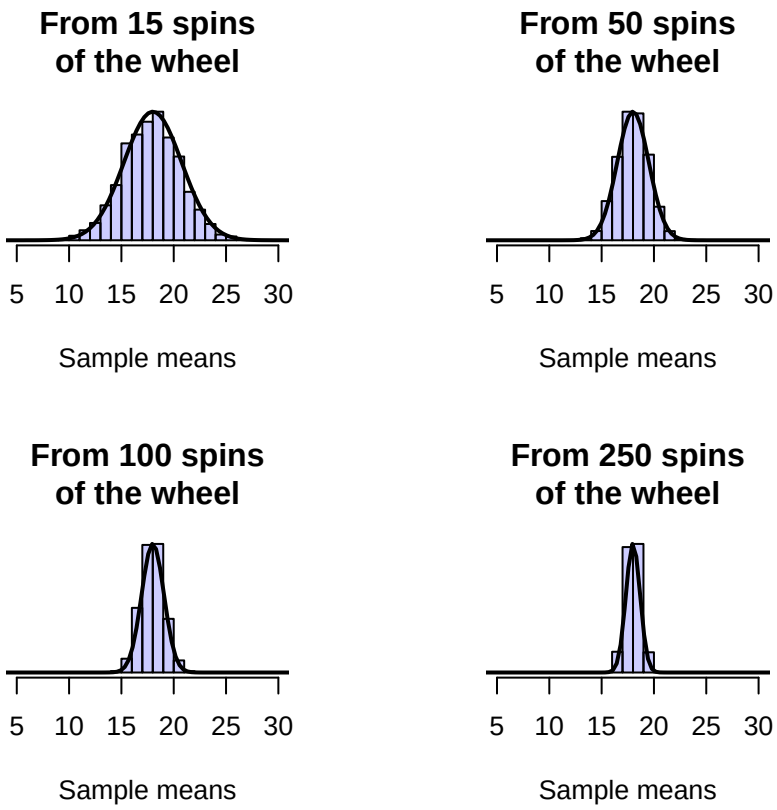


FIGURE 18.3: *Sampling distributions for the mean of the numbers after a roulette wheel spins a certain number of times*

18.4 Standard errors

As we have seen, each sample is likely to be different, so *any* statistic estimated from the sample is likely to be different for each sample. This is called *sampling variation*.

Definition 18.2. Sampling variation refers to how much a sample estimate (a statistic) is likely to vary from sample to sample, because each sample is different.

The value of the sample statistic can vary for every possible sample that we could select, so the actual value of the sample statistic that we observe depends on which sample we have.

That is, all the possible values of sample statistics that we could observe have a distribution (a *sampling distribution*). Perhaps surprisingly, under certain conditions, the sampling distribution is a *normal distribution*.

If the sampling distribution is a normal distribution, it is reasonable to ask what the value of the *standard deviation* of that normal distribution is.

Figs. 18.2 and and 18.3 show that the standard deviation appears to get *smaller* as the sample sizes get *larger*: the sample statistics show less variation for larger n. This makes sense: *larger* samples generally produce more precise estimates. (After all, that's the advantage of using

larger samples: all else being equal, larger samples are preferred as they produce more precise estimates.)

In other words, the sample statistic varies *less* in larger samples: the value of the standard deviation of the sampling distribution is *smaller* for *larger* samples. The standard error is a measure of how precisely the *sample* statistic estimates the *population* parameter.

Example 18.1 (Standard errors)**.** Suppose the sample proportion of odd-spins on the roulette wheel (Sect. 18.2) is estimated as $\hat{p} = 0.51$. If the standard error was 0.01, this estimate is relatively precise: the standard error is very small, which means the value of $\hat{p}$ is not likely to vary greatly from one sample to the next. Any single *estimate* of p is likely to be close to p.

However, if the standard error was 0.2, the estimate of the population proportion is less precise: the standard error is larger, so the value of $\hat{p}$ is likely to vary a lot from one sample to the next. Any single estimate* of p may not be close to p.

Definition 18.3 (Standard error)**.** A *standard error* is the standard deviation of the sampling distribution of a statistic.
Any quantity estimated from a sample has a standard error.

To expand: If every possible sample (found a certain way, and of a given size) was found, and the statistic computed from each sample, the standard deviation of these estimates is the *standard error*.

Recall from Sect. 18.1 that, for many sample statistics, *the variation from sample to sample can be approximately described by a normal distribution* (the *sampling distribution*) if certain conditions are met. Furthermore, the *standard deviation of this normal distribution is the standard error*.

Notice that the standard error is a special type of *standard deviation*; the variation in a sample estimate *from sample to sample*.



The standard error is an unfortunate term: It is not an *error* or *mistake*, or even *standard*. (For example, there is no such thing as a '*non*-standard error.')

18.5 Standard deviation vs. standard error

Even experienced researchers confuse the meaning and the usage of the terms[1] *standard deviation* and *standard error* (Ko et al. 2014), so understanding the difference is important.

The *standard deviation*, in general, quantifies the amount of variation in any variable. Without further qualification, the *standard deviation* quantifies how much individual observations vary from *individual to individual* (for *quantitative* data).

[1] https://retractionwatch.com/2019/12/09/authors-retract-two-studies-on-high-blood-pressure-and-supplements-after-realizing-theyd-made-a-common-error/#more-118562

The *standard error* is a standard deviation that quantifies how much a *sample statistic* varies from *sample to sample*.

Crucially, the standard error *is* a standard deviation, but has a special name to indicate that it is the standard deviation of something very specific.

Any numerical quantity estimated from a sample (a *statistic*) can vary from sample to sample, and so has sampling variation, a sampling distribution, and hence a standard error:

- the *sample* mean $\bar{x}$;
- the *sample* proportion $\hat{p}$;
- the *sample* odds ratio;
- the *sample* median;
- the *sample* standard deviation s;
- etc.



The *standard error* is often abbreviated to 'SE' or 's.e.'

For example, the 'standard error of the sample mean' is written as s.e.$(\bar{x})$, and the 'standard error of the sample proportion' is written as s.e.$(\hat{p})$.

18.6 Summary

A **sampling distribution** describes how a sample statistic is likely to vary from sample to sample. Under certain circumstances, the sampling distribution often can be described by a **normal distribution**. The standard deviation of this normal distribution is called a **standard error**. The standard error is a standard deviation that measures something specific: the variation in the sample statistic *from sample to sample*.

18.7 Quick review questions

1. *Why* is the phrase 'the standard error of the population proportion' inappropriate?
2. *Which* one the following *does not* have a standard error?
3. *Which* one of the following is **true**?
4. True or false: The *standard deviation* is a *standard error* of something quite specific.
5. True or false: Sampling distributions are always *normal* distributions.

18.8 Exercises

Selected answers are available in Sect. D.18.

Exercise 18.1. In the following scenarios, would a *standard deviation* or a *standard error* be the appropriate way to measure the amount of variation? Explain.

1. Researchers are studying the spending habits of customers. They would like to measure the variation in the amount spent by shoppers per transaction at a supermarket.
2. Researchers are studying the time it takes for inner-city office workers to travel to work each morning. They would like to determine the precision with which their estimate (a mean of 47 minutes) has been measured.
3. A study examined the effect of taking a pain-relieving drug on children. The researchers wish to describe the sample they used in the study, including a description of how the ages of the children vary.
4. A study examined the effect of taking a pain-relieving drug in teenagers. The researchers wished to report the percentage of teenagers in the sample that experienced side-effects with some indication of the precision of that estimate.

Exercise 18.2. Which of the following have a *standard error*?

1. The population proportion.
2. The sample median.
3. The sample IQR.
4. The sample standard deviation.
5. The population odds.

Exercise 18.3. A research article made this statement:

> Although [...] samples should always be summarized by the mean and SD [standard deviation], authors often use the standard error of the mean (SEM) to describe the variability of their sample [...] Although the SD and the SEM are related [...], they give two very different types of information.
> — Nagele (2003)

If the standard error of the mean is not used to 'describe the variability of the sample,' then what *is* it used for? How would you explain the difference between the *standard error* and the *standard deviation* to researchers who misuse the terms?

Part VI

Analysis: Confidence intervals

19

Introducing confidence intervals

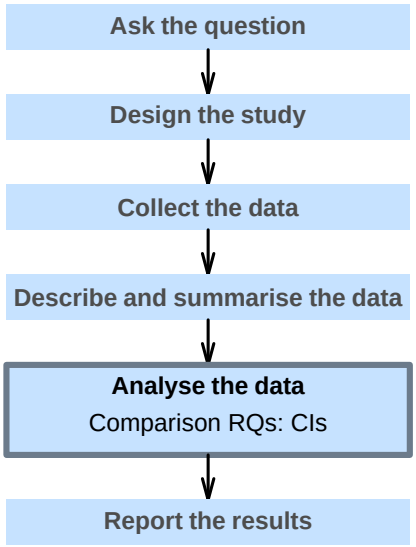
One type of research question, is the estimation-type RQ, when the precision of a *statistic* is of interest. In this Part, answering estimation-type RQs is discussed for:

- Descriptive RQs: Chaps. 20 to 23;
- Relational or interventional RQs with a *comparison*: Chaps. 24 and 25).

Answering estimation-type RQs for relational or interventional RQs with a *connection* is explored later (Chaps. 34 and 35).

The precision of statistics influences decision-type RQ too: when statistics precisely estimate parameters, making decisions about parameters is easier.

The previous chapters, where tools for describing sampling variation were introduced, are used in this part, to understand the precision of statistics.



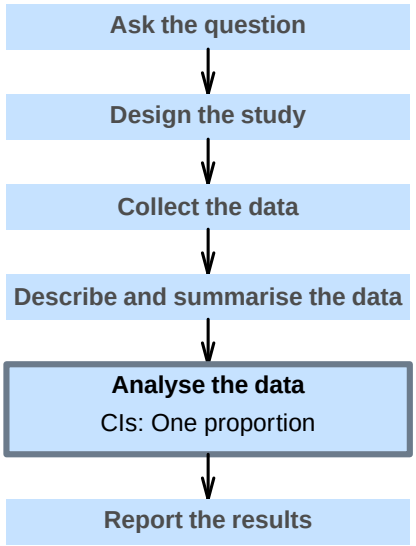
20

Confidence intervals for one proportion

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, and understand the tools of inference.

In this chapter, you will learn about *confidence intervals* for one proportion. You will learn to:

- identify situations where the analysis of one sample proportion is appropriate.
- form confidence intervals for single proportions.
- estimate the sample size needed to estimate a proportion within given constraints.



20.1 Sampling distribution: Known proportion

Suppose a fair, six-sided die[1] is rolled 25 times. What proportion of the rolls will produce an even number? That is, what will be the *sample proportion* of even numbers?

No-one knows exactly what will happen for *any* individual roll, so no-one knows what proportion will be even for any set of 25 rolls.

In addition, the proportion of the 25 rolls that will be even will not be the same for every set of 25 rolls. The sample proportion will *vary*: there is *sampling variation*.

[1] Note the language: two *dice*, but one *die*. *Dice* is plural.

Describing *how* the sample proportion varies from sample to sample is useful (Sect. 15.4.2).

To do this, statistical theory could be used… or thousands of repetitions of a set of 25 rolls could be performed… or a computer could *simulate* many sets of 25 rolls.

Let's simulate rolling a die 25 times, using just 10 sets of 25 rolls; see Fig. 20.1 (the online version has an animation).

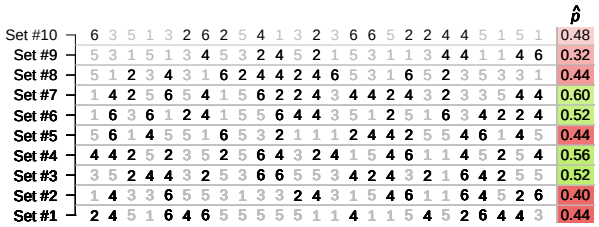


FIGURE 20.1: *The proportion of rolls that are even change from one sample of 25 rolls to the next sample of 25 rolls*

The proportion of rolls that is even varies from set to set. For these 10 sets of $n = 25$ rolls, the percentage of even rolls ranged from $\hat{p} = 0.32$ even rolls to $\hat{p} = 0.60$ even rolls.

The sample proportion of even rolls would be expected to vary around $p = 0.5$, since three of the six faces of the die are even numbers (the *population proportion*), using the classical approach to probability.

Of course, the sample proportion could be very small or very high by chance, but we wouldn't expect to see that very often.

In this example, the *population proportion* of even rolls is known to be $p = 0.5$. Each set of $n = 25$ rolls is a *sample* of all possible sets of $n = 25$ rolls, and the *sample* proportion of even rolls is denoted by $\hat{p}$.

For any set of 25 rolls, the value of $\hat{p}$ will be unknown until we roll the die. The proportion of even rolls is likely to vary from sample to sample; that is, the sample proportions exhibit sampling variation, and the *amount* of sampling variation is quantified using a *standard error*.



p refers to the *population* proportion, and $\hat{p}$ refers to the *sample* proportion.



The symbol $\hat{p}$ is pronounced 'pee-hat.'

Suppose a fair die was rolled 25 times, and this was repeated *thousands* of times (not just 10 sets times, as in Fig. 20.1 above), and the proportion of even rolls was recorded for every set of 25 rolls.

These thousands of sample proportions $\hat{p}$, one from every set of 25 rolls, could be graphed using a histogram; see Fig. 20.2 (the online version has an animation).

The shape of the histogram is roughly a normal distribution. This is no accident: statistical theory says this will happen (when certain conditions are met: see Sect. 20.6).

The possible values of the *sample* proportion $\hat{p}$ have a *sampling distribution* which is roughly a normal distribution; the mean and standard deviation of the normal distribution in Fig. 20.2

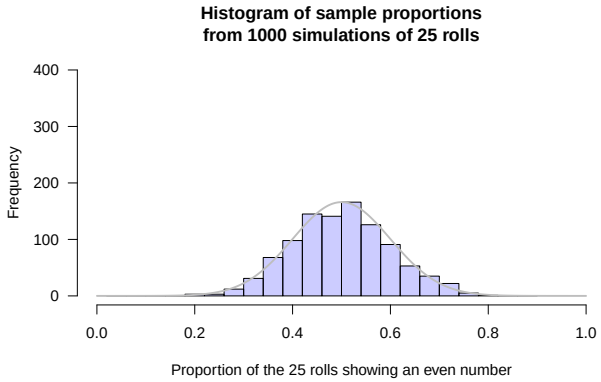


FIGURE 20.2: *The proportion of rolls that are even change from one sample of 25 rolls to the next sample of 25 rolls*

can even be determined. The possible values of the *sample* proportion $\hat{p}$ will have a *sampling distribution*, described by:

- an approximate normal distribution;
- centred around a mean of $p = 0.5$;
- with a standard deviation of 0.1 (where this number comes from will be revealed soon).

This distribution is called a *sampling distribution*, as discussed in Sect. 18.1. The standard deviation of the sampling distribution is called a *standard error*, since it measures how much a sample statistic (in this case, a sample proportion $\hat{p}$) varies from sample to sample.

Since the variation in the sample proportions can be described, a picture of this normal distribution can be drawn (Fig. 20.3). We still don't know *exactly* what we'll find next roll... but we have some idea of *how* the sample proportion is likely to vary in sets of 25 rolls.

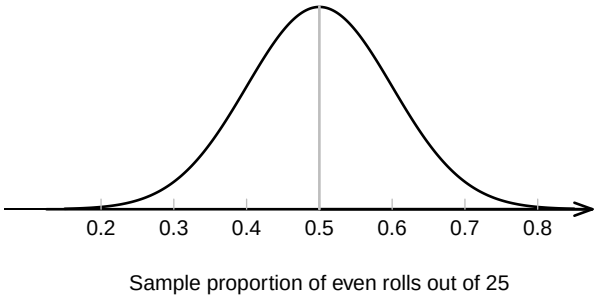


FIGURE 20.3: *The normal distribution, showing how the proportion of even rolls varies when a die is rolled 25 times*



The parameter p and the statistic $\hat{p}$ are both *proportions*, but a *mean* and *standard deviation* are used to describe the *sampling distribution*.

The value of p (the *population* proportion: the proportion of even numbers on the die) remains the same, but the value of $\hat{p}$ (the *sample* proportion: the proportion of even numbers in the sample of 25 rolls) is not the same in every set of 25 rolls. That is, $\hat{p}$ varies, and exhibits *sampling variation*. The variation in $\hat{p}$ from sample to sample is measured by the *standard error of the sample proportion*, written as s.e.$(\hat{p})$.

In general, the **standard error for a sample proportion** when p is known is given by

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{p \times (1-p)}{n}}, \tag{20.1}$$

where n is the number of rolls, and p is the population proportion. For this example, there are $n = 25$ rolls of a die, and the population proportion of even rolls is $p = 0.5$. Then, the standard error of the sample proportion is

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{0.5 \times (1-0.5)}{25}} = 0.1. \tag{20.2}$$

This standard deviation is the standard deviation of the normal distribution in Fig. 20.3.

Recall that the *the standard error is just a special standard deviation*, that measures how much a sample estimate is likely to vary from sample to sample. In that sense, the standard error of the proportion measures how precisely $\hat{p}$ estimates the population proportion p.

Almost always, the value of p is unknown, so when s.e.$(\hat{p})$ is computed, the value of p can't be used. Instead, the best available estimate of p is used, which is $\hat{p}$. This situation is studied from Sect. 20.3 onwards.

Definition 20.1 (Sampling distribution of a sample proportion when p is *known*)**.** When the value of p is *known*, the *sampling distribution of the sample proportion* is described by

- an approximate normal distribution,
- centred around a mean of p,
- with a standard deviation (called the *standard error* of $\hat{p}$) of

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{p \times (1-p)}{n}},$$

when certain conditions are met, where n is the size of the sample, and p is the population proportion.
In general, the approximation gets better as the sample size gets larger.

Think 20.1 (Values for $\hat{p}$)**.** *From the die example, the values of $\hat{p}$ will vary*

- *with an approximate normal distribution;*
- *centred around $p = 0.5$; and*
- *with a standard error of approximately s.e.$(\hat{p}) = 0.1$.*

This distribution is shown in Fig. 20.3. How often would a value of $\hat{p}$ larger than 0.80 be expected?

Answer: Figure 20.3 suggests that, while not impossible, 0.80 or greater will be observed rarely.

20.2 Sampling intervals: Known proportion

The possible values of the sample proportions $\hat{p}$ can be described by an approximate *normal distribution*, as just discussed. This enables the 68–95–99.7 rule to be applied; for example, about 68% of the time with sets of 25 rolls, the sample proportion of even rolls will be between 0.5 give-or-take *one* standard deviation (that is, give-or-take 0.1). So, about 68% of the time, the proportion of even rolls in a set of 25 rolls will be between:

- $0.5 - 0.1 = 0.4$ and
- $0.5 + 0.1 = 0.6$.

Similarly, about 95% of the time, the proportion of even rolls will be between 0.5 give-or-take *two* standard deviations, or between:

- $0.5 - (2 \times 0.1) = 0.3$ and
- $0.5 + (2 \times 0.1) = 0.7$.

This interval tell us what values of $\hat{p}$ are likely to be observed in samples of size 25. Most of the time (i.e., approximately 95% of the time), the value of $\hat{p}$ is expected to be between 0.30 and 0.70. (For instance, in Fig. 20.2, all ten sets of 25 rolls (or 100%) had a sample proportion betweeen 0.30 and 0.70.)

More formally, the sample proportion $\hat{p}$ is likely to lie within the interval

$$p \pm (\text{multiplier} \times \text{s.e.}(\hat{p})),$$

where s.e.$(\hat{p})$ is the *standard error of the sample proportion* (calculated using Eq. (20.1)). The symbol '$\pm$' means 'plus or minus,' or 'give-or-take.'

The *multiplier* depends on how confident we wish to be that the interval contains the value of $\hat{p}$.

For a 95% interval—the most common *level of confidence*—the multiplier is *approximately* 2, based on the 68–95–99.7 rule: Approximately 95% of observations are within *two* standard deviations of the value of p (the mean of the normal distribution in Fig. 20.3).

That is, the *approximate* 95% interval is:

$$p \pm (2 \times \text{s.e.}(\hat{p})).$$

For a 90% interval, either tables or a computer would be used to find the correct multiplier, since the 68–95–99.7 rule isn't helpful.

In practice, 95% intervals are the most common, and we'll use a multiplier of 2 to find an *approximate* 95% interval when computing the interval without using software. Software can be used for any other percentage interval (or for an *exact* 95% interval).

In general, higher confidence means wider intervals (Fig. 20.4), since wider intervals are needed to be more certain that the interval contains $\hat{p}$.

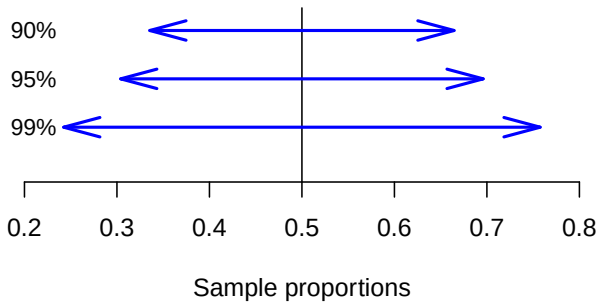


FIGURE 20.4: *To have greater confidence that the interval will include the sample proportion, the interval needs to be wider*

20.3 Sampling distribution: Unknown proportion

In the die example (Sect. 20.1), an equation was given for computing the standard error for the sample proportion for samples of size n, *when the value of* p *was known*.

However, usually the value of p (the *parameter*) is unknown; after all, the reason for taking a sample is to *estimate* the unknown value of p. When p is unknown, the best available estimate can be used, which is $\hat{p}$. *When the value of* p *is unknown*, the standard error of the sample proportion (written s.e.$(\hat{p})$) is approximately

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}.$$

Definition 20.2 (Sampling distribution of a sample proportion when p is *unknown*)**.** When the value of p is *unknown*, the *sampling distribution of the sample proportion* is described by

- an approximate normal distribution,
- centred around the (unknown) mean of p,
- with a standard deviation (called the *standard error* of $\hat{p}$) of

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}, \tag{20.3}$$

when certain conditions are met, where n is the size of the sample, and $\hat{p}$ is the sample proportion.
In general, the approximation gets better as the sample size gets larger.

Let's *pretend* for the moment that the proportion of even rolls of a fair die is unknown (to demonstrate some points). In this case, an *estimate* of the proportion of even rolls can be found by rolling a die $n = 25$ times and computing $\hat{p}$.

Suppose 11 of the $n = 25$ rolls produced an even number, so that $\hat{p} = 11/25 = 0.44$. Then (from Definition 20.2),

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{0.44 \times (1 - 0.44)}{25}} = 0.099277.$$

(This is very similar to the value of 0.1, the value of the standard error when the value of p was known; see Eq. (20.2).)

Hence, the sample proportions would vary with an approximate normal distribution (Fig.20.5), centred around the unknown value of p with a standard deviation of s.e.$(\hat{p}) = 0.099277$.

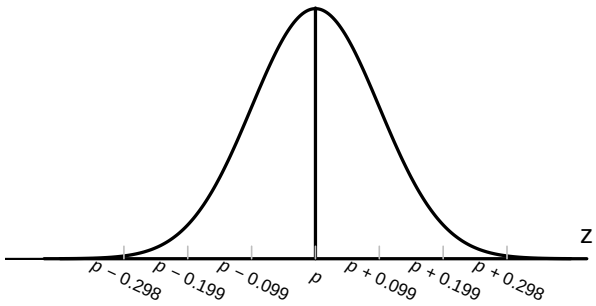


FIGURE 20.5: *The normal distribution, showing how the proportion of even rolls varies when a die is rolled 25 times*

Using the 68–95–99.7 rule again:

> About 95% of the values of $\hat{p}$ are expected to be between $p - 0.199$ and $p + 0.199$.

Though we are pretending the value of p is unknown, the value of $\hat{p}$ is known however. What if the roles of p and $\hat{p}$ were 'reversed?' Then,

> About 95% of the values of p are expected to be between $\hat{p} - 0.199$ and $\hat{p} + 0.199$.

Since $\hat{p} = 0.44$, this is equivalent to:

> About 95% of the values of p are expected to be between 0.24 and 0.64.

This interpretation is not quite correct, but the idea seems reasonable. This is called a *confidence interval* (or CI), based on ideas from Sect. 20.2.

In summary, using $\hat{p} = 0.44$ and s.e.$(\hat{p}) = 0.0993$, the (approximate) 95% CI is

$$0.44 \pm (2 \times 0.0993),$$

or from 0.241 to 0.639. This CI straddles the population proportion of $p = 0.5$, though we would not know this if p truly was unknown.

> ⓘ In this case, we know the value of the population parameter: $p = 0.5$.
>
> Usually we do *not* know the value of the parameter: that's why we are taking a sample.

20.4 Confidence intervals: Unknown proportion

Suppose *thousands* of people rolled a die 25 times, and *each* person found $\hat{p}$ for their sample, and hence computed the CI for their sample of 25 rolls.

Every sample of 25 rolls could produce a different estimate $\hat{p}$, and so a different value for s.e.$(\hat{p})$, and hence a different 95% CI. However, *about 95% of these thousands of confidence intervals from those thousands of repetitions would straddle the true proportion* p.

Since we usually don't know the value of p, and we usually only have one sample (and hence one CI), in general *we never know whether the single CI computed from our single sample straddles* p *or not*.

Again, consider letting the computer *simulate* the situation. Suppose the process of recording the sample proportion of even numbers in $n = 25$ rolls is repeated 50 times, and for each of those 50 sets of 25 rolls a CI is produced (Fig. 20.6; the online version has an animation).

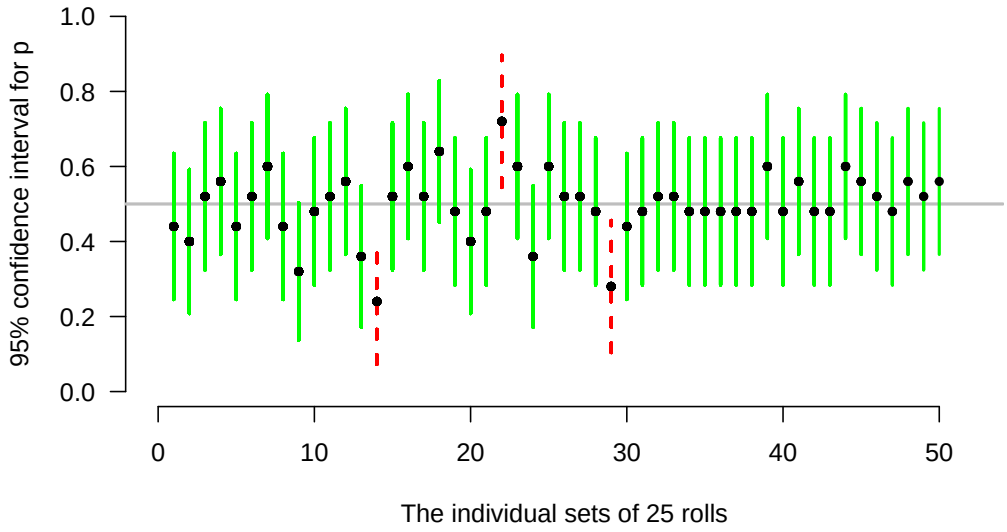


FIGURE 20.6: *About 95% of CIs contain the population proportion*

Most of those CIs straddle the population proportion of $p = 0.5$ (shown as solid lines)... but some do not (shown as dashed lines). Of course, since the value of p is usually unknown, we never know if our CI contains p or not.

Definition 20.3 (Confidence interval)**.** A *confidence interval* is an interval in which the population *parameter* is likely to be contained, if we found many samples the same way.
If a 95% confidence interval (or CI) is computed from each sample, about 95% of the CIs would straddle the *parameter* of interest. This interval is called a *confidence interval*.

In general, a CI for the population proportion p is found using

$$\hat{p} \pm (\text{multiplier} \times \text{s.e.}(\hat{p})),$$

where the multiplier is 2 for an *approximate* 95% CI (based on the 68–95–99.7 rule).

Definition 20.4 (Confidence interval for p)**.** A *confidence interval* (CI) for the unknown value of the parameter p is

$$\hat{p} \pm (\text{multiplier} \times \text{s.e.}(\hat{p})), \tag{20.4}$$

where

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

is the *standard error* of $\hat{p}$, $\hat{p}$ is the sample proportion, and n is the sample size. For an *approximate* 95% CI, the multiplier is 2.

In general, *higher* confidence levels means *wider* intervals: To be *more* confident that the interval straddles the unknown value of p, *wider* intervals are needed (Fig. 20.7) to cover more possibilities.

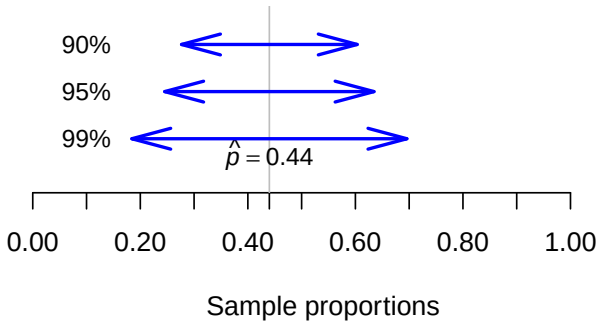


FIGURE 20.7: *To have greater confidence that the interval will straddle the population proportion, the interval needs to be wider*

Example 20.1 (Energy drinks in Canadian youth)**.** A study of young Canadians aged 12–24 (Hammond et al. 2018) found that 365 of the 1516 respondents reported sleeping difficulties after consuming energy drinks.

The unknown parameter is p, the *population* proportion of young Canadians reporting sleeping difficulties.

The sample proportion reporting sleeping difficulties after consuming energy drinks is $\hat{p} = 365/1516 = 0.241$. As usual, the sample proportion would vary from one sample of size $n = 1516$ to another; *sampling variation* exists. The *standard error* (Definition 20.4) quantifies how much the sample proportion is likely to vary from sample to sample:

$$\begin{aligned}\text{s.e.}(\hat{p}) &= \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\ &= \sqrt{\frac{0.241 \times (1 - 0.241)}{1516}} = 0.01098449,\end{aligned}$$

or about 0.011. So, in samples of size 1516, the approximate 95% CI (Definition 20.4) is between

- $0.241 - (2 \times 0.01098449) = 0.2190$ and

- $0.241 + (2 \times 0.01098449) = 0.2627$.

The approximate 95% CI is from 0.219 to 0.263.

This CI may or may not straddle the population proportion p; it is *likely* that the interval straddles the value of p. In other words, it is plausible that the sample proportion of $p = 0.241$ may have come from a population with a proportion somewhere between 0.219 and 0.263.

⚠ Notice that many decimal places are used in the working, but final answers are rounded.

Example 20.2 (Koalas crossing roads)**.** A study of koalas (Dexter et al. 2018) found that 18 of the $n = 51$ koalas studied in a certain area (over 30 months) had crossed at least one road during that time.

The unknown parameter is p, the *population* proportion of koalas that had crossed at least one road over the 30 months.

The sample proportion having crossed a road is $\hat{p} = 18/51 = 0.3529$. The standard error (Definition 20.4) is

$$\begin{aligned} \text{s.e.}(\hat{p}) &= \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} \\ &= \sqrt{\frac{0.3529 \times (1 - 0.3529)}{51}} \\ &= 0.06692. \end{aligned}$$

An approximate 95% CI, then, is $0.3529 \pm (2 \times 0.06692)$, or

$$0.3529 \pm 0.1338.$$

The *margin of error* is 0.1338.

Computing the 'plus' and the 'minus' bits, the approximate 95% CI is from 0.219 to 0.487 (after rounding appropriately).

The approximate 95% CI for the population proportion of koalas that crossed at least one road in the last 30 months is from 0.219 to 0.487. That is, it is plausible that the sample proportion of $\hat{p} = 0.3529$ may have come from a population with a proportion somewhere between 0.219 and 0.487.

The research article reports:

> Of the 51 koalas, 18 (35.3%) crossed at least one road. The [...] probability of a koala crossing at least one road during the study was 35.3% (95% CI = 22–48%).
> — Dexter et al. (2018), p. 70.

This agrees with our calculations.

Example 20.3 (CI)**.** A study of how paramedics administer pain medication (Lord et al. 2009), and to whom, found that

> Forty-five percent of patients reporting pain did not receive analgesia (791/1766) (95% confidence interval [CI], 43%-47%).
> — Lord et al. (2009), p. 525

That is, $\hat{p} = 791/1766 = 0.4479049$ and $n = 1766$.

Hence,

$$\text{s.e.}(\hat{p}) = 0.01183326,$$

so the *approximate* 95% CI is from

$$0.4479049 - (2 \times 0.01183326) = 0.4242384$$

to

$$0.4479049 + (2 \times 0.01183326) = 0.47157134,$$

or from 0.424 to 0.472, which agrees with the article.

(Notice that many decimal places were kept in the working, but the final answers were rounded.)

20.5 Interpretation of a CI

The *correct* interpretation (Definition 20.3) of a 95% CI is the following:

> *If* samples were repeatedly taken many times, and the 95% confidence interval computed for each sample, 95% of these confidence intervals formed would contain the population *parameter*.

However, most people would think of our 95% CI as having a 95% chance of containing the value of population parameter p. This is not strictly correct (our CI either *does* or *does not* contain the value of p), but is common.

More details on interpreting a CI are given in Sect. 21.2.

20.6 Statistical validity conditions

The histogram in Sect. 20.1, shows the proportion of $n = 25$ rolls that were even for many samples; it has an approximate normal distribution. Because of this, the 68–95–99.7 rule could be used to form the approximate 95% CIs.

However, the distribution of the sample proportions only looks like a normal distribution under certain conditions. Certain conditions must be true for the calculations to be sensible, or **statistically valid**.

Definition 20.5 (Statistical validity)**.** A result is *statistically valid* if the conditions for the underlying mathematical calculations and assumptions to be approximately correct are met. Every confidence interval has statistical validity conditions.

Example 20.4 (Statistical validity analogy)**.** Suppose your doctor asks you to get a blood test, after fasting (refraining from eating) for 12 hours before your blood test.

After leaving the doctor, you proceed to a restaurant for dinner. You start the next day with a hearty breakfast, have lunch at a beach-side cafe, and then go for your blood test. Your blood is extracted, the blood is analysed in the pathology lab, and your doctor is emailed the results of the blood test.

However, since you did not fast as required, the results may or may not be valid. The doctor can still learn something... but not as much as if you had followed the instructions.

Similarly, if the conditions for computing the confidence interval are not met, the results may be suspect.

The *statistical validity conditions* for creating CI for a single proportion is that:

- the number of individuals in the group of interest must exceed 5, **and**
- the number of individuals in the group *not* of interest must exceed 5.

These conditions ensure that the sampling distribution of $\hat{p}$ has an approximate normal distribution, so that the 68–95–99.7 rule (approximately) applies. If this condition is not met, the sampling distribution may not have normal distribution, so the 68–95–99.7 rule (used to create the CI) maybe inappropriate, and so the CI may also be inappropriate.

In addition to the statistical validity condition, the CI will be:

- **internally valid** if the study was well designed; and
- **externally valid** if the the sample is a simple random sample and is internally valid.

Example 20.5 (Energy drinks in Canadian youth)**.** In Example 20.1, the approximate 95% CI was from 0.192 to 0.236. This confidence interval for the sample proportion will be *statistically* valid if:

- the number of youth in the sample who experienced sleeping difficulties exceeds 5; **and**
- the number of youth in the sample who *didn't* experience sleeping difficulties exceeds 5.

The number of youth experiencing sleeping difficulties was 365, which is more than five. The number of youth *not* experiencing sleeping difficulties was $1516 - 365 = 1151$, which is also more than five. Hence, the CI is *statistically* valid.

In addition, the CI will be *internally* valid if the study was well designed, and will be *externally* valid if the sample is a simple random sample from the population and is internally valid.

Think 20.2 (Koalas crossing roads)**.** *Consider Example 20.2, about koalas crossing roads. Is the CI likely to be statistically, internally and externally valid?*

Example 20.6 (Statistical validity)**.** Consider an artificial situation to estimate the proportion of die rolls that show as a **one**. The population proportion (using the classical approach to probability) is 1/6, or about 0.167.

If we repeatedly rolled a die in sets of $n = 20$ rolls, say 5000 times, the proportion of rolls that showed as **one** could be recorded for each set of 20 rolls. Then, a histogram of the sample proportions could be produced. Using a computer to simulate this, a histogram of the sample proportions is shown in the top panel of Fig. 20.8. The normal distribution does a poor job of describing the sampling distribution (the distribution is not even symmetric). The statistical validity conditions do *not* seem satisfied.

Alternatively, we could repeatedly roll a die in sets of $n = 60$ rolls, say 5000 times, and record the proportion of rolls that show as **one** for each set of 60 rolls. Then, a histogram of the proportion of **ones** for those sets of 60 rolls could be produced. Using a computer to simulate this, a histogram of these proportions is shown in the bottom panel of Fig. 20.8. The normal distribution does a reasonable job of describing the sampling distribution. The statistical validity conditions seem satisfied.

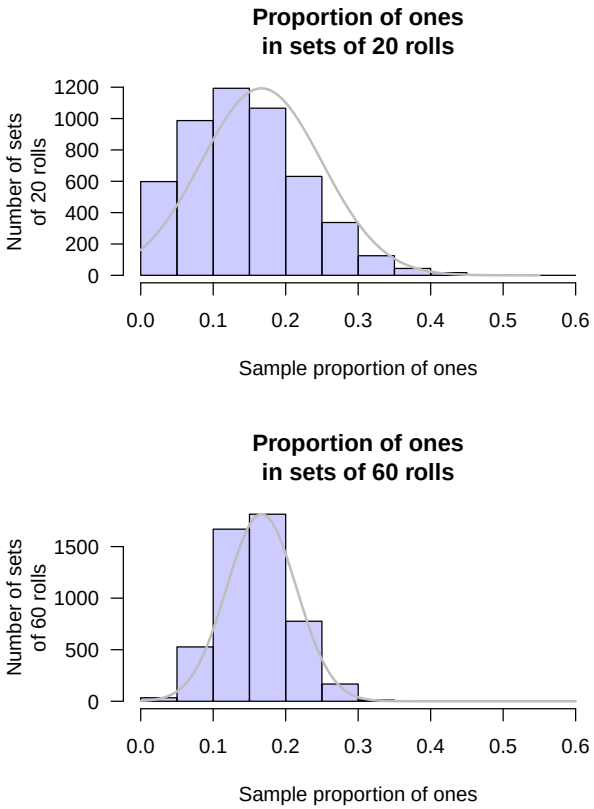


FIGURE 20.8: *The sampling distribution of the proportion of ones rolled, for sets of 20 rolls (top panel) and sets of 60 rolls (bottom panel)*

20.7 Summary: Finding a CI for p

The procedure for computing a confidence interval (CI) for a proportion is:

- Compute the sample proportion, $\hat{p}$, and identify the sample size n.
- Compute the standard error, which quantifies how much the value of $\hat{p}$ varies from one sample to the next:

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}.$$

- Find the multiplier: this is 2 for an approximate 95% CI using the 68–95–99.7 rule. (Note: (Multiplier$\times$standard error) is called the *margin of error*.)
- Compute:

$$\hat{p} \pm (\text{Multiplier} \times \text{standard error}) .$$

- Check the statistical validity conditions are satisfied.



You must use *proportions* in this formula, **not** *percentages* (that is, values like 0.23 and **not** 23%).

Example 20.7 (NHANES data)**.** For the NHANES data, first seen in Sect. 12.10, the unknown parameter is p, the *population* proportion of Americans that currently smoke.

In the study, 1466 out of the 3211 respondents who reported their smoking status said they currently smoked: $\hat{p} = 1466 \div 3211 = 0.4566$.

What is the *population* proportion p that currently smoke? We don't know, and the estimate of p from every sample is likely to be different. The standard error is $\text{s.e.}(\hat{p}) = 0.00879$, so the approximate 95% CI for p is 0.4566 ± 0.01758, or from 0.439 to 0.474. *(Check the calculations!)*

For the conclusions to be statistically valid, the number of smokers must exceed 5, **and** the number of non-smokers must exceed 5. Both are true. The CI appears to be statistically valid.

We write:

> Based on the *sample*, we are approximately 95% confident that the interval from from 0.429 to 0.474 straddles the *population* proportion of smokers in the USA.

20.8 Estimating sample sizes: one proportion

For a given level of confidence, the width of a CI depends on the size of the sample. In general, larger samples produce more precise estimates of the parameter (Sect. 5.2), and hence narrower CIs.

Suppose we want our 95% CI for the proportion of smokers (Example 20.7) to be precise to give-or-take 0.01 (rather than the ± 0.018 found from the sample): what size sample is needed? Since we seek a *more* precise estimate, we'd expect to need a *larger* sample... but how much larger?

Conservatively, the sample size for a 95% CI needed is *at least*

$$\frac{1}{(\text{Margin of error})^2}.$$

That is, a sample size of at least $\dfrac{1}{0.01^2} = 10\,000$ Americans is needed.

Example 20.8 (Sample size calculations)**.** To estimate the population proportion of Australians that smoke, to within 0.07 with 95% confidence, a sample size of at least

$$\frac{1}{(\text{Margin of error})^2} = \frac{1}{0.07^2}$$

is needed; *at least* $n = 204.0816$ people.

In practice, *at least* 205 people are needed to achieve this desired level of precision (that is, **always round up** in sample size calulations).

⚠ Always **round up** the result of the sample size calculation.

20.9 Example: Female coffee drinkers

A study of 360 female college students in the United States (Kelpin et al. 2018) found that 61 drank coffee daily.

The unknown parameter is p, the *population* proportion of female college students in the United States that drink coffee daily.

The sample size is $n = 360$, and the *sample* proportion of daily coffee drinkers is $\hat{p} = 61/360 = 0.16944$. Another sample of 360 students from the same population is likely to produce a different sample proportion $\hat{p}$ of daily coffee drinkers: the sample proportion has *sampling variation*. The size of this sampling variation is quantified using a *standard error*; from (20.3):

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{0.16944 \times (1 - 0.16944)}{360}} = 0.01977.$$

An *approximate* 95% CI is $0.1694 \pm (2 \times 0.01977)$, or 0.1694 ± 0.03954. That is, the margin of error is 0.03954.

Computing the 'plus' and the 'minus' bits, the approximate 95% CI is from $0.1694 - 0.03954 = 0.12986$ to $0.1694 + 0.03954 = 0.20894$. Round appropriately, the approximate 95% CI is from 0.130 to 0.209.

The plausible values for p that may have led to this value of $\hat{p} = 0.1694$ are between 0.130 and 0.209. (This CI may or may not contain the true proportion p.)

This CI is *statistically* valid. We cannot comment on the internal validity: we would need details of how the study was conducted.

The CI is *externally* valid if the sample is simple random sample of some population, and the study is internally valid. The CI is approximately *externally* valid if the sample is somewhat representative of some population, and the study is internally valid.

20.10 Quick review questions

1. True or false: p is called a *parameter*.
2. True or false: The value of p will vary from sample to sample.
3. True or false: The *standard error* refers to the sampling variation in p.
4. Suppose $n = 50$ and $\hat{p} = 0.4$. What is the standard error?

20.11 Exercises

Selected answers are available in Sect. D.19.

Exercise 20.1. A study of salt intake in the United Kingdom (Sutherland et al. 2012) found that 2,182 out of the 6,882 people sampled in 2007 'generally added salt at the table.'

Find an approximate 95% CI for the true proportion of Britons that generally add salt at the table.

Exercise 20.2. A study of the eating habits of university students in Canada (Mann and Blotnicky 2017) found that 8 students out of 154 met the recommendation for eating a sufficient number of servings of grains each day.

1. Find an approximate 95% CI for the true proportion of Canadian students that meet the recommendation for eating a sufficient number of servings of grains each day.

2. Would these results be likely to apply to Australian university students? Why or why not?

Exercise 20.3. A meta-study of hiccups (Lee et al. 2016a) found that, of 864 patients examined (across many different studies) who had hiccups, 708 were male.

1. Find an approximate 95% CI for the true proportion of people with hiccups who are male.
2. Check if the statistical validity conditions are met or not.
3. Draw a sketch of how the sample proportion varies from sample to sample for samples of size 864.

Exercise 20.4. We wish to estimate the population proportion of Australians that smoke.

1. Suppose we wish our 95% CI to be give-or-take 0.05. How many Australians would we need to survey?
2. Suppose we wish our 95% CI to be give-or-take 0.025; that is, we wish to *halve* the width of the interval above. How many Australians would we need to survey?
3. How many *times* as many Australians are needed to *halve* the width of the interval?

Exercise 20.5. A study of turbine failures (Nelson 1982; Myers et al. 2002) ran 42 turbines for around 3000 hours, and found that nine developed fissures (small cracks). Find a 95% CI for the true proportion of turbines that would develop fissures after 3000 hours of use. Are the statistical validity conditions satisfied?

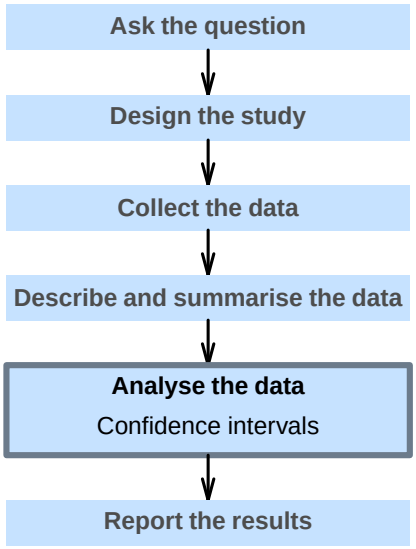
The study also ran 39 turbines for around 400 hours, and found that zero developed fissures. Find a 95% CI for the true proportion of turbines that would develop fissures after 400 hours of use. Are the statistical validity conditions satisfied?

21

More about forming CIs

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, and understand the tools of inference.

In this chapter, you will learn more about *confidence intervals*.



21.1 General comments

The previous chapter discussed forming a confidence interval (CI) for one proportion. We will also study CIs in other contexts too.

The following applies to *all* CIs:

- CIs are formed for the unknown *population* parameter (such as the population proportion p), based on a *sample* statistic (such as the sample proportion $\hat{p}$).
- CIs give an interval in which the sample statistic is likely to lie, over repeated sampling.
- Loosely speaking, this is usually interpreted as the CIs giving an interval which is likely to straddle the value of the unknown *population* quantity. That is, the CI gives an interval of plausible values of the population parameter that may have produced the observed sample statistic.

- Most CIs have the form

$$\text{Statistic} \pm \overbrace{(\text{Multiplier} \times \text{standard error})}^{\text{Called the 'margin of error'}}.$$

- The *multiplier* is *approximately* 2 for a 95% CI (from the 68–95–99.7 rule).
- The *margin of error* is (Multiplier × standard error).
- The statistical conditions should always be checked to see if the CI is (at least approximately) statistically valid.

21.2 Interpretation of a CI

Interpreting CIs correctly is tricky. The *correct* interpretation (Definition 20.3) of a 95% CI is the following:

> *If* samples were repeatedly taken many times, and the 95% confidence interval computed for each sample, 95% of these confidence intervals formed would contain the population *parameter*.

This is the idea shown in Fig. 20.6. In practice, this definition is unsatisfying, since we almost always have only *one* sample. And since the value of the parameter is unknown (after all, we went to the bother of taking a sample so we could *estimate* the value of the parameter), we don't know if *our* CI includes the population parameter or not.

A reasonable alternative interpretation is:

> The interval gives a range of values of the parameter that could plausibly (with 95% confidence) have given rise to our observed value of the statistic.

Or we might say that:

> There is a 95% chance that our computed CI straddles the value of the population parameter.

These alternatives are not absolutely correct, but are reasonable interpretations.

Many people will write—and you will see it written in many places—that the CI means that there is a 95% chance that the CI contains the population parameter. This is not strictly correct, but is common (probably because it is easier to understand).

I use this analogy: Most people say the sun rises in the east. This is incorrect: the sun doesn't *rise* at all. It *appears* to rise in the east because the earth rotates on its axis. But almost everyone says that the 'sun rises in the east,' and for most circumstances this is fine and serviceable, even though technically incorrect.

Similarly, most people use the final interpretation above for a CI in practice, even though it is technically incorrect.

> **Example 21.1** (Energy drinks in Canadian youth)**.** In Example 20.1, the approximate 95% CI was from 0.192 to 0.236 The correct interpretation is:

> If we took many samples of 1516 Canadian youth, and computed the approximate 95% CI for each one, about 95% of those CIs would contain the population proportion.

We don't know if *our* CI includes the value of p, however. We might say:

> This 95% CI is likely to straddle the actual value of p.

or

> The range of values of p that could plausibly (with 95% confidence) have produced $\hat{p} = 0.241$ is between 0.192 and 0.236.

In practice, the CI is usually interpreted as saying:

> There is a 95% chance that the population proportion of Canadian youth who have experienced sleeping difficulties after consuming energy drinks is between 0.192 to 0.236.

This is not strictly correct, but is commonly used.

Think 21.1 (Interpretation of a CI)**.** *In Example 20.2 about koalas crossing roads, the approximate 95% CI was from 0.130 to 0.209.*

What is the correct interpretation of this CI?

21.3 Validity and confidence intervals

When constructing confidence intervals, certain *statistical validity conditions* must be true; these ensure that the sampling distribution is sufficiently close to a normal distribution for the 68–95–99.7 rule rule to apply.

If these conditions are *not* met, the sampling distribution may not be normally distributed, so the 68–95–99.7 rule (on which the CI is based) maybe inappropriate, so the CI itself may also be inappropriate.

In addition to the statistical validity condition, the *internal validity* and *external validity* of the study should be discussed also (Fig. 21.1).

Regarding *external validity*, all the CI computations in this book assume a *simple random sample*. If the sample is from a *random* sampling method, but not from a *simple random sample*, then methods exist for producing CIs that are externally valid, but are more complicated than those described in this book.

If the sample is a non-random sample, then the CI may be reasonable for the quite specific population that *is* represented by the sample; however, the sample probably does not represent the more general population that is probably intended.

Externally validity requires that a study is also internally valid. *Internal validity* can only be discussed if details are known about the study design.

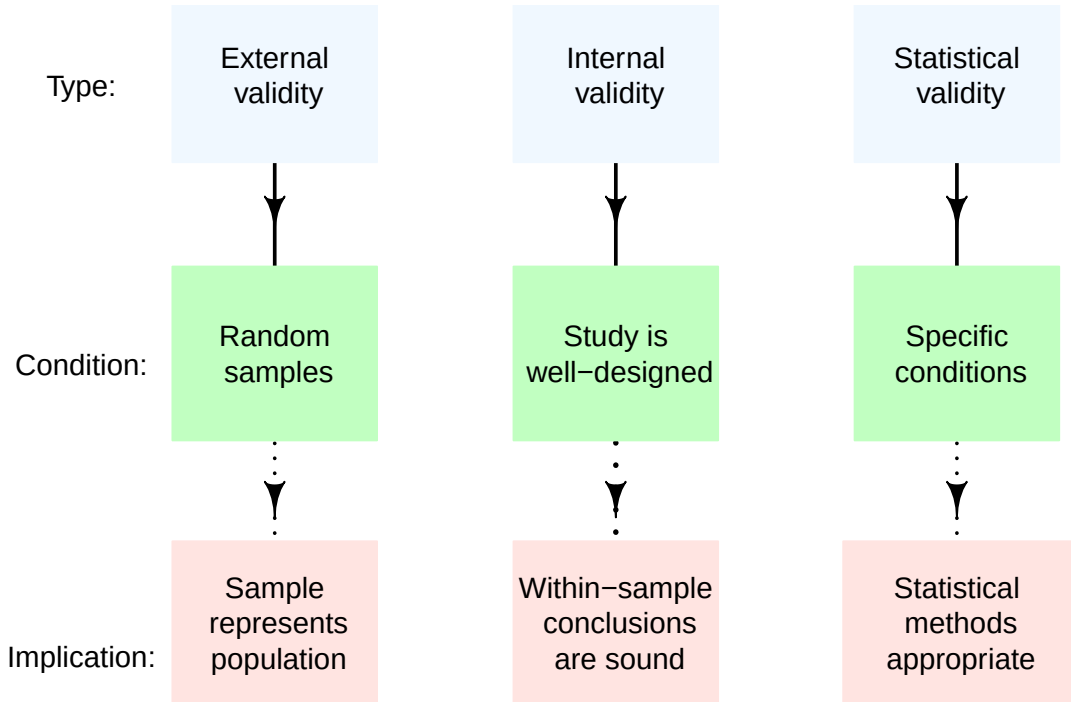


FIGURE 21.1: *Three types of validities for studies.*

In addition, CIs also require that the sample size is less than 10% of the population size; however this is almost always the case.

21.4 Quick revision exercises

1. True or false: CIs always have 95% confidence.
2. True or false: The statistical validity conditions concern *external* validity.
3. True or false: CIs give intervals in which the value of a *population* parameter will fall.
4. True or false: All other things being equal, a 95% CI is *wider* than a 90% CI.
5. The multiplier times the standard error is called the
6. A CI gives an interval in which we are fairly sure that the value of the

is within.

21.5 Exercises

Selected answers are available in Sect. D.20.

Exercise 21.1. A researcher was computing a 95% CI for a single proportion to estimate the proportion of trees with apple scabe (Hirst and Stedman 1962), and found that $\hat{p} = 0.314$ and s.e.$(\hat{p}) = 0.091$.

What is wrong with the following conclusion that the researcher made?

> The approximate 95% CI for the sample proportion is between 0.223 and 0.405.

Exercise 21.2. A researcher was computing a 95% CI for a single proportion to estimate the proportion of trees with apple scabe (Hirst and Stedman 1962), and found that $\hat{p} = 0.314$ and s.e.$(\hat{p}) = 0.091$.

What is wrong with the following conclusion that the researcher made?

> This CI means we are 95% confident that between 22.3 and 40.5 trees are infected with apple scab.

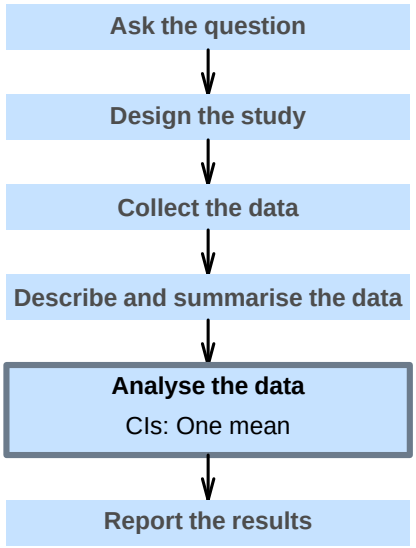
22

Confidence intervals for one mean

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, and understand the tools of inference.

In this chapter, you will learn about *confidence intervals* for one mean. You will learn to:

- produce confidence intervals for one mean.
- determine whether the conditions for using the confidence intervals apply in a given situation.
- compute sample size estimates in this situation.



22.1 Sampling distribution: One mean with population standard deviation known

In this chapter, we study the situation where a population mean μ (the parameter) is estimated by a sample mean $\bar{x}$ (the statistic).

Of course, every sample is likely to be different, and is likely to produce a different sample mean $\bar{x}$. That is, the value of the sample mean will vary from sample to sample and exhibit *sampling variation* (which can be quantified using the *standard error*).

Consider rolling dice again. Suppose a die is rolled $n = 25$ times, and the *mean* of the 25

numbers that are rolled is recorded. What will be the sample mean of the numbers in the 25 rolls?

The sample mean will vary from sample to sample, (*sampling variation*). Since every face of the die is equally likely to appear on any one roll, the population mean of all possible rolls is $\mu = 3.5$ (in the middle of the numbers on the faces of the die, which is also the *median*).

An example of the mean after repeatedly rolling a die 25 times is shown in Fig. 22.1 for 10 sets of 25 rolls. (The online version has an animation.) The mean of the 25 rolls clearly varies. In the simulation, the sample mean of 25 rolls was as low as 3.08 and as high as 3.76.

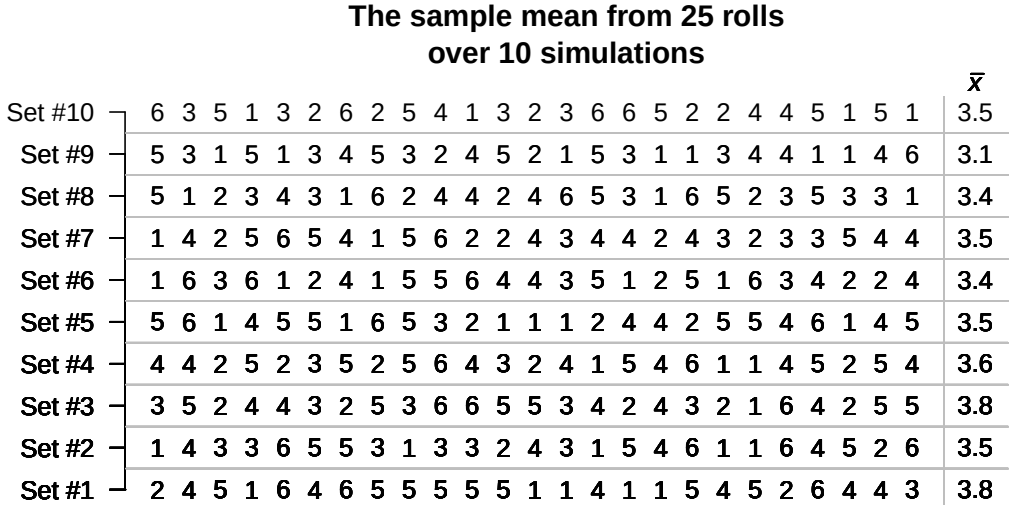


FIGURE 22.1: *Rolling dice: The average of 25 rolls, for 10 sets of 25 rolls*

The mean for any single *sample* of $n = 25$ rolls will sometimes be higher than $\mu = 3.5$, and sometimes lower than $\mu = 3.5$, but most of the time the mean should be close to 3.5.

If many people made a set of 25 rolls, and computed the mean for their set, every person would have a sample mean for their set of 25 rolls, and we could produce a histogram of all these sample means; see Fig. 22.2. (The online version has an animation.)

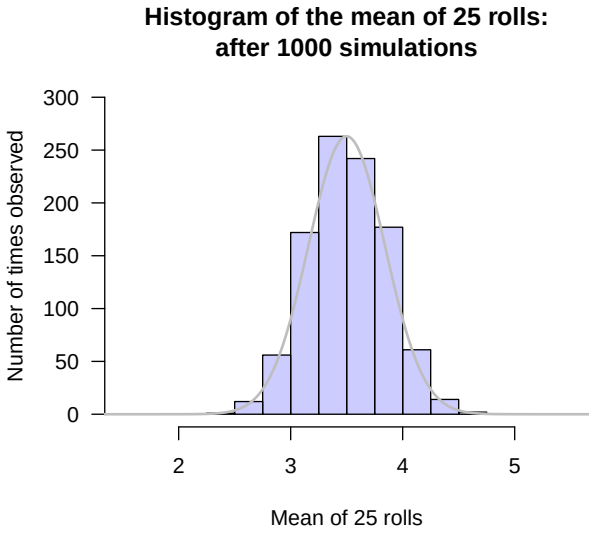


FIGURE 22.2: *Rolling dice: The mean of 25 rolls, for 1000 repetitions*

From Fig. 22.2, the sample means appear to vary with an approximate normal distribution (as we saw with the sample proportions). This normal distribution is centred around the population mean μ. The standard deviation of the normal distribution is the *standard error* of the sample mean $\bar{x}$, written as s.e.$(\bar{x})$.

When the *population* standard deviation σ is *known*, then

$$\text{s.e.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

So the possible values of the sample means have a *sampling distribution* described by:

- an approximate normal distribution,
- with mean μ, and
- a standard deviation, called the standard error, of s.e.$(\bar{x}) = \sigma/\sqrt{n}$.

Usually the population mean and the population standard deviation are *unknown*. Nonetheless, because the sampling distribution has an approximate normal distribution, the 68–95–99.7 rule can be applied: approximately 95% of the sample means are expected to be within two standard errors of μ.

22.2 Sampling distribution: One mean with population standard deviation unknown

When a sample mean is used to estimate a population mean, the sample mean will vary from sample to sample: sampling variation exists, as we saw in the previous section.

When we do not know the *population* standard deviation σ (which is almost always the case), we estimate it using the *sample* standard deviation s. Then, the *standard error* of the sample mean is s.e.$(\bar{x}) = \frac{s}{\sqrt{n}}$. With this information, we can describe the *sampling distribution of the sample mean*.

Definition 22.1 (Sampling distribution of a sample mean)**.** When the *population* standard deviation is unknown, the *sampling distribution of the sample mean* is described by:

- an approximate a normal distribution,
- centred around μ,
- with a standard deviation (called the *standard error of the mean*) of

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}}, \tag{22.1}$$

when certain conditions are met, where n is the size of the sample, and s is the standard deviation of the individual observations in the sample (that is, the *sample* standard deviation).

22.3 Confidence intervals: One mean

We don't know the value of μ (the paremeter), the population mean, but we have an *estimate*: the value of $\bar{x}$, the sample mean (the statistic). The actual value of μ might be a bit larger than $\bar{x}$, or a bit smaller than $\bar{x}$; that is, μ is probably about $\bar{x}$, give-or-take a bit.

Furthermore, we have seen that the values of $\bar{x}$ vary from sample to sample (*sampling variation*), and noted that they vary with an approximate normal distribution. So, using the 68–95–99.7 rule, we could create an approximate 95% interval for the plausible values of μ that may have given the observed values of the sample mean. This is a *confidence interval*.

A confidence interval (CI) for the population mean is an interval surrounding a sample mean. In general, an approximate 95% confidence interval (CI) for μ is $\bar{x}$ give-or-take about two standard errors. In general, the confidence interval (CI) for μ is

$$\bar{x} \pm \overbrace{(\text{Multiplier} \times \text{s.e.}(\bar{x}))}^{\text{Called the 'margin of error'}}.$$

For an approximate 95% CI, the multiplier is, as usual, about 2 (since about 95% of values are within two standard deviations of the mean from the 68–95–99.7 rule).

We often find 95% CIs, but we can find a CI with *any* level of confidence: we just need a different multiplier. We'll just use a multiplier of 2 (and hence find *approximate* 95% CIs), and otherwise use software. Commonly, CIs are computed at 90%, 95% and 99% confidence levels.

The multiplier of 2 is not a z-score here. The multiplier would be a z-score *if* we knew the value of σ; since we don't, the multiplier is a t-score and not a z-score The t- and z-multipliers are very similar, and (except for very small sample sizes) using an approximate multiplier of 2 is reasonable for computing *approximate* 95% CIs in either case. We'll let software handle the specifics.

If we collected many samples of a specific size, $\bar{x}$ and s would be different for each sample, so the calculated CI would be different for each. Some CIs would straddle the population mean μ, and some would not; and we never know if the CI computed from our single sample straddles μ or not.

Loosely speaking, there is a 95% chance that our 95% CI straddles μ. For a CI computed from a single sample, we don't know if our CI includes the value of μ or not. The CI could also be interpreted as the range of plausible values of μ that could have produced the observed value of $\bar{x}$.

Example 22.1 (School bags)**.** A study of the school bags that 586 children (in Grades 6–8 in Tabriz, Iran) take to school found that the mean weight was $\bar{x} = 2.8$ kg with a standard deviation of $s = 0.94$ kg (Dianat et al. 2014).

The parameter is the population mean weight of school bags for Iranian children in Grades 6–8.

Of course, another sample of 586 children would produce a different sample mean: the sample mean varies from sample to sample.

The *standard error* of the sample mean is

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{0.94}{\sqrt{586}} = 0.03883;$$

see Fig. 22.3. The approximate 95% CI for the population mean school-bag weight is

$$2.8 \pm (2 \times 0.03883),$$

or 2.8 ± 0.07766. (The *margin of error* is 0.07766.) This is equivalent to an approximate 95% CI from 2.72 kg to 2.88 kg. This CI has a 95% chance of straddling the population mean bag weight.

Think 22.1 (Width of CI)**.** *Would a 99% CI for μ be* wider *or* narrower *than the 95% CI? Why?*

Answer: The answer is given in the online book.

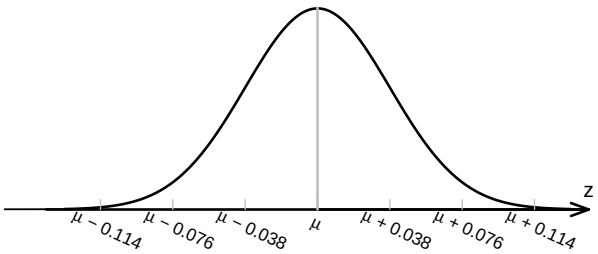


FIGURE 22.3: *The normal distribution, showing how the sample mean bag weight varies in samples of size $n = 586$*

Example 22.2 (Black bears)**.** A study of American black bears (Bartareau 2017) found that the mean weight of the $n = 185$ male bears in their study was $\bar{x} = 84.9$ kg, with a standard deviation of $s = 51.1$ kg.

The *parameter* of interest is the population mean weight of an American black bear, μ.

Using the sample information, the standard error of the mean is

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{51.1}{\sqrt{185}} = 3.756947,$$

so the approximate 95% CI is from

$$84.9 - (2 \times 3.756947) = 77.38611$$

to

$$84.9 + (2 \times 3.756947) = 92.41389.$$

The *approximate* 95% CI is from 77.4 to 92.4 kg. We would say:

> We are approximately 95% confidence that the *population mean* weight of male American black bears is between 77.4 and 92.4 kg.

The article gives the *exact* CI as 77.4 to 99.5 kg, agreeing with the CI we calculated.

22.4 Statistical validity conditions: One mean

As with any inference procedure, the underlying mathematics requires certain conditions to be met so that the results are statistically valid. The CI for one mean, will be *statistical valid* if *one* of these is true:

1. The sample size is at least 25, *or*
2. The sample size is smaller than 25 *and* the *population* data has an approximate normal distribution.

The sample size of 25 is a rough figure here, and some books give other values (such as 30). This condition ensures that the *distribution of the sample means has an approximate normal distribution* so that the 68–95–99.7 rule can be used.

Provided the sample size is larger than about 25, this will be approximately true *even if* the distribution of the individuals in the population does not have a normal distribution. That is, when $n > 25$ the sample means generally have an approximate normal distribution, even if the data themselves don't have a normal distribution.

In addition to the statistical validity condition, the CI will be

- **internally valid** if the study was well designed; and
- **externally validity** if the the sample is a simple random sample and is internally valid.

⚠ When $n > 25$ approximately, we do *not* require that the *data* has a normal distribution. We require that the *sample means* have a normal distribution, which is approximately true if the statistical validity condition is true.

This is one reason why means are used to describe samples: under certain conditions, sample means have an approximate normal distribution (so the 68–95–99.7 rule applies). In contrast, the distribution of sample medians is far more complicated to describe.

To determine if assuming the *population* has an approximate normal distribution in the statistical validity condition, the histogram of the *sample* can be constructed. However, we can't really be sure about the distribution of the *population* from the distribution of the *sample*.

All we can reasonably do is to identify (from the sample) populations that likely to be very non-normal (when the CI would be not valid).

Example 22.3 (Assumptions)**.** A study (Silverman et al. 1999; Zou et al. 2003) to examine exposure to radiation for CT scans in the abdomen assessed $n = 17$ patients. A histogram of the total radiation dose received is shown in Fig. 22.4; the sample mean dose is 26.86 rads.

A CI for the mean radiation dose received could be formed. However, as the sample size is 'small' (less than 25), the *population* must have a normal distribution for the CI to be statistically valid. Even though the histogram is from *sample* data, it seems improbable that the data in the sample would have come from a *population* with a normal distribution: the histogram of the sample data doesn't look normally distributed at all.

Computing a CI for the mean of these data will probably be statistically invalid. Other methods (beyond the scope of this course) are possible for computing a confidence interval for the mean.

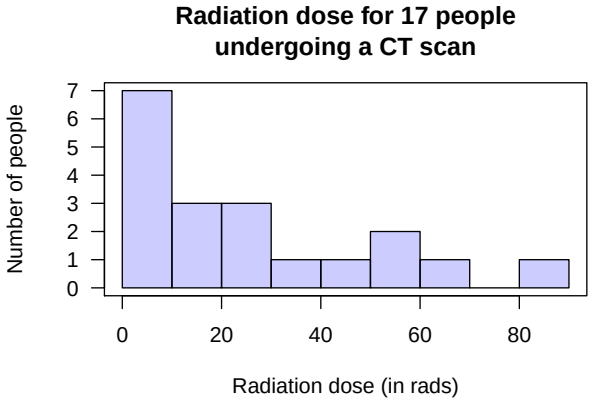


FIGURE 22.4: *The radiation doses from CT scans for 17 people*

Example 22.4 (School bags)**.** In Example 22.1, an approximate 95% CI was formed for the mean weight of school bags for Iranian children.

Since the sample size was $n = 586$, the CI is statistically valid.

We *do not* have to assume that the distribution of school bag weights has a normal distribution in the population, as the sample size is (much) larger than 25.

Example 22.5 (Black bears)**.** In Example 22.2, the approximate 95% CI was formed from a sample of size $n = 185$ male bears.

This CI is statistically valid, since the sample size is much larger than 25.

We *do not* have to assume that the distribution of the weights of male black bears has a normal distribution in the population, as the sample size is (much) larger than 25.

22.5 Example: NHANES

Previously, we asked this question about the NHANES data:

> Among Americans, is the mean direct HDL cholesterol different for current smokers and non-smokers?

The response variable is direct HDL cholesterol concentration. The parameter is μ, the population mean HDL cholesterol concentration.

What is the *population* mean direct HDL cholesterol concentration?

From the data (using jamovi or SPSS), the sample mean is $\bar{x} = 1.3649$ mmol/L; the standard deviation is $s = 0.39926$ mmol/L; and the sample size is $n = 8474$.

The value of $\bar{x}$ will vary from sample to sample; sampling variation exists. The standard error is:

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{0.39926}{\sqrt{8474}} = 0.00434 \text{ mmol/L}.$$

The approximate 95% CI uses a multiplier of 2, so the margin-of-error is

$$2 \times 0.0043 = 0.00867.$$

The approximate 95% CI is 1.365, give-or-take 0.00867; or from 1.356 to 1.374 mmol/L.

> Based on the sample of size $n = 8474$, a 95% CI for the population mean direct HDL cholesterol levels of Americans is between 1.356 and 1.374 mmol/L.

If many samples of the same size were found in the same way, and computed the CI from each, about 95% of the CIs would contain μ (but *this* particular CI may or may not contain the value of μ). We could also say that the CI gives a range of plausible values for μ, or that we are about 95% confident that this CI straddles the value of μ.

The statistical validity condition should also be checked to ensure the CI is statistically valid.

Since the sample size is much larger than 25, this CI for mean direct HDL cholesterol is statistically valid, *even though* the histogram of direct HDL cholesterol for individuals is skewed right (Fig. 22.5). Recall: the distribution of the *sample means* should be normally distributed, *not* the distribution of the data.

22.6 Example: Cadmium in peanuts

A study of peanuts from the United States (Blair and Lamb 2017) found the sample mean cadmium concentration was 0.0768 ppm with a standard deviation of 0.0460 ppm, from a

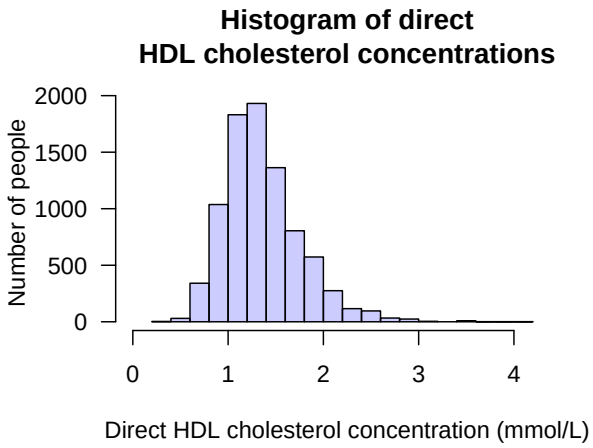


FIGURE 22.5: *Histogram of direct HDL cholesterol concentration*

sample of size 290 peanuts gathered from a variety of regions at various times (attempting to find a representative sample).

The parameter is μ, the population mean cadmium concentration in peanuts.

Every sample of $n = 290$ peanuts is likely to produce a different sample mean; that is, the sample means show *sampling variation*. The sampling variation can be measured using the standard error:

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{0.0460}{\sqrt{290}} = 0.002701 \text{ ppm.}$$

The approximate 95% CI is

$$0.0768 \pm (2 \times 0.002701),$$

or 0.0768 ± 0.00540, which is from 0.0714 to 0.0822 ppm. (The *margin of error* is 0.00540.)

If we repeatedly took samples of size 290 from this population, about 95% of the 95% CIs would contain the population mean (but *this* CI may or may not contain the value of μ).

The plausible values of μ that could have produced $\bar{x} = 0.0768$ are between 0.0714 and 0.0822ppm. Alternatively, we are about 95% confident that the CI of 0.0714 to 0.0822 ppm straddles the population mean.

Since the sample size is larger than 25, the CI is statistically valid.

22.7 Estimating sample sizes: one mean

For a specified level of confidence, the width of a CI depends on the size of the sample. In general, larger samples produce more precise estimates of the parameter (Sect. 5.2), and hence narrower CIs.

To determine the sample size needed to estimate a *sample mean* with a given precision for a 95% CI is *at least*

$$\left(\frac{2 \times s}{\text{Margin of error}}\right)^2.$$

⚠ Always **round up** the results of a sample size calculation.

Example 22.6 (Sample size estimation)**.** For the NHANES data, What size sample is needed to estimate the direct HDL cholesterol levels to within 0.02 mmol/L, with 95% confidence?

Since we would like to estimate the population mean give-or-take 0.02 mmol/L, the 'margin of error' that we would like is 0.02. So, using $s = 0.39926$, the required sample size is *at least*

$$\left(\frac{2 \times 0.39926}{0.02}\right)^2 = 1594.085;$$

at least 1595 Americans are needed. (Remember to always *round up* in sample size calculations.)

22.8 Quick review questions

1. True or false: The value of $\bar{x}$ varies from sample to sample.
2. True or false: A CI for μ is statistically valid only if the histogram of the *data* has an approximate normal distribution.
3. Suppose $s = 8$ and $n = 20$. Which *one* of the following is **true**?

22.9 Exercises

Selected answers are available in Sect. D.21.

Exercise 22.1. A study (Tager et al. 1979; Kahn 2005) of the lung capacity of children in East Boston measured the forced expiratory volume (FEV) of children in the area.

The sample contained $n = 45$ eleven-year-old girls. For these children, the mean lung capacity was $\bar{x} = 2.85$ litres and the standard deviation was $s = 0.43$ litres.

Find an approximate 95% CI for the population mean lung capacity of eleven-year-old females from East Boston.

Exercise 22.2. A study of lead smelter emissions near children's public playgrounds (Taylor et al. 2013) found the mean lead concentration at one playground (Memorial Park, Port Pirie, in South Australia) to be 6956.41 micrograms per square metre, with a standard deviation of 7571.74 micrograms of lead per square metre, from a sample of $n = 58$ wipes taken

over a seven-day period. (As a reference, the Western Australian government recommends a maximum of 400 micrograms of lead per square metre.)

Find an approximate 95% CI for the mean lead concentration at this playground. Would these results apply to other playgrounds?

Exercise 22.3. A study (Macgregor and Rugg-Gunn 1985) of the brushing time for 60 young adults (aged 18–22 years old) found the mean brushing time was 33.0 seconds, with a standard deviation of 12.0 seconds.

Find an approximate 95% CI for the mean brushing time for young adults.

Exercise 22.4. A study of paramedics (Williams and Boyle 2007) asked participants ($n = 199$) to estimate the amount of blood loss on four different surfaces. When the actual amount of blood spill on concrete was 1000 ml, the mean guess was 846.4 ml (with a standard deviation of 651.1 ml).

1. What is the approximate 95% CI for the mean guess of blood loss?
2. Are the participants good at estimating the amount of blood loss on concrete?
3. Is this CI likely to be valid?
4. How many paramedics would be needed if the mean guess was to be estimated with an precision of give-or-take 50 ml?
5. How many paramedics would be needed if the mean guess was to be estimated with an precision of give-or-take 25 ml?
6. How many times greater does the sample size need to be to *halve* the width of the margin of error?

Exercise 22.5. In Sect. 22.5, the approximate 95% CI for the mean direct HDL cholesterol was given as 1.356 to 1.374 mmol/L. Which (if any) of these interpretations are acceptable? Explain why are the other interpretations are *incorrect*.

1. In the *sample*, about 95% of individuals have a direct HDL concentration between 1.356 to 1.374 mmol/L.
2. In the *population*, about 95% of individuals have a direct HDL concentration between 1.356 to 1.374 mmol/L.
3. About 95% of the *samples* are between 1.356 to 1.374 mmol/L.
4. About 95% of the *populations* are between 1.356 to 1.374 mmol/L.
5. The *population* mean varies so that it is between 1.356 to 1.374 mmol/L about 95% of the time.
6. We are about 95% sure that *sample* mean is between 1.356 to 1.374 mmol/L.
7. It is plausible that the *sample* mean is between 1.356 to 1.374 mmol/L.

Exercise 22.6. An article (Grabosky and Bassuk 2016) describes the diameter of *Quercus bicolor* trees planted in a lawn as having a mean of 25.8 cm, with a standard error of 0.64 cm, from a sample of 19 trees. Which (if any) of the following is correct?

1. About 95% of the trees in the *sample* will have a diameter between $25.8 - (2 \times 0.64)$ and $25.8 + (2 \times 0.64)$ (based on using the 68–95–99.7 rule).
2. About 95% of these types of trees in the *population* will have a diameter between $25.8 - (2 \times 0.64)$ and $25.8 + (2 \times 0.64)$ (based on using the 68–95–99.7 rule)?

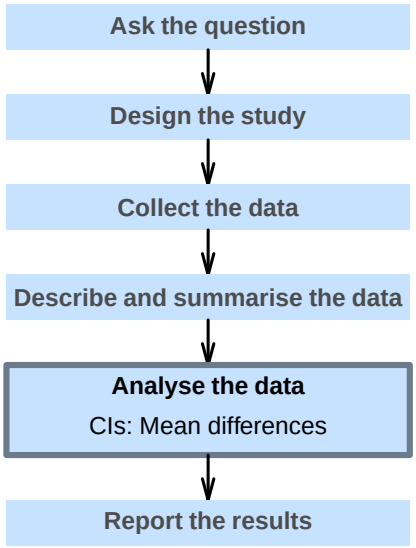
23

Confidence intervals for mean differences (paired data)

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, and understand the tools of inference.

In this chapter, you will learn about *confidence intervals* for mean differences (i.e., for *paired* data). You will learn to:

- produce a confidence interval for a mean difference.
- determine whether the conditions for using the confidence interval apply in a given situation.
- compute sample size estimates in these this situation.



23.1 Mean differences

House insulation is important for saving energy, particularly in cold climates.

Consider a study to estimate the average energy savings made by using a new type of house insulation. Different study designs could be used to address this.

One approach is to take a sample of homes, and measure the energy consumption *before* adding the insulation, and then *after* adding the insulation for the same houses. Each home gets *two* observations: the energy consumption *before* and *after* adding the insulation.

This is a *descriptive RQ*: the *Outcome* is the mean energy *saving*, and the response variable is the energy saving for each house. There is *no* Comparison: units of analysis that have been treated differently are not compared.

Alternatively, the researchers could take a sample of homes *without* the insulation, and measure their energy consumption; then take a *different* sample of homes with the insulation, and measure their energy consumption.

This is a *relational RQ*: the *Outcome* is the mean energy consumption, and the response variable is the energy consumption for each house. The Comparison is between units of analysis *with* the insulation, and units of analysis *without* the insulation.

Either study is possible, and each has advantages and disadvantages (Zimmerman 1997). Here the *first* (Descriptive) design would seem superior (why?). In the first design, each home gets a *pair* of energy consumption measurements: this is *paired data*, which is the subject of this chapter. The second (Relational) design requires the means of two different groups of homes to be compared, which is the topic of the next chapter.

Definition 23.1 (Paired data)**.** Data are *paired* when two observations about the same variable are recorded for each unit of analysis.

Since each unit of analysis has two observations about energy consumption, the *change* (or the *difference*, or the *reduction*) in energy consumption can be computed for each house. Then, questions can be asked about the *population mean difference*, which is not the same as *difference between two separate population means* (the subject of the next chapter). In paired data, finding the difference between the two measurements for each individual unit of analysis makes sense: each unit of analysis (each house) has two related observations.

Think 23.1 (Paired situations)**.** *Which of these are paired situations?*

1. *The mean difference between blood pressure for 36 people, before and after taking a drug.*
2. *The difference between the mean HDL cholesterol levels for 22 males and 19 females.*
3. *The mean protein levels were compared in sea turtles before and after being rehabilitated (March et al. 2018).*

Answer: Only situations **1** and **3** are paired.

23.2 Mean differences: An example

The *Electricity Council* in Bristol wanted to determine if a certain type of wall-cavity insulation reduced energy consumption in winter (The Open University 1983). Their (Descriptive) RQ was:

What is the *mean reduction* in energy consumption after adding home insulation?

The parameter is μ_d, the population mean *reduction* in energy consumption.

For the collected data (Table 23.1) the same variable (energy consumption) is measured twice for each unit of analysis (the house): energy consumption *before* adding insulation and *after* adding insulation.

Finding the *difference* in energy consumption for each house seems sensible, as the data are *paired*. Once the differences are computed, the process for computing a CI is the same as in Chap. 22, where these changes (or differences) are used as the data.



Be clear about *how* the differences are computed. Differences could be computed as *Before* minus *After* (the energy consumption *saving*), or *After* minus *Before* (the energy consumption *increase*).

Either is fine, as long as you are consistent throughout. The meaning of any conclusions will be the same.

Here, discussing energy *savings* seems most natural, so we compute the differences as energy *savings*: *Before* minus *After*.

TABLE 23.1: *The house insulation data: Energy consumption before and after adding insulation, and the energy saving (all in MWh)*

Before	After	Energy savings
12.1	12.0	0.1
11.0	10.6	0.4
14.1	13.4	0.7
13.8	11.2	2.6
15.5	15.3	0.2
12.2	13.6	-1.4
12.8	12.6	0.2
9.9	8.8	1.1
10.8	9.6	1.2
12.7	12.4	0.3



One energy saving value is *negative*. This does *not* mean negative energy usage: the values are *differences* (more specifically, energy *reductions* or *savings*).

The differences are computed as *Before* minus *After*, so a negative value means that the *After* value is greater than the *Before* value: an *increase* in energy consumption.

As always, begin by understanding the data: producing appropriate graphical and numerical summaries.

23.3 Notation: Mean differences

The notation used for paired data reflects how we work with the *differences* (Table 23.2). Apart from that, the notation is similar to that used in Chap. 22.

TABLE 23.2: *The notation used for mean differences (paired data) compared to the notation used for one sample mean*

	One sample mean	**Mean of paired data**
The observations:	Values: x	Differences: d
Sample mean:	$\bar{x}$	$\bar{d}$
Standard deviation:	s	s_d
Standard error of sample mean:	s.e.$(\bar{x}) = \dfrac{s}{\sqrt{n}}$	s.e.$(\bar{d}) = \dfrac{s_d}{\sqrt{n}}$
Sample size:	Number of *observations*: n	Number of *differences*: n

23.4 Graphical summaries: Mean differences

Since the data are the differences (quantitative), the appropriate graph is *a histogram* (or a dot plot, or a stem-and-leaf plot) of the *differences* (Fig. 23.1).

Graphing the *Before* and *After* data may also be useful too, but a graph of the differences is *crucial*, as the RQ is about the differences.

A case-profile plot (Sect. 12.8.2) is also useful, but is sometimes harder to produce in software, and difficult to read when the sample size is large (the graph contains a line for each unit of analysis).

23.5 Numerical summaries: Mean differences

Since the data are differences, a *numerical* should summarises the *differences*. Summarising the *Before* and *After* data is useful too, but summarising the differences is *crucial* because the RQ is about the differences (Table 23.3).

For the house insulation data, the appropriate *numerical summary* for paired data *summarises the differences* using means, standard deviations, and so on, as appropriate. (While a mean or a median may be appropriate for describing the *data*, the CI is about the *mean*, since the sampling distribution for the sample mean (under certain conditions) has a *normal distribution* which is best described using a mean.) A numerical summary of the energy savings from a calculator (Statistics Mode) or computer software gives the sample mean of the differences as $\bar{d} = 0.54$, and the standard deviation of the differences as $s_d = 1.015655$.

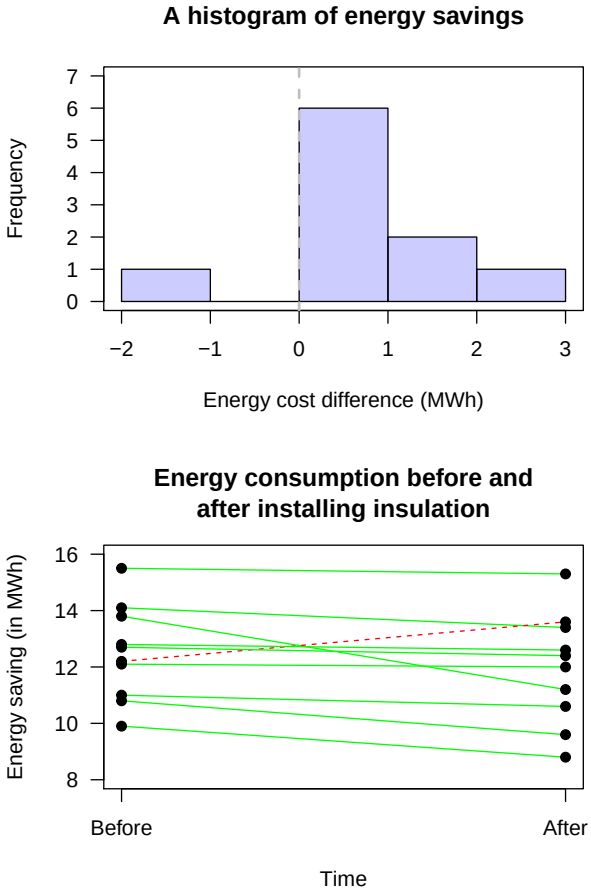


FIGURE 23.1: *A plot of the energy savings from the insulation data. Top panel: A histogram (the vertical grey line represents no energy saving). Bottom panel: Case-profile plot (a dashed line represents an energy increase)*

A formal numerical summary table is shown in Table 23.3.

TABLE 23.3: *The mean, median, standard deviation and IQR for the energy consumption data (in MWh)*

	Mean	Median	Std dev	IQR
Before	12.49	12.45	1.68	2.28
After	11.95	12.20	1.96	2.45
Energy savings	0.54	0.35	1.02	0.80

23.6 Sampling distribution: Means differences

The study concerns the mean energy *saving* (the mean *difference*). Every sample of $n = 10$ houses is likely to comprise different houses, and hence different before and after energy consumptions will be recorded, and hence different energy savings will be recorded. As a result, the *sample mean energy differences* will vary from sample to sample. That is, the mean differences have a *sampling distribution*, and a *standard error*.

Since the differences are like a single sample of data (Chap. 22), the sampling distribution for the differences will have a similar sampling distribution to the mean of a single sample $\bar{x}$ (provided the conditions are met; Sect. 23.9).

Definition 23.2 (Sampling distribution of a sample mean difference)**.** The *sampling distribution of a sample mean difference* is described by:

- an approximate normal distribution;
- centred around μ_d (the population mean *difference*);
- with a standard deviation of s.e.$(\bar{d}) = \dfrac{s_d}{\sqrt{n_d}}$,

when certain conditions are met, where n is the size of the sample, and s_d is the standard deviation of the individual differences in the sample.

For the home insulation data, the variation in the sample mean differences $\bar{d}$ can be described by

- approximate normal distribution;
- centred around μ_d;
- with a standard deviation of s.e.$(\bar{d}) = \dfrac{1.015655}{\sqrt{10}} = 0.3211784$, called the *standard error of the differences*.

Notice that many decimal places are used in the working here; results will be rounded when reported.

23.7 Confidence intervals: Mean differences

The CI for the mean difference has the same form as for a single mean (Chap. 22), so an approximate 95% confidence interval (CI) for μ_d is

$$\bar{d} \pm 2 \times \text{s.e.}(\bar{d}).$$

This is the same as the CI for $\bar{x}$ if the differences are considered as the data.

For the insulation data:

$$0.54 \pm (2 \times 0.3211784),$$

or 0.54 ± 0.642. This CI is equivalent to $0.54 - 0.642 = -0.102$, up to $0.54 + 0.642 = 1.182$. We write:

> Based on the sample, an *approximate* 95% CI for the population mean energy *saving* after adding the wall cavity insulation is from -0.10 to 1.18MWh.

The negative number is *not* an energy consumption value; it is a negative mean amount of energy *saved*. Saving a *negative* amount is like using *more* energy. So the 95% CI is saying that we are reasonably confident that, after adding the insulation, the mean energy-use difference is between using 0.10MWh *more* energy to using 1.18MWh *less* energy. Alternatively, the plausible values for the mean energy savings are between -0.10 to 1.18MWh.

Example 23.1 (COVID lockdown)**.** A study of $n = 213$ Spanish health students (Romero-Blanco et al. 2020) measured (among other things) the number of minutes of vigorous physical activity (PA) performed by students *before* and *during* the COVID-19 lockdown (from March to April 2020 in Spain).

Since the *before* and *during* lockdown were both measured on *each* participant, the data are *paired*. The data are summarised below.

	Mean (minutes)	Standard deviation (minutes)
Before	28.47	54.13
During	30.66	30.04
Difference	-2.68	51.30

Notice that the *differences* are defined as *Before* minus *During*. A *positive* difference therefore means the *Before* value is higher; hence, the differences tell us how much longer the student spent doing vigorous PA *before* the COVID lockdown. Similarly, a *negative* value means that the *During* value is higher.

In this situation, the *parameter* of interest is the population mean difference μ_d, the mean amount that students spent in vigorous PA *before* the lockown compared to *during* the lockdown.

Also notice that the standard deviation of the difference (s.e.$(\bar{d}) = 51.30$) is **not** $54.13 - 30.04$, or $30.04 - 54.13$. Those calculations would find the difference between the two standard deviations... not the standard deviation of the list of differences.

Every sample would contain different students, and hence would produce different pre- and during-COVID mean amounts of PA, so those means would have standard error.

Likewise, the mean of each individuals' *difference* would vary from sample to sample, so the *mean difference* would vary and hence have a standard error:

$$\text{s.e.}(\bar{d}) = \frac{s_d}{n} = \frac{51.30}{\sqrt{213}} = 3.515018.$$

The approximate 95% CI for the population mean *difference* is from

$$-2.68 - (2 \times 3.515018) = -9.710036$$

to

$$-2.68 + (2 \times 3.789094) = 4.350036,$$

so the approximate 95% CI for the population mean *difference* is from -9.71 to 4.35 minutes.

Notice that one of the values is *negative*. This does **not** mean a negative amount of PA (which would make no sense); the CI is for the population mean *difference*. So, a negative value means that the *During* values are higher than the *Before* values on average.

So, the CI means:

> In the population, the mean difference between the amount of vigorous PA by Spanish health students is between 9.71 minutes more *during* lockdown, and 4.35 minutes more *before* lockdown.

23.8 Using software: CIs for mean differences

Software (such as jamovi or SPSS) can be used to produce *exact* 95% CIs, which will may be slightly different than the *approximate* 95% CI (since the 68–95–99.7 rule is an *approximation*). The *approximate* and *exact* 95% CIs are similar when the sample size is not small; here the sample size *is* small ($n = 10$). From the jamovi (Fig. 23.2) or SPSS output (Fig. 23.3):

Based on the sample, a 95% CI is for the population mean energy *saving* because of the wall cavity insulation is from -0.19 to 1.27MWh.

Paired Samples T-Test

Paired Samples T-Test

						95% Confidence Interval	
			statistic	df	p	Lower	Upper
Before	After	Student's t	1.68	9.00	0.127	−0.187	1.27

FIGURE 23.2: *The insulation data: jamovi output*

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Energy consumption (Before) - Energy consumption (After)	.5400	1.0157	.3212	-.1866	1.2666	1.681	9	.127

FIGURE 23.3: *The insulation data: SPSS output*

As expected, this 95% CI is slightly different than the CI computed by hand, since the sample size is small. For our purposes, however, using the approximate multiplier of 2 is sufficient when not using software.

23.9 Statistical validity conditions: Mean differences

As with any inferential procedure, these results apply under certain conditions. The conditions under which the CI is statistically valid for paired data are similar to those for one sample mean, rephrased for differences.

The CI computed above is statistically valid if *one* of these conditions is true:

1. The sample size of differences is at least 25; **or**
2. The sample size of differences is smaller than 25, **and** the *population* of *differences* has an approximate normal distribution.

The sample size of 25 is a rough figure here, and some books give other values (such as 30). This condition ensures that the *distribution of the sample means has an approximate normal distribution* so that the 68–95–99.7 rule is used. Provided the sample size is larger than about 25, this will be approximately true *even if* the distribution of the individuals in the population

does not have a normal distribution. That is, when $n > 25$ the sample means generally have an approximate normal distribution, even if the data themselves don't have a normal distribution.

In addition to the statistical validity condition, the CI will be

- **internally valid** if the study was well designed; and
- **externally valid** if the sample is a simple random sample and is internally valid.

Example 23.2 (Statistical validity)**.** For the insulation data, the sample size is small, so we require that the differences *in the population* follow a normal distribution. We don't know if they do (the data, graphed in Fig. 23.1, don't seems to identify any obvious doubts). So the CI is possibly statistically valid, but we aren't sure.

In this case then, the results may not be valid; that is, the CI limits that we calculated will be approximately correct only. (This doesn't mean the CI is useless!)

23.10 Example: Blood pressure

A US study (Schorling et al. 1997; Willems et al. 1997) examined how CHD risk factors were assessed among parts of the population with diabetes. Subjects reported to the clinic on multiple occasions. Consider this RQ:

What is the mean difference in diastolic blood pressure from the first to the second visit?

Each person has a *pair* of diastolic blood pressure (DBP) measurements: One each from their first and second visits. The data (some shown in Table 23.5) are from the 141 people for whom *both* measurements are available (some data are missing). The differences could be computed as:

- The first visit DBP minus the second visit DBP: the *reduction* in DBP; or
- The second visit DBP minus the first visit DBP: the *increase* in DBP.

Either way is fine, provided the order is used consistently. Here, the observation from the *second* visit will be used, so that the differences represent the *reduction* in DBP from the first to second visit.

The parameter is μ_d, the population mean *reduction* in DBP.

Since the data set is large, the appropriate graphical summary is a histogram of differences (Fig. 23.4). The numerical summary can summarise both the first and second visit observations, but *must* summarise the *differences*. Numerical summaries can be computed using software, then reported in a suitable table (Table 23.6).

The *standard error* of the sample mean is

$$\text{s.e.}(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{8.02614}{\sqrt{141}} = 0.67592.$$

Using an approximate multiplier of 2, the margin of error is:

TABLE 23.5: *The first six observations (from the $n = 141$ available) from the diabetes study for people with both measurements: Diastolic blood pressure (DBP) for the first and second visits, and the decrease in DBP, all in mm Hg*

DBP: First visit	DBP: Second visit	Reduction in DBP
92	92	0
112	112	0
80	86	-6
90	90	0
90	96	-6
88	84	4

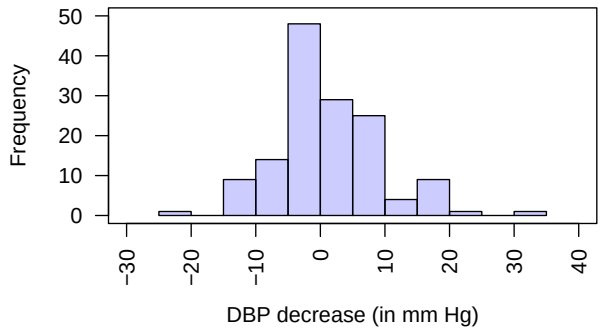


FIGURE 23.4: *Histogram of the decrease in DBP between the first and second visits*

$$2 \times 0.67592 = 1.3518,$$

so an approximate 95% CI for the decrease in DBP is

$$1.9504 \pm 1.3518,$$

or from 0.60 to 3.30 mm Hg, after rounding sensibly. We write:

> Based on the sample, an *approximate* 95% CI for the mean *decrease* in DBP is from 0.60 to 3.30 mm Hg.

The *exact* 95% CI from jamovi (Fig. 23.5) or SPSS (Fig. 23.6), using an exact t-multiplier rather than an approximate multiplier of 2, is similar since the sample size is large. After rounding, write:

TABLE 23.6: *The numerical summary for the diabetes data (in mm Hg). The differences are the second visit value minus the first visit value: the decreases in diastolic blood pressure from the first to second visit*

	Mean	Standard deviation	Standard error	Sample size
DBP: First visit	94.48	11.473	0.966	141
DBP: Second visit	92.52	11.555	0.973	141
Decrease in DBP	1.95	8.026	0.676	141

Based on the sample, an exact 95% CI for the decrease in DBP is from 0.61 to 3.29 mm Hg.

The wording ('for the *decrease* in DBP') implies which reading is the higher reading on average: the first.

Paired Samples T-Test

Paired Samples T-Test

						95% Confidence Interval	
			statistic	df	p	Lower	Upper
bp.1d	bp.2d	Student's t	2.89	140	0.005	0.614	3.29

FIGURE 23.5: *jamovi output for the blood pressure data, including the exact 95% CI*

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Diastolic blood pressure 1 – Diastolic blood pressure 2	1.950	8.026	.676	.614	3.287	2.885	140	.005

FIGURE 23.6: *SPSS output for the blood pressure data, including the exact 95% CI*

⚠ Be clear in your conclusion about how the differences are computed.

The CI is statistically valid as the sample size is larger than 25. (The *data* do not need to follow a normal distribution.)

Think 23.2 (Understanding samples)**.** *Is there a mean difference in DBP* in the population*?*

23.11 Quick review questions

1. True or false: For paired data, the mean of the *differences* is treated like the mean of a single variable.
2. True or false: The appropriate graph for displaying paired data is often a histogram of the differences.
3. True or false: The population mean difference is denoted by μ_d.
4. True or false: The standard error of the sample mean difference is denoted by s_d.

23.12 Exercises

Selected answers are available in Sect. D.22.

Exercise 23.1. People often struggle to eat the recommended intake of vegetables. In one study exploring ways to increase vegetable intake in teens (Fritts et al. 2018), teens rated the taste of raw broccoli, and raw broccoli served with a specially-made dip.

Each teen ($n = 101$) had a *pair* of measurements: the taste rating of the broccoli *with* and *without* dip. Taste was assessed using a '100 mm visual analog scale,' where a *higher* score means a *better* taste. In summary:

- For raw broccoli, the mean taste rating was 56.0 (with a standard deviation of 26.6);
- For raw broccoli served with dip, the mean taste rating was 61.2 (with a standard deviation of 28.7).

Because the data are paired, the *differences* are the best way to describe the data. The mean difference in the ratings was 5.2, with s.e.$(\bar{d}) = 3.06$. From this information:

1. Construct a suitable numerical summary table.
2. Compute the approximate 95% CI for the mean difference in taste ratings.

Exercise 23.2. In a study of hypertension (MacGregor et al. 1979; Hand et al. 1996), 15 patients were given a drug (Captopril) and their systolic blood pressure measured immediately before and two hours after being given the drug.

1. Explain why it is sensible to compute differences as the *Before* minus the *After* measurements. What do the differences *mean* when computed this way?
2. Compute the differences.
3. Compute an *approximate* 95% CI for the mean difference.
4. Write down the *exact* 95% CI using the computer output (jamovi: Fig. 23.7; SPSS: Fig. 23.8).
5. Why are the two CIs different?

TABLE 23.7: *The Captopril data: before after after systolic blood pressures (in mm Hg)*

Before	After	Before	After
210	201	173	147
169	165	146	136
187	166	174	151
160	157	201	168
167	147	198	179
176	145	148	129
185	168	154	131
206	180	NA	NA

Exercise 23.3. A study (Allen et al. 2018) examined the effect of exercise on smoking. Men and women were assessed on a range of measures, including the 'intention to smoke.'

'Intention to smoke' was assessed both before and after exercise for each subject, using the

Paired Samples T-Test

Paired Samples T-Test

						95% Confidence Interval	
			statistic	df	p	Lower	Upper
Before	After	Student's t	8.12	14.0	< .001	13.9	23.9

FIGURE 23.7: *jamovi output for the Captoril data*

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Before blood pressure (in mm Hg) - After blood pressure (in mm Hg)	18.933	9.027	2.331	13.934	23.93	8.123	14	.000

FIGURE 23.8: *The Captoril data: SPSS output*

10-item quantitative *Questionnaire of Smoking Urges – Brief*[1] scale (Cox et al. 2001), and the quantitative *Minnesota Nicotine Withdrawal Scale*[2] (Shiffman et al. 2004).

Smokers (people smoking at least five cigarettes per day) aged 18 to 40 were enrolled for the study. For the 23 women in the study, the mean intention to smoke after exercise *reduced* by 0.66 (with a standard error of 0.37).

1. Find a 95% confidence interval for the population mean reduction in intention to smoke for women after exercising.
2. Is this CI statistically valid?

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2527734/

[2] http://www.med.uvm.edu/behaviorandhealth/research/minnesota-tobacco-withdrawal-scale

24

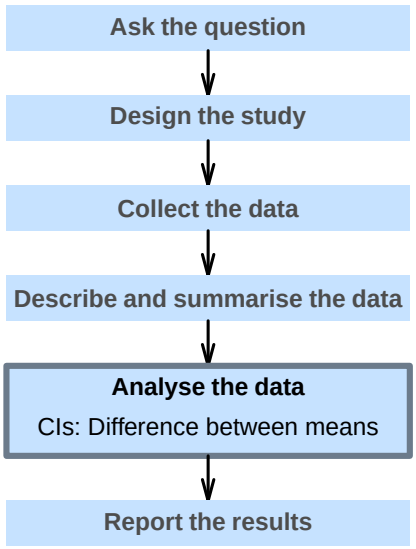
Confidence intervals for two independent means



So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, and understand the tools of inference.

In this chapter, you will learn about *confidence intervals* for the differences between two means. You will learn to:

- produce confidence intervals for two independent means.
- determine whether the conditions for using the confidence intervals apply in a given situation.



24.1 Means of two independent samples

A study (Strayer and Johnston 2001; Agresti and Franklin 2007) examined the reaction times of students while driving.

In one study, two different groups of students were used: one group *used* a mobile phone, and a different group *did not use* a mobile phone. The reaction time for each student was measured in a driving simulator.

The study uses two groups with different treatments: one group using a mobile phone while driving, and a different group *not* using a mobile phone while driving.

The data are not paired; instead, the means of two separate (or independent) samples are being compared. (The data would be paired if *each* student was measured twice: once using a phone, and once without using a phone.)

Consider the RQ:

> For students, what is the difference between the mean reaction time while driving when using a mobile phone and the mean reaction time while driving when *not* using a mobile phone?

Part of the data are shown in Table 24.1.

Think 24.1 (POCI)**.** *What are P, O, C and I in this study?*

Answer: The answer is given in the online book.

TABLE 24.1: *Reaction times (in milliseconds) for students using, and not using, mobile phones. The first ten observations are shown, but 32 students are in each group*

Using phone	Not using phone
636	557
623	572
615	457
672	489
601	532
600	506
542	648
554	485
543	610
520	444

24.2 Graphical summary: Two independent means

To compare two *quantitative* variables, a suitable graphical summary may be a boxplot (Fig. 24.1) or (when samples sizes aren't too large) a dot chart.

For the reaction-time data, the boxplot shows that the sample medians are a little different, but the IQR about the same; one large outlier is present for the phone-using group.

24.3 Notation: Two independent means

Since two groups are being compared, distinguishing between the statistics for the two groups (say, Group A and Group B) is important. One way is to use subscripts (Table 24.2).

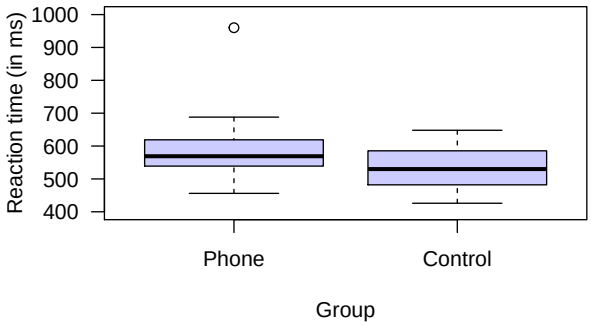


FIGURE 24.1: *Plots of the reaction times (in milliseconds) for students using, and not using, mobile phones.*

TABLE 24.2: *Notation used to distinguish between the two independent groups*

	Group A	**Group B**
Population means:	μ_A	μ_B
Sample means:	$\bar{x}_A$	$\bar{x}_B$
Standard deviations:	s_A	s_B
Standard errors:	$\text{s.e.}(\bar{x}_A) = \frac{s_A}{\sqrt{n_A}}$	$\text{s.e.}(\bar{x}_B) = \frac{s_B}{\sqrt{n_B}}$
Sample sizes:	n_A	n_B

Using this notation, the difference between population means, the parameter of interest, is $\mu_A - \mu_B$. As usual, the population values are unknown, so this parameter is estimated using the statistic $\bar{x}_A - \bar{x}_B$.

Notice that Table 24.2 does not include a standard deviation or a sample size for the *difference between means*; they make no sense in this context.

For example, if Group A has 15 individuals, and Group B has 45 individuals, and we wish to study the difference $\bar{x}_A - \bar{x}_B$. what is the sample size be? Certain not $15 - 45 = -30$.

On the other hand, the *standard error* of the difference between the means does make sense: it measures how much the value of $\bar{x}_A - \bar{x}_B$ varies from sample to sample.

For the reaction-time data, we will use the subscripts P for phone-users group, and C for the control group. That means that the two sample means would be denoted as $\bar{x}_P$ and $\bar{x}_C$, and the difference between them as $\bar{x}_P - \bar{x}_C$.

24.4 Numerical summary: Two independent means

The numerical summary should summarise both groups, but *must* summarise the differences between the *means* (since the RQ is about this difference). All this information can be found using jamovi (Fig. 24.2) or SPSS (Fig. 24.3), then compiled into a table (Table 24.3).

Each time the study is repeated, the means for each group are likely to be different, and so the difference between the means is likely to be different. That is, the difference between

Independent Samples T-Test

Independent Samples T-Test

					95% Confidence Interval	
		statistic	df	p	Lower	Upper
Reaction	Student's t	2.63	62.0	0.011	12.4	90.8
	Welch's t	2.63	56.7	0.011	12.3	90.9

FIGURE 24.2: *jamovi output for the phone reaction time data*

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Reaction time (in ms)	Phone	32	585.19	89.646	15.847
	Control	32	533.59	65.360	11.554

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Reaction time (in ms)	Equal variances assumed	.077	.783	2.631	62	.011	51.594	19.612	12.390	90.798
	Equal variances not assumed			2.631	56.70	.011	51.594	19.612	12.317	90.871

FIGURE 24.3: *SPSS output for the phone reaction time data*

the means has *sampling variation* and hence a *standard error*, since it varies from sample to sample.

Think 24.2 (Standard deviation and standard error)**.** *For those using a phone, what is the difference between the* standard deviation *and the* standard error *in the context of the reaction-time study?*

Answer: *Standard deviation*: how much the individual reactions times vary from *person to person*. *Standard error*: how much the *difference between sample mean reaction times* varies from *sample to sample*.

TABLE 24.3: *The mean, median, standard deviation and standard error for the driving data (in ms)*

	Mean	Sample size	Standard deviation	Standard error
Using phone	585.19	32	89.65	15.847
Not using phone	533.59	32	65.36	11.554
Difference	51.59			19.612

24.5 Sampling distribution: Two independent means

Since the difference between the population means is unknown (that's why the study was done), the difference is estimated using the sample means. For the reaction time data, we will use subscripts such as P for phone-users group, and C for the control group. Then the difference between the two sample means (the *statistic*) is $\bar{x}_P - \bar{x}_C$.

The parameter is $\mu_P - \mu_C$, the difference between the two population means (using a phone, minus *not* using a phone).

The differences could be compute in the opposite direction ($\bar{x}_C - \bar{x}_P$). However, for the reaction-time data, computing differences as the reaction time for phone users, *minus* the reaction time for non-phone users (controls) probably makes more sense: the differences then refer to how much greater (on average) the reaction times are when students are using phones,



Making clear how the differences are computed is important! Therefore, carefully defining the *parameter* is important.

The differences could be computed as:

- the reaction time for phone users, *minus* the reaction time for non-phone users (how much slower the phone users are, on average); or
- the reaction time for non-phone users, *minus* the reaction time for phone users (how much slower the non-phone users are, on average).

Either is fine, provided you are consistent, and clear about how the difference are computed. The meaning of any conclusions will be the same.

Each sample of students will comprise different students, and will give different reaction times while driving. The means for each group will differ from sample to sample, and the *difference* between the means will be different for each sample. The *difference* between the sample means varies from sample to sample, and so has a sampling distribution and *standard error*.

Definition 24.1 (Sampling distribution of the difference between two sample means)**.** The *sampling distribution of the difference between two sample means* is described by:

- an approximate normal distribution;
- centred around $\mu_A - \mu_B$ (the *differences* between the means);
- with a standard deviation of s.e.$(\bar{x}_A - \bar{x}_B)$,

when the appropriate conditions are met.
We don't give a formula for finding the standard error s.e.$(\bar{x}_A - \bar{x}_B)$, so the value of this standard error will be *given*.



A formula exists for finding the standard error of the difference between two means, but is complicated and we won't provide it. (It is not necessary for our purposes anyway; software can handle details.) We will provide the standard error of the difference between two means, or expect you to find it on computer output.

For the reaction-time data, the differences between the sample means will have:

- an approximate normal distribution;
- centred around $\mu_P - \mu_C$ (the *differences* between the means in the two *populations*);
- with a standard deviation, called the *standard error* of the difference, of s.e.$(\bar{x}_P - \bar{x}_C) = 19.61$.

We can draw this sampling distribution (Fig. 24.4).

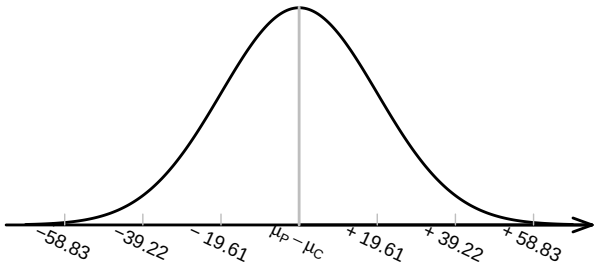


FIGURE 24.4: *The sampling distribution of the difference between the reaction times in the phone and control groups (phone, minus control)*

Think 24.3 (Understanding differences)**.** *What does a* negative *difference mean?*

24.6 Confidence intervals: Two independent means

Being able to describe the sampling distribution implies that we have some idea of how the values of
$\bar{x}_P - \bar{x}_C$ are likely to vary from sample to sample. Then, finding an approximate 95% CI for the difference between the mean reaction times is similar to the process used in Chap. 22. Approximate 95% CIs all have the same form:

$$\text{statistic} \pm (2 \times \text{s.e.}(\text{statistic})).$$

When the statistic is $\bar{x}_P - \bar{x}_C$, the approximate 95% CI is

$$(\bar{x}_P - \bar{x}_C) \pm (2 \times \text{s.e.}(\bar{x}_P - \bar{x}_C)).$$

In this case (using more decimal places than in the summary table in Table 24.3), the CI is

$$51.59375 \pm (2 \times 19.61213),$$

or 51.59375 ± 19.61213. After rounding appropriately, an approximate 95% CI for the difference is from 12.37 to 90.82 milliseconds. We write:

Based on the sample, an *approximate* 95% CI for the difference in reaction time while driving, for those using a phone and those not using a phone, is from 12.37 to 90.82 milliseconds (higher for those using a phone).

The plausible values for the difference between the two population means are between 12.37 to 90.82 milliseconds.

Stating the CI is insufficient; you must also state the *direction* in which the differences were calculated, so readers know which group had the higher mean.

Example 24.1 (Gray whales)**.** A study of gray whales (*Eschrichtius robustus*) measured (among other things) the length of whales at birth (Agbayani et al. 2020). The data are shown below.

Sex	Mean (in m)	Standard deviation (in m)	Sample size
Female	4.66	0.379	26
Male	4.60	0.305	30

How much longer are female gray whales than males, on average?

Let's define the *difference* as the mean length of female gray whales *minus* the mean length of male gray whales. Then we wish to estimate the difference $\mu_F - \mu_M$, where F and M represent female and male gray whales respectively. In this situation, this is the *parameter* of interest. The best estimate of this difference is $\bar{x}_F - \bar{x}_M = 4.66 - 4.60 = 0.06$.

We know that this value is likely to vary from sample to sample, and hence it has a standard error.

We cannot easily determine the standard error of this difference from the above information (though it is possible), so we must be *given* this information: s.e.$(\bar{x}_F - \bar{x}_M) = 0.0929$.

Then the approximate 95% CI is from

$$0.06 - (2 \times 0.0929) = -0.125747$$

to

$$0.06 + (2 \times 0.0929) = 0.245747,$$

so the CI is from -0.12 m to 0.25 m.

Notice that one of these limits is a *negative* value. This does not mean a *negative* length for a whale; that would be silly. Remember that this CI is for the *difference* between the mean lengths, and a *negative* length just says the mean length for males is greater than the mean length for females.

So we could say:

> The population mean difference between the length of female and male gray whales at birth has a 95% chance of being between 0.12 m longer for male whales to 0.25 m longer for female whales.

24.7 Using software: CIs for two independent means

The jamovi output (Fig. 24.5) and the SPSS output (Fig. 24.6) both show *two* CIs. *We will use the results from the second row* in both cases, as this row of output is more general (and makes fewer assumptions).

⚠ jamovi and SPSS give *two* confidence intervals. In this book, we will use the *second row* of information (the Welch's t row in jamovi; the 'Equal variance not assumed' row in SPSS) because it is more general and makes fewer assumptions.

Independent Samples T-Test

Independent Samples T-Test

					95% Confidence Interval	
		statistic	df	p	Lower	Upper
Reaction	Student's t	2.63	62.0	0.011	12.4	90.8
	Welch's t	2.63	56.7	0.011	12.3	90.9

FIGURE 24.5: *The jamovi output for the phone-reaction data*

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Reaction time (in ms)	Phone	32	585.19	89.646	15.847
	Control	32	533.59	65.360	11.554

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Reaction time (in ms)	Equal variances assumed	.077	.783	2.631	62	.011	51.594	19.612	12.390	90.798
	Equal variances not assumed			2.631	56.70	.011	51.594	19.612	12.317	90.871

FIGURE 24.6: *The SPSS output for the phone-reaction data*

From the SPSS output, the standard error is s.e.$(\bar{x}_P - \bar{x}_C) = 19.612$. From the jamovi or SPSS output, the exact 95% CI is from 12.3 to 90.9.

The *approximate* CI and the *exact* (from SPSS) CIs are only slightly different, as SPSS uses an *exact* multiplier (whereas manually an approximate t-multiplier of 2 is used, based on the 68–95–99.7 rule), and the sample sizes aren't too small.

24.8 Statistical validity conditions: Two independent means

As usual, these results apply under certain conditions. The CI computed above is statistically valid if *one* of these conditions is true:

1. *Both* sample sizes are at least 25; *or*
2. Either sample size is smaller than 25, **and** the *populations* corresponding to both comparison groups have an approximate normal distribution.

The sample size of 25 is a rough figure here, and some books give other values (such as 30). We can explore the histograms of the *samples* to determine if normality of the *populations* seems reasonable.

In addition to the statistical validity condition, the CI will be

- **internally valid** if the study was well designed; and
- **externally valid** if the sample is a simple random sample and is internally valid.

Example 24.2 (Statistical validity)**.** For the reaction-time data, both samples are larger than 25, so the CI will be statisticall valid.

Example 24.3 (Statistical validity)**.** In the whales examples of Example 24.1, the two sample sizes are 26 (for females) and 30 (for males).

Since both samples are larger than 25, the CI will be statistically valid.

24.9 Error bar charts

A useful way to display the CIs from two (or more) groups is with an *error bar chart*, which displays the *CIs* (or sometimes the *standard errors*) for each group being compared. (A boxplot displays the *data*.)

Error bars charts display the expected variation *in the sample means* from sample to sample, while boxplots display the variation *in the individual observations* and show the median.

For the reaction time data, the error bar chart (Fig. 24.7) shows the 95% CI for each group (the mean has been added as a dot).

Think 24.4 (Comparing graphs)**.** *What is different about the information displayed in the error bar chart in (Fig. 24.7) and the boxplot (Fig. 24.1)?*

Example 24.4 (Error bar charts)**.** A study (Aloy et al. 2011) examined the impact of plastic

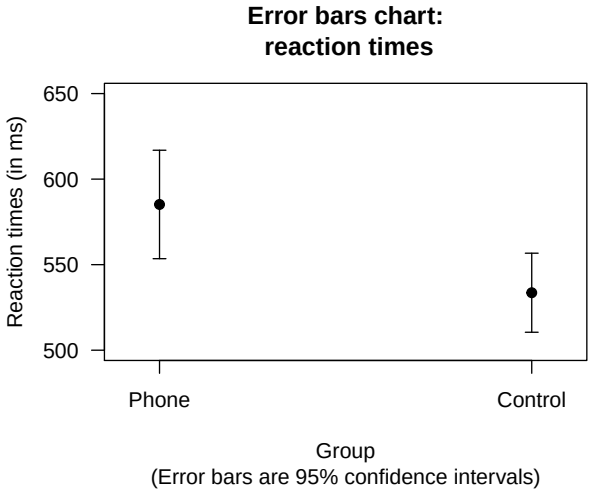


FIGURE 24.7: *Error bar chart comparing the mean reaction time for students using a mobile phone and not using a mobile phone (control)*

litter on the shoreline at Talim Bay (Batangas, Philippines) during various seasons, and the impact on the gastropod *Nassarius pullus*. The error bar chart (Fig. 24.8) shows that summer seems different—in terms of average value (mean) and the amount of variation—than the other seasons.

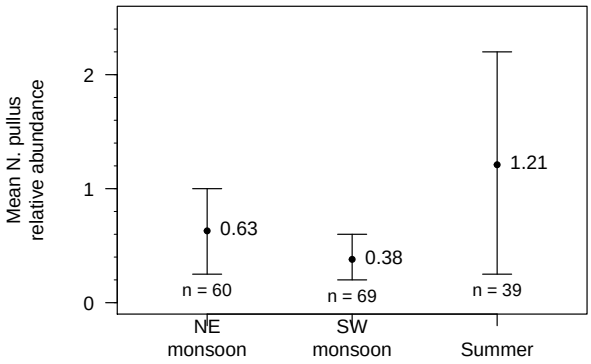


FIGURE 24.8: *Relative abundance of a gastropod from random quadrat surveys conducted over prevalent monsoon types in all study areas in Talim Bay. Error bars represent 95% confidence intervals.*

Example 24.5 (Error bar charts)**.** A study (Schepaschenko et al. 2017) examined the foliage biomass of small-leaved lime trees from three sources were studied: coppices; natural; planted.

Two graphical summaries are shown in Fig. 24.9: a boxplot (showing the variation in *individual* trees) and an error bar chart (showing the variation in the *sample means*). Using a better scale for the error-bar plot is helpful (Fig. 24.10).

Example 24.6 (Whales)**.** Using the data about gray whales from Example 24.1, the error bar chart in Fig. 24.11 can be constructed.

The plot seems to suggest little difference between the mean length of female and male gray whales at birth.

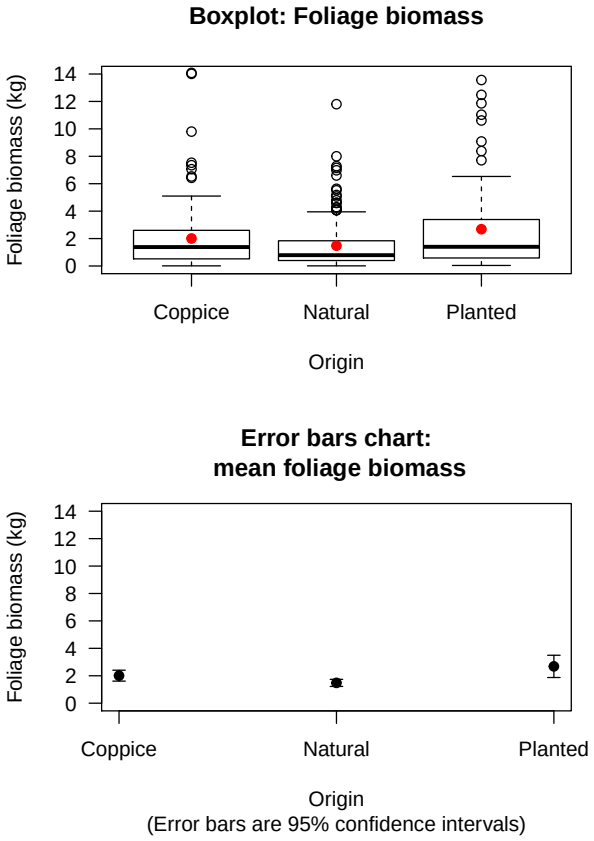


FIGURE 24.9: *Boxplot and error bar chart comparing the mean foliage biomass for small-leaved lime trees from three sources, using the same vertical scale. The solid dots in the boxplot are shown the mean of the distributions*

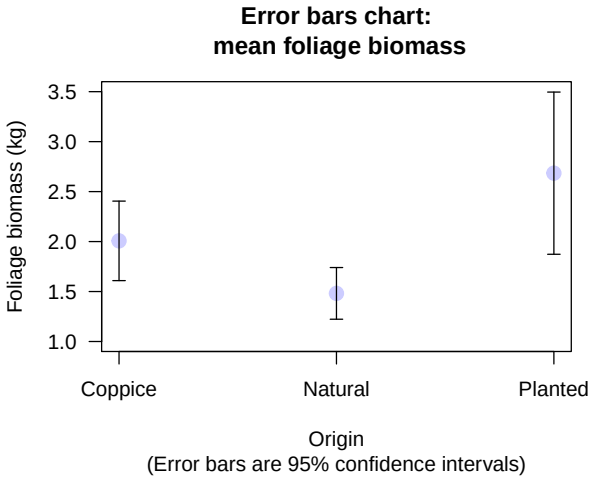


FIGURE 24.10: *Error bar chart comparing the mean foliage biomass for small-leaved lime trees from three sources, but with a more sensible scale on the vertical axis*



FIGURE 24.11: *Error bar chart comparing the lengths of female and male gray whales*

24.10 Example: Health Promotion services

A study (Becker et al. 1991) compared the access to health promotion (HP) services for people with and without a disability.

Access was measured using the quantitative *Barriers to Health Promoting Activities for Disabled Persons* (BHADP) scale[1]. *Higher* scores mean *greater* barriers to health promotion services.

The RQ is:

> What is the difference between the mean BHADP scores, for people with and without a disability?

The parameter is $\mu_D - \mu_{ND}$, the difference between the two population means (disability, minus non-disability).

In this case, only summary data is available (Table 24.5): the data is not available. Nonetheless, a useful graphical summary (an error bar chart) can be produced by computing the CI for each group manually (Fig. 24.12).

The best estimate of the difference between the population means is the difference between sample means: $(\bar{x}_D - \bar{x}_{ND}) = 6.76$. The standard error for estimating this *difference* is s.e.$(\bar{x}_D - \bar{x}_{ND}) = 0.80285$, as *given* in the table.

> The *standard error is given*; you *cannot* easily calculate this from the other information. You are not expected to do so.

Based on the sample, an approximate 95% CI for the difference in population mean BHADP scores between people with and without a disability is

[1] http://www.utexas.edu/nursing/chpr/resources/bhadp.html

TABLE 24.5: *The BHADP data summary*

	Sample mean	Std deviation	Sample size	Std error
Disability	31.83	7.73	132	0.6728
No disability	25.07	4.8	137	0.4101
Difference	6.76			0.80285

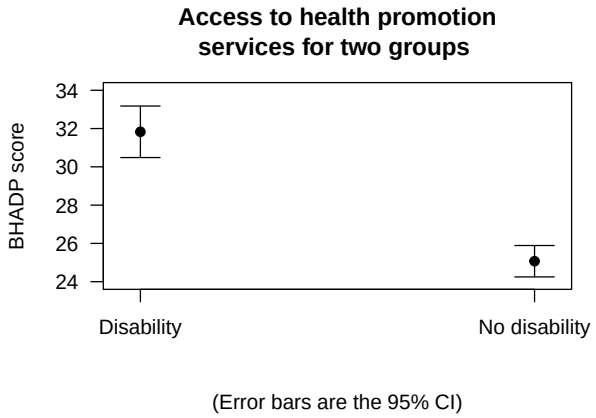


FIGURE 24.12: *Error bar chart showing the mean BHADP score for people with and without a disability, and the 95% CIs*

$$6.76 \pm (2 \times 0.80285),$$

or from 5.15 to 8.37, higher for those with a disability.

This means that, if many samples of size 132 and 137 were found, and the difference between the mean BHADP scores were found, about 95% of the CIs would contain the population difference ($\mu_D - \mu_{ND}$). Loosely speaking, there is a 95% chance that our CI straddles the difference in the population means ($\mu_D - \mu_{ND}$).

Using the validity conditions, the CI is statistically valid.

⚠ Remember: clearly state which mean is larger.

24.11 Example: Face-plant study

A study (Wojcik et al. 1999) compared the lean-forward angle in younger and older women. An elaborate set-up was constructed to measure this angle, using a harnesses.

Consider the RQ:

> Among healthy women, what is difference between the mean lean-forward angle for younger women compared to older women?

The parameter is $\mu_Y - \mu_O$, the difference between the two population means (younger, minus older).

The data are shown in Table 24.6. An appropriate graph for displaying the *data* is a *boxplot* or *dotplot* (since the sample sizes are small).

The appropriate numerical summary for the means of two independent samples summarises both groups, and (most importantly) the *difference* (Table 24.7). Summarising the *difference* is important, as the RQ is about those differences.

The error bar chart is the best plot for comparing the *mean* of the two groups (Fig. 24.13).

TABLE 24.6: *Lean-forward angles for older and younger women*

Younger women		Older women
29	34	18
32	27	15
34	32	23
31	28	13
33	27	12

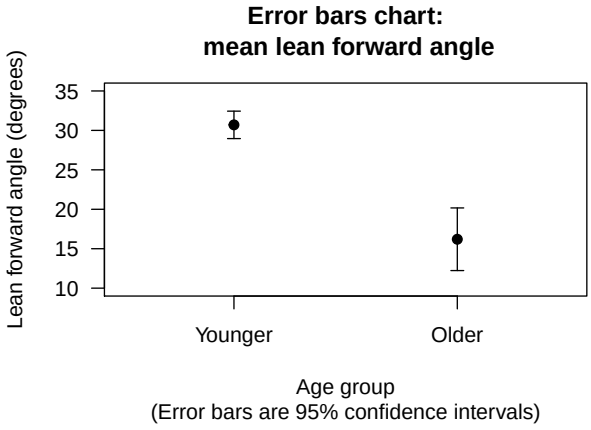


FIGURE 24.13: *Plot of the face-plant data*

TABLE 24.7: *Numerical summary for the face-plant data (means, standard deviations and standard errors are in degrees)*

	Mean	Standard deviation	Standard error	Sample size
Younger women	30.7	2.75	0.87	10
Older women	16.2	4.44	1.98	5
Difference	14.5		2.17	

The *second row* of the jamovi output (Fig. 24.14) and SPSS output (Fig. 24.15) show that the 95% CI is from 9.10 to 19.90. (We could also compute the *approximate* 95% CI manually.) After rounding the numbers:

> Based on the sample, a 95% CI for the difference between population mean one-step fall-recovery angle for heathy women is between 9.1 and 19.9 degrees *greater* for younger women than for older women (two independent samples).

Independent Samples T-Test

Independent Samples T-Test

					95% Confidence Interval	
		statistic	df	p	Lower	Upper
LeanAngle	Student's t	7.88	13.0	< .001	10.5	18.5
	Welch's t	6.69	5.59	< .001	9.10	19.9

FIGURE 24.14: *jamovi output for the face-plant data*

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Maximum lean angle (in degrees)	Equal variances assumed	1.578	.231	7.875	13	.000	14.500	1.841	10.522	18.478
	Equal variances not assumed			6.691	5.592	.001	14.500	2.167	9.102	19.898

FIGURE 24.15: *SPSS output for the face-plant data*

The statement clearly states which group has the higher mean (younger women). This CI tells us that if we found many samples (of sizes 10 and 5) in the same way, and computed the CI for the difference between the mean from each sample, about 95% of the CIs would contain the difference between the means in the population: $\mu_Y - \mu_O$. Loosely speaking: There is a 95% chance that our CI straddles $\mu_Y - \mu_O$.

The CI may not be statistically valid (as the sample sizes are not large), so the CIs may not be accurate.

24.12 Quick review questions

1. The appropriate graph for displaying quantitative *data* for two separate groups is a:
2. True or false: The difference in population means could be denoted by $\mu_A - \mu_B$.
3. True or false: The standard error of the difference between the sample means is denoted by $\text{s.e.}(\bar{x}_A) - \text{s.e.}(\bar{x}_B)$.

24.13 Exercises

Selected answers are available in Sect. D.23.

Exercise 24.1. Earlier, we studied the NHANES study (Sect. 12.10), and this RQ:

Among Americans, is the mean direct HDL cholesterol different for current smokers and non-smokers?

Use the SPSS output (Fig. 24.16) to answer these questions.

1. Construct an appropriate table showing the numerical summary.
2. Determine, and suitably communicate, the 95% CI for the difference between the direct HDL cholesterol values between current smokers and non-smokers.

T-Test

Group Statistics

	SmokeNow	N	Mean	Std. Deviation	Std. Error Mean
DirectChol	No	1668	1.3924	.42792	.01048
	Yes	1388	1.3077	.42353	.01137

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Direct Chol	Equal variances assumed	.204	.652	5.473	3054	.000	.08470	.01547	.05435	.11504
	Equal variances not assumed			5.478	2964.437	.000	.08470	.01546	.05438	.11501

FIGURE 24.16: *SPSS output for the NHANES data*

Exercise 24.2. A study (Barrett et al. 2010) of the effectiveness of echinacea to treat the common cold compared, among other things, the duration of the cold for participants treated with echinaca or a placebo. Participants were blinded to the treatment, and allocated to the groups randomly. A summary of the data is given in Table 24.8.

1. Compute the standard error for the mean duration of symptoms for each group.
2. Compute an approximate 95% CI for the *difference* between the mean durations for the two groups.
3. In which direction is the difference computed? What does it *mean* when the difference is calculated in this way?
4. Compute an approximate 95% CI for the population mean duration of symptoms for those treated with echinacea.
5. Are the CIs likely to be statistically valid?

TABLE 24.8: *Numerical summary of duration (in days) of common cold symptoms, for blinded patients taking echinacea or a placebo*

	Mean	Std deviation	Std error	Sample size
Placebo	6.87	3.62		176
Echinacea	6.34	3.31		183
Difference	0.53		0.367	

Exercise 24.3. Carpal tunnel syndrome (CTS) is pain experienced in the wrists. One study (Schmid et al. 2012) compared two different treatments: night splinting, or gliding exercises. Participants were *randomly allocated* to one of the two groups. Pain intensity (measured using a quantitative visual analog scale; *larger* values mean *greater* pain) were recorded after one week of treatment. The data are summarised in Table 24.9.

1. Compute the standard error for the mean pain intensity for each group.
2. In which direction is the difference computed? What does it *mean* when the difference is calculated in this way?
3. Compute an approximate 95% CI for the *difference* in the mean pain intensity for the treatments.
4. Compute an approximate 95% CI for the population mean pain intensity for those treated with splinting.
5. Are the CIs likely to be statistically valid?

TABLE 24.9: *Numerical summary of pain intensity for two different treatments of carpal tunnel syndrome*

	Mean	**Std deviation**	**Std error**	**Sample size**
Exercise	0.8	1.4		10
Splinting	1.1	1.1		10
Difference	0.3		0.563	

Exercise 24.4. A study (Woodward and Walker 1994) examined the sugar consumption in industrialised (mean: 41.8 kg/person/year) and non-industrialised (mean: 24.6 kg/person/year) countries. Using the jamovi output (Fig. 24.17), write down and interpret the CI.

Independent Samples T-Test

Independent Samples T-Test

							95% Confidence Interval	
		statistic	df	p	Mean difference	SE difference	Lower	Upper
Sugar	Student's t	−5.25 [a]	88.0	< .001	−17.2	3.29	−23.8	−10.7
	Welch's t	−6.47	87.2	< .001	−17.2	2.66	−22.5	−11.9

[a] Levene's test is significant (p < .05), suggesting a violation of the assumption of equal variances

Group Descriptives

	Group	N	Mean	Median	SD	SE
Sugar	No	61	24.6	24.2	16.6	2.13
	Yes	29	41.8	44.0	8.63	1.60

FIGURE 24.17: *jamovi output for the sugar-consumption data*

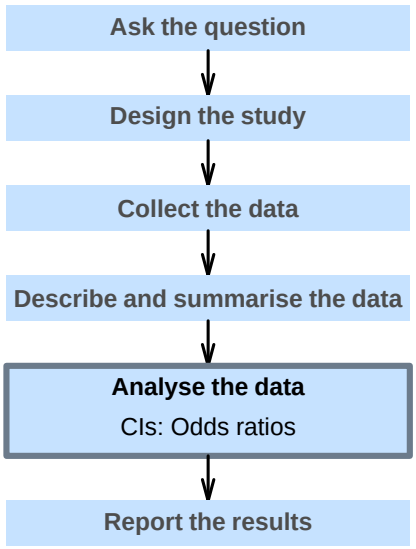
25

Confidence intervals for odds ratios

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, and understand the tools of inference.

In this chapter, you will learn about *confidence intervals* for odds ratios. You will learn to:

- produce confidence intervals for odds ratios using jamovi and SPSS output.
- determine whether the conditions for using the confidence intervals apply in a given situation.



25.1 Introduction: Odds ratios

A study (Mann and Blotnicky 2017) examined the eating habits of university students. One issue studied was the relationship between eating on-campus, and where the student lived.

The researchers cross-classified the $n = 183$ students into groups: each student (the unit of analysis) was observed on two *qualitative* variables:

- Where they lived: With their parents, or *not* with their parents;
- Whether they ate most meals *off-campus*, or most meals *on-campus*.

Since both variables are *qualitative*, **means are not appropriate** for summarising the data.

However, the students can be classified into a two-way table of counts (Table 25.1), called a *contingency table*. Both qualitative variables have two levels, so the table is a 2×2 table.

TABLE 25.1: *Where university students live and eat*

	Lives with parents	Doesn't live with parents	Total
Most meals off-campus	52	105	**157**
Most meals on-campus	2	24	**26**
Total	**54**	**129**	**183**

The purpose of the research is to study the odds (or proportion) of students who eat most meals off-campus, comparing those who live with their parents and those who do *not* live with their parents?

The parameter of interest could be the difference between the **proportions** (or **percentages**) in each group, a comparison between the **odds** in each group, or the **odds ratio**.

For reasons that we can't delve into, usually the **odds ratio** (OR)[1] is used as the parameter. One important reason is that software produces output related to the sample OR.

⚠ To compare two groups with regard to another qualitative variable, software usually works with **odds** rather than percentages or proportions.

For this reason, writing the RQ in terms of odds is also most appropriate.

Using the OR, the RQ could be written as

> Among university students, what is the odds ratio of students eating most meals off-campus, comparing those who *do* and *do not* live with their parents?

The parameter is the population OR, comparing the odds of eating most meals *off*-campus for students living with their parents to students *not* living with their parents.

⚠ Take care in defining the odds ratios in the parameter!

Recall (Sect. 14.2 that software usually compares Row 1 to Row 2, and Columnn 1 to Column 2

Think 25.1 (POCI)**.** *What is P, O, C and I for this RQs?*

[1] `OddsRatio`

25.2 Numerical and graphical summaries: Comparing odds

With two qualitative variables, an appropriate numerical summary includes the odds and percentages from each comparison group for the outcome of interest, and the sample sizes.

From these data, the odds[2] of eating most meals *off-campus* is:

- $52 \div 2 = 26$ for students *living with their parents*.
- $105 \div 24 = 4.375$ for students *not living with their parents*.

So the *odds ratio* (OR)[3] of eating most meals *off-campus*, comparing students living with parents to students *not* living with parents, is $26 \div 4.375 = 5.943$.

The numerical summary (Table 25.2) shows the percentage and odds of eating most meals off-campus, comparing students living at home and those not living at home.



Understanding how software computes the odds ratio is important for understanding the output. In jamovi and SPSS, the odds ratio can be interpreted in *either* of these two ways:

- The *odds* are the odds of eating most meals *off-campus* (Row 1 of Table 25.1).
  Then, the odds ratio compares these odds for students living with their parents (Column 1 of Table 25.1) to those *not* living with their parents (Column 2 of Table 25.1).
  That is, the odds are $52/2 = 26$ (for those living with parents) and $105/24 = 4.375$ (for those not living with parents), so the OR is then $26/4.375 = 5.943$, as in the output (Fig. 25.2).
- The *odds* are the odds of living with parents (Column 1 of Table 25.1).
  Then, the odds ratio compares these odds for students eating most meals off-campus (Row 1 of Table 25.1) to the odds of students eating most meals on-campus (Row 2 of Table 25.1).
  That is, the odds of living with parents are $52/105 = 0.49524$ (for those eating most meals off-campus) and $2/24 = 0.083333$ (for those eating most meals on-campus), so the OR is then $0.49524/0.083333 = 5.943$, as in the output (Fig. 25.2).

In other words, the odds and odds ratios are relative to the **first row** or **first column**.

TABLE 25.2: *The odds and percentage of university students eating most meals off-campus*

	Odds of having most meals off-campus	Percentage having most meals off-campus	Sample size
Living with parents	0.4952	16.6	54
Not living with parents	0.0833	3.8	129
Odds ratio	5.943		

An appropriate graph (Fig. 25.1) is a *side-by-side* bar chart or a *stacked* bar chart. For comparing the *odds*, the side-by-side bar chart is better. (A *stacked* bar chart is better for comparing *proportions*, but either is fine.)

[2] `Odds`
[3] `OddsRatio`

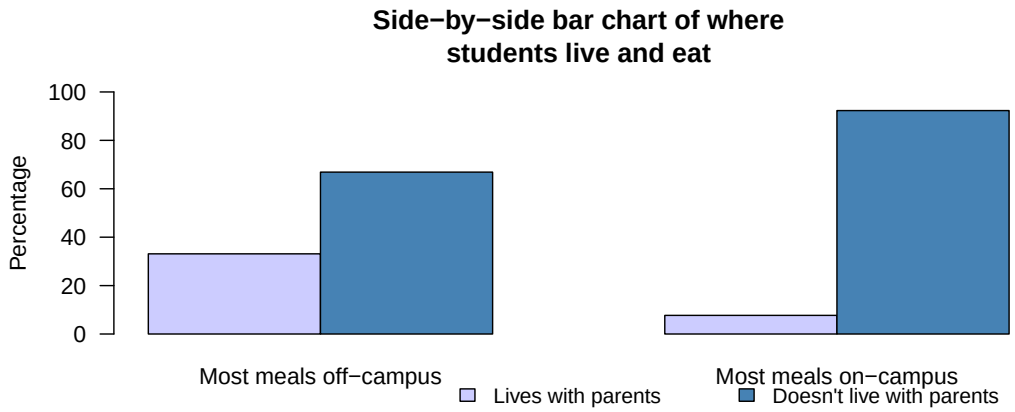


FIGURE 25.1: *A plot of the uni-student eating data: A side-by-side bar chart*

25.3 Sampling distribution: Comparing odds

From the numerical summary table (Table 25.2), the odds of a student eating most meals *off-campus* is:

- 26 for students *living with their parents*.
- 4.375 for students *not living with their parents*.

So the OR of eating most meals *off-campus*, comparing students living with parents to students *not* living with parents, is $26 \div 4.375 = 5.943$. The odds are different in each group, and hence the OR is not one. The OR means that the odds of eating most meals off-campus is 5.943 times larger for students living *with* their parents.

Of course, every sample of students is likey to be different, so the OR *varies* from sample to sample, so there is *sampling variation*. This means that the odds ratio has a *sampling distribution* and a *standard error*.

Unfortunately, the sampling distribution of the sample OR is not a normal distribution.[4] Fortunately, a simple transformation to the sample OR has a normal distribution. For this reason, we will use software output for finding the CI for the odds ratio, and not discuss the sampling distribution directly. In other words, we will rely on software to find CIs for odds ratios.

25.4 Confidence intervals: Comparing odds

As noted, we rely on software to find the CI for the odds ratio, such as jamovi (Fig. 25.2) and the *second table* of the SPSS output (labelled **Risk Estimate**; Fig. 25.3). Both show that the

[4]For those who want to know (this is *optional*): The OR is only defined for *non-negative* values so a normal distribution is inappropriate. However, the *logarithm* of the OR has an approximate normal distribution under certain conditions.

sample OR is 5.94, and the (exact) 95% CI is from 1.35 to 26.1. (The SPSS output shows other information too, some of which will be useful later.)

Comparative Measures

		95% Confidence Intervals	
	Value	Lower	Upper
Odds ratio	5.94	1.35	26.1

FIGURE 25.2: *The jamovi output for computing a CI*

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.934[a]	1	.008		
Continuity Correction[b]	5.765	1	.016		
Likelihood Ratio	8.528	1	.003		
Fisher's Exact Test				.009	.005
Linear-by-Linear Association	6.896	1	.009		
N of Valid Cases	183				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.67.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Meals (Most off-campus / Most on-campus)	5.943	1.352	26.114
For cohort Live = Living with parents	4.306	1.116	16.608
For cohort Live = Not living with parents	.725	.620	.847
N of Valid Cases	183		

FIGURE 25.3: *The SPSS output for computing a CI*



Recall that jamovi and SPSS compute the odds ratio as either

- 'Row 1 to Row 2, comparing Column 1 to Column 2,' or
- 'Column 1 to Column 2, comparing Row 1 to Row 2.'

The OR can be interpreted either way.

We write:

> Based on the sample, a 95% CI for the OR comparing the odds of eating most meals off-campus is from 1.35 to 26.1 (living with parents, compared to *not* living with parents).

This means there is a 95% chance that this CI straddles the population OR.

Notice that the meaning of the OR is explained in the conclusions: the odds of eating most meals *off*-campus, and comparing students living with parents to *not* living with parents.

The CI for an OR is not symmetrical, like the others we have seen[5].

ⓘ Interpreting ORs can be confusing, so take care!

Example 25.1 (Crashes in China)**.** A study of car crashes in a rural, mountainous county in western China (Wang et al. 2020) recorded the data in the table below.

Type of crash	2011	2015
Involving pedestrians	15	37
Involving vehicles	35	85

Clearly the *number* of crashes is larger in 2015. However, the interest is in comparing the *odds* (or percentage) of crashes involving pedestrians in 2011 and 2015. (Of course, comparing the odds (or percentages) involving *vehicles* is also possible.)

The data can be summarised as shown below.

Year	Percentage involving pedestrians	Odds involving pedestrians	Sample size
In 2011	30.0	0.429	50
In 2015	30.3	0.435	122
	Odds ratio:	**0.985**	

In this table, the *odds* are the odds that a crash involves a pedestrian.

The *odds ratio* is the odds of a crash involving pedestrians in 2011, compared to the odds of a crash involving pedestrians in 2015. In this situation, this is the *parameter* of interest.

Both the percentage and odds columns, and the odds ratio, suggest that the relative proportion of crashes involving pedestrians is very similar in 2011 and 2015.

The odds ratio is 0.986, but this value would change from sample to sample. From software, the 95% CI for the odds ratio is from 0.480 to 2.018. We would write

> The population odds ratio for a crash involving pedestrians (comparing 2011 to 2015) has a 95% chance of being between 0.480 and 2.018.

[5] This is because the OR has no upper limit, but the lower limit is zero. (The *logarithm* of the limits of the CI form a symmetric interval.)

25.5 Statistical validity conditions: Comparing odds

As usual, these results hold under certain conditions. The CI computed above is statistically valid if

- All *expected* counts are at least five.

Some books may give other (but similar) conditions.

In addition to the statistical validity condition, the CI will be

- **internally valid** if the study was well designed; and
- **externally valid** if the sample is a simple random sample and is internally valid.

The statistical validity condition is a bit tricky to understand (but is explained further in Sect. 31.3). SPSS will let you know if the expected count condition is not met, underneath the *first* output table in Fig. 25.3. In jamovi, the *expected* counts must be explicitly requested to see if this condition is satisfied.

Example 25.2 (Statistical validity)**.** In Fig. 25.3 (for the uni-students data), the text under the *first table* table of SPSS output (labelled **Chi-Square Tests**) says

> 0 cells (0.0%) have expected count less than 5.

That is, *all* the cells have expected counts of at least five, so the statistical validity condition is satisfied. Notice from Table 25.1 that the *observed* counts are not all greater than five (one cell has a count of 2). The statistical validity condition is about the *expected* counts though, not the *observed* counts.

In jamovi, the expected counts must be requested explicitly (Fig. 25.4), but again none are less than five.

In either case, the conclusion is statistically valid.

Contingency Tables

Meals		Live	
		Living with parents	Not living with parents
Most on-campus	Expected	46.33	110.7
Most off-campus	Expected	7.67	18.3
Total	Expected	54.00	129.0

FIGURE 25.4: *The expected counts in jamovi, for the uni-students data*

Example 25.3 (Car crashes in China)**.** In Example 25.1, all the *observed* counts are larger than five.

The *expected* counts are shown below. Since all *expected* counts are larger than five, the CI will be statistically valid.

Type of crash	2011	2015
Involving pedestrians	15.11	36.88
Involving vehicles	34.88	85.12

These counts are what we would *expected* to find if there was no relationship between the type of crash in 2011 and 2015; that is, if the proportion of crashes involving pedestrians was the same in 2011 and 2015.

The observed counts are *very* close to these *expected* counts, meaning that what we observe is very close to what we expected if there was no relatiionship.

25.6 Example: Pet birds

A study examined people with lung cancer, and a matched set of controls who did not have lung cancer, and compared the proportion in each group that kept pet birds (Kohlmeier et al. 1992). One RQ of the study was:

> What is the odds ratio of keeping a pet bird, comparing people *with* lung cancer (cases) compared to people *without* lung cancer (controls)?

The parameter is the population OR, comparing the odds of keeping a pet bird, for adults with lung cancer to adults who do not have lung cancer.

The data, compiled in a 2×2 *contingency table*, are given in Table 25.6. The numerical summary (Table 25.7) contains percentages, odds and the odds ratios; some of these may need to be computed *manually* from the data. The graphical summary (Fig. 25.5) shows a difference between the two groups *in the sample*.

Software computes the CI for the *population* odds ratio (jamovi: Fig. 25.6; SPSS: Fig. 25.7) based on the sample. The *sample* OR is 2.257, and the 95% CI is from 1.605 to 3.174.

We write:

> Based on the sample, a 95% CI for the OR of keeping a pet bird is from 1.605 to 3.174 (comparing people *with* lung cancer to those *without* lung cancer).

That is, the plausible values for the population OR that could have produced the sample OR are between 1.605 and 3.174.

The CI will be statistically valid if the sample is somewhat representative of some population. We see that the text under the **first** table of SPSS output (Fig. 25.7) indicates that the expected-counts condition is met.

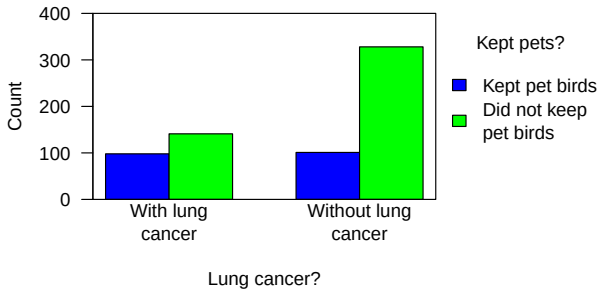


FIGURE 25.5: *A plot of the pet-birds data*

Comparative Measures

		95% Confidence Intervals	
	Value	Lower	Upper
Odds ratio	2.26	1.61	3.17

FIGURE 25.6: *jamovi output for the pet-birds data*

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	22.374[a]	1	.000		
Continuity Correction[b]	21.547	1	.000		
Likelihood Ratio	21.924	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	22.341	1	.000		
N of Valid Cases	668				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 71.20.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval Lower	Upper
Odds Ratio for Pets (Kept pet birds / Did not keep pet birds)	2.257	1.605	3.174
For cohort LC = Adults with lung cancer	1.638	1.345	1.995
For cohort LC = Adults without lung cancer	.726	.625	.842
N of Valid Cases	668		

FIGURE 25.7: *SPSS output for the pet-birds data*

TABLE 25.6: *The pet bird data*

	Adults with lung cancer	Adults without lung cancer
Kept pet birds	98	101
Did not keep pet birds	141	328

TABLE 25.7: *The odds and percentage of subjects keeping pet birds*

	Odds of keeping pet bird	**Percentage keeping pet bird**	**Sample size**
With lung cancer:	0.6950	41.0%	238
Without lung cancer:	0.3079	25.5%	429
Odds ratio:	2.26		

25.7 Example: B12 deficiency

A study in New Zealand (Gammon et al. 2012) examined B12 deficiencies in 'predominantly overweight/obese women of South Asian origin living in Auckland,' some of whom were on a vegetarian diet and some of whom were on a non-vegetarian diet. One RQ was:

> What is the odds ratio of these women being B12 deficient, comparing vegetarians to non-vegetarians?

The parameter is the population OR, comparing the odds of being B12 deficient, for vegetarians to non-vegetarians.

The data appear in Table 25.8. From the jamovi output (Fig. 25.9) or SPSS output (Fig. 25.10), the OR (and 95% CI) is 3.15 (1.08 to 9.24). The numerical summary table (Table 25.9) and graphical summary (Fig. 25.8). can hence be constructed.

TABLE 25.8: *The number of vegetarian and non-vegetarian women who are (and are not) B12 deficient*

	B12 deficient	**Not B12 deficient**	**Total**
Vegetarians	8	26	**34**
Non-vegetarians	8	82	**90**
Total	**16**	**108**	**124**

To check if these results statistically valid, notice that the text under the **first** table of SPSS output (Fig. 25.10) says:

> 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.39.

This a warning that one *expected* count is less than 5. Nonetheless, only *one* cell has an expected count less than five, and only *just* under 5, so we shouldn't be too concerned about statistical validity (but it should be noted).

We write:

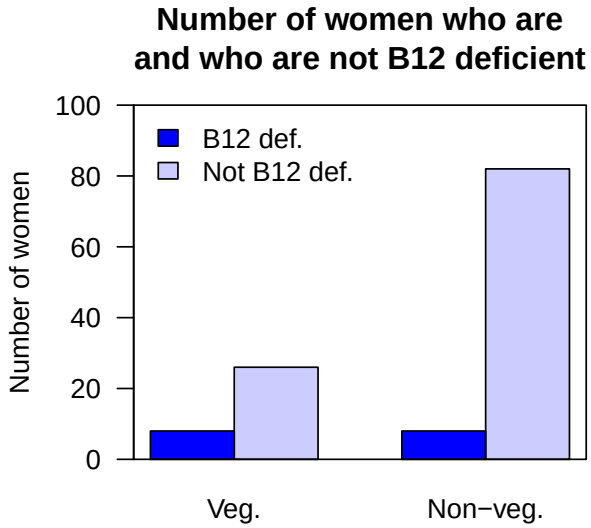


FIGURE 25.8: *A side-by-side barchart comparing the number of women B12 deficient*

TABLE 25.9: *The odds and percentage of subjects that are B12 deficient*

	Odds B12 deficient	**Percentage B12 deficient**	**Sample size**
Vegetarians:	0.3077	23.5%	34
Non-vegetarians:	0.0976	8.9%	90
Odds ratio:	3.15		

Based on the sample, a 95% CI for the OR of being B12 deficient is from 1.08 to 9.24 (comparing vegetarians to *non*-vegetarians).

25.8 Quick review questions

A study (Egbue et al. 2017) of the adoption of electric vehicle (EVs) by a certain group of professional Americans (Example 5.14) compiled the data in Table 25.10. Output from using jamovi is shown in Fig. 25.11.

Comparative Measures

		95% Confidence Intervals	
	Value	Lower	Upper
Odds ratio	3.15	1.08	9.24

FIGURE 25.9: *jamovi output for the B12 data*

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.707[a]	1	.030		
Continuity Correction[b]	3.494	1	.062		
Likelihood Ratio	4.273	1	.039		
Fisher's Exact Test				.039	.035
Linear-by-Linear Association	4.669	1	.031		
N of Valid Cases	124				

a. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.39.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval Lower	Upper
Odds Ratio for Diet (Vegetarian / Non-vegetarian)	3.154	1.077	9.238
For cohort B12 Deficiency = B12 deficient	2.647	1.079	6.492
For cohort B12 Deficiency = Not B12 deficient	.839	.689	1.022
N of Valid Cases	124		

FIGURE 25.10: *SPSS output for the B12 data*

TABLE 25.10: *Responses to the question 'Would you purchase an electric vehicle in the next 10 years?' by education*

	Yes	No
No post-grad	24	8
Post-grad study	51	29

Comparative Measures

	Value	95% Confidence Intervals Lower	Upper
Odds ratio	1.71	0.68	4.28

FIGURE 25.11: *jamovi output for the EV study*

1. The *percentage* of people without post-grad study who would buy an EV in the next 10 years is (**do not** add the percentage symbol):
2. The *odds* that a person without post-grad study would buy an EV in the next 10 years is:
3. Using the output, what is the OR of buying an electric vehicle in the next 10 years, comparing those *without* post-grad study to those *with* post-grad study?
4. True or false: The CI means that the sample OR is likely to be between 0.68 and 4.28.
5. True or false: The analysis is likely to be statistically valid?

Answer:

1. The number *without* post-grad study: $24 + 8 = 32$. The *percentage* of people without post-grad study who would buy an EV in the next 10 years is $24/32 = 0.75$, or 75%.
2. The people without post-grad study are in the *top* row. The *odds* of people without post-grad study who would buy an EV in the next 10 years is $24/8 = 3$.
3. The odds of people *without* post-grad study who would by an electric vehicle is $24/8 = 3$.
   The odds of people *with* post-grad study who would by an electric vehicle is $51/29 = 1.7586$.
   So the OR is $3/1.7586 = 1.706$.
4. Not at all. We know *exactly* what the sample OR is (it is 1.706). CIs always give an interval in which the *population parameter* is likely to be within.
5. The CI is statistically valid if all the *expected* counts exceed 5. So we don't really know for sure from the given information. But the *observed* counts are all reasonably large, so it is *very probably* statistically valid.

25.9 Exercises

Selected answers are available in Sect. D.24.

Exercise 25.1. A prospective observational study (Wallace et al. 2017) compared the heights of scars from burns received in Western Australia (Table 25.11). jamovi was used to analyse the data (Fig. 25.12).

1. Compute the *odds* of having a smooth scar (that is, height is 0mm) for women.
2. Compute the *odds* of having a smooth scar (that is, height is 0mm) for men.
3. Compute the *odds ratio* of having a smooth scar, comparing women to men.
4. Interpret what this odds ratio means.
5. Sketch a suitable graph to display the data.
6. Construct an appropriate numerical summary table for the data.
7. Write down the CI.
8. Carefully interpret what this CI means.

Comparative Measures

		95% Confidence Intervals	
	Value	Lower	Upper
Odds ratio	0.850	0.576	1.26

FIGURE 25.12: *jamovi output for the scar-height data*

TABLE 25.11: *The number of men and women, with scars of different heights*

	Women	**Men**
Scar height 0mm (smooth)	99	216
Scar height more than 0mm, less than 1mm	62	115

Exercise 25.2. A study of ear infections in Sydney swimmers (Smyth 2010) recorded whether people reported an ear infection or not, and where they usually swam.

The SPSS output is shown in Fig. 25.13. Explain carefully the meaning of the OR and the corresponding CI.

Had ear infection? * Usual swimming location Crosstabulation

Count

		Usual swimming location		
		Beach	NonBeach	Total
Had ear infection?	No	90	61	151
	Yes	57	79	136
Total		147	140	287

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8.963[a]	1	.003		
Continuity Correction[b]	8.269	1	.004		
Likelihood Ratio	9.008	1	.003		
Fisher's Exact Test				.003	.002
N of Valid Cases	287				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 66.34.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Had ear infection? (No / Yes)	2.045	1.277	3.274
For cohort Usual swimming location = Beach	1.422	1.121	1.803
For cohort Usual swimming location = NonBeach	.695	.547	.885
N of Valid Cases	287		

FIGURE 25.13: *SPSS output for the ear-infection data*

Exercise 25.3. A study of turbine failures (Nelson 1982; Myers et al. 2002) ran 73 turbines for around 1800 hours, and found that seven developed fissures (small cracks). They also ran a different set of 42 turbines for about 3000 hours, and found that nine developed fissures.

1. Construct the two-way table for the data.
2. Use the jamovi output (Fig. 25.14) to construct a 95% CI for the odds ratio.
3. Compute, then carefully interpret, the OR.
4. Write down, then carefully interpret, the CI for the OR.

5. Is the CI likely to be statistically valid (Fig. 25.15)?

χ^2 Tests

	Value	df	p
χ^2	3.12	1	0.077
N	115		

Comparative Measures

		95% Confidence Intervals	
	Value	Lower	Upper
Odds ratio	0.39	0.13	1.14

FIGURE 25.14: *jamovi output for the turbine data: output*

Contingency Tables

		Fissures	
Hours		Yes	No
"1800"	Expected	10.16	62.84
"3000"	Expected	5.84	36.16
Total	Expected	16	99

FIGURE 25.15: *jamovi output for the turbine data: expected counts*

Exercise 25.4. The *Southern Oscillation Index* (SOI) is a standardised measure of the air pressure difference between Tahiti and Darwin, and is related to rainfall in some parts of the world (Stone et al. 1996), and especially Queensland (Stone and Auliciems 1992; Dunn 2001).

The rainfall at Emerald (Queensland) was recorded for Augusts between 1889 to 2002 inclusive (Dunn and Smyth 2018), where the monthly average SOI was positive, and when the SOI was non-positive (that is, zero or negative), as shown in Table 25.12.

Using the jamovi output in Fig. 25.16:

1. Find a 95% CI for the OR.
2. Carefuly explain what this OR means.

TABLE 25.12: *The SOI, and whether rainfall was recorded in Augusts between 1889 and 2002 inclusive*

	Non-positive SOI	**Positive SOI**
No rainfall recorded	14	7
Rainfall recorded	40	53

Comparative Measures

	Value	95% Confidence Intervals	
		Lower	Upper
Odds ratio	2.65	0.979	7.17

FIGURE 25.16: *jamovi output for the Emerald-rain data*

Exercise 25.5. A research study conducted in Brisbane (Dexter et al. 2019) recorded the number of people at the foot of the Goodwill Bridge, Southbank, who wore sunglasses and hats between 11:30am to 12:30pm. Table 25.13 records the number of females and males wearing hats.

Using the SPSS output in Fig. 25.17, find a 95% CI for the OR, and carefully explain what OR this CI applies to. Also, construct the numerical summary table.

TABLE 25.13: *The number of people wearing hats, for males and females*

	No hat	**Hat**
Male	307	79
Female	344	22

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	33.763[a]	1	.000		
Continuity Correction[b]	32.531	1	.000		
Likelihood Ratio	35.712	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	33.718	1	.000		
N of Valid Cases	752				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 49.16.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Hat (No / Yes)	.249	.151	.408
For cohort Gender = Male	.603	.529	.687
For cohort Gender = Female	2.426	1.665	3.535
N of Valid Cases	752		

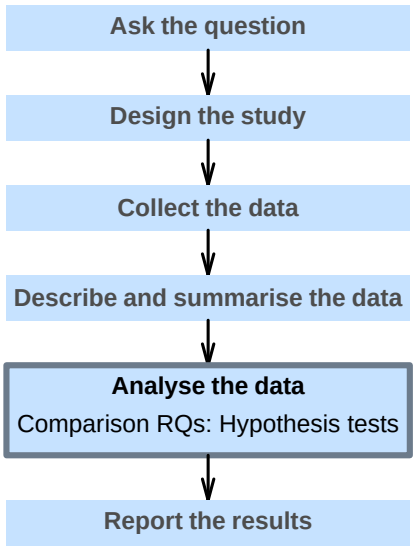
FIGURE 25.17: *SPSS output for the hats data*

Part VII

Analysis: Hypothesis testing

26

Introducing hypothesis tests



We have studied forming *confidence intervals*, which answer estimation-type RQs, and indicate the precision with which a statistic estimates a parameter.

Now, we begin looking at decision-type RQs, which help us make decisions about the value of unknown parameters based on the value of the statistic (Table 26.1). The decision-making process (Chap. 15) we discussed was:

1. **Assumption**: Make an assumption about the *population*.
2. **Expectation**: Based on this assumption, the distribution of the possible values of the sample statistic can be described.
3. **Observation**: If *sample* information is observed that is:
    - unlikely to happen by chance, it is *contrary to* that assumption about the *population parameter*, and the assumption is probably **wrong**. There is *evidence* to suggest that the assumption is wrong (but it is not *certainly* wrong).
    - *likely* to happen by chance, it is **consistent with** that assumption about the *population parameter*, and the assumption may be **correct**. There is no *evidence* to suggest the assumption is wrong (though it may be wrong).

In this Part, we explore decision-type relational or interventional RQs with a *comparison*. (Decision-type RQs with a *connection* are discussed in Chaps. 34 and 35.)

TABLE 26.1: *Confidence interrvals and hypothesis tests for different situations*

	Estimation type (CI)	Decision type (Tests)
Descriptive RQs		
Proportions for one sample	Chap. 20	
Means for one sample	Chap. 22	Chap. 27
Mean differences (for paired data)	Chap. 23	Chap. 29
Relational/Interventional RQs with a *Comparison*		
Means for two samples	Chap. 24	Chap. 30
Odds for two samples (ORs)	Chap. 25	Chap. 31
Relational/Interventional RQs with a *Connection*		
Correlation		Sect. 34.4
Regression	Sect. 35.7	Sect. 35.7

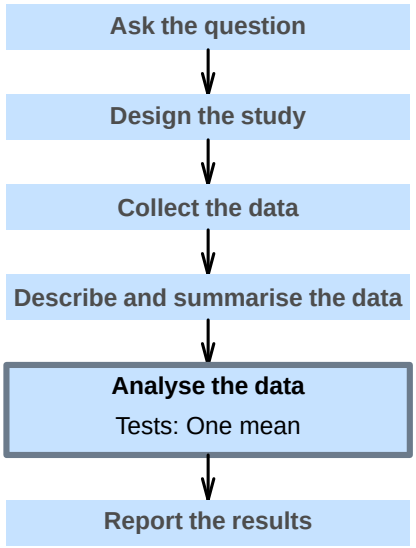
27

Hypothesis tests for one mean

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, understand the tools of inference, and to form *confidence intervals*.

In this chapter, you will learn about *hypothesis tests* for one mean. You will learn to:

- conduct hypothesis tests for one sample mean, using a t-test.
- determine whether the conditions for using these methods apply in a given situation.



27.1 Introduction: Body temperatures

The average internal body temperature is commonly believed to be $\mu = 37.0°\text{C}$, a guideline based on data over 150 years old (Wunderlich 1868). More recently, researchers wanted to re-examine this claim (Mackowiak et al. 1992) to see if this benchmark is still appropriate.

In this example, a decision is sought about the value of the *population* mean body temperature μ. The value of μ will never be known: the internal body temperature of every person alive would need to be measured... and even those not yet born.

The parameter is μ, the population mean internal body temperature.

However, a *sample* of people can be taken to determine whether or not there is evidence that the *population* mean internal body temperature is still $37.0°\text{C}$.

To make this decision, the decision-making process (Sect. 15.3) is used. Begin by **assuming** that $\mu = 37.0°\text{C}$ (as there is no evidence that this accepted standard is wrong), and then determine if the evidence supports this claim or not. The RQ could be stated as:

> Is the *population* mean internal body temperature $37.0°\text{C}$?

27.2 Hypotheses and notation: One mean

The decision making process begins by **assuming** that the population mean internal body temperature is $37.0°\text{C}$.

The sample mean $\bar{x}$ is likely to be different for every sample (*sampling variation*). The *sampling distribution* of $\bar{x}$ describes how the value of $\bar{x}$ varies from sample to sample. Because $\bar{x}$ varies, the *sample* mean $\bar{x}$ probably won't be exactly $37.0°\text{C}$, *even if* μ is $37.0°\text{C}$.

If $\bar{x}$ is not $37.0°\text{C}$, two broad reasons could explain why:

1. The *population* mean body temperature *is* $37.0°\text{C}$, but $\bar{x}$ isn't exactly $37.0°\text{C}$ due to sampling variation (that is, the sample mean varies and is likely to be different in every sample); or
2. The *population* mean body temperature *is not* $37.0°\text{C}$, and the *sample* mean body temperature reflects this.

These two possible explanations are called *hypotheses*. More formally, the two hypotheses above are:

1. The *null hypothesis* (H_0): $\mu = 37.0°\text{C}$; the *population* mean body temperature *is* $37.0°\text{C}$; and
2. The *alternative hypothesis* (H_1): $\mu \neq 37.0°\text{C}$; the *population* mean body temperature *is not* $37.0°\text{C}$.

Since the null hypothesis is assumed true, the evidence is evaluated to determine if it is supported by the data, or not.

Note that the alternative hypothesis asks if μ is $37.0°\text{C}$ or not: the value of μ may be smaller or larger than $37.0°\text{C}$. Two possibilities are considered: for this reason, this alternative hypothesis is called a *two-tailed* alternative hypothesis.

27.3 Sampling distribution: One mean

A RQ is answered using data (this is partly what is meant by *evidence-based* research). Fortunately, for the body-temperature study, data are available from a comprehensive American study (Shoemaker 1996).

Summarising the data is important, because the data are the means by which the RQ is answered (Table 27.1).

A graphical summary (Fig. 27.1) shows that the internal body temperature of individuals varies from person to person: this is *natural variation*. A numerical summary (from software) shows that:

- The *sample* mean is $\bar{x} = 36.8051°\text{C}$;
- The *sample* standard deviation is $s = 0.40732°\text{C}$;
- The sample size is $n = 130$.

The sample mean is *less* than the assumed value of $\mu = 37°\text{C}$... The question is *why*: can the difference reasonably be explained by sampling variation, or not?

A 95% CI can also be computed (using software or manually): the 95% CI for μ is from $36.73°$ to $36.88°\text{C}$. This CI is narrow, implying that μ has been estimated with precision, so detecting even small deviations of μ from $37°$ should be possible.

TABLE 27.1: *The body temperature data: The first 5 and the last 5 of the 130 observations*

Gender	Body temp (deg. C)
35.72	37.22
35.94	37.28
36.06	37.28
36.11	37.33
36.17	37.33
36.17	37.39
36.17	37.44
36.22	37.72
36.28	37.78
36.33	38.22

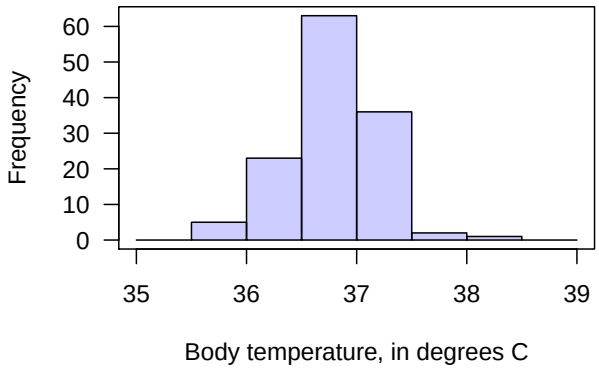


FIGURE 27.1: *The histogram of the body temperature data*

The decision-making process **assumes** that the population mean temperature is $\mu = 37.0°\text{C}$, as stated in the null hypothesis. Because of sampling variation, the value of $\bar{x}$ sometimes would be smaller than $37.0°\text{C}$ and sometimes greater than $37.0°\text{C}$.

How much variation in the value of $\bar{x}$ could be expected, simply due to sampling variation, when $\mu = 37.0°\text{C}$? This variation is described by the *sampling distribution*.

The sampling distribution of $\bar{x}$ was discussed in Sect. 22.2 (and Def. 22.1 specifically). From

this, if μ really was 37.0°C and if certain conditions are true, the possible values of the sample means can be described using:

- An approximate normal distribution;
- With mean 37.0°C (from H_0);
- With standard deviation of s.e.$(\bar{x}) = \dfrac{s}{\sqrt{n}} = \dfrac{0.40732}{\sqrt{130}} = 0.035724$. This is the *standard error* of the sample means.

A picture of this sampling distribution (Fig. 27.2) shows *how the sample mean varies when* $n = 130$*, simply due to sampling variation, when* $\mu = 37°C$. This enables questions to be asked about the likely values of $\bar{x}$ that would be found in the sample, when the population mean is $\mu = 37$°C.

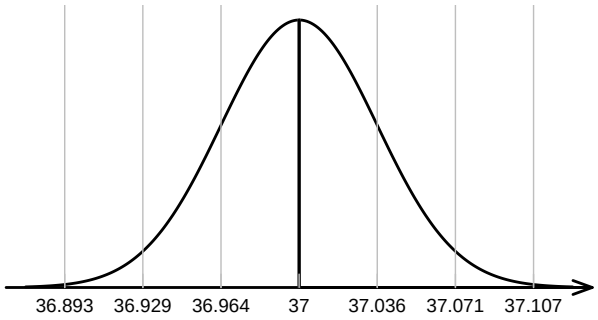


FIGURE 27.2: *The distribution of sample mean body temperatures, if the population mean is* $37°C$ *and* $n = 130$*. The grey vertical lines are 1, 2 and 3 standard deviations from the mean.*

Think 27.1 (Values of $\bar{x}$)**.** *Given the sampling distribution shown in Fig. 27.2, use the 68–95–99.7 rule to determine how often will* $\bar{x}$ *be* ***larger*** *than 37.036 degrees C just because of sampling variation, if* μ *really is* $37°C$*.*

Answer: About 16% of the time.

27.4 The test statistic and t-scores: One mean

The sampling distributions describes what to **expect** from the sample mean, **assuming** $\mu = 37.0$°C. The value of $\bar{x}$ that is **observed**, however, is $\bar{x} = 36.8051°$ How likely is it that such a value could occur by chance?

The value of the observed sample mean can be located the picture of the sampling distribution[1] (Fig. 27.3). The value $\bar{x} = 36.8051$°C is unusually small. About how many standard deviations is $\bar{x}$ away from $\mu = 37$? A lot…

[1] `fig:BodyTempSamplingDist`

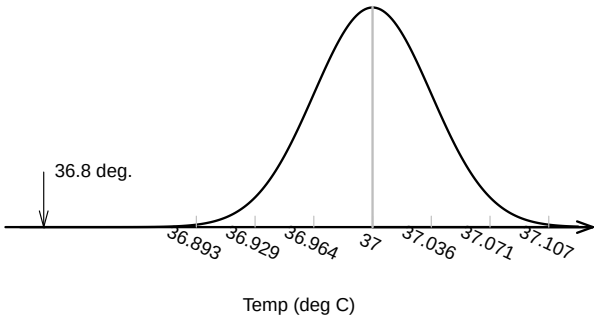


FIGURE 27.3: *The sample mean of $\bar{x} = 36.8041°C$ is very unlikely to have been observed if the poulation mean really was $37°C$, and $n = 130$*

Relatively speaking, the *distance* that the observed sample mean (of $\bar{x} = 36.8051$) is from the mean of the sampling distribution (Fig. 27.3). is found by computing *how many* standard deviations the value of $\bar{x}$ is from the mean of the distribution; that is, computing something like a z-score. (Remember that the standard deviation in Fig. 27.3 is the the *standard error*: the amount of variation in the sample means.)

Since the mean and standard deviation (i.e., the *standard error*) of this normal distribution are known, the number of standard deviations that $\bar{x} = 36.8051$ is from the mean is

$$\frac{36.8051 - 37.0}{0.035724} = -5.453.$$

This value is like a z-score. However, this is actually called a t-score because it has been computed when *the population standard deviation is unknown*, and the best estimate (the *sample* standard deviation) is used when s.e.$(\bar{x})$ was computed.

Both t and z scores measure *the number of standard deviations that an observation is from the mean*: z-scores use σ and t-scores use s. Here, the distribution of the *sample statistic* is relevant, so the appropriate standard deviation is the standard deviation of the sampling distribution: the *standard error*.



Like z-scores, t-scores measure the number of standard deviations that an observation is from the mean. z-scores are calculated using the *population* standard deviation, and t-scores are calculated using the *sample* standard deviation.

In hypothesis testing, t-scores are more commonly used than z-scores, because almost always the population standard deviation is unknown, and the sample standard deviation is used instead.

In this course, it is sufficient to think of z-scores and t-scores as approximately the same. Unless sample sizes are small, this is a reasonable approximation.

So the calculation is:

$$t = \frac{36.8051 - 37.0}{0.035724} = -5.453;$$

the observed sample mean is *more than five standard deviation below the population mean*. This is *highly* unusual based on the 68–95–99.7 rule, as seen in Fig. 27.3.

In general, a t-score in hypothesis testing is

$$t = \frac{\text{sample statistic} - \text{assumed population parameter}}{\text{standard error of the sample statistic}}. \quad (27.1)$$

27.5 P-values: One mean

This is the decision-making progress so far:

1. **Assume** that the population mean is $37.0°$C (this is H_0).
2. Based on this assumption, describe what to **expect** from the sample means (Fig. 27.2).
3. The **observed statistic** is computed, relative to what is expected using a t-score (Fig. 27.3): $t = -5.453$.

The value of the t-score shows that the value of $\bar{x}$ is highly unusual. *How* unusual can be assessed more precisely using a P-*value*, which is used widely in scientific research. The P-value is a way of measuring how unusual an observation is (if H_0 is true).

P values can be *approximated* using the 68–95–99.7 rule and a diagram (Sect. 27.5.1), but more commonly by using software (Sect. 27.5.2).

27.5.1 Approximating P-values using the 68–95–99.7 rule

The P-value is the area *more extreme* than the calculated t-score. For example:

- *If* the calculated t-score was $t = -1$, the *two-tailed* P-value would be the shaded area in Fig. 27.4 (top panel): About 32%, based on the 68–95–99.7 rule. Because the alternative hypothesis is *two-tailed*, *both* sides of the mean are considered: the P-value would be the same if $t = +1$.
- *If* the calculated t-score was $t = -2$, the *two-tailed* P-value would be the shaded area shown in Fig. 27.4 (bottom panel): About 5%, based on the 68–95–99.7 rule. Because the alternative hypothesis is *two-tailed*, *both* sides of the mean are considered: the P-value would be the same if $t = +2$.

Clearly, from what the P-value means, a P-value is always between 0 and 1.

Think 27.2 (P-values)**.** *What do you think the P-value will be for $t = -5.45$ (using Fig. 27.3)?*

Answer: Based on the 68–95–99.7 rule, the P-value will be *extremely* small.

27.5.2 Finding P-values using sofware

Software computes the t-score and a precise P-value (jamovi: Fig. 27.5; SPSS: Fig. 27.6). The output (in jamovi, under the heading `p`; in SPSS, under the heading `Sig. (2-tailed)`) shows that the P-value is indeed very small. Although SPSS reports the P-value as 0.000,

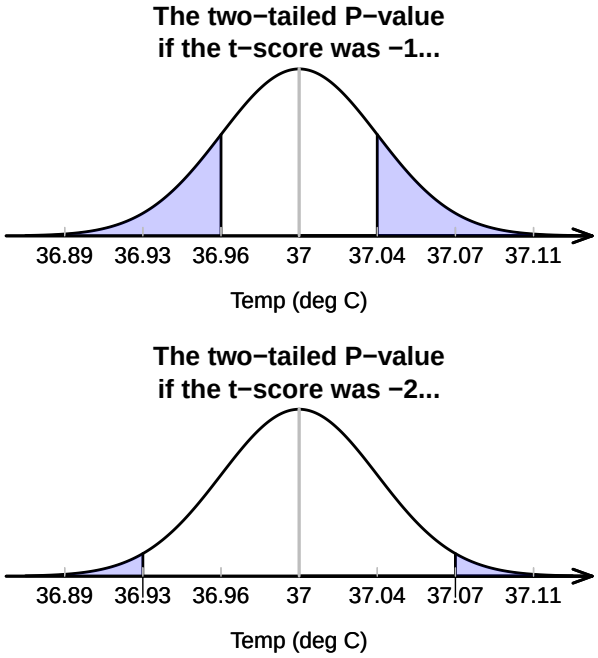


FIGURE 27.4: *Computing P-values for the body temperature data*

P-values can never be *exactly* zero, so we interpret this as 'zero to three decimal places,' or that P is less than 0.001 (written as $P < 0.001$, as jamovi reports).

When software reports a P-value of `0.000`, it really means (and we should write) $P < 0.001$: That is, the P-value is *smaller* than 0.001.

This P-value means that, assumpting $\mu = 37.0°\text{C}$, observing a sample mean as low as $36.8051°\text{C}$ just through sampling variation (from a sample size of $n = 130$) is almost *impossible*. And yet, we did...

Using the decision-making process, this implies that the initial assumption (the null hypothesis) is contradicted by the data: The evidence suggests that the *population* mean body temperature is *not* $37.0°\text{C}$.

One Sample T-Test

One Sample T-Test

		statistic	df	p
BodyTempC	Student's t	−5.45	129	< .001

Note. H_a population mean ≠ 37

Descriptives

	N	Mean	Median	SD	SE
BodyTempC	130	36.8	36.8	0.407	0.0357

FIGURE 27.5: *jamovi output for conducting the t-test for the body temperature data*

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
BodyTempC	130	36.8051	.40732	.03572

One-Sample Test

Test Value = 37

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
BodyTempC	-5.455	129	.000	-.19487	-.2656	-.1242

FIGURE 27.6: *SPSS output for conducting the t-test for the body temperature data*



SPSS always produces **two-tailed** P-values, calls then *Significance values*, and labels them as `Sig.`

jamovi can produce one- or two-tailed P-values.

27.6 Making decisions with P-values

P-values tells us the likelihood of observing the sample statistic (or something more extreme), based on the **assumption** about the population parameter being true. In this context, the P-value tells us the likelihood of observing the value of $\bar{x}$ (or something more extreme), just through sampling variation (chance) if $\mu = 37$. The P-value is a probability, albeit a probability of something quite specific, so it is a value between 0 and 1. Then:

- 'Big' P-values mean that the sample statistic (i.e., $\bar{x}$) could reasonably have occurred through sampling variation, if the assumption about the parameter (stated in H_0) was true (Fig. 27.7, top panel): The data **do not contradict** the assumption (H_0).
- 'Small' P-values mean that the sample statistic (i.e., $\bar{x}$) is unlikely to have occurred through sampling variation, if the assumption about the parameter (stated in H_0) was true: (Fig. 27.7, bottom panel): The data **contradict** the assumption.

What is meant by 'small' and 'big?' It is *arbitrary*: no definitive rules exist. Commonly, a P-value smaller than 1% (that is, smaller than 0.01) is usually considered 'small,' and a P-value larger than 10% (that is, larger than 0.10) is usually considered 'big.' Between the values of 1% and 10% is often a 'grey area.'

Traditionally, a P-value is 'small' if it is less than 5% (less than 0.05), and 'big' if greater than 5% (greater than 0.05). However, again this is *arbitrary*, and binary decision making (*either* big *or* small) is unreasonable. More reasonably, P-values should be interpreted as providing varying strength of evidence in support of the alternative hypothesis H_1 (Table 28.1). These

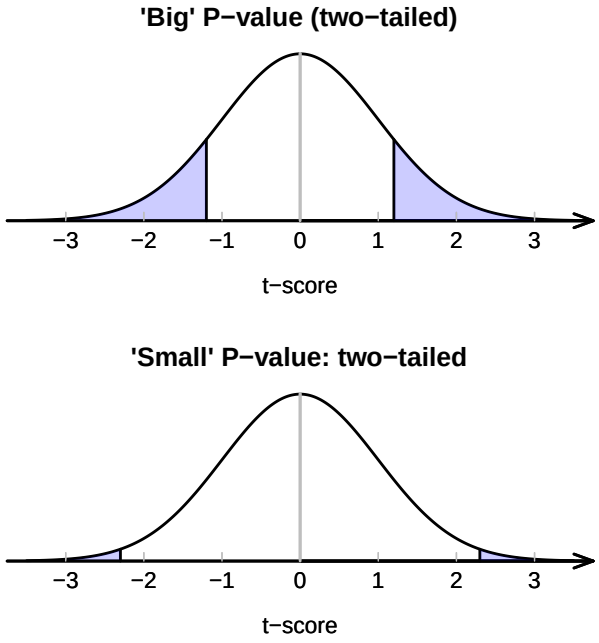


FIGURE 27.7: *A picture of large (top) and small (bottom) P-value situations*

are not definitive, but are only guidelines. Of course, conclusions should be written in the context of the problem.

⚠ For *one-tailed tests*, the P-value is *half* the value of the two-tailed P-value.

SPSS always produces two-tailed P-values, usually calls them 'Significance values,' and labels them as `Sig.`, and sometimes explicitly notes that they are two-tailed.

For the body-temperature data then, where $P < 0.001$, the P-value is *very* small, so there is *very strong evidence* that the population mean body temperature is not 37.0°C.

27.7 Communicating results: One mean

In general, to communicate the results of any hypothesis test, report:

- An *answer to the RQ*;
- The *evidence* used to reach that conclusion (such as the t-score and P-value—including if it is a one- or two-tailed P-value); and
- Some *sample summary information*, including a CI, summarising the data used to make the decision.

So write:

> The sample provides very strong evidence ($t = -5.45$; two-tailed $P < 0.001$) that the population mean body temperature is not 37.0°C ($\bar{x} = 36.81$; $n = 130$; 95% CI from 36.73°C to 36.88°C).

The components are:

- The *answer to the RQ*: 'The sample provides very strong evidence... that the population mean body temperature is not 37.0°C.'
- The *evidence* used to reach the conclusion: '$t = -5.45$; two-tailed $P < 0.001$.'
- Some *sample summary information* (including a CI): '$\bar{x} = 36.81$; $n = 130$; 95% CI from 36.73°C to 36.88°C.'

Notice how the conclusion is worded: There is *evidence* to support the alternative hypothesis. In fact, the alternative hypothesis *may* or *may not* be true... but the evidence (data) available supports the alternative hypothesis.

27.8 Hypothesis testing for one mean: A summary

Let's recap the decision-making process seen earlier, in this context about body temperatures:

- **Step 1: Assumption**: Write the *null hypothesis* about the parameter (based on the RQ): H_0: $\mu = 37.0$°C. In addition, write the *alternative hypothesis* H_1: $\mu \neq 37.0$°C. (This alternative hypothesis is two-tailed.)
- **Step 2: Expectation**: The *sampling distribution* describes what to expect from the sample statistic *if* the null hypothesis is true: under certain circumstances, the sample means will vary with an approximate normal distribution around a mean of $\mu = 37.0$°C with a standard deviation of s.e.$(\bar{x}) = 0.03572$ (Fig. 27.3).
- **Step 3: Observation**: Compute the t-score: $t = -5.45$. The t-score can be computed by software, or using the general equation (27.1).
- **Step 4: Consistency?**: Determine if the data are *consistent* with the assumption, by computing the P-value. Here, the P-value is much smaller than 0.001. The P-value can be computed by software, or approximated using the 68–95–99.7 rule.

The **conclusion** is that there is very strong evidence that μ is not 37.0°C:

Example 27.1 (Mean driving speeds)**.** A study of driving speeds in Malaysia (Azwari and Hamsa 2021) recorded the speeds of vehicles on various roads.

One RQ of interest was whether the mean speed of cars on one road was the posted speed limit of 90 km.h^{-1}, or whether it was higher. The *parameter* of interest is μ, the mean speed in the *population*.

The hypotheses are:

- H_0: $\mu = 90$; and
- H_1: $\mu > 90$ (since the researchers were interested in whether the mean speed was *higher* than the posted speed limit).

The researchers recorded the speed of $n = 400$ vehicles on this road, and found $\bar{x} = 96.56$, but this value is likely to vary from sample to sample. The sample standard deviation was $s = 13.874$, so that

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{13.874}{\sqrt{400}} = 0.6937.$$

Hence, the test statistic is

$$t = \frac{\bar{x} - \mu}{\text{s.e.}(\bar{x})} = \frac{96.56 - 90}{0.6937} = 9.46,$$

where (as usual) the value of μ is taken from the null hypothesis (which we always assume to be true).

This is a *huge* value, suggesting that the (one-tailed) P-value is very small.

We write (remembering the alternative hypothesis is one-tailed):

> There is very strong evidence ($t = 9.46$; one-tailed $P < 0.001$) that the mean speed of vehicles on this road (sample mean: 96.56; standard deviation: 13.874) is greater than 90 km.h^{-1}.

Of course, this statement refers to the *mean* speed; there may be individual vehicles travelling below the speed limit.

27.9 Statistical validity conditions: One mean

As with any inference procedure, the underlying mathematics requires certain conditions to be met so that the results are statistically valid. For a hypothesis test for one mean, these conditions are the same as for the CI for one mean (Sect. 22.4).

The test will be statistically valid if *one* of these is true:

1. The sample size is at least 25, *or*
2. The sample size is smaller than 25 *and* the *population* data has an approximate normal distribution.

The sample size of 25 is a rough figure here, and some books give other values (such as 30).

This condition ensures that the *distribution of the sample means has an approximate normal distribution* so that the 68–95–99.7 rule can be used. Provided the sample size is larger than about 25, this will be approximately true *even if* the distribution of the individuals in the population does not have a normal distribution. That is, when $n > 25$ the sample means generally have an approximate normal distribution, even if the data themselves don't have a normal distribution.

In addition to the statistical validity condition, the test will be

- **internally valid** if the study was well designed; and
- **externally valid** if the sample is a simple random sample and is internally valid.

Example 27.2 (Statistical validity)**.** The hypothesis test regarding body temperature is statistically valid since the sample size is large ($n = 130$).

Since the sample size is large, we *do not* require the data to come from a population with a normal distribution.

Example 27.3 (Driving speeds)**.** In Example 27.1 about mean driving speeds, the sample size was 400, much larger than 25.

The test will be statistically valid.

27.10 Example: Recovery times

Seventeen patients were treated for medial collateral ligament (MCL) and anterior cruciate ligament (ACL) tears using a new treatment method (Altman 1991; Nakamura et al. 2000). The current existing treatment has an average recovery time of 15 days. The RQ is:

For patients with this type of injury, does the new treatment method lead to *shorter* mean recovery times?

The parameter is μ, the population mean recovery time.

The hypotheses (**Step 1: Assumption**) about the parameter are, from the RQ:

- H_0: $\mu = 15$ (the population mean is 15, but $\bar{x}$ is not 15 due to sampling variation): This is the initial **assumption**.
- H_1: $\mu < 15$ (μ is not 15; it really does produce *shorter* recovery times, on average).

This test is *one-tailed*: the RQ only asks if the new method produces *shorter* recovery times.

The evidence (Table 27.2) can be summarised numerically, using software or (since the data set is small) a calculator. Either way, $\bar{x} = 13.29$ and $s = 8.887$.

TABLE 27.2: *The recovery times (in days) for a new treatment*

14	18	12	10	8	28	24	3	9
9	26	0	4	21	24	2	14	

If the null hypothesis is true (and $\mu = 15$), the values of the sample mean that are likely to occur through sampling variation can be described (**Step 2: Expectation**). The sample means are likely to vary with an approximate normal distribution (under certain assumptions), with mean $\mu = 15$ and a standard deviation of

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{8.887}{\sqrt{17}} = 2.155.$$

This describes what values of $\bar{x}$ we should *expect* in the sample if the mean recovery time μ really was 15 days (Fig. 27.10).

The sample mean is $\bar{x} = 13.29$, so the t-score to determine where the sample mean is located (**Step 3: Observation**), relative to what is expected, is

$$t = \frac{\bar{x} - \mu}{\text{s.e.}(\bar{x})} = \frac{13.29 - 15}{2.155} = -0.79.$$

Software could also be used (jamovi: Fig. 27.8; SPSS: Fig. 27.9), but in either case, $t = -0.79$.

A z-score of -0.79 is not unusual, and (since t-scores are like z-scores) this t-score is not unusual either (**Step 4: Consistency**). The P-value in the jamovi output[2] or SPSS output[3] confirms this: the *two*-tailed P-value is 0.440, so the *one*-tailed P-value is $0.440 \div 2 = 0.220$.

One Sample T-Test

One Sample T-Test

		statistic	df	p
RecTime	Student's t	−0.791	16.0	0.440

Note. H_a population mean ≠ 15

Descriptives

	N	Mean	Median	SD	SE
RecTime	17	13.3	12.0	8.89	2.16

FIGURE 27.8: *jamovi output for the t-test for the recovery-times data*

One-Sample Test

	Test Value = 15					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
RecTime	-.791	16	.440	-1.706	-6.27	2.86

FIGURE 27.9: *SPSS output for the t-test for the recovery-times data*

⚠ Recall: For *one-tailed tests*, the P-value is *half* the value of the two-tailed P-value.

This 'large' P-value suggests that a sample mean of 13.29 could reasonably have been observed just through sampling variation: there is no evidence to support the alternative hypothesis H_1. If μ really was 15, then about 22% of the time $\bar{x}$ would be less than 13.29 just through sampling variation alone.

To summarise:

[2] `fig:RecoveryTimestestjamovi`
[3] `fig:RecoveryTimestestSPSS`

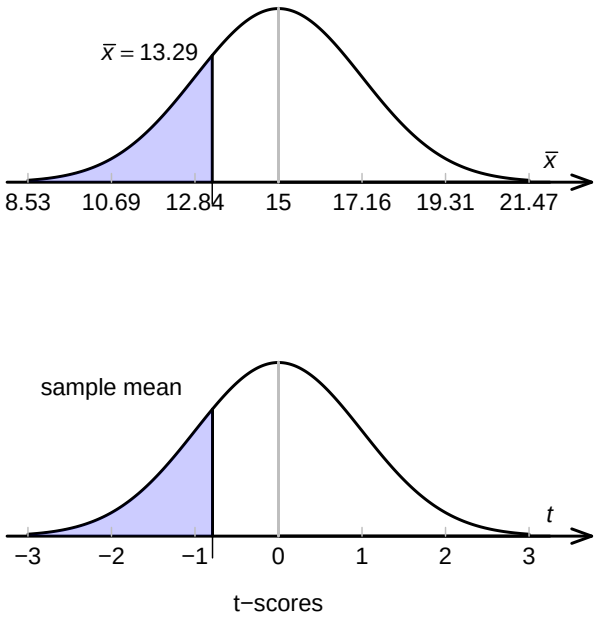


FIGURE 27.10: *The sampling distribution for the recovery-times data*

- **Step 1 (Assumption)**:
  - H_0: $\mu = 15$, the initial assumption;
  - H_1: $\mu < 15$ (note: *one*-tailed)
- **Step 2 (Expectation)**: The sample means will vary, and this sampling variation is described by an approximate normal distribution with mean 15 and standard deviation s.e.$(\bar{x}) = 2.155$.
- **Step 3 (Observation)**: $t = -0.791$.
- **Step 4 (Consistency?)**: The *one-tailed* P-value is 0.220: The data are consistent with H_0, so there is no evidence to support the alternative hypothesis.

To write a conclusion, include an *answer* to the question, *evidence* leading to the conclusion, and some *sample summary information*:

> No evidence exists in the sample (one sample $t = -0.79$; one-tailed $P = 0.220$) that the population mean recovery time is less than 15 days (mean 13.29 days; $n = 17$; 95% CI from 8.73 to 17.86 days) using the new treatment method.

(The CI is found using the ideas in Sect. 22.3, or manually.) Notice the wording: The new method *may* be better, but no evidence exists of this in the sample. The onus is on the new method to demonstrate that it is better than the current method.

The sample size is small ($n = 17$), so the test may not be statistically valid (but the P-value is so large that it probably won't affect the conclusions).

27.11 Summary

To test a hypothesis about a population mean μ, initially **assume** the value of μ in the null hypothesis to be true. Then, describe the **sampling distribution**, which describes what to **expect** from the sample statistic based on this assumption: under certain statistical validity conditions, the sample mean varies with an approximate normal distribution centered around the hypothesised value of μ, with a standard deviation of

$$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}}.$$

The **observations** are then summarised, and *test statistic* computed:

$$t = \frac{\bar{x} - \mu}{\text{s.e.}(\bar{x})},$$

where μ is the hypothesised value given in the null hypothesis. The t-value is like a z-score, and so an approximate P**-value** can be estimated using the 68–95–99.7 rule, or found using software.

27.12 Quick review questions

A study (Imtiaz et al. 2017) compared the nutritional intake of $n = 50$ anaemic infants in Lahore (Pakistan) with the amounts recommended.

The mean daily protein intake in the sample was 14g, with a standard deviation of 3g. The recommended protein intake was 13g.

The researchers wanted to see if the mean intake met the recommendation, or not.

1. The standard error of the mean (to four decimal places) is
2. The null hypothesis is:
3. The test statistic (to two decimal places) is
4. The two-tailed P-value is
5. True or false? We accept the *null* hypothesis.
6. There is

to support the *alternative* hypothesis (that the mean daily protein intake is *not* 13g).

27.13 Exercises

Selected answers are available in Sect. D.25.

Exercise 27.1. The recommended daily energy intake for women is 7725kJ (for a particular cohort, in a particular country; Altman (1991)). The daily energy intake for 11 women was measured to see if this is being adhered to. The RQ is

> For this group of women, is the population mean daily energy intake 7725kJ?

The data collected are shown in Table 27.3.

1. Write the hypotheses for answering this RQ.
2. Use the jamovi output (Fig. 27.11), write down the value of $\bar{x}$ and s.e.$(\bar{x})$.
3. Using this output, write down the t-value and the P-value.
4. Write a suitable conclusion.
5. Is the test statistically valid?
6. Sketch the sampling distribution of $\bar{x}$.

TABLE 27.3: *Energy consumptions (in kJ) for women*

5260	5640	6390	6515	7515	8770
5470	6180	7515	6805	8230	

One Sample T-Test

One Sample T-Test

		statistic	df	p
Energy	Student's t	−2.82	10.0	0.018

Note. H_a population mean ≠ 7725

Descriptives

	N	Mean	Median	SD	SE
Energy	11	6754	6515	1142	344

FIGURE 27.11: *jamovi output for the energy-intake data*

Exercise 27.2. Most dental associations[4] recommend brushing teeth for two minutes. A study (Macgregor and Rugg-Gunn 1979) of the brushing time for 85 uninstructed school children from England (11 to 13 years old) found the mean brushing time was 60.3 seconds, with a standard deviation of 23.8 seconds.

[4]Such as the *American Dental Association* and the *Australian Dental Association*.

1. Is there evidence that the mean brushing time for schoolchildren from England is two minutes (as recommended)?
2. Sketch the sampling distribution of the sample mean.

Exercise 27.3. A study (Greenlee et al. 2018) of human-automation interaction with automated vehicles aimed to

> ... determine whether monitoring the roadway for hazards during automated driving results in a vigilance decrement.
> — Greenlee et al. (2018), p. 465

(A 'decrement' is a *reduction*.) That is, they were interested in whether the average mental demand of 'drivers' of automated vehicles was higher than the average mental demand for ordinary tasks.

In the study, the $n = 22$ participants 'drove,' in a simulator, an automated vehicle for 40 minutes. While driving, the drivers monitored the road for hazards. The researchers assessed the 'mental demand' placed on these drivers, where scores of 50 over 'typically indicate substantial levels of workload' (p. 471). For the sample, the mean score was 84.00 with a standard deviation of 22.05.

Is there evidence of a 'substantial workload' associated with monitoring roadways while 'driving' automated vehicles?

Exercise 27.4. A study explored the quality of life of patients receiving cavopulmonary shunts (Steele et al. 2016). 'Quality of life' was assessed using a 36-question health survey, where the scale is standardised so that the mean of the general population is 50.

For the 14 patients in the study, the sample mean for the 'Physical component' of the survey was 47.2 (with a standard deviation of 8.2). The sample mean for the 'Mental component' of the survey was 52.7 (with a standard deviation of 5.6).

Is there evidence that the patients are different, on average, to the general population on the basis of the results?

Exercise 27.5. A *Cherry Ripe* is a popular chocolate bar in Australia. In 2017 and 2018, I 'sampled' some *Cherry Ripe* Fun Size bars. The packet claimed that the Fun Size bars weigh 12 g (on average). Use the SPSS summary of the data (Fig. 27.12) to perform a hypothesis test to determine if the mean weight really is 12 g or not.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Bar weight (in grams)	43	12.85	15.98	14.9577	.60652
Valid N (listwise)	43				

FIGURE 27.12: *jamovi output for the Cherry Ripes data*

Exercise 27.6. A study of paramedics (Williams and Boyle 2007) asked participants ($n = 199$) to estimate the amount of blood on four different surfaces. When the actual amount of blood

spilt on concrete was 1000 ml, the mean guess was 846.4 ml (with a standard deviation of 651.1 ml).

Is there evidence that the mean guess really is 1000 ml (the true amount)? Is this test likely to be valid?

Exercise 27.7. A quality-control study (Feng et al. 2017) assessed the accuracy of two instruments from a clinical laboratory, by comparing the reported luteotropichormone (LH) concentrations to known pre-determined values (data below). q Perform a series of tests to determine how well the two instruments perform, for both high- and mid-level LH concentrations (using Table 27.5).

TABLE 27.4: *The quality-control data: LH levels (in mIU/mL) for two instruments (only the first ten observations shown)*

High level (Inst. 1)	Mid level (Inst. 1)	High level (Inst. 2)	Mid level (Inst. 2)
61.63	18.36	62.64	19.12
63.11	18.77	64.36	19.07
66.88	18.98	66.06	19.58
62.56	17.97	65.39	19.35
66.12	19.69	66.85	19.65
65.34	19.63	65.56	19.13
64.83	19.50	66.60	19.55
64.22	19.39	66.90	19.85
65.54	19.87	65.50	19.45
65.33	19.66	65.92	19.87

TABLE 27.5: *Summary of the quality-control data for LH levels (in mIU/mL) for two instruments*

	High level (Inst. 1)	**Mid level (Inst. 1)**	**High level (Inst. 2)**	**Mid level (Inst. 2)**
Mean of data	64.31	19.240	64.970	19.400
Std. dev. of data	1.70	0.588	1.029	0.413
Pre-determined target	64.22	19.010	65.050	19.450

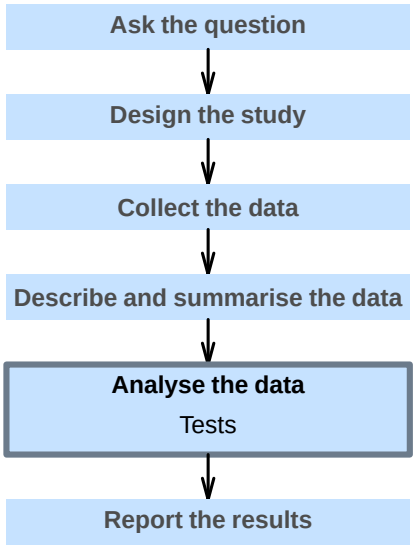
28

More about hypothesis testing

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, understand the tools of inference, and to form *confidence intervals*.

In this chapter,
you will learn about *hypothesis tests*. You will learn to:

- communicate the results of hypothesis tests.
- interpret P-values.



28.1 Introduction

In Chap. 27, hypothesis tests for one mean were studied. In later chapters, hypothesis tests are discussed in other contexts, too.

The general approach to hypothesis testing is the same for any hypothesis test, and so some general ideas are discussed in this chapter. All hypothesis tests answer questions about unknown *population* quantities (such as the population mean μ), based on *sample* statistics (such as the sample mean $\bar{x}$).

The sections that follow discuss:

- The **assumptions** and forming hypotheses (Sect. 28.2).

- The sampling distribution, and the **expectations** (Sect. 28.3).
- The **observations** and the test statistic (Sect. 28.4).
- Weighing the evidence for **consistency**: P-values (Sect. 28.6).
- Wording **conclusions** (Sect. 28.7).

When raw data are provided, begin by producing graphical and numerical summaries of the data. The statistical validity conditions, which vary for different hypothesis tests, should always be checked to see if the test is statistically valid.

28.2 About hypotheses and assumptions

Two hypotheses are made about the population parameter:

- The null hypothesis H_0; and
- The alternative hypothesis H_1.

28.2.1 Null hypotheses

Hypotheses *always concern a population parameter*. Hypothesising, for example, that the *sample* mean body temperature is equal to $37.0°$C is pointless, because it clearly isn't: the sample mean is $36.8051°$C. Besides, the RQ is about the unknown *population*: the **P** in **P**OCI stands for **P**opulation.

The **null hypothesis** H_0 offers one possible reason why the value of the sample statistic (such as the sample mean) is not the same as the value of the proposed population parameter (such as the population mean): *sampling variation*. Every sample is different, and so the *sample statistic* will vary from sample to sample; it may not be equal to the *population parameter*, just because of the sample used by chance. Null hypotheses always have an 'equals' in them (for example, the population mean *equals* 100, is *less than or equal to* 100, or is *more than or equal to* 100), because (as part of the decision making process), something specific must be assumed for the population parameter.

The parameter can take many different forms, depending on the context. The null hypothesis about the parameter is the *default* value of that parameter; for example,

- there is *no difference* between the parameter value in two (or more) groups;
- there is *no change* in the parameter value; or
- there is *no relationship* as measured by a parameter value.

⚠ Hypothesis testing starts by assuming that the null hypothesis is true.

The onus is on the data to provide evidence to refute this default position.

The null hypothesis is always about a population parameter, and always has the form 'no difference, no change, no relationship.'

Definition 28.1 (Null hypothesis)**.** The **null hypothesis** proposes that *sampling variation* explains the difference between the proposed value of the parameter, and the observed value of the statistic.

28.2.2 Alternative hypotheses

The other hypothesis is called the **alternative hypothesis** H_1. The alternative hypothesis offers another possible reason why the value of the sample statistic (such as the sample mean) is not the same as the value of the proposed population parameter (such as the population mean). The alternative hypothesis proposes that the value of the population parameter really is not the value claimed in the null hypothesis.

Definition 28.2 (Alternative hypothesis)**.** The **alternative hypothesis** proposes that the difference between the proposed value of the parameter and the observed value of the statistic cannot be explained by *sampling variation*: the proposed value of the parameter is probably not true.“‘

Alternative hypotheses can be *one-tailed* or *two-tailed*. A *two*-tailed alternative hypothesis means, for example, that the population mean could be either smaller *or* larger than what is claimed. A *one*-tailed alternative hypothesis admits only one of those two possibilities. Most (but not all) hypothesis tests are two-tailed.

The decision about whether the alternative hypothesis is one- or two-tailed is made by reading the RQ (*not* by looking at the data). Indeed, *the RQ and hypotheses should (in principle) be formed before the data are obtained*, or at least before looking at the data if the data are already collected.

The ideas are the same whether the alternative hypothesis is one- or two-tailed: based on the data and the sample statistic, a decision is to be made about whether the alternative hypotheses is supported by the data.

Example 28.1 (Alternative hypotheses)**.** For the body-temperature study, the alternative hypothesis is *two-tailed*: The RQ asks if the population mean is $37.0°$C or *not*. That is, two possibilities are considered: that μ could be either larger *or* smaller than $37.0°$C.

A *one-tailed alternative hypothesis* would be appropriate if the RQ was: ‘Is the *population* mean internal body temperature *greater* than $37.0°$C?’ or Is the *population* mean internal body temperature *smaller* than $37.0°$C?.

⚠ Important points about forming hypotheses:

- Hypotheses always concern a *population* parameter.
- Null hypotheses always contain an ‘equals.’
- Alternative hypothesis are one-tailed or two-tailed, depending on the RQ.
- Hypotheses emerge from the RQ (not the data): The RQ and the hypotheses could be written down *before* collecting the data.

28.3 About sampling distributions and expectations

The *sampling distribution* describes, approximately, how the sample statistic (such as $\bar{x}$) is likely to vary from sample to sample over many repeated samples, when H_0 is true: it describes the *sampling variation*. Under certain circumstances, sampling distributions often have an approximate normal distribution, which is the basis for computing P-values (or approximating P-values using the 68–95–99.7 rule).

When the sampling distribution is described by a normal distribution, the *mean* of the normal distribution is the parameter value given in the *assumption* (H_0), and the *standard deviation* of the normal distribution is called the *standard error*.

In some cases, the sample statistic may not have a normal distribution, but a quantity easily derived from the sample statistic does have a normal distribution (for example, the odds ratio[1]).

28.4 About observations and the test statistic

The sampling distribution describes what values the sample statistic can reasonably be expected to have, over many repeated samples. Since the sampling distribution of the statistic has an approximate normal distribution under certain conditions, the observed value of the sample statistic can be expressed as a something like a z-score (called a t-score when the *population* standard deviation is unknown). In general, t-scores always have the same form:

$$\text{statistic} = \frac{\text{sample statistic} - \text{assumed population parameter}}{\text{measure of variation of the sample statistic}}.$$

The t-score here is the *test statistic*, since it is based on sample data ('a statistic') and used in a hypothesis test.



A t-score is similar to a z-score; both the z- and t-scores have the same form:

$$\frac{\text{sample value} - \text{population value}}{\text{measure of variation of the sample value}}.$$

Then:

- If the 'sample value' refers to an *individual* observation x, the measure of variation is the standard deviation, because the standard deviation measures the variation in the individual observations.
- If the 'sample value' is a *sample statistic*, the measure of variation is a *standard error*, because the standard deviation measures the variation in the sample statistic.

In both cases, if the measure of variation uses a known *population* value, a z-score is found; if the measure of variation uses a *sample* value, a t-score is found.

[1] In this case, the *logarithm* of the odds ratio has an approximate normal distribution.

28.5 About finding P-values

As demonstrated in Sect. 27.5.1, often P-values can be *approximated* by using the the 68–95–99.7 rule and using a diagram of a normal distribution. The P-value is the area *more extreme* than the calculated t-score; the 68–95–99.7 rule can be used to approximate this tail area.

For **two-tailed** tests, the P-value is the *combined* area in the left and right tails. For **one-tailed** tests, the P-value is the area in just the left or right tail.



When software reports *two-tailed P-values*, a one-tailed P is found by *halving the two-tailed P-value*.

More accurate estimates of the P-value can be found using z-tables, though we do not demonstrate this in this book. Even more precise estimates of P-values can be found using specially-prepared t-tables. Again, we do not do so in this book.

For more precise P-values, we will take the P-values from software output.



When using software to obtain P-values, be sure to check if the software reports one- or two-tailed P-values.

For example, some software (such as SPSS) always reports two-tailed P-values.

28.6 About interpreting P-values

A P-value is the likelihood of observing the sample statistic (or something even more extreme) over repeated sampling, under the assumption that the null hypothesis about the population parameter is true.

P-values can be computed because the sampling distribution often has an approximate normal distribution.

TABLE 28.1: *A guideline for interpreting P-values. P-values should be interpreted in context.*

If the P-value is...	Write the conclusion as...
Larger than 0.10	*Insufficient* evidence to support H_1
Between 0.05 and 0.10	*Slight* evidence to support H_1
Between 0.01 and 0.05	*Moderate* evidence to support H_1
Between 0.001 and 0.01	*Strong* evidence to support H_1
Smaller than 0.001	*Very strong* evidence to support H_1

⚠ Conclusion are **always** about the population values.

No-one needs P-values to see if the sample values are the same: We can just look at them, and see.

P-values are needed to determine what we learn about the unknown **population** values, based on what we see in the **sample** values.

Commonly, a P-value smaller than 5% is considered 'small,' but this is *arbitrary*. More reasonably, P-values should be interpreted as giving varying degrees of evidence in support of the alternative hypothesis (Table 28.1), but these are only guidelines. Conclusions should be written in the context of the problem. Sometimes, authors will write that the results are 'statistically significant' when $P < 0.05$.

Definition 28.3 (P-value)**.** A P-value is the likelihood of observing the sample statistic (or something more extreme) over repeated sampling, under the assumption that the null hypothesis about the population parameter is true.

⚠ P-values are never exactly zero. When SPSS reports that '$P = 0.000$,' it means that the P-value is less than 0.001, which we write as '$P < 0.001$.'

jamovi usually reports very small P-values as '$P < 0.001$.'

P-values are commonly used in research, but they need to be used and interpreted correctly (Greenland et al. 2016). Specifically:

- A P-value **is not** the probability that the null hypothesis is true.
- A P-value **does not** *prove* anything.
- A big P-value **does not** mean that the null hypothesis H_0 is true, or that H_1 is false.
- A small P-value **does not** mean that the null hypothesis H_0 is false, or that H_0 is true.
- A small P-value **does not** indicate that the results are practically important (Sect. 28.8).
- A small P-value does not mean a large difference between the statistic and parameter; it means that the difference could not reasonably be attributed to *sampling variation* (chance).

⚠ Sometimes, the results from hypothesis tests are called "significant" or "statistically significant."

This means that the P-value is small (traditionally, but arbitrarily, $P < 0.05$), and hence the evidence supports the alternative hypothesis.

To avoid confusion, the word "significant" should be avoided in writing about research unless "statistical significance" is what is actually what is meant. In other situations, consider using words like "substantial."

28.7 About writing conclusions

When reporting a conclusion, three things should be included:

1. The *answer to the RQ*;
2. The *evidence* used to reach that conclusion (such as the t-score and P-value, clarifying if the P-value is *one-tailed* or *two-tailed*); and
3. Some *sample summary statistics* (such as sample means and sample sizes), including a CI (which indicates the precision with which the statistic has been estimated).

Conclusions can never be made with *certainty* from one sample. Partly this is because a *sample* has been studied, while the RQ asks about the whole *population*: The entire population wasn't studied.

For this reason, care must be taken when answering the RQ. A hypothesis test *never proves*. anything: It might conclude that evidence exists (perhaps weak evidence; perhaps strong evidence) to support the alternative hypothesis. Of course, there may be no evidence to support the alternative hypothesis either.

Since the value of the parameter in the null hypothesis is assumed true, the onus is on the data to provide evidence to refute this default position. For this reason, *conclusions are worded in terms of the level of support for the alternative hypothesis*.

⚠ Conclusions are always made in terms of how much evidence supports the *alternative* hypothesis. Hypothesis tests assume the null hypothesis is true, so the onus is on the data to provide evidence in support of the alternative hypothesis.

Think 28.1 (Conclusions)**.** *What is wrong with the following conclusion?*

The evidence proves that the mean internal body temperature has changed.

28.8 About practical importance and statistical significance

Hypothesis tests assess *statistical significance*, which answers the question: 'Is there evidence of a difference between the value of the statistic and the value of the assumed parameter?' Even very small differences between the sample statistic and the population parameter can be *statistically* different if the sample size is large enough.

In contrast, *practical importance* asks the question:

Is the difference between the value of the statistic and the value of the assumed parameter of any *practical* importance?

'Practical importance' and 'statistical significance' are two separate (but both important) issues. Whether a results is of practical importance depends upon the context: what the data are being used for, by whom, and for what purpose.

Example 28.2 (Practical importance)**.** In the body-temperature study, very strong evidence exists that the mean body temperature had changed ('statistical significance').

But the change was so small, that for most purposes it has no practical importance. (There may be other (e.g., medical) situations where it *does* have practical importance however.)

Practical importance depends on the context in which the results will be used.

Example 28.3 (Practical importance)**.** A study of some herbal medicines (Maunder et al. 2020) for weight loss found:

> *Phaseolus vulgaris* resulted in a statistically significant weight loss compared to placebo, although this was not considered clinically significant.

In other words, although the difference in weight loss between placebo and *Phaseolus vulgaris* was unikely to be explained by chance ($P < 0.001$, which is 'statistical significant'), the difference was so small in size (a mean weight loss of just 1.61 kg) that it was unlikely to be of any use in practice ('practical importance').

In this context, a weight loss of at least 2.5 kg was considered to be of practical importance.

28.9 Validity and hypothesis testing

When performing hypothesis tests, certain *statistical validity conditions* must be true. These conditions ensure that the sampling distribution is sufficiently close to a normal distribution for the 68–95–99.7 rule rule to apply and hence for P-values to be computed[2].

If these conditions are *not* met, the sampling distribution may not be normally distributed, so the P-values (and hence conclusions) maybe inappropriate.

In addition to the statistical validity condition, the *internal validity* and *external validity* of the study should be discussed also (Fig. 28.1). These are usually (but not always) the same as for CIs (Sect. 21.3).

Regarding *external validity*, all the computations in this book assume a *simple random sample*. If the sample is from a *random* sampling method, but not from a *simple random sample*, then methods exist for conducting hypothesis tests that are externally valid, but are more complicated than those described in this book.

[2]Not all sample statistics have normal distributions, but all the sample statistics in this book are either normally distributed or are closely related to normal distributions.

If the sample is a non-random sample, then the hypothesis test may be reasonable for the quite specific population that *is* represented by the sample; however, the sample probably does not represent the more general population that is probably intended.

Externally validity requires that a study is also internally valid. *Internal validity* can only be discussed if details are known about the study design.

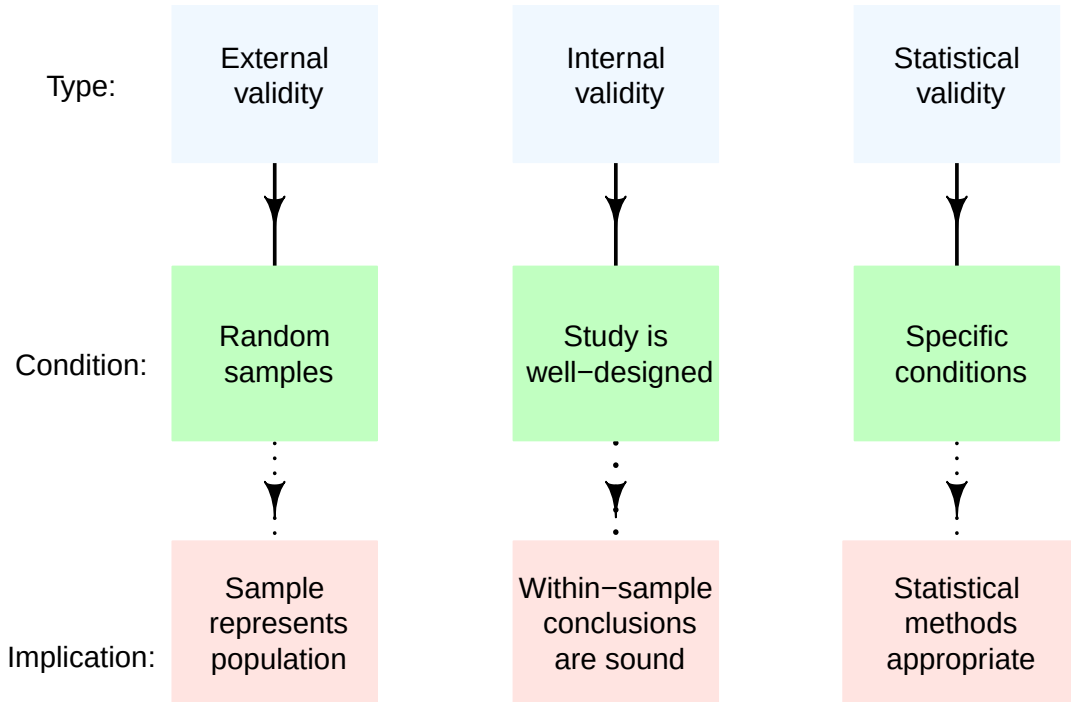


FIGURE 28.1: *Three types of validities for studies.*

In addition, hypothesis tests also require that the sample size is less than 10% of the population size; however this is almost always the case.

28.10 Summary

Hypothesis testing formalises the steps of the decision-making process. Starting with an **assumption** about a population parameter of interest, a description of what values the sample statistic might take (based on this assumption) is produced: this describes what values the statistic is **expected** to take, just through sampling variation. This sampling distribution is often a normal distribution, or related to a normal distribution.

The sample statistic (the *estimate*) is then **observed**, and a *test statistic*, which often is a t-score, is computed to describe this sample statistic. Using a P-value, a decision is made about whether the sample evidence supports or contradicts the initial assumption, and hence a **conclusion** is made. Since t-scores are like z-scores, P-values can often be approximated using the 68–95–99.7 rule.

28.11 Quick review questions

1. True or false? When a P-value is very small, a very large difference exists between the statistic and parameter.
2. True or false? The alternative hypothesis is one-tailed if the sample statistic is larger than the hypothesised population mean.
3. What is wrong (if anything) with this null hypothesis: $H_0 = 37$?
4. True or false: When the sampling distribution is a normal distribution, the standard deviation of this normal distribution is called the *standard error*.
5. True or false? Both z-scores and t-scores *can* be test statistics.
6. True or false? P-values can never be exactly zero.
7. True or false? A P-value is the probability that the null hypothesis is true.

28.12 Exercises

Selected answers are available in Sect. D.26.

Exercise 28.1. Use the 68–95–99.7 rule to approximate the *two*-tailed P-value if:

1. the t-score is 3.4.
2. the t-score is -2.9.
3. the t-score is 1.2.
4. the t-score is -0.95.
5. the t-score is -0.2.
6. the t-score is 6.7.

Exercise 28.2. Consider the t-scores in Exercise 28.1. Use the 68–95–99.7 rule to approximate the *one*-tailed P-values in each case.

Exercise 28.3. Suppose a hypothesis test results in a P-value of 0.0501. What would we conclude? What about if the P-value was 0.0499?

Exercise 28.4. Consider again the study to determine the mean body temperature, where $\bar{x} = 36.8051°$C. What, if anything, is wrong with these hypotheses? Explain.

1. H_0: $\bar{x} = 36$ and H_1: $\bar{x} \neq 36$.
2. H_0: $\bar{x} = 36.8051$ and H_1: $\bar{x} > 36.8051$.
3. H_0: $\mu = 36.8051$ and H_1: $\mu \neq 36.8051$.
4. H_0: $\mu = 36$ and H_1: $\mu = 36.8051$.
5. H_0: $\mu > 36$ and H_1: $\bar{x} > 36$.
6. H_0: $\mu = 36$ and H_1: $\mu > 36$.

Exercise 28.5. The recommended daily energy intake for women is 7725kJ (for a particular cohort, in a particular country; Altman (1991)). The daily energy intake for 11 women was measured to see if this is being adhered to. The RQ was

> Is the population mean daily energy intake 7725kJ?

The test produced $P = 0.018$. What, if anything, is wrong with these conclusions after completing the hypothesis test?

1. There is moderate evidence ($P = 0.018$) that the energy intake is not meeting the recommended daily energy intake.
2. There is moderate evidence ($P = 0.018$) that the sample mean energy intake is not meeting the recommended daily energy intake.
3. There is moderate evidence ($P = 0.018$) that the population energy intake is not meeting the recommended daily energy intake.

Exercise 28.6. A study compared ALDI batteries to another brand of battery. In one test comparing the length of time it takes for 1.5 volt AA batteries to reach 1.1 volts, the ALDI brand battery took 5.73 hours, and the other brand (Energizer) took 5.44 hours (Dunn 2013).

1. The P-value for comparing these two means is about $P = 0.70$. What does this mean?
2. Is this difference likely to be of any practical importance? Explain.
3. What would be a useful, but correct, conclusion for ALDI to report from the study? Explain.
4. What else would be useful to know in comparing the two brands of batteries?

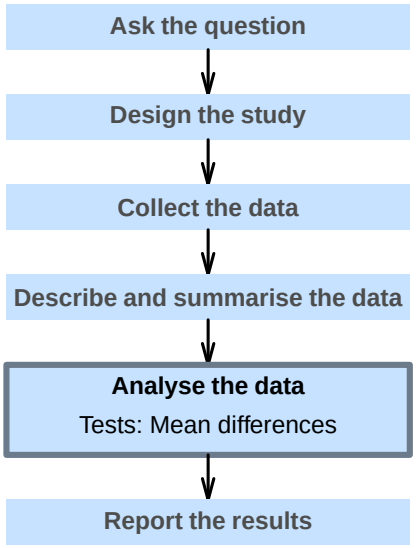
29

Hypothesis tests for the mean difference (paired data)

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, understand the tools of inference, and to form *confidence intervals*.

In this chapter, you will learn about *hypothesis tests* for the mean difference (i.e., for *paired data*). You will learn to:

- conduct hypothesis tests for the mean difference with paired data
- determine whether the conditions for using these methods apply in a given situation.



29.1 Introduction: Insulation

The Electricity Council in Bristol wanted to determine if a certain type of wall-cavity insulation reduced energy consumption in winter, on average (The Open University 1983). Their RQ was:

> Is there a *mean saving* in energy consumption due to adding insulation?

The data collected are shown in Table 29.1. These data were used in Sect. 23, where a CI was constructed for the mean energy saving.

For these data, finding the *difference* in energy consumption for each house seems sensible.

The data are *paired* (Sect. 23.1): the same unit of analysis is measured twice on the same variable (*before* and *after*), and the *mean* change is of interest. Pairing the values for each house makes sense; hence finding the *difference* in energy consumption at each house makes sense.

The parameter is μ_d, the population mean *saving* in energy consumption.



Making clear *how* the differences are computed is important. Here, the differences could be computed as the *Before* minus *After* (the energy consumption *saving*), or the *After* minus *Before* (the energy consumption *increase*). Either is fine, as long as you are consistent throughout. The meaning of any conclusions will be the same.

In this case, discussing energy *savings* seems most natural, so we compute the differences as *Before* minus *After*.

TABLE 29.1: *The house insulation data: Energy consumption before and after adding insulation, and the energy saving (all in MWh)*

Before	After	Energy savings
12.1	12.0	0.1
11.0	10.6	0.4
14.1	13.4	0.7
13.8	11.2	2.6
15.5	15.3	0.2
12.2	13.6	-1.4
12.8	12.6	0.2
9.9	8.8	1.1
10.8	9.6	1.2
12.7	12.4	0.3

29.2 Hypotheses and notation: Mean differences

The RQ asks if the mean energy saving *in the population* is zero or not. The *parameter* is the *population mean difference*. To make things clear, notation is needed (recapping Sect. 23.3):

- μ_d: The *population* mean *difference*.
- $\bar{d}$: The *sample* mean *difference*.
- s_d: The sample standard deviation of the *differences*.
- n: The number of *differences*.
- s.e.$(\bar{d})$: The standard error of the mean *differences*, where s.e.$(\bar{d}) = \dfrac{s_d}{\sqrt{n}}$.

The hypotheses, therefore, can be written in terms of the parameter μ_d. The *null* hypothesis is 'there is *no change* in the energy consumption, in the population':

- H_0: $\mu_d = 0$.

As noted in Sect. 28.2, the null hypothesis states that there is 'no difference, no change, no

relationship,' as measured by a parameter value. This hypothesis, the initial **assumption**, postulates that the mean difference may not be zero in the *sample* due to sampling variation.

Since the RQ asks specifically if the insulation *saves* energy, the alternative hypothesis will be a *one-tailed* hypothesis:

- H_1: $\mu_d > 0$ (one-tailed).

This hypothesis says that the mean energy saving in the population is *greater than* zero. The alternative hypothesis is *one*-tailed because of the wording of the RQ. Recall that the differences are defined as energy *savings*.

29.3 Sampling distribution: Mean differences

Initially, assume that $\mu_d = 0$. However, the *sample* mean energy saving will vary depending on which sample is randomly obtained, *even if* the mean saving in the population is zero: the sample mean energy saving has *sampling variation* and hence a *standard error*. The *sampling distribution* of $\bar{d}$ can be described, which describes what values of the statistic might be **expected** in the sample if the $\mu_d = 0$.

Answering the RQ requires data. The data should be summarised numerically (using your calculator or software, such as jamovi (Fig. 29.1) or SPSS (Fig. 29.2)), and graphically (Fig. 23.1).

Paired Samples T-Test

Paired Samples T-Test

			statistic	df	p	Mean difference	SE difference
Before	After	Student's t	1.68	9.00	0.127	0.540	0.321

Descriptives

	N	Mean	Median	SD	SE
Before	10	12.5	12.4	1.68	0.533
After	10	11.9	12.2	1.96	0.619

FIGURE 29.1: *jamovi output for the insulation data*

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Energy consumption (Before) in MWh	10	9.9	15.5	12.490	1.6842
Energy consumption (After) in MWh	10	8.8	15.3	11.950	1.9580
Energy saving in MWh	10	-1.40	2.60	.5400	1.01566
Valid N (listwise)	10				

FIGURE 29.2: *SPSS output for the insulation data*

The sample mean difference is $\bar{d} = 0.5400\ldots$ but this value of $\bar{d}$ will vary from sample

to sample even if $\mu_d = 0$. The amount of variation in $\bar{d}$ is quantified using the *standard error*. More precisely, the possible values of the sample mean differences can, under certain conditions, described using

- an approximate normal distribution; with
- a mean of $\mu_d = 0$ (taken from H_0); and
- a standard deviation (called the *standard error*) of s.e.$(\bar{d}) = s_d/\sqrt{n} = 0.3212$, where s_d is the *standard deviation* of the differences.

This describes what can be **expected** from the possible values of $\bar{d}$ (Fig. 29.3), just through sampling variation (chance) if $\mu_d = 0$.

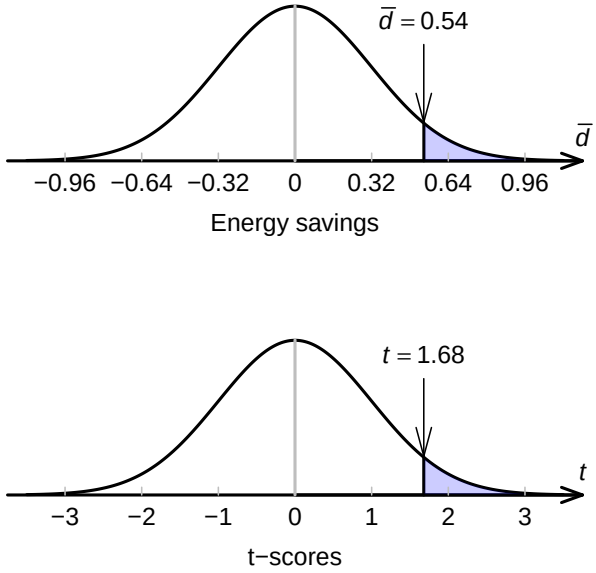


FIGURE 29.3: *The sampling distribution of sample means, if the energy saving in the population really was zero*

29.4 The test statistic: Mean differences

The sample mean difference can be located on the sampling distribution (Fig. 29.3) by computing the t-score:

$$t = \frac{0.54 - 0}{0.3212} = 1.681,$$

following the ideas in Equation (27.1). Software computes the t-score too (jamovi: Fig. 29.4; SPSS: Fig. 29.5). The t-score places the **observed** sample statistic on the sampling distribution.

29.5 P-values: Mean differences

A P-value determines if the sample data are consistent with the assumption (Table 28.1), Since $t = 1.681$, the *one-tailed* P-value is between 2.5% and 16% based on the 68–95–99.7 rule. This is a wide, and inconclusive, interval. Software gives a more precise P-value (jamovi: Fig. 29.4; SPSS: Fig. 29.5): the *two-tailed* P-value is 0.127, so the *one-tailed* P-value is $0.127/2 = 0.0635$.

Paired Samples T-Test

Paired Samples T-Test

						95% Confidence Interval	
			statistic	df	p	Lower	Upper
Before	After	Student's t	1.68	9.00	0.127	−0.187	1.27

FIGURE 29.4: *jamovi output for the insulation data*

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Energy consumption (Before) – Energy consumption (After)	.5400	1.0157	.3212	-.1866	1.2666	1.681	9	.127

FIGURE 29.5: *SPSS output for the insulation data*

⚠ The software clarifies *how* the differences have been computed:

- **jamovi**: At the left of the output (Fig. 29.4), the order implies the differences are found as *Before minus After*.
- **SPSS**: At the left of the output (Fig. 29.5), the difference is described as *Before − After*.

29.6 Conclusions: Mean differences

The *one-tailed* P-value is 0.0635, suggesting only slight evidence[1] supporting H_1. To write a conclusion, an *answer to the RQ* is needed, plus *evidence* leading to that conclusion; and some *summary statistics*, including a CI (indicating the precision of the statistic):

[1] `tab:PvaluesInterpretation`

Slight evidence exists in the sample (paired $t = 1.68$; one-tailed $P = 0.0635$) of a mean energy saving in the population (mean saving: 0.54 MWh; $n = 10$; 95% CI from -0.19 to 1.27 MWh) after adding the insulation.

The wording implies the direction of the differences (by talking of 'savings'). Of course, statistically validity shoud be checked; this was done in Sect. 23.9, but the validity conditions are given again in the next section, for completeness.

Example 29.1 (COVID lockdown)**.** A study of $n = 213$ Spanish health students (Romero-Blanco et al. 2020) measured (among other things) the number of minutes of vigorous physical activity (PA) performed by students *before* and *during* the COVID-19 lockdown (from March to April 2020 in Spain).

These numerical summary of the data are shown in Example 23.1, so we do not repeat it here. We define the *differences* as the number of minutes of vigorous PA *before* the COVID lockdown, minus the number of minutes of vigorous PA *during* the COVID lockdown. A difference is computed for each participant, so the data are *paired*.

Using this definition, a *positive* difference means the *Before* value is higher; hence, the differences tell us how much longer the student spent doing vigorous PA *before* the COVID lockdown. Similarly, a *negative* value means that the *During* value is higher.

The RQ is

> For Spanish health students, is there a mean change in the amount of vigorous PA during and before the COVID lockdown?

In this situation, the *parameter* of interest is the population mean difference μ_d, the mean amount that students spent in vigorous PA *before* the lockdown compared to *during* the lockdown. The hypotheses are:

- H_0: $\mu_d = 0$
- H_1: $\mu_d \neq 0$ (i.e., two-tailed)

The mean *difference* is $\bar{d} = -2.68$ minutes, with a standard deviation of $s_d = 51.30$ minutes. However, we know that the sample mean difference could vary from sample to sample, so has a standard error:

$$\text{s.e.}(\bar{d}) = \frac{s_d}{n} = \frac{51.30}{\sqrt{213}} = 3.515018.$$

The test statistic is

$$t = \frac{\bar{d} - \mu_d}{\text{s.e.}(\bar{d})} = \frac{-2.68 - 0}{3.515018} = -0.76.$$

This is a very small value, so (using the 68-95-99.7 rule) the P-value will be very large: a sample mean difference of -2.68 minutes could easily have happened by chance even if the population mean difference was zero.

We write:

There is no evidence (paired $t = -0.76$, $P > 0.10$) of a mean change in the amount of vigorous PA *before* and *during* lockdown (sample mean 2.68 minutes greater *during lockdown; standard deviation: 51.30 minutes).

29.7 Statistical validity conditions: Mean differences

As with any inferential procedure, these results apply under certain conditions. For a hypothesis test for the mean of paired data, these conditions are the same as for the CI for the mean difference for paired data (Sect. 23.9), and similar to those for one sample mean.

The test above is statistically valid if *one* of these conditions is true:

1. The sample size of differences is at least 25; **or**
2. The sample size of differences is smaller than 25, **and** the *population* of *differences* has an approximate normal distribution.

The sample size of 25 is a rough figure here, and some books give other values (such as 30). This condition ensures that the *distribution of the sample means has an approximate normal distribution* so that we can use the 68–95–99.7 rule.

Provided the sample size is larger than about 25, this will be approximately true *even if* the distribution of the individuals in the population does not have a normal distribution. That is, when $n > 25$ the sample means generally have an approximate normal distribution, even if the data themselves don't have a normal distribution.

In addition to the statistical validity condition, the test will be

- **internally valid** if the study was well designed; and
- **externally valid** if the sample is a simple random sample and is internally valid.

Example 29.2 (Statistical validity)**.** For the insulation data used above, the sample size is small, so the test will be statistically valid if the differences in the *population* follow a normal distribution.

We don't know if they do, though the sample data (Fig. 23.1) don't identify any obvious doubts. So the test is possibly statistically valid, but we can't be sure.

Example 29.3 (COVID lockdown)**.** In Example 29.1 concerning COVID lockdowns, the sample size was 213 Spanish health students.

Since the sample size is muich larger than 25,, the test is statistically valid.

29.8 Example: Endangered species

A study of endangered species (Harnish and Nataraajan 2020) examined

> ...whether perceived physical attractiveness of a species impacted participants' attitudes toward supporting and protecting the species...
> — Harnish and Nataraajan (2020), p. 1703

To do so, 210 undergraduate students were surveyed about 14 animals on various aspects of supporting and protecting them. Part of the data are summarised below, for two animals when asked about 'support to protect the animal from illicit trade.' *Larger* values means *greater* support for protecting the animal from illicit trade.

Species	Mean score	Standard deviation
Bay Checkerspot Butterfly	3.10	1.06
Valley Elderberry Longhorn Beetle	2.33	1.13
Difference	0.77	1.07

(Notice that the standard deviation of the difference is **not** the difference between the two given values of the standard deviation.)

The *difference* is defined as each student's score for the butterfly (deemed more attractive) *minus* their score for the beetle (deemed less attractive). A positive value therefore means more support (on average) for the butterfly.

The RQ is whether there is a mean difference between support for each animal, so the parameter is μ_d, the population mean difference.

The researchers wished to test if

> ...animals perceived as more physically attractive (i.e., the butterfly) compared to those which are perceived as less physically attractive (i.e., the beetle) will receive relatively more support to prevent the species from illicit trade
> — Harnish and Nataraajan (2020), p. 1704

Given how the difference are defined, the hypotheses are:

- H_0: $\mu_d = 0$
- H_1: $\mu_d > 0$ (i.e., one-tailed, based on the researchers' purpose)

The mean difference is $\bar{d} = 0.77$ and $s_d = 1.07$. The value of $\bar{d}$ will vary from sample to sample, so has a standard error:

$$\text{s.e.}(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{1.07}{\sqrt{210}} = 0.073837.$$

The value of the test statistic is

$$t = \frac{\bar{d} - \mu_d}{\text{s.e.}(\bar{d})} = \frac{0.77 - 0}{0.073837} = 10.43,$$

which is a *very* large value. Hence, the P-value will be very small, certainly less than 0.05.

Since the sample size is much larger than 25, the test will be statistically valid.

We write:

> There is very strong evidence ($t = 10.43$; one-tailed $P < 0.001$) that the mean difference in support for protecting the Bay Checkerspot Butterfly from illicit trade is greater than support for protecting the Valley Elderberry Longhorn Beetle from illict trade (mean difference: 0.77; standard deviation: 1.07; 95% CI for the difference: 0.62 to 0.92).

29.9 Example: Blood pressure

A US study (Schorling et al. 1997; Willems et al. 1997) was conducted to determine how CHD risk factors were assessed among parts of the population. Subjects were required to report to the clinic on multiple occasions.

One RQ of interest is:

> Is there a mean difference in blood pressure measurements between the first and second visits?

The parameter is μ_d, the population mean *reduction* in blood pressure.

Each person has a *pair* of diastolic blood pressure (DBP) measurements: One each from their first and second visits. The data (Table 23.5) are from 141 people. These data were shown in Sect. 23.10. The differences could be computed in one of two ways:

- The observation from the first visit, minus the observation from the second visit: the *reduction* in BP; or
- The observation from the second visit, minus the observation from the first visit: the *increase* in BP.

Either way is fine, as long as the order remains consistent, and the direction is made clear. Here, the observations from the *first* visit minus the observation from the *second* visit will be used, so that the differences represent the *decrease* in BP from the first to second measurement.

The appropriate graphical summary is a histogram of differences (Fig. 23.4); the numerical summary is shown in Table 29.4. Notice that having the information about the differences is essential, as the RQ is about the differences.

As always (Sect. 28.2), the null hypothesis is the 'no difference, no change, no relationship' position, proposing that the mean difference in the population is non-zero due to sampling variation:

- H_0: $\mu_d = 0$ (differences: first − second);
- H_1: $\mu_d \neq 0$.

TABLE 29.3: *The first ten observations (from the $n = 141$ available) from the diabetes study for people with both measurements: Diastolic blood pressure (DBP) for the first and second visits, and the decrease in DBP, all in mm Hg*

DBP: First visit	DBP: Second visit	Reduction in DBP
92	92	0
112	112	0
80	86	-6
90	90	0
90	96	-6
88	84	4
100	110	-10
88	88	0
70	70	0
122	112	10

TABLE 29.4: *The numerical summary for the diabetes data (in mm Hg). The differences are the second visit value minus the first visit value: the decreases in diastolic blood pressure from the first to second visit*

	Mean	Standard deviation	Standard error	Sample size
DBP: First visit	94.48	11.473	0.966	141
DBP: Second visit	92.52	11.555	0.973	141
Decrease in DBP	1.95	8.026	0.676	141

The alternative hypothesis is *two-tailed* because of the wording of the RQ. As usual, **assume** that H_0 is true, and then the evidence is evaluated to determine if it contradicts this assertion.

The sampling distribution describes how the sample mean difference is **expected** to vary from sample to sample due to sampling variation, when $\mu_d = 0$. Under certain circumstances, the sample mean differences are likely to vary with a normal distribution, with a mean of 0 (from H_0) and a standard deviation of s.e.$(\bar{d}) = 0.676$.

The relative value of the **observed** sample statistic is found by computing a t-score, using software (jamovi: Fig. 29.6; SPSS: Fig. 29.7), or manually (Eq. (27.1), using the information in Table 29.4):

$$t = \frac{\text{sample statistic} - \text{assumed population parameter}}{\text{standard error of the statistic}}$$
$$= \frac{1.950 - 0}{0.676} = 2.885.$$

Either way, the t-score is the same.

A P-value is then needed to decide if the sample is **consistent** with the assumption. Using the 68–95–99.7 rule, the approximate two-tailed P-value is much smaller than 0.05. Alternatively, the software output (Fig. 29.6; Fig. 29.7) reports the two-tailed P-value as $P = 0.005$.

We conclude:

Paired Samples T-Test

Paired Samples T-Test

								95% Confidence Interval	
			statistic	df	p	Mean difference	SE difference	Lower	Upper
bp.1d	bp.2d	Student's t	2.89	140	0.005	1.95	0.676	0.614	3.29

Descriptives

	N	Mean	Median	SD	SE
bp.1d	141	94.5	94.0	11.5	0.966
bp.2d	141	92.5	92.0	11.6	0.973

FIGURE 29.6: *jamovi output for the diabetes data*

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Diastolic blood pressure 1	94.48	141	11.473	.966
	Diastolic blood pressure 2	92.52	141	11.555	.973

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Diastolic blood pressure 1 & Diastolic blood pressure 2	141	.757	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Diastolic blood pressure 1 – Diastolic blood pressure 2	1.950	8.026	.676	.614	3.287	2.885	140	.005

FIGURE 29.7: *SPSS output for the diabetes data*

Strong evidence exists in the sample (paired $t = 2.855$; two-tailed $P = 0.005$) of a population mean difference between the first and second DBP readings (mean difference 1.95 mm Hg higher for first reading; 95% CI from 0.61 to 3.3 mm Hg; $n = 141$).

Since $n > 25$, the results are statistically valid.

⚠ Just saying 'there is evidence of a difference' is insufficient; it is also important to say *which* measurement is, on average higher (that is, what the differences *mean*).

29.10 Summary

Consider testing a hypothesis about a population mean difference μ_d, based on the value of the sample mean difference $\bar{d}$. Under certain statistical validity conditions, the sample mean difference varies with an approximate normal distribution centered around the hypothesised value of μ_d, with a standard deviation of

$$\text{s.e.}(\bar{d}) = \frac{s_d}{\sqrt{n}}.$$

This distribution describes what values of the sample mean difference could be **expected** if the value of μ_d in the null hypothesis was true. The *test statistic* is

$$t = \frac{\bar{d} - \mu_d}{\text{s.e.}(\bar{d})},$$

where μ_d is the hypothesised value in the null hypothesis. The t-score describes what value of $\bar{d}$ was **observed** in the sample, relative to what was expected. The t-value is like a z-score, so an approximate P**-value** can be estimated using the 68–95–99.7 rule, or is found using software. The P-values helps determine if the sample evidence is consistent with the assumption, or contradicts the assumption.

29.11 Quick review questions

A study (Bacho et al. 2019) compared joint pain in stroke patients before and after a supervised exercise treatment. Participants ($n = 34$) were assessed *before* and *after* treatment. The mean *change* in joint pain after 13 weeks was 1.27 (with a standard error of 0.57) using a standardised tool.

1. True or false? Only 'before and after' studies can be paired.
2. True or false? The null hypothesis is about the population mean *difference*.
3. The value of the test statistic (to two decimal places) is
4. The two-tailed P-value will be

29.12 Exercises

Selected answers are available in Sect. D.27.

Exercise 29.1. People often struggle to eat the recommended intake of vegetables. In one study exploring ways to increase vegetable intake in teens (Fritts et al. 2018), teens rated the

taste of raw broccoli, and raw broccoli served with a specially-made dip. (These data were also seen in Exercise 23.1.)

Each teen ($n = 101$) had a *pair* of measurements: the taste rating of the broccoli *with* and *without* dip. Taste was assessed using a '100 mm visual analog scale,' where a higher score means a better taste. In summary:

- For raw broccoli, the mean taste rating was 56.0 (with a standard deviation of 26.6);
- For raw broccoli served with dip, the mean taste rating was 61.2 (with a standard deviation of 28.7).

Because the data are paired, the *differences* are the best way to describe the data. The mean difference in the ratings was 5.2, with standard error of 3.06.

Perform a hypothesis test to see if the use of dip increases the mean taste rating.

Exercise 29.2. In a study of hypertension (MacGregor et al. 1979; Hand et al. 1996), patients were given a drug (Captopril) and their systolic blood pressure measured immediately before and two hours after being given the drug (data shown). The aim is to see if there is evidence of a *reduction* in blood pressure after taking Captopril. (This study was also seen in Exercise 23.2.)

Using these data and the software output (jamovi: Fig. 29.8; SPSS: Fig. 29.9):

1. Explain why it is probably more sensible to compute differences as the *Before* minus the *After* measurements. What do the differences *mean* when computed this way?
2. Compute the differences.
3. Construct a suitable graph for the differences.
4. Write down the hypotheses.
5. Write down the t-score.
6. Write down the P-value.
7. Write a conclusion.

TABLE 29.5: *The Captopril data: before after after systolic blood pressures (in mm Hg)*

Before	After	Before	After
210	201	173	147
169	165	146	136
187	166	174	151
160	157	201	168
167	147	198	179
176	145	148	129
185	168	154	131
206	180	NA	NA

Exercise 29.3. A study (Allen et al. 2018) examined the effect of exercise on smoking. Men and women were assessed on a range of measures, including the 'intention to smoke.' (This study was also seen in Exercise 23.3.) 'Intention to smoke,' and other measures, were assessed both before and after exercise for each subject, using the 10-item quantitative *Questionnaire*

Paired Samples T-Test

Paired Samples T-Test

						95% Confidence Interval	
			statistic	df	p	Lower	Upper
Before	After	Student's t	8.12	14.0	< .001	13.9	23.9

FIGURE 29.8: *jamovi output for the Captoril data*

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Before blood pressure (in mm Hg) - After blood pressure (in mm Hg)	18.933	9.027	2.331	13.934	23.93	8.123	14	.000

FIGURE 29.9: *SPSS output for the Captoril data*

of Smoking Urges – Brief[2] scale (Cox et al. 2001), and the quantitative *Minnesota Nicotine Withdrawal Scale*[3] (Shiffman et al. 2004).

Smokers (defined as people smoking at least five cigarettes per day) aged 18 to 40 were enrolled for the study. For the 23 women in the study, the mean intention to smoke after exercise *reduced* by 0.66 (with a standard error of 0.37). Perform a hypothesis test to determine if there is evidence of a population mean reduction in intention to smoke for women after exercising.

Exercise 29.4. In a study (Cressie et al. 1984) conducted at the Adelaide Children's Hospital:

> . . . a group of beta thalassemia patients [. . .] were treated by a continuous infusion of desferrioxamine, in order to *reduce* their ferritin content. . .
> — Cressie et al. (1984), p. 107; emphasis added

Using the data (shown below), conduct a hypothesis test to determine if there is evidence that the treatment reduces the ferritin content, as intended.

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2527734/

[3] http://www.med.uvm.edu/behaviorandhealth/research/minnesota-tobacco-withdrawal-scale

TABLE 29.6: *The ferritin content (in μg/L) for 20 thalassemia patients at the Adelaide Children's Hospital*

September	March	Reduction
6630	5100	1530
4590	3510	1080
3510	6600	-3090
6375	8000	-1625
2500	2800	-300
1400	2860	-1460
4580	3640	940
6885	9030	-2145
4200	4420	-220
5600	7910	-2310
5360	6780	-1420
6110	7250	-1140
5300	6000	-700
3120	4300	-1180
3300	4680	-1380
11400	8500	2900
3100	3735	-635
2800	2730	70
3500	6600	-3100
12700	7000	5700

30

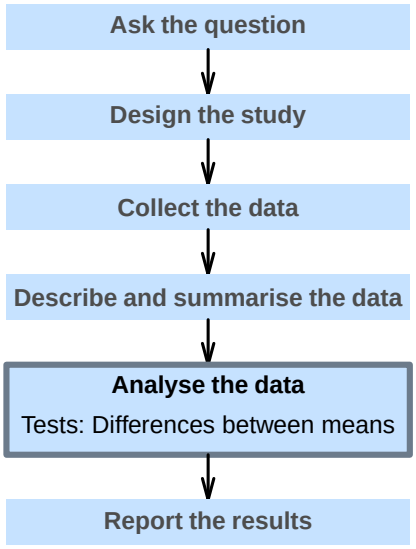
Hypothesis tests for means of two independent groups



So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, understand the tools of inference, and to form *confidence intervals*.

In this chapter, you will learn about *hypothesis tests* for the difference between two means. You will learn to:

- conduct hypothesis tests for comparing two means.
- determine whether the conditions for using these methods apply in a given situation.



30.1 Introduction: Reaction times

A study (Strayer and Johnston 2001), examined the reaction times of students while driving.

In one study, two different groups of students were used: one group *used* a mobile phone, and a different group *did not use* a mobile phone. Their reaction times were measured in a driving simulator. These data were seen previously in Sect. 24.1.

The two groups receive different treatments: one group used a mobile phone while driving, and a different group *did not* use a mobile phone while driving.

The data are not paired; instead, the means of two separate (or independent) samples are being

compared. (The data would be paired if *each* student was measured twice: once using a phone, and once without using a phone.)

Consider the RQ:

> For students, is there a difference between the mean reaction time while driving, between students who *are using* a mobile phone and students who are *not using* a mobile phone?

Part of the data are shown in Table 23.1.

TABLE 30.1: *Reaction times (in milliseconds) for students using, and not using, mobile phones. The first ten observations are shown, but 32 students are in each group*

Using phone	Not using phone
636	557
623	572
615	457
672	489
601	532
600	506
542	648
554	485
543	610
520	444

30.2 Hypotheses and notation: Two independent means

Since two groups are being compared, distinguishing between the statistics for the two groups (say, Group A and Group B) is important (recapping Sect. 24.3).

One way is to use subscripts (see Table 30.2). Using this notation, the *parameter* in the RQ is the difference between population means: $\mu_A - \mu_B$.

As usual, the population values are unknown, so this is estimated using the statistic $\bar{x}_A - \bar{x}_B$.

TABLE 30.2: *Notation used to distinguish between the two independent groups*

	Group A	Group B
Population means:	μ_A	μ_B
Sample means:	$\bar{x}_A$	$\bar{x}_B$
Standard deviations:	s_A	s_B
Standard errors:	$\text{s.e.}(\bar{x}_A) = \dfrac{s_A}{\sqrt{n_A}}$	$\text{s.e.}(\bar{x}_B) = \dfrac{s_B}{\sqrt{n_B}}$
Sample sizes:	n_A	n_B

For the reaction-time data, the differences are computed as the mean reaction time for phone users, *minus* the mean reaction time for non-phone users: $\mu_P - \mu_C$. By this definition, the

differences refer to how much greater (on average) the reaction times are when students are using phones.

The parameter is $\mu_P - \mu_C$, the difference between the population mean reaction times (using phone, minus *not* using a phone).

⚠ Here the difference is computed as the mean reaction time for phone users, *minus* the mean reaction time for non-phone users. Computing the difference as the mean reaction time for non-phone users, *minus* the mean reaction time for phone users is also correct; you need to be clear about how the difference is computed, and be consistent throughout.

As always (Sect. 28.2), the null hypothesis is the default 'no difference, no change, no relationship' position; hence the null hypothesis is that there is 'no difference' between the population means of the two groups:

- H_0: $\mu_P - \mu_C = 0$ (or $\mu_P = \mu_C$).

This hypothesis proposes that any difference between the *sample* means is due to *sampling variation*. This becomes the initial **assumption**.

From the RQ, the alternative hypothesis will be *two*-tailed:

- H_1: $\mu_P - \mu_C \neq 0$ (or $\mu_P \neq \mu_C$).

30.3 Sampling distribution: Two independent means

The data for testing the hypothesis are shown in Table 30.1. The numerical summary (Sect. 24.1) *must* summarise the difference between the means (since the RQ is about the difference), and should summarise each group. All this information is found using software (jamovi: Fig. 30.1; SPSS: Fig. 30.2), and can be compiled into a table (Table 30.3).

The appropriate summary for graphically summarising the *data* is a boxplot (though a dot chart is also acceptable). An error bar chart (Fig. 24.7), which allows the *sample means* to be compared, should also be produced.

The difference between the sample means is 51.59 ms… but this value will vary from sample to sample; that is, there is *sampling variation*. The sampling variation (**expectation**) for the values of $\bar{x}_A - \bar{x}_B$ can be described as having:

- an approximate normal distribution;
- centred around $\mu_P - \mu_C = 0$ (from H_0);
- with a standard deviation of s.e.$(\bar{x}_P - \bar{x}_C)$, called the *standard error for the difference between the means*.

⚠ jamovi and SPSS give results from *two* similar hypothesis tests. In this book, we will always use the second row of information (the "Welch's t" row in jamovi; the "Equal variance not assumed" row in SPSS) because it is more general and makes fewer assumptions.

Independent Samples T-Test

Independent Samples T-Test

							95% Confidence Interval	
		statistic	df	p	Mean difference	SE difference	Lower	Upper
Reaction	Student's t	2.63	62.0	0.011	51.6	19.6	12.4	90.8
	Welch's t	2.63	56.7	0.011	51.6	19.6	12.3	90.9

Group Descriptives

	Group	N	Mean	Median	SD	SE
Reaction	Phone	32	585	569	89.6	15.8
	Control	32	534	530	65.4	11.6

FIGURE 30.1: *jamovi output for the phone reaction time data*

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Reaction time (in ms)	Phone	32	585.19	89.646	15.847
	Control	32	533.59	65.360	11.554

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Reaction time (in ms)	Equal variances assumed	.077	.783	2.631	62	.011	51.594	19.612	12.390	90.798
	Equal variances not assumed			2.631	56.70	.011	51.594	19.612	12.317	90.871

FIGURE 30.2: *SPSS output for the phone reaction time data*

30.4 The test statistic: Two independent means

The observed sample mean difference (**observations**), relative to what was expected, is found by computing the test statistic, in this case, a t-score. The software output (jamovi: Fig. 30.1; SPSS: Fig. 30.2) can be used, but the t-score can also be computed manually:

TABLE 30.3: *Numerical summaries of the reaction-time data (in milliseconds)*

	Mean	Sample size	Standard deviation	Standard error
Using phone	585.1875	32	89.6460558116231	15.8473334927566
Not using phone	533.59375	32	65.3599756065616	11.5541204923968
All students	51.59375			19.6121309189865

$$
\begin{aligned}
t &= \frac{\text{sample statistic} - \text{assumed population parameter, from } H_0}{\text{standard error for sample statistic}} \\
&= \frac{(\bar{x}_P - \bar{x}_C) - (\mu_P - \mu_C)}{\text{s.e.}(\bar{x}_P - \bar{x}_C)} \\
&= \frac{51.594 - 0}{19.612} = 2.631,
\end{aligned}
$$

as in the software output.

30.5 P-values: Two independent means

A P-value is needed to determine if the sample statistic is consistent with the assumption. Since the t-score is large, the P-value will be small using the 68–95–99.7 rule. This is confirmed by the software (jamovi: Fig. 30.1; SPSS: Fig. 30.2): the two-tailed P-value is 0.011. The small P-value suggests that the observations are *inconsistent* with the assumption (Table 28.1).

30.6 Conclusions: Two independent means

In conclusion we write:

> Moderate evidence exists in the sample (two independent samples $t = 2.631$; two-tailed $P = 0.011$) that the population mean reaction time is different for students using a mobile phone (mean: 585.19 ms; $n = 32$) and students *not* using a mobile phone (mean: 533.59ms; $n = 32$; 95% CI for the difference: 12.4 to 90.9ms longer for phone users).

Again, the conclusions contains an *answer to the RQ*, the *evidence* leading to this conclusion ($t = 2.631$; two-tailed $P = 0.011$), and some *sample summary statistics*, including a CI.

30.7 Statistical validity conditions: Two independent means

As usual, these results apply under certain conditions, which are the same as those for forming a CI for the *difference* between two means.

The test above is statistically valid if *one* of these conditions is true:

1. *Both* sample sizes are at least 25; *or*
2. Either sample size is smaller than 25, **and** both *populations* have an approximate normal distribution.

The sample size of 25 is a rough figure here, and some books give other values (such as 30).

We can explore the histograms of the *samples* to determine if normality of the *populations* seems reasonable.

In addition to the statistical validity condition, the test will be

- **internally valid** if the study was well designed; and
- **externally valid** if the sample is a simple random sample and is internally valid.

Example 30.1 (Statistical validity)**.** For the reaction-time data, both samples sizes are $n = 32$. This means that the results will be statistically valid.

Explicitly, the data in each group do not need be normally distributed, since both sample sizes are larger than 25.

Example 30.2 (Gray whales)**.** A study of gray whales (*Eschrichtius robustus*) measured (among other things) the length of adult whales (Agbayani et al. 2020). The data are shown below.

Sex	Mean (in m)	Standard deviation (in m)	Sample size
Female	12.70	0.611	260
Male	12.07	0.705	139

Are adult female gray whales longer than males, on average?

Let's define the *difference* as the mean length of female gray whales *minus* the mean length of male gray whales. Then we wish to estimate the difference $\mu_F - \mu_M$, where F and M represent female and male gray whales respectively; this is the *parameter* of interest. The best estimate of this difference is $\bar{x}_F - \bar{x}_M = 12.70 - 12.07 = 0.63$ m.

The hypotheses are:

- H_0: $\mu_F - \mu_M = 0$
- H_1: $\mu_F - \mu_M \neq 0$

We know that the difference between the sample means is likely to vary from sample to sample, and hence it has a standard error.

We cannot easily determine the standard error of this difference from the above information (though it is possible), so we must be *given* this information: s.e.$(\bar{x}_F - \bar{x}_M) = 0.07079$.

The test statistic is

$$t = \frac{(\bar{x}_F - \bar{x}_M) - (\mu_F - \mu_M)}{\text{s.e.}(\bar{x}_F - \bar{x}_M)} = \frac{0.63 - 0}{0.07079} = 8.90,$$

which is *very* large. This means that the P-value will be very small (using the 68-95–99.7 rule).

We write:

> There is very strong evidence ($t = 8.90$; two-tailed $P < 0.001$) that the mean length of adult gray whales is different for females (mean: 12.70 m; standard deviation: 0.611 m) and males (mean: 12.07 m; standard deviation: 0.705 m; 95% CI for the difference: 0.48 m to 0.77 m).

Since both sample sizes are large, the test is statistically valid.

(Check that you can compute the correct CI!)

30.8 Example: Health Promotion services

A study (Becker et al. 1991) compared the access to health promotion (HP) services for people with and without a disability. (This study was seen in Sect. 24.10.) Access was measured using the *Barriers to Health Promoting Activities for Disabled Persons*, where higher scores mean greater barriers. The RQ is:

Is the mean BHADP score the same for people with and without a disability?

The parameter is $\mu_D - \mu_N$, the difference between the population mean BHADP score (people with disabilities, minus people without disabilities).

In this case, only numerical summary data is available (Table 30.5), not the original data. (An appropriate graphical summary, an error bar chart, can be constructed from the summary information (Fig. 24.12, though a boxplot cannot be constructed from the information.) Denoting those with and without a disability with subscripts D and N respectively, the hypotheses are:

- H_0: $\mu_D - \mu_N = 0$: There *is no* difference in the population mean BHADP scores
- H_1: $\mu_D - \mu_N \neq 0$: There *is a* difference in the population mean BHADP scores

TABLE 30.5: *The BHADP data summary*

	Sample mean	Std deviation	Sample size	Std error
Disability	31.83	7.73	132	0.6728
No disability	25.07	4.8	137	0.4101
Difference	6.76			0.80285

The best estimate of the difference in *population* means is the difference between the *sample* means: $(\bar{x}_D - \bar{x}_{ND}) = 6.76$. The table also gives the standard error for estimating this *difference* as s.e.$(\bar{x}_D - \bar{x}_{ND}) = 0.80285$ (as *given* in the article).



The *standard error is given here*; you **cannot** easily calculate this from the given information. You are not expected to do so.

Using the summary information in Table 30.5, the t-score is computed using Equation (27.1):

$$t = \frac{6.76 - 0}{0.80285} = 8.42.$$

(Recall that $\mu_D - \mu_N = 0$ from the null hypothesis.) Using the 68–95–99.7 rule, this *very* large t-score implies the P-value will be *very* small. We conclude:

> Strong evidence exists in the sample ($t = 8.42$; two-tailed $P < 0.001$) that people with a disability (mean: 31.83; $n = 132$; standard deviation: 7.73) and people without a disability (mean: 25.07; $n = 137$; standard deviation: 4.80) have different population mean access to health promotion services (95% CI for the difference: 5.17 to 8.35).

30.9 Example: Face-plant study

A study (Wojcik et al. 1999) compared the lean-forward angle in younger and older women. (This study was seen in Sect. 24.11.) An elaborate set-up was constructed to measure this lean-forward angle, using harnesses.

Consider this RQ:

> Among healthy women, is the mean lean-forward angle *greater* for younger women compared to older women?

The parameter is $\mu_Y - \mu_O$, the difference between the population mean lean-forward angle (younger women, minus older women).

This is a *one-tailed* RQ. Denoting the younger and older women with subscrpts Y and O respectively, the hypotheses are:

- H_0: $\mu_\text{Y} - \mu_\text{O} = 0$ (or $\mu_\text{Y} = \mu_\text{O}$): There *is no* difference in the population mean lean-forward angle between the two age groups (the **assumption**);
- H_1: $\mu_\text{Y} - \mu_\text{O} > 0$ (or $\mu_\text{Y} > \mu_\text{O}$): There *is a* difference in the population mean lean-forward angle between the two groups.

The data (Table 24.6), numerical summary (Table 24.7). and error bar chart (Fig. 24.13) were shown in Sect. 24.11.

Using the sampling distribution (**expectation**), the t-score can be found on the the software output (jamovi: Fig. 30.3; SPSS: Fig. 30.3), or manually:

$$t = \frac{14.5 - 0}{2.167} = 6.691$$

(**observation**). The two-tailed P-value is 0.001, so the *one-tailed* P-value is $0.001 \div 2 = 0.0005$. This is very small, so we **conclude**:

> Very strong evidence exists in the sample ($t = 6.691$; one-tailed $P = 0.0005$) that the population mean one-step fall recovery angle for healthy women is *greater* for young women (mean: $30.7°$; std. dev.: $2.58°$; $n = 10$) compared to older women (mean: $16.20°$; std. dev.: $4.44°$; $n = 5$; 95% CI for the difference: $9.1°$ to $19.9°$).

Independent Samples T-Test

Independent Samples T-Test

							95% Confidence Interval	
		statistic	df	p	Mean difference	SE difference	Lower	Upper
LeanAngle	Student's t	7.88	13.0	< .001	14.5	1.84	10.5	18.5
	Welch's t	6.69	5.59	< .001	14.5	2.17	9.10	19.9

Group Descriptives

	Group	N	Mean	Median	SD	SE
LeanAngle	Younger females	10	30.7	31.5	2.75	0.870
	Older females	5	16.2	15.0	4.44	1.98

FIGURE 30.3: *jamovi output for the face-plant data*

The sample sizes are both small, so the test may not be statistical valid. However, since the P-value is so small, the conclusion is unlikely to change substantially.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Maximum lean angle (in degrees)	Equal variances assumed	1.578	.231	7.875	13	.000	14.500	1.841	10.522	18.478
	Equal variances not assumed			6.691	5.592	.001	14.500	2.167	9.102	19.898

FIGURE 30.4: *SPSS output for the face-plant data*

30.10 Summary

To test a hypothesis about a difference between two population means $\mu_1 - \mu_2$, based on the value of the difference between two sample mean $\bar{x}_1 - \bar{x}_2$, **assume** the value of $\mu_1 - \mu_2$ in the null hypothesis to be true (usually zero). Then, the difference between the sample means varies from sample to sample and, under certain statistical validity conditions, varies with an approximate normal distribution centered around the hypothesised value of $\mu_1 - \mu_2$, with a standard deviation of s.e.$(\bar{x}_1 - \bar{x}_2)$. This distribution describes what values of the sample mean could be **expected** in the sample if the value of μ in the null hypothesis was true. The *test statistic* is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\text{s.e.}(\bar{x}_1 - \bar{x}_2)},$$

where $\mu_1 - \mu_2$ is the hypothesised value given in the null hypothesis. This describes the **observations**. The t-value is like a z-score, and so an approximate P**-value** is estimated using the 68–95–99.7 rule, which is how we weigh the evidence to determine if it is **consistent** with the assumption.

30.11 Quick review questions

A study (Lee et al. 2016b) compared using a vegan ($n = 46$) and a conventional ($n = 47$) diet for 12 weeks, for a group of Koreans with Type II diabetes. A summary of the data for iron levels are shown in Table 30.6.

TABLE 30.6: *Comparing the iron levels (mg) for subjects using a vegan or conventional diet for 12 weeks*

	Mean	**Std. dev**	**n**
Vegan diet	13.9	2.3	46
Conventional diet	15	2.7	47
Difference	1.1		

1. The sample size is missing from the 'Difference' row. What should the sample size in this row be?
2. What is the *standard deviation* for the difference?
3. What is the *standard error* for the difference?
4. The two-tailed P-value for the comparison is given as $P = 0.046$. What does this mean?

30.12 Exercises

Selected answers are available in Sect. D.28.

Exercise 30.1. Earlier, the NHANES study (Sect. 12.10; Exercise 24.1), was used to address this RQ:

> Among Americans, is the mean direct HDL cholesterol different for current smokers and non-smokers?

Use the SPSS output in Fig. 30.5 to perform a hypothesis test to answer the RQ.

T-Test

Group Statistics

	SmokeNow	N	Mean	Std. Deviation	Std. Error Mean
DirectChol	No	1668	1.3924	.42792	.01048
	Yes	1388	1.3077	.42353	.01137

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Direct Chol	Equal variances assumed	.204	.652	5.473	3054	.000	.08470	.01547	.05435	.11504
	Equal variances not assumed			5.478	2964.437	.000	.08470	.01546	.05438	.11501

FIGURE 30.5: *SPSS output for the NHANES data*

Exercise 30.2. A study of male paramedics in Western Australia compared conventional paramedics with special operations paramedics (Chapman et al. 2007). Some information comparing their physical profiles is shown in Table 30.7.

1. Compute the missing standard errors.
2. Consider comparing the mean grip strength for the two groups of paramedics. The *standard error for the difference between the means* is 3.3044.
   - Carefully write down the hypotheses.
   - Compute the t-score for testing if a difference exists between the two types of paramedics.
   - Approximate the P-value using the 68–95–99.7 rule.

- Discuss the conditions required for statistical validity in this context.
- Make a conclusion.

3. Consider comparing the mean number of push-ups completed in one minute. The *standard error for the difference between the means* is 4.0689.
   - Carefully write down the hypotheses.
   - Compute the t-score for testing if a difference exists between the two types of paramedics.
   - Approximate the P-value using the 68–95–99.7 rule.
   - Discuss the conditions required for statistical validity in this context.
   - Make a conclusion.

TABLE 30.7: *The physical profile of conventional ($n = 18$) and special operation ($n = 11$) paramedics in Western Australia*

	Conventional	Special Operations
Grip strength (in kg)		
Mean	51	56
Std deviation	8	9
Std error		
Push-ups (per minutes)		
Mean	36	47
Std deviation	10	11
Std error		

Exercise 30.3. Consider again the body temperature data from Sect. 27.1. The researchers also recorded the gender of the patients, as they also wanted to compare the mean internal body temperatures for males and females.

Use the jamovi output in Fig. 30.6 to perform this test and make a conclusion. Also comment on the practical significance of your results.

Independent Samples T-Test

Independent Samples T-Test

							95% Confidence Interval	
		statistic	df	p	Mean difference	SE difference	Lower	Upper
BodyTempC	Student's t	−2.29	128	0.024	−0.161	0.0703	−0.300	−0.0216
	Welch's t	−2.29	128	0.024	−0.161	0.0703	−0.300	−0.0216

Group Descriptives

	Group	N	Mean	Median	SD	SE
BodyTempC	Male	65	36.7	36.7	0.388	0.0481
	Female	65	36.9	36.9	0.413	0.0512

FIGURE 30.6: *jamovi output for the body-temperature data*

Exercise 30.4. A study (Woodward and Walker 1994) examined the sugar consumption in industrialised (mean: 41.8 kg/person/year) and non-industrialised (mean: 24.6 kg/person/year) countries. The jamovi output is shown in Fig. 30.7.

1. Write the hypotheses.
2. Write down and interpret the CI.
3. Write a conclusion for the hypothesis test.

Independent Samples T-Test

Independent Samples T-Test

							95% Confidence Interval	
		statistic	df	p	Mean difference	SE difference	Lower	Upper
Sugar	Student's t	−5.25 [a]	88.0	< .001	−17.2	3.29	−23.8	−10.7
	Welch's t	−6.47	87.2	< .001	−17.2	2.66	−22.5	−11.9

[a] Levene's test is significant (p < .05), suggesting a violation of the assumption of equal variances

Group Descriptives

	Group	N	Mean	Median	SD	SE
Sugar	No	61	24.6	24.2	16.6	2.13
	Yes	29	41.8	44.0	8.63	1.60

FIGURE 30.7: *jamovi output for the sugar-consumption data*

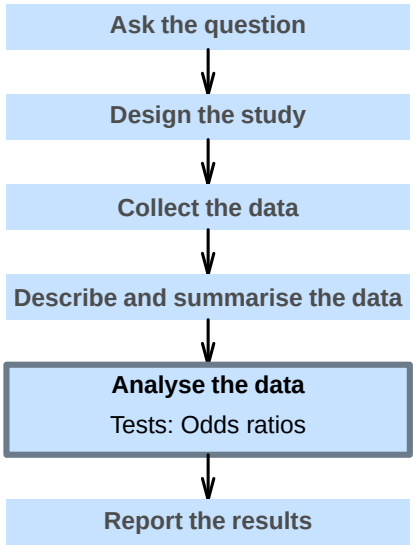
31

Hypothesis tests for comparing odds

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, understand the tools of inference, and to form *confidence intervals*.

In this chapter, you will learn about *hypothesis tests* for odds ratios. You will learn to:

- conduct hypothesis tests for an OR (i.e., comparing two proportions, or comparing two odds), using chi-square tests using jamovi and SPSS output.
- determine whether the conditions for using these methods apply in a given situation.



31.1 Introduction: Meals on-campus

In Sect. 25.1 a study (Mann and Blotnicky 2017) was introduced to examine the eating habits of university students. Researchers cross-classified? $n = 183$ students into groups according to two qualitative variables:

- Where they live: With their parents, or not with their parents;
- Whether they eat most of their meals *off-campus*, or most of their meals *on-campus*.

Since both variables observed on each student (the unit of analysis) are qualitative, means are not appropriate. However, the data can be compiled into a two-way table of counts (Table 31.1).

TABLE 31.1: *Where university students live and eat*

	Lives with parents	**Doesn't live with parents**	**Total**
Most off-campus	52	105	**157**
Most on-campus	2	24	**26**
Total	**54**	**129**	**183**

Since both qualitative variables have two levels, the table is a 2×2 table. A graphical summary is shown in Fig. 25.1, and a numerical summary in Table 31.2. (The details of the computations appear in Sect. 25.1).

TABLE 31.2: *The odds and percentage of university students eating most meals off-campus*

	Odds of having most meals off-campus	Percentage having most **meals off-campus**	**Sample size**
Living with parents	26	96.3	54
Not living with parents	4.375	81.4	129
Odds ratio	5.943		

The parameter is the population OR, comparing the odds of eating most meals *off*-campus for students living with their parents to students *not* living with their parents.



Understanding how software computes the odds ratio is important for understanding the output. In jamovi and SPSS, the odds ratio can be interpreted in *either* of these two ways:

- The *odds* are the odds of eating most meals *off-campus* (Row 1 of Table 31.1). Then, the odds ratio compares these odds for students living with their parents (Column 1 of Table 31.1) to those *not* living with their parents (Row 2 of Table 31.1). That is, the odds are $52/2 = 26$ (for those living with parents) and $105/24 = 4.375$ (for those not living with parents), so the OR is then $26/4.375 = 5.943$, as in the output (jamovi: Fig. 31.1; SPSS: Fig. 31.2).
- The *odds* are the odds of living with parents (Column 1 of Table 31.1). Then, the odds ratio compares these odds for students eating most meals off-campus (Row 1 of Table 31.1) to the odds of students eating most meals on-campus (Row 2 of Table 31.1). That is, the odds of living with parents are $52/105 = 0.49524$ (for those eating most meals off-campus) and $2/24 = 0.083333$ (for those eating most meals on-campus), so the OR is then $0.49524/0.083333 = 5.943$, as in the output (jamovi: Fig. 31.1; SPSS: Fig. 31.2).

In other words, the odds and odds ratios are relative to the first row or first column.

Unlike the previous decision-making RQs, this RQ does not concerns means. Instead, the RQ can be written in terms of comparing proportions, odds, or odds ratios.

For reasons that we can't delve into, usually the odds ratio (OR) is used as the parameter. One important reason is that software produces output related to testing the OR. Using the OR, the RQ could be written as

> Is the *population odds ratio* of eating most meals off-campus, comparing students who live *with their* parents to students *not living with* their parents, equal to one?

Alternatively, but probably easier to understand, is to write the RQ in terms of comparing the odds in the two groups explicitly:

> Are the *population odds* of students eating most meals off-campus the same for students *living with* their parents and for students *not living with* their parents?

The RQ can also be worded as comparing the percentages (or proportions) of students eating meals off-campus in each group. This is equivalent to the RQs above, but is not directly related to the software output, which works with odds ratios.

Another alternative, which sounds less direct but is useful for two-way tables larger than 2×2 (see Sect. 31.10), is worded in terms of *relationships* or *associations* between the variables:

> Is there a relationship (or association) between where students eat most of their meals and whether or not the student lives with their parents?

All of these are equivalent. Usually, for 2×2 tables, working with *odds* or *odds ratios* is best, because most software (including jamovi and SPSS) readily produces CIs for the odds ratio.

31.2 Hypotheses and notation: Comparing odds

For two-way tables of counts, the *parameter* is the population odds ratio. As usual, the null hypothesis is the 'no difference, no change, no relationship' position. So in this context:

- H_0: The *population* OR is one; or (equivalently):
  The *population* odds are the same in each group.

This hypothesis proposes that the *sample* odds are not the same due to sampling variation. This is the initial **assumption**.

The alternative hypothesis is

- H_1: The *population* OR is not one; or (equivalently):
  The *population* odds are *not* the same in each group.

The alternative hypothesis is *always* two-tailed for analysing two-way tables of counts.

> ⚠ For analysing two-way tables of counts, the alternative hypotheses **are always two-tailed**.

The hypotheses can also be written in terms of differences in percentages (or proportions), though the software output is usually expressed in terms of odds. The hypotheses can also be written in terms of associations:

- H_0: In the *population*, there is *no association* between the two variables
- H_1: In the *population*, there is *an association* between the two variables



The RQ and hypothesis only needs to be given in one of these ways. The RQ and hypotheses should be consistent (for example, if the RQ is written in terms of odds, the hypotheses should be written in terms of odds).

As usual, following the decision-making process, start by **assuming** that the null hypothesis is true: that the *population* odds ratio is one.

31.3 Expected values: Comparing odds

Assuming that the odds of having most meals off-campus is the same for both groups (that is, the population OR is one), how would the sample OR be **expected** to vary from sample to sample just because of *sampling variation*?

If the population OR was one, the odds are the same in both groups; equivalently, the percentages are the same in both groups. That is, the percentage of students eating most meals off-campus is the same for students *living with* and *not living with* their parents.

Let's consider the implication. From Table 31.1, 157 students out of 183 ate most meals off-campus; that is,

$$\frac{157}{183} \times 100 = 85.79\%$$

of the students in the entire sample ate most of their meals off-campus.

If the percentage of students who eat most of their meals off-campus is the *same* for those who live with their parents and those who don't, then we'd **expect** 85.79% of students in *both* groups to be equal to this value. That is, we would expect

- 85.79% of the 54 students (that is, 46.33) who *live with their parents* to eat most meals off-campus; and
- 85.79% of the 129 students (that is, 110.67) who *don't live with their parents* to eat most meals off-campus.

That is, the percentage (and hence the odds) is the same in each group. Those are the numbers that are *expected* to appear if the percentage was exactly the same in each group (Table 31.3), if the null hypothesis (the assumption) was true.

Think 31.1 (OR for expected counts)**.** *Consider the expected counts in Table 31.3.*

Confirm that the odds *of having most meals off-campus is the same for students living with their parents, and for students not living with their parents.*

How do those *expected values* compare to what was *observed*? For example:

- 46.33 of the 54 students who *live with their parents* are **expected** to eat most meals off-campus; yet we observed 52.
- 110.67 of the 129 students who *don't live with their parents* are **expected** to eat most meals off-campus; yet we observed 105.

The observed and expected counts are similar, but not the exactly same. This is no surprise: each sample will produce slightly different observed counts (*sampling variation*). The difference between what the observed and expected counts may be explained by sampling variation (that is, the null hypothesis explanation).



You do not have to compute the expected values when you answer one of these types of RQs (software does it for you). However, seeing how the decision-making process works in this context is helpful.

When discussing previous hypothesis tests, the *sampling distribution* of the sample statistic (in this case, the sampling distribution of the sample odds ratio) was described, and this sampling distribution had an approximate normal distribution (whose standard deviation is called the *standard error*). However, the sampling distribution of the odds ratio is more involved[1] so will not be presented.

TABLE 31.3: *Where university students live and eat: Expected counts*

	Lives with parents	Doesn't live with parents	Total
Most off-campus	46.33	110.67	**157**
Most on-campus	7.67	18.33	**26**
Total	**54.00**	**129.00**	**183**

31.4 The test statistic: Comparing odds

The decision-making process compares what is *expected* from the sample statistic if the null hypothesis about the parameter is true (Table 31.3) to what is **observe** in the sample (Table 31.1). Previously, when the summary statistics were means, t-tests were used. However, these data are not summarised by means, and a different test statistic is used.

Rather than using a t-score as the *test-statistic*, the test-statistic here is a 'chi-squared' statistic, written χ^2. A χ^2 statistic measures the overall size of the differences between the expected counts and observed counts, over the entire table.



The Greek letter χ is pronounced 'ki,' as in **ki**te.

The test statistic χ^2 is pronounced as 'chi-squared.'

From the software (jamovi: Fig. 31.1; SPSS: Fig. 31.2), $\chi^2 = 6.934$. In a 2×2 table of counts

[1] For those who wish to know: The *logarithm* of the sample ORs have an approximate normal distribution, and a *standard error*.

(when the 'degrees of freedom,' or `df`, is equal to 1, as shown in the computer output), the *square root* of the χ^2 value is approximately equivalent to a z-score. So here, the equivalent z-score is about $\sqrt{6.934} = 2.63$, which is fairly large: a small P-value is expected.

More generally, for two-way tables of any size,

$$\sqrt{\frac{\chi^2}{\text{df}}}$$

is like a z-score, where df is the 'degrees of freedom' (related to the size of the table[2]), as shown in the software output. This allows a P-value to be estimated using the 68–95–99.7 rule from the value of the χ^2 statistic.

Contingency Tables

Contingency Tables

Meals	Live		Total
	Living with parents	Not living with parents	
Most on-campus	52	105	157
Most off-campus	2	24	26
Total	54	129	183

χ^2 Tests

	Value	df	p
χ^2	6.93	1	0.008
N	183		

Comparative Measures

	Value	95% Confidence Intervals	
		Lower	Upper
Odds ratio	5.94	1.35	26.1

FIGURE 31.1: *The jamovi output for computing a CI and conducting a test*

In a chi-squared test, with a given number of 'degrees of freedom' (written `df` in the software output), the value of

$$\sqrt{\frac{\chi^2}{\text{df}}}$$

is like a z-score. This allows the P-value to be estimated using the 68–95–99.7 rule.

[2]For those who want to know: the degrees of freedom in a two-way table is the number of rows of data minus one, times the number of columns of data minus one. So, for a two-way table, the degrees of freedom is $(2-1) \times (2-1) = 1$.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.934[a]	1	.008		
Continuity Correction[b]	5.765	1	.016		
Likelihood Ratio	8.528	1	.003		
Fisher's Exact Test				.009	.005
Linear-by-Linear Association	6.896	1	.009		
N of Valid Cases	183				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.67.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval Lower	Upper
Odds Ratio for Meals (Most off-campus / Most on-campus)	5.943	1.352	26.114
For cohort Live = Living with parents	4.306	1.116	16.608
For cohort Live = Not living with parents	.725	.620	.847
N of Valid Cases	183		

FIGURE 31.2: *The SPSS output for computing a CI and conducting a test*

31.5 P-values: Comparing odds

The differences between the observated sample statistic (the sample OR) and the hypothesised population parameter (the population OR of one) is summarised by $\chi^2 = 6.934$ (approximately equivalent to $z = 2.63$). Using the 68–95–99.7 rule, a small P-value is expected.

The corresponding two-tailed P-value reported by jamovi (Fig. 31.1, under the `p` column) and SPSS (Fig. 31.2, in the `Asymptotic Significance (2-sided)` column and `Pearson Chi-Square` row) is very small (0.008 to three decimals).



Recall that, for two-way tables of counts, the alternative hypotheses *are always two-tailed*, so a two-tailed P-value is always reported.

31.6 Conclusions: Comparing odds

As usual, a very small P-value (0.008 to three decimals) means there is very strong evidence[3] supporting H_1: the evidence suggests a difference in the *population* odds in the two groups. We write:

[3] `tab:PvaluesInterpretation`

The *sample* provides strong evidence ($\chi^2 = 6.934$; two-tailed $P = 0.008$) that the odds in the *population* of having most meals off-campus is different for students living with their parents (odds: 26) and students *not* living with their parents (odds: 4.375; OR: 5.94; 95% CI from 1.35 to 26.1).

Again, as seen in Sect. 28.7, the conclusion includes three components: The *answer to the RQ*; the *evidence* used to reach that conclusion ('$\chi^2 = 6.934$; two-tailed $P = 0.008$'); and some *sample summary statistics* (inclding the 95% CI for the odds ratio).

The conclusion also makes clear what the odds and the odds ratio *mean*. The odds are describing as the 'odds… of having most meals off-campus,' and the OR as then comparing these odds between 'students living with their parents… and students *not* living with their parents.'

⚠ For two-way tables, RQs are best framed in terms of ORs or odds (but can be framed in terms of proportions or percentages, or associations or relationships).

For consistency: if the RQ is about the odds ratio, the hypotheses and conclusion should be about the odds ratio; if the RQ is about odds, the hypotheses and conclusion should be about the odds; and so on.

31.7 Statistical validity conditions

As usual, these results hold under certain conditions. The test above is statistically valid if

- All *expected* counts are at least five.

Some books may give other (but similar) conditions.

In addition to the statistical validity condition, the test will be

- **internally valid** if the study was well designed; and
- **externally valid** if the sample is a simple random sample and is internally valid.

The statistical validity condition refers to the *expected* (not the *observed*) counts. SPSS tells us if a problem exists with the expected count condition, underneath the *first* output table in Fig. 25.3. In jamovi, the *expected* counts must be explicitly requested to see if this condition is satisfied (Fig. 31.3).

For the student-eating data, the smallest *observed* count is 2 (living with parents; most meals off-campus), but the smallest *expected* count is 7.67, which is greater than five. The size of the *expected* counts is important for the statistical validity condition.

Example 31.1 (Statistical validity)**.** For the university-student eating data, *all* the cells have an expected count of at least five so the statistical validity condition is satisfied.

Contingency Tables

Meals		Live: Living with parents	Live: Not living with parents	Total
Most on-campus	Observed	52	105	157
	Expected	46.33	110.7	
Most off-campus	Observed	2	24	26
	Expected	7.67	18.3	
Total	Observed	54	129	183
	Expected	54.00	129.0	

FIGURE 31.3: *The expected values, as computed in jamovi*

31.8 Example: Pet birds

A study examined people with lung cancer, and a matched set of similar controls who did not have lung cancer, and compared the proportion in each group that had pet birds (Kohlmeier et al. 1992).

These data were studied in Sect. 25.6; the data are shown again in Table 31.4, and the numerical summary in Table 31.5 (the computations are shown in Sect. 25.6).

TABLE 31.4: *The pet bird data*

	Adults with lung cancer	**Adults without lung cancer**	**Total**
Kept pet birds	98	101	**199**
Did not keep pet birds	141	328	**469**
Total	**239**	**429**	**668**

One RQ in the study was:

> Are the **odds** of having a pet bird the same for people *with* lung cancer (cases) and for people *without* lung cancer (controls)?

The parameter is the population OR, comparing the odds of keeping a pet bird, for adults with lung cancer to adults who do not have lung cancer.

> **Extra information:** The RQ could also be written as:
>
> - Is the **percentage** of people having a pet bird the same for people *with* lung cancer (cases) and for people *without* lung cancer (controls)?
> - Is the **odds ratio** of people having a pet bird, comparing people *with* lung cancer (cases) and for people *without* lung cancer (controls), equal to one?
> - Is there a **relationship** between having a pet bird and having lung cancer?
>
> Of these, the first is probably the easiest to understand.

From this RQ (which is written in terms of *odds*), the hypotheses could be written as:

- H_0: The *odds* of having a pet bird is *the same* for people *with* lung cancer (cases) and for people *without* lung cancer (controls).

- H_1: The *odds* of having a pet bird is *not the same* for people *with* lung cancer (cases) and for people *without* lung cancer (controls).

Extra information: The null hypothesis could also be written as:

- The **percentage** of people having a pet bird is *the same* for people *with* lung cancer (cases) and for people *without* lung cancer (controls).
- The **odds ratio** of people having a pet bird, comparing people *with* lung cancer (cases) and for people *without* lung cancer (controls), is equal to one.
- There is **no relationship** between having a pet bird and having lung cancer.

Of these, the first is probably the easiest to understand.

Begin by **assuming** the null hypothesis is true: no difference exists between the odds in the *population*. Based on this assumption, the **expected** counts can be found.

From the data (Table 31.4), overall $199 \div 668 = 29.79\%$ of people own a pet bird. If there really was no difference in the odds (or the percentages) of owning a pet bird between those with and without lung cancer, about 29.79% of the people in *both* lung cancer groups are **expected** to own a pet bird.

TABLE 31.5: *The odds and percentage of subjects keeping pet birds*

	Odds of keeping pet bird	**Percentage keeping pet bird**	**Sample size**
With lung cancer:	0.6950	41.0%	239
Without lung cancer:	0.3079	25.5%	429
Odds ratio:	2.26		

About 29.79% of the 239 lung-cancer cases (or 71.20) would be expected to have a pet bird, and about 29.79% of the 429 non-lung-cancer cases (or 127.80) would be expected to have a pet bird. A table of these *expected counts* (Table 31.6). shows that all expected counts are greater than five. In practice, you do not need to compute the expecte counts; software does this automatically.

TABLE 31.6: *The expected counts for the pet bird data, if the proportion owning pet birds was the same for lung cancer cases and non-lung-cancer cases*

	Adults with lung cancer	Adults without lung cancer	Total
Kept pet birds	71.2	127.8	**199**
Did not keep pet birds	167.8	301.2	**469**
Total	**239.0**	**429.0**	**668**

The numbers in Table 31.6 are what is *expected*, *if* the percentage of people owning a pet bird is the same for lung cancer and non-lung cancer cases. How close are the expected and observed counts (in Table 31.4)?

To compare the sample statistic (what we **observed**) with the hypothesised population parameter, software is used to compute the value of χ^2 (jamovi: Fig. 31.4; SPSS: Fig. 31.5): $\chi^2 = 22.374$, approximately equivalent to a z-score of

$$\sqrt{22.374/1} = 4.730,$$

which is very large. Hence, a small P-value is expected.

The software shows that the P-value is very small ($P < 0.001$). As usual, a small P-value means that there is very strong evidence[4] supporting H_1, if H_0 is assumed true. That is, the evidence suggests there is a *difference* in the odds in the *population*. We write:

> The *sample* provides very strong evidence ($\chi^2 = 22.374$; two-tailed $P < 0.001$) that the odds in the *population* of having a pet bird is not the same for people with lung cancer (odds: 0.695) and for people without lung cancer (odds: 0.308; OR: 2.26; 95% CI from 1.6 to 3.2).

Contingency Tables

Contingency Tables

Pets		LC: Adults with lung cancer	LC: Adults without lung cancer	Total
Kept pet birds	Observed	98	101	199
	Expected	71.2	128	
Did not keep pet birds	Observed	141	328	469
	Expected	167.8	301	
Total	Observed	239	429	668
	Expected	239.0	429	

χ^2 Tests

	Value	df	p
χ^2	22.4	1	< .001
N	668		

Comparative Measures

	Value	95% Confidence Intervals: Lower	95% Confidence Intervals: Upper
Odds ratio	2.26	1.61	3.17

FIGURE 31.4: *jamovi output for the pet-birds data*

> **Think 31.2** (Interpretation)**.** *This doesn't imply that owning a pet bird* causes *lung cancer. Why not?*

Answer: The answer is given in the online book.

31.9 Example: B12 deficiency

A study in New Zealand (Gammon et al. 2012) asked:

[4] `tab:PvaluesInterpretation`

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	22.374[a]	1	.000		
Continuity Correction[b]	21.547	1	.000		
Likelihood Ratio	21.924	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	22.341	1	.000		
N of Valid Cases	668				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 71.20.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval Lower	Upper
Odds Ratio for Pets (Kept pet birds / Did not keep pet birds)	2.257	1.605	3.174
For cohort LC = Adults with lung cancer	1.638	1.345	1.995
For cohort LC = Adults without lung cancer	.726	.625	.842
N of Valid Cases	668		

FIGURE 31.5: *SPSS output for the pet-birds data*

> Among a certain group of women, are the odds of being vitamin B12 deficient different for women on a vegetarian diet compared to women on a non-vegetarian diet?

The population was 'predominantly overweight/obese women of South Asian origin living in Auckland.' The RQ could be worded in terms of odds ratios or proportions, too.

To test the claim, the hypotheses are:

- H_0: population odds for vegetarians $=$ population odds for non-vegetarians: The odds of B12 deficiency are the same in both groups.
- H_1: population odds for vegetarians $\neq$ population odds for non-vegetarians: The odds of B12 deficiency are *not* the same in both groups.

The parameter is the population OR, comparing the odds of being B12 deficient, for vegetarians to non-vegetarians.

> **Extra information:** Again (see Sect. 31.8), the RQ and the hypotheses could be written in terms of comparing the **percentages**, in terms of the **odds ratio** being equal to one, or in terms of **relationships** between the variables.

Here, the odds refer to the odds of a woman being B12 deficient. As with the RQ, the hypotheses could be worded in terms of odds ratios, proportions (or percentages), or relationships.

The data are shown in Table 31.7, and the numerical summary in Table 31.8. Since the RQ is about odds, a side-by-side bar chart is produced (Fig. 31.6) as the graphical summary.

TABLE 31.7: *The number of vegetarian and non-vegetarian women who are (and are not) B12 deficient*

	B12 deficient	Not B12 deficient	Total
Vegetarians	8	26	**34**
Non-vegetarians	8	82	**90**
Total	**16**	**108**	**124**

TABLE 31.8: *The odds and percentage of subjects that are B12 deficient*

	Odds B12 deficient	Percentage B12 deficient	Sample size
Vegetarians:	0.3077	23.5%	34
Non-vegetarians:	0.0976	8.9%	90
Odds ratio:	3.15		

The software output (jamovi: Fig. 31.7; SPSS: Fig. 31.8) shows that the OR (and 95% CI) is 3.154 (1.077 to 9.238). The chi-square value is 4.707, approximately equivalent to z-score of

$$\sqrt{\frac{4.707}{1}} = 2.17;$$

a small P-value is expected using the 68–95–99.7 rule. The software output shows that the two-tailed P-value is 0.030, which is indeed 'small.'

We conclude:

> The sample provides *moderate evidence* ($\chi^2 = 4.707$; $P = 0.030$) that the odds in the population of being vitamin B12 deficient is different for vegetarian women (odds: 0.3077) compared to non-vegetarian women (odds: 0.0976; OR: 3.2; 95% CI: 1.1 to 9.2).

The statistically valid shoud be checked. The jamovi output (Fig. 31.7) shows that the smallest expected count is 4.39. Likewise, the text under the *first* table of SPSS output in Fig. 31.8 says that

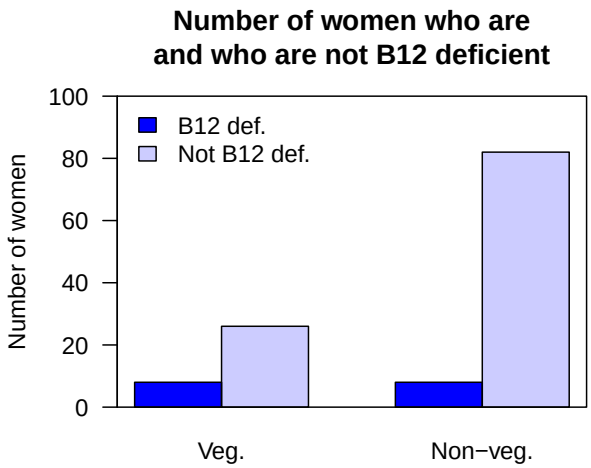


FIGURE 31.6: *A side-by-side barchart comparing the number of women B12 deficient*

Contingency Tables

Contingency Tables

Diet		B12 deficient	Not B12 deficient	Total
Vegetarian	Observed	8	26	34
	Expected	4.39	29.6	
Non-vegetarian	Observed	8	82	90
	Expected	11.61	78.4	
Total	Observed	16	108	124
	Expected	16.00	108.0	

χ^2 Tests

	Value	df	p
χ^2	4.71	1	0.030
N	124		

Comparative Measures

	Value	95% Confidence Intervals	
		Lower	Upper
Odds ratio	3.15	1.08	9.24

FIGURE 31.7: *jamovi output for the B12 data*

```
1 cells (25.0%) have expected count less than 5.  The minimum
expected count is 4.39.
```

The smallest expected count is *smaller* than five, so the results may be statistically invalid. Nonetheless, only *one* cell has an expected count less than five, and only *just* under 5, so we shouldn't be too concerned (but it should be noted).

31.10 Example: Kerbside dumping

A study of dumping households goods on the kerbside in Brisbane (Comerford et al. 2018) asked people about their opinions on the dumping. All participants were from Brisbane suburbs where a high level of kerbside dumping occurred.

The data are summarised in Table 31.9. Notice that this is a 2×3 table of counts, so it is more difficult to define a parameter.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.707[a]	1	.030		
Continuity Correction[b]	3.494	1	.062		
Likelihood Ratio	4.273	1	.039		
Fisher's Exact Test				.039	.035
Linear-by-Linear Association	4.669	1	.031		
N of Valid Cases	124				

a. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.39.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval Lower	Upper
Odds Ratio for Diet (Vegetarian / Non-vegetarian)	3.154	1.077	9.238
For cohort B12 Deficiency = B12 deficient	2.647	1.079	6.492
For cohort B12 Deficiency = Not B12 deficient	.839	.689	1.022
N of Valid Cases	124		

FIGURE 31.8: *SPSS output for the B12 data*

TABLE 31.9: *The opinion of Brisbane residents about kerbside dumping*

	Acceptable	Not acceptable	Conditionally acceptable
Reuseable	22	18	15
Non-reuseable	6	36	1

The software output is shown in Fig. 31.9 (for jamovi), and Fig. 31.10 (for SPSS); a graphical summary in Fig. 31.11.

Most of the numerical summary must be produced manually (Table 31.10), since software only produces odds ratios for 2×2 tables.

TABLE 31.10: *Odds and percentage that the opinion listed is held for reuseable rubbish; the odds ratios are computed relative to the 'Acceptable' opinion*

	Odds	**Odds ratio**	**Percentage**	**Sample size**
Acceptable:	3.667	(Reference)	78.6%	28
Not acceptable:	0.5	0.136	33.3%	54
Conditionally acceptable:	15	4.09	93.4%	16

In Table 31.10, the odds are that the given opinion refers to **Reusable** goods. Here are some of the details of these calculations:

- For **Acceptable** goods: the **odds** that these are reusable goods is $22/6 = 3.667$.

Contingency Tables

Contingency Tables

Type		Opinion			Total
		Acceptable	Not acceptable	Conditionally acceptable	
Reuseable	Observed	22	18	15	55
	Expected	15.7	30.3	8.98	
Non-reuseable	Observed	6	36	1	43
	Expected	12.3	23.7	7.02	
Total	Observed	28	54	16	98
	Expected	28.0	54.0	16.00	

χ^2 Tests

	Value	df	p
χ^2	26.3	2	< .001
N	98		

Comparative Measures

		95% Confidence Intervals	
	Value	Lower	Upper
Odds ratio	NaN [a]		

[a] Available for 2x2 tables only

FIGURE 31.9: *jamovi output for the kerbside-dumping data*

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	26.318[a]	2	.000
Likelihood Ratio	29.062	2	.000
Linear-by-Linear Association	.007	1	.935
N of Valid Cases	98		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.02.

FIGURE 31.10: *SPSS output for the kerbside-dumping data*

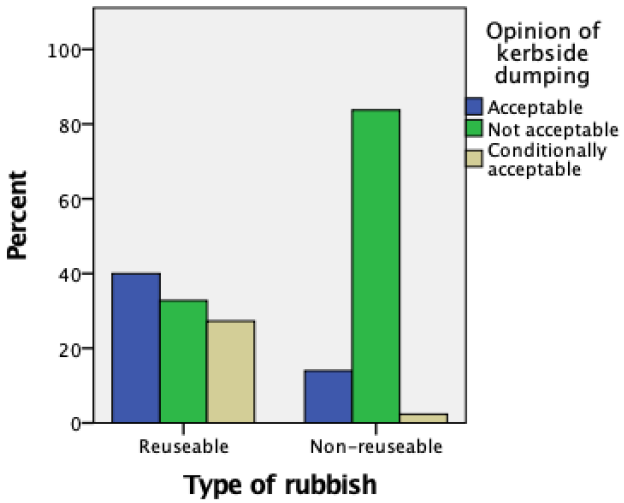


FIGURE 31.11: *A side-by-side bar chart for the kerbside-dumping data*

- For **Not acceptable** goods: the **odds** that these are reusable goods is $18/36 = 0.5$.
- For **Conditionally acceptable** goods: the **odds** that these are reusable goods is $15/1 = 15$.

Then the **odds ratios** can be computed:

- Comparing the odds of **Not acceptable** to **Acceptable**: $0.5/3.667 = 0.136$.
- Comparing the odds of **Conditionally acceptable** to **Acceptable**: $15/3.667 = 4.09$.

Note that Table 31.10 has *three* groups to compare, so three odds calculations. However, the summary has $3 - 1 = 2$ odds ratios, since odds *ratios* compare two odds. The level to which the other two are compared is called the **Reference level**. In Table 31.10, the reference level is 'Acceptable.'

(In a 2×2 table, with *two* groups to compare, the summary has only $2 - 1 = 1$ odds ratio.)

The hypotheses can be expressed in many ways (in terms of odds, odds ratio, or percentages), but perhaps the easiest approach with two-way tables larger than 2×2 is worded in terms of *relationships* or *associations* between the two variables:

- H_0: There *is no* association between the type of rubbish and the opinion of kerbside duming.
- H_1: There *is an* association between the type of rubbish and the opinion of kerbside duming.

From the software output, the χ^2-value is 26.318 and the degrees of freedom is two, so this χ^2 value is approximately equivalent to a z-score of

$$\sqrt{\frac{26.318}{2}} = 3.63.$$

This is a large z-score so, using the 68–95–99.7 rule, a very small P-value is expected; indeed, the software output reports $P < 0.001$. This suggests very strong evidence in the sample that opinions are not the same for reuseable and non-reuseable rubbish.

The conclusion could be written as

The *sample* provides very strong evidence ($\chi^2 = 26.318$; df $= 2$) that there is a relationship in the *population* between opinions about kerbside dumping and the type of rubbish.

While sample summary information could be added, the conclusion statements then become cumbersome. Instead, pointing readers to the numerical summary (Table 31.10) is probably better. Furthermore, CIs are not reported since the software does not produce CIs for tables larger than 2×2.

All *expected* values all exceed 5 (as in the jamovi (Fig. 31.9) and SPSS output (Fig. 31.10)), even though one *observed* count is less than five. The results are statistically valid.

31.11 Summary

To test a hypothesis about a population odds ratio, based on the value of the sample odds ratio, initially **assume** the value of the population odds ratio in the null hypothesis (usually one) to be true. Then, **expected counts (Step 2)** can be computed. Since the sample odds ratio varies from sample to sample, under certain statistical validity conditions a quantity closely-related to the sample odds ratio varies with an approximate normal distribution. This distribution describes what values of the sample odds ratio could be **expected** in the sample if the value of the populations odds ratio in the null hypothesis was true. The *test statistic* is a χ^2 statistic, which compares the expected and observed counts. (The value of $\sqrt{\chi^2/\text{df}}$ is like a z-score, where 'df' is the 'degrees of freedom' reported by software, and so an approximate P**-value** can be estimated using the 68–95–99.7 rule.) Software reports the P-value to assess whether the data are **consistent (Step 4)** with the assumption.

31.12 Quick review questions

A study (Egbue et al. 2017) of the adoption of electric vehicle (EVs) by a certain group of professional Americans (Example 5.14) compiled the data in Table 31.11. Output from using jamovi is shown in Fig. 31.12.

TABLE 31.11: *Responses to the question 'Would you purchase an electric vehicle in the next 10 years?' by education*

	Yes	**No**
No post-grad	24	8
Post-grad study	51	29

1. The χ^2 value is:
2. The approximately-equivalent z-score (to two decimal places) is:

3. Using the 68–95–99.5 rule, the P-value is:
4. From the software output, the P-value is:
5. The alternative hypothesis will be:
6. True or false: There is *no* evidence of a difference in the odds of buying a car in the next 10 years, between those with and without post-graduate study.

χ^2 Tests

	Value	df	p
χ^2	1.31	1	0.253
N	112		

Comparative Measures

		95% Confidence Intervals	
	Value	Lower	Upper
Odds ratio	1.71	0.68	4.28

FIGURE 31.12: *jamovi output for the EV study*

31.13 Exercises

Selected answers are available in Sect. D.29.

Exercise 31.1. Researchers (Christensen et al. 1972) studied the number of sandflies caught in light traps set at 3 and 35 feet above ground in eastern Panama. They asked:

> In eastern Panama, are the odds of finding a male sandfly the same at 3 feet above ground as at 35 feet above ground?

The data are compiled into a table (Table 31.12), and summarised numerically (Table 31.13; partially edited) and graphically (Fig. 31.13). Use the jamovi output (Fig. 31.14) to evaluate the evidence, complete Table 31.13, and write a conclusion.

TABLE 31.12: *The sex of sandflies at two heights*

	3 feet above ground	**35 feet above ground**
Males	173	125
Females	150	73

TABLE 31.13: *Odds and percentages of male sandflies at two heights above ground level*

	Odds	Percentage	Sample size
3 feet:	??	??	298
35 feet:	1.71	67.3%	223
Odds ratio:	0.67		

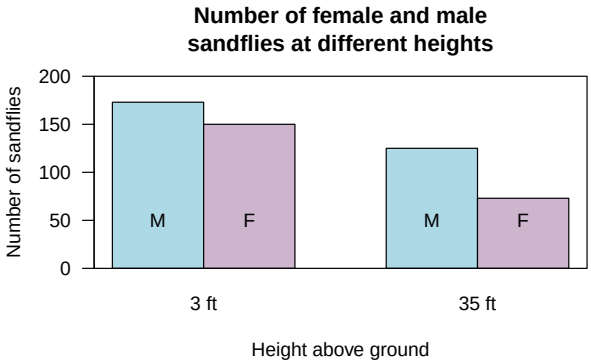


FIGURE 31.13: *A side-by-side barchart of the sandflies data*

Contingency Tables

Contingency Tables

Sex		Height: 3 feet above ground	Height: 35 feet above ground	Total
Male	Observed	173	125	298
	Expected	185	113.3	
Female	Observed	150	73	223
	Expected	138	84.7	
Total	Observed	323	198	521
	Expected	323	198.0	

χ^2 Tests

	Value	df	p
χ^2	4.59	1	0.032
N	521		

Comparative Measures

	Value	95% Confidence Intervals: Lower	95% Confidence Intervals: Upper
Odds ratio	0.674	0.469	0.968

FIGURE 31.14: *Using jamovi to compute a CI for the sandflies data*

Exercise 31.2. A prospective observational study in Western Australia compared the heights of scars from burns received (Wallace et al. 2017). The data are shown in Table 31.14. SPSS was used to analyse the data (Fig. 31.15). (This study also appeared in Exercise 25.1, where the odds ratio, and the CI for the odds ratio, were computed.)

1. Perform a hypothesis test to determine if the odds of having a smooth scar are the same for women and men.
2. Write down the conclusion.
3. Is the test statistically valid?

TABLE 31.14: *The number of men and women, with scars of different heights*

	Women	**Men**
Scar height 0mm (smooth)	99	216
Scar height more than 0mm, less than 1mm	62	115

Contingency Tables

Contingency Tables

ScarHt		Gender: Women	Gender: Men	Total
0mm	Observed	99	216	315
	Expected	103.1	212	
Between 0mm and 1mm	Observed	62	115	177
	Expected	57.9	119	
Total	Observed	161	331	492
	Expected	161.0	331	

χ^2 Tests

	Value	df	p
χ^2	0.667	1	0.414
N	492		

Comparative Measures

		95% Confidence Intervals	
	Value	Lower	Upper
Odds ratio	0.850	0.576	1.26

FIGURE 31.15: *Using jamovi to compute a CI for the scar-height data*

Exercise 31.3. In a study of turbine failures (Nelson 1982; Myers et al. 2002), 73 turbines were run for around 1800 hours, and seven developed fissures (small cracks). Forty-two different turbines were run for about 3000 hours, and nine developed fissures.

1. Use the jamovi output (Fig. 31.16) to test for a relationship.

2. Compute, then carefully interpret, the OR.
3. Write down, then carefully interpret, the test results.
4. Is the CI likely to be statistically valid (Fig. 31.17)?

Contingency Tables

Contingency Tables

Hours	Fissures: Yes	Fissures: No	Total
"1800"	7	66	73
"3000"	9	33	42
Total	16	99	115

χ^2 Tests

	Value	df	p
χ^2	3.12	1	0.077
N	115		

Comparative Measures

	Value	95% Confidence Intervals: Lower	95% Confidence Intervals: Upper
Odds ratio	0.39	0.13	1.14

FIGURE 31.16: *jamovi output for the turbine data*

Contingency Tables

Hours		Fissures: Yes	Fissures: No
"1800"	Expected	10.16	62.84
"3000"	Expected	5.84	36.16
Total	Expected	16	99

FIGURE 31.17: *jamovi output for the turbine data: expected counts*

Exercise 31.4. The *Southern Oscillation Index* (SOI) is a standardised measure of the air pressure difference between Tahiti and Darwin, and has been shown to be related to rainfall in some parts of the world (Stone et al. 1996), and especially Queensland (Stone and Auliciems 1992; Dunn 2001). As an example (Dunn and Smyth 2018), the rainfall at Emerald (Queensland) was recorded for Augusts between 1889 to 2002 inclusive, in Augusts when the monthly average SOI was positive, and when the SOI was non-positive (that is, zero or negative), as shown in Table 31.15. (This study also appeared in Exercise 25.4.)

1. Using the jamovi output in Fig. 31.18, perform a hypothesis test to determine if the odds of having no rain is the same Augusts with non-positive and negative SOI.
2. Write down the conclusion.
3. Is the test statistically valid?

TABLE 31.15: *The SOI, and whether rainfall was recorded in Augusts between 1889 and 2002 inclusive*

	Non-positive SOI	Positive SOI
No rainfall recorded	14	7
Rainfall recorded	40	53

Contingency Tables

Contingency Tables

RainRecorded		SOIPositive: Non-positive SOI	SOIPositive: Positive SOI	Total
No rainfall recorded	Observed	14	7	21
	Expected	9.95	11.1	
Rainfall recorded	Observed	40	53	93
	Expected	44.05	48.9	
Total	Observed	54	60	114
	Expected	54.00	60.0	

χ^2 Tests

	Value	df	p
χ^2	3.85	1	0.050
N	114		

Comparative Measures

	Value	95% Confidence Intervals: Lower	95% Confidence Intervals: Upper
Odds ratio	2.65	0.979	7.17

FIGURE 31.18: *jamovi output for the Emerald-rain data*

Exercise 31.5. A research study conducted in Brisbane (Dexter et al. 2019) recorded the number of people at the foot of the Goodwill Bridge, Southbank, who wore sunglasses and hats. The data were recorded between 11:30am to 12:30pm. Table 31.16 records the number of females and males wearing hats.

1. Compute the percentages of females wearing a hat.
2. Compute the percentages of males wearing a hat.
3. Compute the odds of a female wearing a hat.
4. Compute the odds of a male wearing a hat.
5. Compute the odds ratio of wearing a hat, comparing females to males.

6. Compute the odds ratio of wearing a hat, comparing males to females.
7. Find the 95% CI for the appropriate OR.
8. Using the SPSS output in Fig. 31.19, perform a hypothesis test to determine if the odds of wearing a hat is the same for females and males.
9. Write down the conclusion.
10. Is the test statistically valid?

TABLE 31.16: *The number of people wearing hats, for males and females*

	Not wearing hat	Wearing hat
Male	307	79
Female	344	22

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	33.763[a]	1	.000		
Continuity Correction[b]	32.531	1	.000		
Likelihood Ratio	35.712	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	33.718	1	.000		
N of Valid Cases	752				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 49.16.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval Lower	95% Confidence Interval Upper
Odds Ratio for Hat (No / Yes)	.249	.151	.408
For cohort Gender = Male	.603	.529	.687
For cohort Gender = Female	2.426	1.665	3.535
N of Valid Cases	752		

FIGURE 31.19: *SPSS output for the hats data*

Exercise 31.6. A study (Lennon et al. 2017) asked people about their mobile-phone interactions while crossing the road as pedestrians. Part of the data are summarised in Table 31.17.

1. Compute the column percentages.
2. Compute the odds of low exposure to each behaviour.
3. Write the hypothesis for conducting a hypothesis test.
4. Compute the *expected counts*.
5. After analysis in jamovi, the value of χ^2 is 20.923 with two degrees of freedom. What is the approximately-equivalent z-score? Would you expect a large or small P-value?
6. The P-value is given as $P < 0.000$. Write a conclusion.

TABLE 31.17: *Mobile-phone behaviour of pedestrians. ('Low exposure' means the behaviour was displayed less than once per week; 'High exposure' means the behaviour was displayed one per week or more.)*

	Answer call	Respond to text	Reply to email
Low exposure	263	259	302
High exposure	94	98	51

32

Selecting a hypothesis testing

Selecting the correct hypothesis test (or confidence interval) can be tricky… and in this book only a small number of hypothesis tests were described. (Literally hundreds of tests exist (Kanji 2006).)

For the tests studied in this book, determining if the response and explanatory *variables* are qualitative or quantitative is important (Table 32.1). So far, only situations with a *qualitative* explanatory variable have been considered. In the next chapters, cases where both the response and explanatory variables are *quantitative* are studied.

TABLE 32.1: *Four different scenarios studied so far*

Graphical summary	Numerical summary	Hypothesis test	Confidence interval
Mean of one sample			
Histogram; stem-and-leaf plot; dot chart	Means, medians; Std. dev., IQR; etc.	One-sample t	CI for one mean
Mean of differences (paired data)			
Histogram of differences; case-profile	Mean, std. dev. etc. of differences	t-test for mean differences	CI for mean difference
Comparing odds/percentages in two groups			
Error bar chart	Mean and std. error of the difference; mean, std. dev. etc. of each group	t-test comparing the difference between two means	CI of the difference between means
Comparing means in two groups			
Side-by-side bar chart; stacked bar chart	Odds; OR; percentages	Chi-square test	CI for OR

Part VIII

Connection RQs: Regression and Correlation

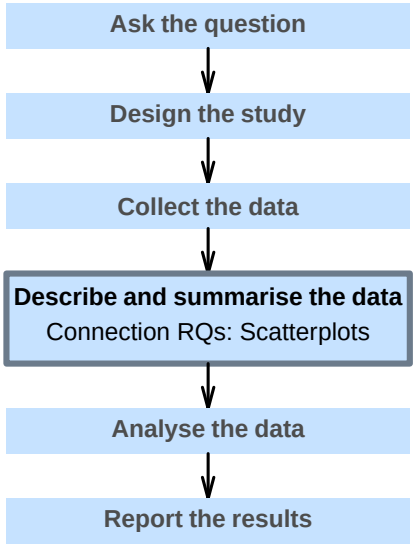
33

Relationships between two quantitative variables

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, understand the tools of inference, to form *confidence intervals*, and to perform *hypothesis tests*.

In this chapter, you will learn about relationships between two quantitative variables. You will learn to:

- describe the relationships between two quantitative variables.



33.1 Introduction: The red deer data

Consider a study (Holgate 1965) that examined the relationship between the age of $n = 78$ male red deer and the weight of their molars (Table 33.1).

33.2 Two quantitative variables: Graphical summaries

For the red deer data, both variables are *quantitative*, so the appropriate graphical summary (Sect. 12.5) is a *scatterplot* (Fig. 33.1).

TABLE 33.1: *The first 10 observations of the male red deer data*

Age (in years)	Molar weight (in g)
4.4	2.42
4.4	4.45
4.4	5.24
4.4	3.19
4.4	3.90
4.4	3.26
4.4	3.07
4.8	4.48
4.8	3.18
5.4	3.36

In the graph, the *response* variable is graphed on the *vertical* axis, and denoted y; the *explanatory* variable is graphed on the *horizontal* axis, and denoted x. The *explanatory* variable (potentially) *influences* the response variable, so in this example:

- The *explanatory variable* (x) is the age of the deer (in years), and
- The *response variable* (y) is the weight of molars (in grams).

In other words, the age of the deer would seem likely to influence the weight of the molars. (In some cases, it doesn't matter which is x and which is y, such as exploring the relationship between height and weight of red deer.)

Each row in the data set (and each point on the scatterplot) correspond to a single deer (that is, the individual deer are the units of analysis), and two different variables (age; molar weight) are measured on each deer.

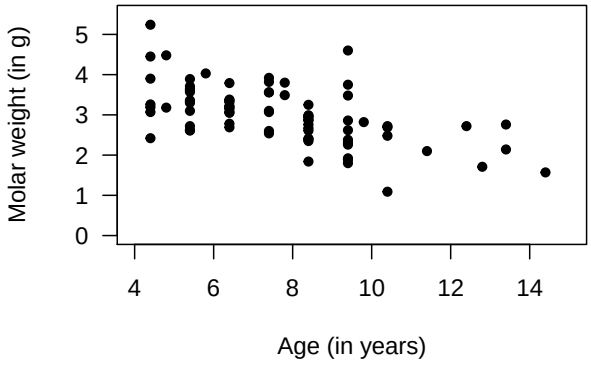


FIGURE 33.1: *Molar weight verses age for the red deer data*

33.3 Understanding scatterplots

The purpose of a graph is to help us *understand* the data (Sect. 12.1). To understand the data displayed in a scatterplot, the *form*, *direction*, and *variation* (or the *strength*) are described:

1. *Form*: Identify the overall *form* or structure of the relationship (e.g., linear; curved upwards; etc.).
2. *Direction*: Identify the *direction* of the relationship (sometimes not relevant if the relationship is non-linear):
   - The variables are *positively* associated if high values of one variable accompany *high* values of the other variable, in general.
   - The variables are *negatively* associated if high values of one variable accompany *low* values of the other variable, in general.
3. *Variation*: The amount of *variation* in the relationship. A small amount of variation in the response variable for given values of the explanatory variable means the relationship is strong; a lot of variation in the response variable for given values of the explanatory variable means the relationship is less strong.

Anything unusual or noteworthy should also be discussed. These three features help us understand the type of relationship (*form* and *direction*), and the strength of that relationship (*variation*).

To demonstrate the use of these descriptions, see the example scatterplots in Fig. 33.2. (The online version has extra examples.)

Example 33.1 (Describing scatterplots)**.** A study (Tager et al. 1979; Kahn 2005) examined the lung capacity of children in Boston (measured using the forced expiratory volume (FEV)). The scatterplot (Fig. 33.3) could be described as curved (*form*), where older children have larger FEVs, in general (*direction*). The *variation* gets larger for taller youth.

Think 33.1 (Scatterplots)**.** *Describe the scatterplot of* diastolic *BP against age (Fig. 33.4), from the NHANES data.*

Answer: *Form*: curved. *Direction*: not relevant (up, then down). *Variation*: large.

Example 33.2 (Scatterplots)**.** For the red deer data (Fig. 33.1), the scatterplot could be described as approximately linear (form), with a negative direction (*older* deer generally have *less heavy* teeth); the *variation* is... perhaps moderate.

33.4 Summary

A **scatterplot** is used to show the relationship between two quantitative variables (the response denoted y; the explanatory denoted x). The relationship can be described by the **form** (linear, or otherwise), the **direction** of the relationship (sometimes not relevant if the graph is not linear), and the **variation** in the relationship (or the **strength** of the relationship).

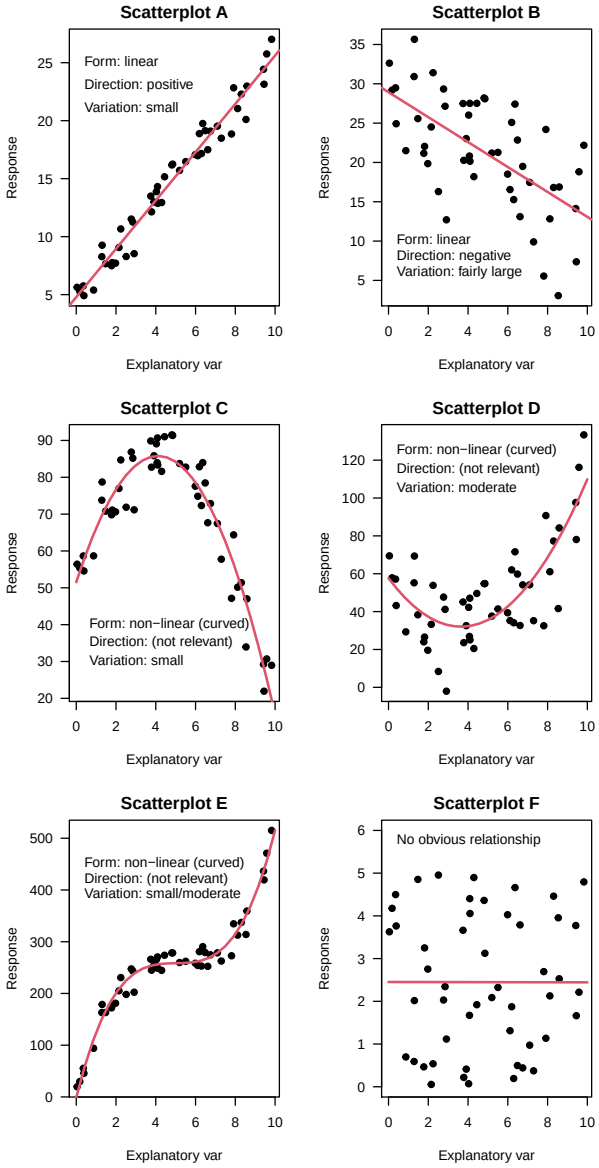


FIGURE 33.2: *Some example scatterplots*

33.5 Quick review questions

A study of onion growth (Mead 1970) produced the scatterplot shown in Fig. 33.5.

1. The x-variable is
2. The *form* is best described as
3. The *direction* is best described as
4. The *variation* is best described as

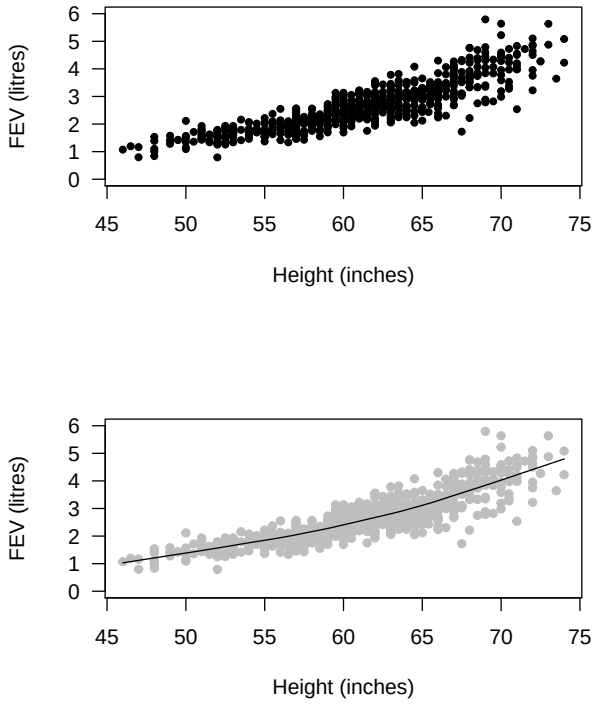


FIGURE 33.3: *FEV plotted against height for children in Boston*

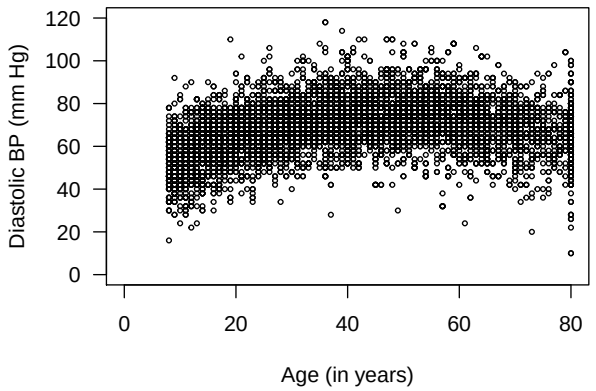


FIGURE 33.4: *Diastolic blood pressure plotted against age for the NHANES data*

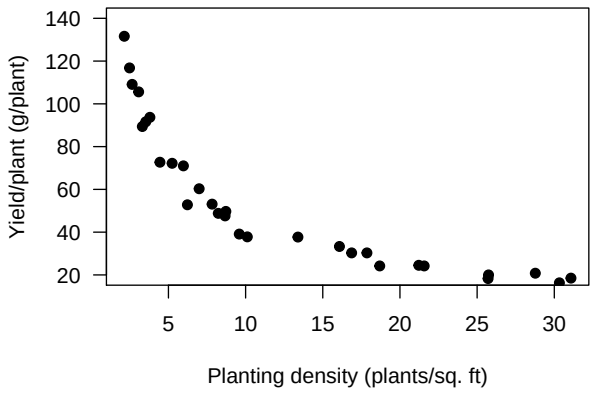


FIGURE 33.5: *Onion yield plotted against planting density*

33.6 Exercises

Selected answers are available in Sect. D.30.

Exercise 33.1. A study evaluated various food mixtures for sheep (Moir 1961). Describe the scatterplot (Fig. 33.6) in terms of the *form* of the relationship, the *direction* of the relationship (if relevant), and the *variation* in the relationship.

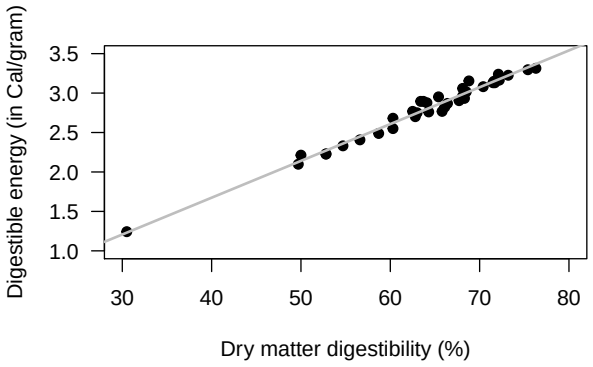


FIGURE 33.6: *Scatterplots for the sheep-food data*

Exercise 33.2. A study examined the direct current generated by a windmill and its association with wind speed (Joglekar et al. 1989; Hand et al. 1996). Describe the relationship (Fig. 33.7).

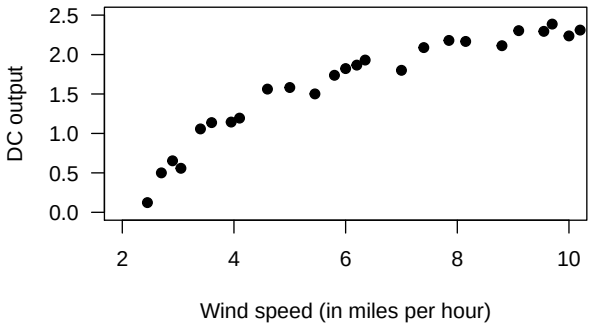


FIGURE 33.7: *The relationship between DC output and wind speed*

Exercise 33.3. A study examined the mandible length and gestational age for 167 foetuses from the 12th week of gestation onward (Royston and Altman 1994). How would you describe the relationship (Fig. 33.8)?

Exercise 33.4. A study examined the time taken to deliver soft drinks to vending machines (Montgomery and Peck 1992). How would you describe the relationship (Fig. 33.9)?

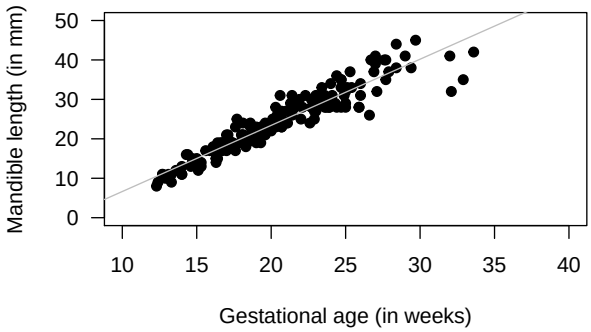


FIGURE 33.8: *The relationship between gestational age and mandible length*

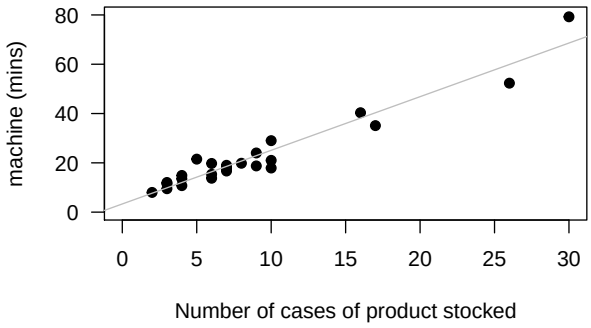


FIGURE 33.9: *The time taken to deliver soft drinks to vending machines*

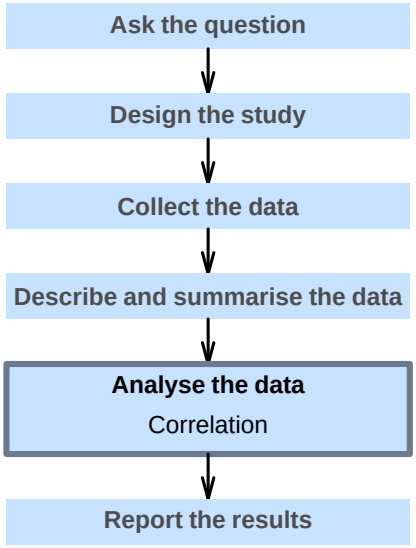
34

Correlation

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, understand the tools of inference, to form *confidence intervals*, and to perform *hypothesis tests*.

In this chapter, you will learn about *correlation*. You will learn to:

- produce correlation coefficients for exploring the relationship between two quantitative variables.
- produce and interpret R^2.
- conduct hypothesis tests for correlation coefficients.



34.1 Correlation coefficients

Describing the *linear* relationship between two *quantitative* variables, requires a description of the form, direction and variation. A *correlation coefficient* is a single number encapsulating all this information.

In the *population*, the unknown value of the correlation coefficient is denoted ρ ('rho'); in the *sample* the value of the correlation coefficient is denoted r. As usual, r (the *statistic*) is an estimate of ρ (the *parameter*), and the value of r is likely to be different in every sample (that is, *sampling variation* exists).



The symbol ρ is the Greek letter 'rho,' pronounced 'row,' as in 'row your boat'[1].

Correlation coefficients only apply if the form is approximately *linear*, so checking if the relationship is linear first (using a scatterplot) is important. Here, the *Pearson* correlation coefficient is discussed, which is suitable for describing linear relationships between quantitative data[2].



The Pearson correlation coefficient only make sense if the relationship is approximately linear.

The values of ρ and r are *always* between -1 and $+1$. The *sign* indicates whether the relationship has a positive or negative linear association, and the *value* of the correlation coefficient tells us the strength of the relationship:

- $r = 0$ means *no linear relationship* between the two variables: Knowing how the value of x changes tells us nothing about how the value of y changes.
- $r = +1$ means a *perfect, positive* relationship: knowing the value of x means we can perfectly predict the value of y (and *larger* values of y are associated with *larger* values of x, in general).
- $r = -1$ means a *perfect, negative* relationship: knowing the value of x means we can perfectly predict the value of y (and *larger* values of y are associated with *smaller* values of x, in general).

Numerous example scatterplots were shown in Sect. 33.3; a correlation coefficient is not relevant for Plots C, D or E, as those relationships are not linear. In Plot A, the correlation coefficient will be *positive*, and reasonably close to one. In Plot B, the correlation coefficient will be *negative*, but not that close to -1. In Plot F, the correlation coefficient will close to zero.

Example 34.1 (Correlation coefficients)**.** For the red deer data (Fig. 33.1), $r = -0.584$. The value of r is *negative*, because, in general, *older* deer (x) are associated with *smaller* weight molars (y).

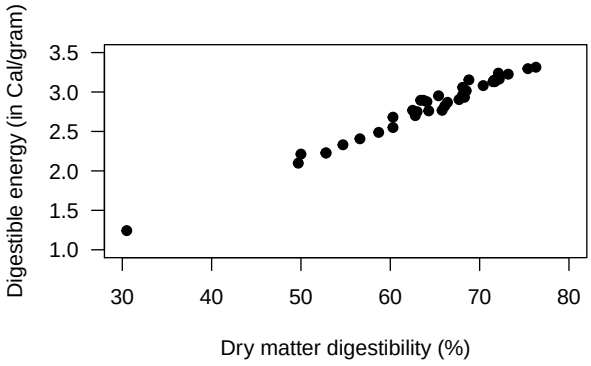


FIGURE 34.1: *Scatterplot for the sheep-food data*

[2]Other types of correlation coefficients also exist, such as the *Spearman* correlation, which may be used for monotonic, non-linear relationships.

Example 34.2 (Correlation coefficients)**.** Consider the plot in Fig. 34.2 from the NHANES data. This scatterplot of *diastolic* BP against age is not linear, so a correlation coefficient is *not appropriate*.

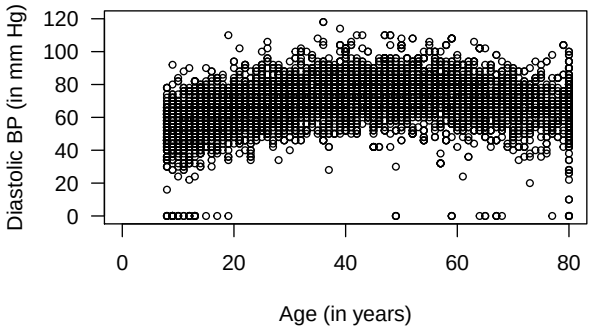


FIGURE 34.2: *A scatterplot of the diastolic blood pressure against age for the NHANES data*

Example 34.3 (Correlation coefficients)**.** Consider the plot in Fig. 34.3 from the NHANES data. This scatterplot of *systolic* BP against age is approximately linear, so a correlation coefficient is *appropriate*. The correlation coefficient is $r = 0.532$.

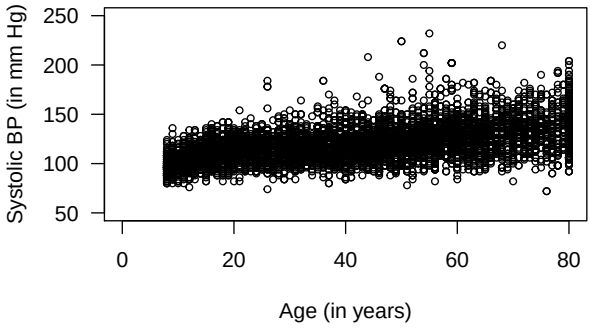


FIGURE 34.3: *A scatterplot of the systolic blood pressure against age for the NHANES data*

Think 34.1 (Estimate r)**.** *A study evaluated various food mixtures for sheep (Moir 1961). One combination of variables that was assessed is shown in Fig. 34.1.*

Estimate the value of r.

Answer: The answer is given in the online book.

Think 34.2 (Guess the value of r)**.** *Earlier, we looked at the NHANES data to explore the relationship between direct HDL cholesterol and current smoking status. The NHANES project is an observational study, so confounding is a potential issue. For this reason, relationships between the response and extraneous variables, and between explanatory and extraneous variables, should be examined.*

For example, the relationship between Age (an extraneous variables) and direct HDL cholesterol (the response variable) is shown in Fig. 34.4.

How would you describe the relationship? What do you guess for the value of r?

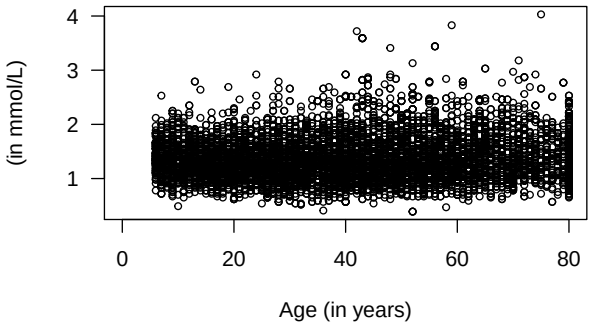


FIGURE 34.4: *Direct HDL cholesterol plotted against age for the NHANES data*

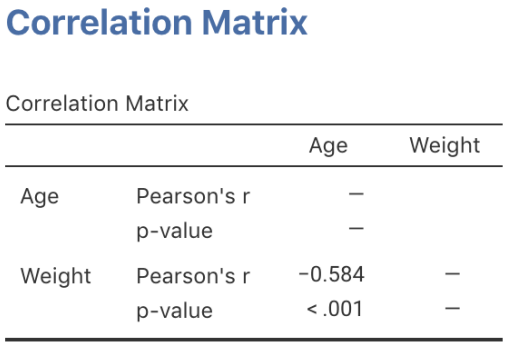
Answer: The answer is given in the online book.

The web page http://guessthecorrelation.com makes a game out of trying to guess the correlation coefficient!

34.2 Using software

Software is used to compute the value of r. For the red deer data (Fig. 33.1), the relationship is approximately linear, and the jamovi output (Fig. 34.5) and SPSS output (Fig. 34.6) show that $r = -0.584$:

- Direction: The *sign* of r indicates the direction. Here we see a *negative* relationship: Higher ages are associated with lighter molars (in general), which makes sense.
- Variation: The *value* of r indicates the strength of the relationship. Here, perhaps we could describe the variation as moderate.



Correlation Matrix

Correlation Matrix

		Age	Weight
Age	Pearson's r	—	
	p-value	—	
Weight	Pearson's r	-0.584	—
	p-value	< .001	—

FIGURE 34.5: *jamovi correlation output for the red deer data*

Correlations

		Age (in years)	Molar weight (in grams)
Age (in years)	Pearson Correlation	1	-.584**
	Sig. (2-tailed)		.000
	N	78	78
Molar weight (in grams)	Pearson Correlation	-.584**	1
	Sig. (2-tailed)	.000	
	N	78	78

**. Correlation is significant at the 0.01 level (2-tailed).

FIGURE 34.6: *SPSS correlation output for the red deer data*

34.3 R-squared (R^2)

While using r tells us about the strength and direction of the linear relationship, knowing exactly what the value *means* is tricky. Interpretation is easier using R^2, or 'R-squared': the square of the value of r.



The value of R^2 is *never* negative, and is usually expressed as a percentage.



The value of R^2 is never negative. However, you need to be careful when using your calculator!

With most calculators, entering `-0.5^2` will return `-0.25`. This is correct, because the calculator interprets your input as meaning `-(0.25^2)`.

You need to enter `(-0.5)^2`. This will give you the expected answer of `0.25`.

The value of R^2 is the percentage reduction in the unknown variation of y because the value of x is known. In other words, it is the percentage of the variation in y explained by using the linear relationship, rather than just the mean value of y.

Example 34.4 (Values of R^2)**.** For the red deer data (Fig. 33.1), the value of R^2 from the software output (Fig. 34.5; Fig. 34.6) is $R^2 = (-0.584)^2 = 0.341$, usually written as a percentage: 34.1%.

The value of R^2 is positive, even though the value of r is negative.

For the red deer data, R^2 means that about 34.1% of the variation in molar weights can be explained by variation in the age of the deer. The rest of the variation in molar weights is due to extraneous variables, such as weight, diet, amount of exercise, genetics, etc.

Think 34.3 (Interpreting R^2)**.** *From Example 34.3, the correlation coefficient between the systolic blood pressure and age in the NHANES data is* $r = 0.532$.

What is the value of R^2*? What does it mean?*

Answer: $R^2 = (0.532)^2 = 0.283$: about 28.3% of the variation in systolic BP is due to age; extraneous variable (weight, gender, amount of exercise, genetics, etc.) explain the remaining 71.7% of the variation in SBP values.

34.4 Hypothesis testing

34.4.1 Introduction

For the red deer data (Sect. 33.2; Sect. 34.1), the population correlation coefficient between the weight of molars y and age of the deer x is unknown and denoted by ρ. The sample correlation coefficient is $r = -0.584$, but the value of r varies from sample to sample (there is *sampling variation*).

The size of the sampling variation is measured with a *standard error*. However, there is a complication for correlation coefficients[3]; so we will not produce CIs for the correlation coefficient.

34.4.2 Hypothesis testing details

As usual, questions can be asked about the relationship between the variables, as measured by the unknown *population* correlation coefficient:

Is the *population* correlation coefficient zero, or not?

In the context of the red deer data:

[3] For those who want to know: the value of r only varies between -1 and 1, so the sampling distribution is not a normal distribution. Instead, a transformation of the correlation coefficient has an approximate normal distribution and *standard error*. Software automatically does this transforming.

In male red deer, is there a correlation between age and the weight of molars?

The RQ is about the population parameter ρ. Clearly, the *sample* correlation coefficient r is not zero, and the RQ is asking whether this could be attributed to sampling variation. The null hypotheses is:

- H_0: $\rho = 0$

The parameter is ρ, the population correlation between the age and molar weight in the red deer.

This is the usual 'no relationship' position, which proposes that the population correlation coefficient is zero. The alternative hypothesis is:

- H_1: $\rho \neq 0$

This is a *two-tailed* test here, based on the RQ.

The approach is to **assume** that $\rho = 0$ (from H_0), then describe what values of r could be **expected**, under that assumption, just through sampling variation. Then the **observed** value of r is compared to the expected values to determine if the valuew of r supports or contradicts the assumption.

Software is used to test the hypotheses; the output in Figs. 34.5 (jamovi) and 34.6 (SPSS) contains the relevant P-value (twice in the SPSS output!). The two-tailed P-value for the test (labelled `Sig.` by SPSS) is less than 0.001 (`0.000` in SPSS). That is, the P-value is zero *to three decimal places*, so there is *very strong evidence* to support H_1 (that the correlation in the population is not zero). We write:

The sample presents very strong evidence (two-tailed $P < 0.001$) of a correlation between molar weight and the age of the male red deer ($r = -0.584$; $n = 78$) in the population.

Notice the three features of writing conclusions again: An *answer to the RQ*; evidence to support the conclusion ('two-tailed $P < 0.001$'; no test statistic is given); and some *sample summary information* ('$r = -0.584$; $n = 78$').

The evidence suggests that the correlation is not zero (in the population). However, a *non-zero* correlation doesn't necessarily mean a *strong* correlation.

The correlation may be weak in the population (as estimated by the value of r), but there is evidence that it is not zero in the *population*.

This may be a useful analogy: If a rain forecast says 'there is a very high chance of rain tomorrow,' it doesn't mean there will be a *lot* of rain, just a high chance of *some* rain.

34.4.3 Statistical validity conditions

As usual, these results hold under certain conditions to be met. The conditions for which the test is statistically valid are:

1. The relationship is approximately linear.
2. The variation in the response variable is approximately constant for all values of the explanatory variable.
3. The sample size is at least 25.

The sample size of 25 is a rough figure here, and some books give other values.

In addition to the statistical validity condition, the test will be **externally valid** if the the sample is a simple random sample from the population. The test will also be **internally valid** if the study was well designed.

Example 34.5 (Statistical validity)**.** For the red deer data, the scatterplot (Fig. 33.1) shows that the relationship is approximately linear, and the variation in molar weights doesn't seem to be obviously getting larger or smaller for older deer, so correlations are sensible. The sample size is also greater than 25.

The test in Sect. 34.4 will be statistically valid.

Example 34.6 (Statistical validity)**.** A study (Schepaschenko et al. 2017; Dunn and Smyth 2018) examined the foliage biomass of small-leaved lime trees. A plot of the foliage biomass against diameter (Fig. 34.7) shows that the relationship is non-linear. In addition, the variation in foliage biomass *increases* for larger diameters (for values of x near 10, the values of y do not vary much at all, but for values of x near 30, the values of y vary greatly).

Both of these issues mean that correlations are not appropriate. A hypothesis test similar to that in Sect. 34.4 is inappropriate.

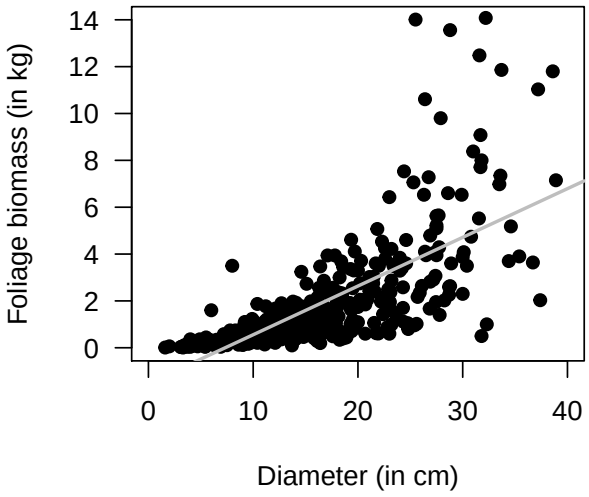


FIGURE 34.7: *Foliage biomass plotted against diameter for small-leaved lime trees*

Example 34.7 (Phu Quoc ridgeback dogs)**.** A study of Phu Quoc Ridgeback dogs (*Canis familiaris*) recorded many measurements of the dogs, including body length and body height (Quan et al. 2017).

The scatterplot (Fig. 34.8) shows an approximate linear relationship. We know that each sample

could produce a different sample correlation coefficient. We expect that taller dogs would also be longer, so we may ask:

> For these dogs, are longer dogs also taller dogs, in general?

The hypotheses are:

- H_0: $\rho = 0$
- H_1: $\rho > 0$ (i.e., one-tailed)

The correlation co-efficient is $r = 0.837$ and software notes that the two-tailed $P < 0.001$, based on $n = 30$ dogs.

We write:

> There is very strong evidence that longer Phu Quoc ridgeback dogs are also taller ($r = 0.837$; one-tailed $P < 0.001$; $n = 30$).

Since (a) the sample size is larger than 25; (b) the relationship is approximately linear; and (c) the variation in heights do not seem to differ for different lengths, the test is statistically valid.

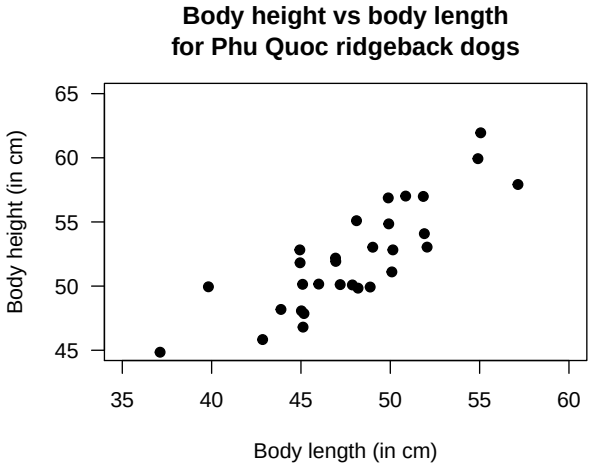


FIGURE 34.8: *Scatterplot of the body height vs body length for Phu Quoc ridgeback dogs*

Example 34.8 (Drug calculations)**.** A study of $n = 30$ paramedicine students examined (among other things) the relationship between the amount of stress experienced (measured using the State–Trait Anxiety Inventory (STAI) while performing drug-dose calculation, and length of work experience (LeBlanc et al. 2005).

The hypotheses are:

- H_0: $\rho = 0$
- H_1: $\rho \neq 0$

The correlation co-efficient is given as $r = 0.346$ and $P = 0.18$.

No scatterplot is provided, so the test is statistically valid only if the relationship is approximately

linear and that the variation in STAI scores does not vary for different levels of work experience. The sample size is larger than 25, however.

We write:

> There is no evidence ($r = 0.346$; two-tailed $P = 0.18$) that the length of work experience is associated with STAI stress levels when performing drug-dose calculations.

34.5 Example: Removal efficiency

In wastewater treatment facilities, air from biofiltration is passed through a membrane and dissolved in water, and is transformed into harmless byproducts. The removal efficiency y (in %) may depend on the inlet temperature (in °C; x).

The RQ is

> In treating biofiltation wastewater, is the removal efficiency associated with the inlet temperature?

The population parameter is ρ, the correlation between the removal efficiency and inlet temperature.

A scatterplot of $n = 32$ samples (Fig. 34.9) suggests an approximately linear relationship (Chitwood and Devinny 2001; Devore and Berk 2007). The output (jamovi: Fig. 34.10; SPSS: Fig. 34.11) shows that the sample correlation coefficient is $r = 0.891$, and so $R^2 = (0.891)^2 = 79.4\%$. This means that about 79.4% of the variation in removal efficiency can be explained by knowing the inlet temperature.

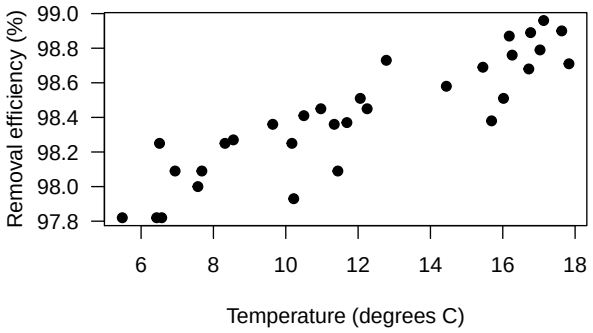


FIGURE 34.9: *The relationship between removal efficiency and inlet temperature*

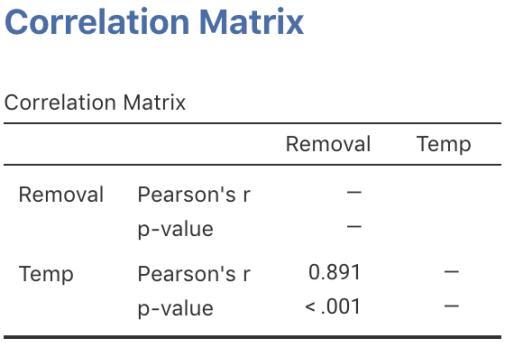
To test if a relationship exists in the population, write:

- H_0: $\rho = 0$;
- H_1: $\rho \neq 0$: Two-tailed (as implied by the RQ).

The software output (jamovi: Fig. 34.10; SPSS: Fig. 34.11) shows that $P < 0.001$ (which is what $P = 0.000$ in SPSS means). We conclude:

> The sample presents very strong evidence (two-tailed $P < 0.001$) that removal efficiency depends on the inlet temperature ($r = 0.891$; $n = 32$) in the population.

The relationship is approximately linear and there is no obvious non-constant variance, and the sample size is larger than 25, so the hypothesis test results will be statistically valid.



Correlation Matrix

Correlation Matrix

		Removal	Temp
Removal	Pearson's r	—	
	p-value	—	
Temp	Pearson's r	0.891	—
	p-value	< .001	—

FIGURE 34.10: *jamovi output for the removal-efficiency data*

Correlations

		Removal efficiency (in %)	Inlet temperature (in deg C)
Removal efficiency (in %)	Pearson Correlation	1	.891**
	Sig. (2-tailed)		.000
	N	32	32
Inlet temperature (in deg C)	Pearson Correlation	.891**	1
	Sig. (2-tailed)	.000	
	N	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

FIGURE 34.11: *SPSS output for the removal-efficiency data*

34.6 Summary

In this chapter, **correlation** was used to describe the *strength* and direction of *linear* relationships between two *quantitative* variables.

Correlation coefficients (denoted r in the sample; ρ in the population) are always between -1 and $+1$.

Positive values denote *positive* relationships between the two variables: as one values gets larger, the other tends to get larger too.

Negative values denote *negative* relationships between the two variables: as one values gets larger, the other tends to get *smaller*. Values close to -1 or $+1$ are very strong relationships;

values near zero shows very little linear relationship between the variables. Hypothesis tests for r can be conducted using software.

Sometimes, R^2 is used to describe the relationship: it indicates what percentage of the variation in the response variable can be explained by knowing the value of the explanatory variables.

34.7 Quick review questions

A study of Chinese paedatric patients (Wong et al. 2018) studied the relationship between the 6-minute walk distance (6MWD) and maximum oxygen uptake ($VO2_{max}$) for $n = 29$ patients. The correlation coefficient is reported as $r = 0.457$, and the corresponding P-value as $P = 0.013$.

1. The x-variable is
2. True or false: Since the P-value is small, the correlation must be quite strong.
3. The relationship is best described as
4. The value of R^2 (to one decimal place, expressed as a percentage) is:
5. For statistical validity, we need to *assume* that:

34.8 Exercises

Selected answers are available in Sect. D.31.

Exercise 34.1. Draw a scatterplot with:

1. A negative correlation coefficient, with the value of r very close to (but not equal to) -1.
2. A positive correlation coefficient, with the value of r very close to (but not equal to) $+1$.
3. A correlation coefficient very close to 0.

Exercise 34.2. A study (Myers (1990), p. 75) of American footballers measured the right-leg strengths x of 13 players (using a weight lifting test), and the distance y they punted a football (with their right leg) (Fig. 34.12).

1. The value of the correlation coefficient is 0.881. Compute the value of R^2, and explain what this means.
2. jamovi was used to study the correlation (Fig. 34.13). Using this output, perform a hypothesis test to determine if a correlation exists between punting distance and right-leg strength.

Exercise 34.3. A study examined the time taken to deliver soft drinks to vending machines

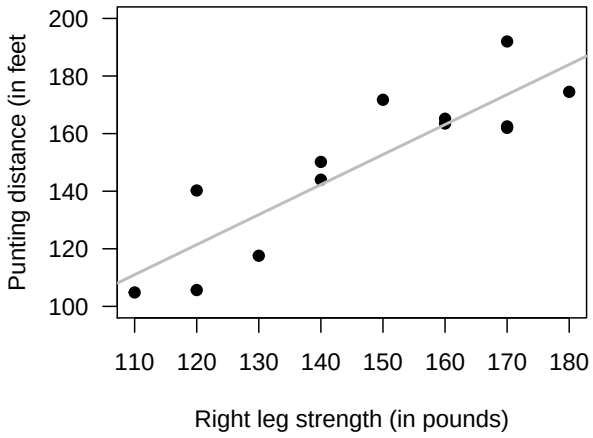


FIGURE 34.12: *Punting distance and right leg strength*

Correlation Matrix

Correlation Matrix

		Punt	Right
Punt	Pearson's r	—	
	p-value	—	
Right	Pearson's r	0.881	—
	p-value	< .001	—

FIGURE 34.13: *jamovi output for the punting data*

(Montgomery and Peck 1992) using a sample of size $n = 25$ (Fig. 34.14). To perform a test of the correlation coefficient, are the statistical validity conditions met?

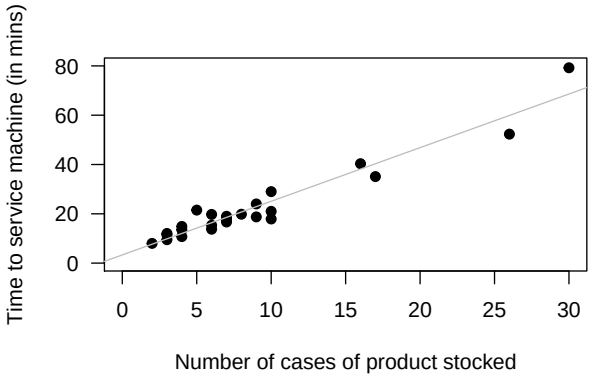


FIGURE 34.14: *The time taken to deliver soft drinks to vending machines*

Exercise 34.4. A study of hot mix asphalt (Panda et al. 2018) created $n = 42$ samples of asphalt and measured the volume of air voids and the bitumen content by weight (Fig. 34.15).

1. Using the plot, estimate the value of r.
2. The value of R^2 is 99.29%. What is the value of r? (Hint: Be careful!)
3. Would you expect the P-value testing H_0: $\rho = 0$ to be small or large? Explain.
4. Would the test be statistically valid?

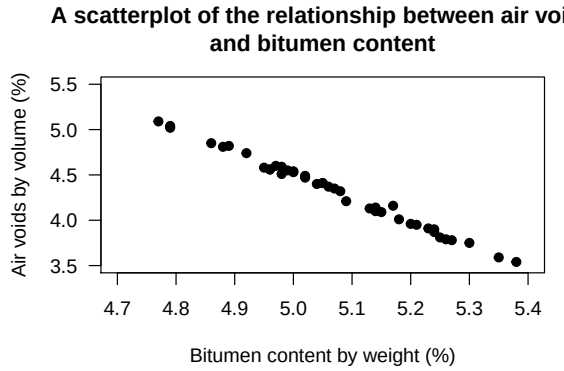


FIGURE 34.15: *Air voids in bitumen samples*

Exercise 34.5. The *California Bearing Ratio* (CBR) value is used to describe soil-sub grade for flexible pavements (such as in the design of air field runways).

One study (Talukdar 2014) examined the relationship between CBR and other properties of soil, including the plasticity index (PI, a measure of the plasticity of the soil).

The scatterplot from 16 different soil samples from Assam, India, is shown in Fig. 34.16.

1. Using the plot, estimate the value of r.
2. The value of R^2 is 67.07%. What is the value of r? (Hint: Be careful!)
3. Would you expect the P-value testing H_0: $\rho = 0$ to be small or large? Explain.
4. Would the test be statistically valid?

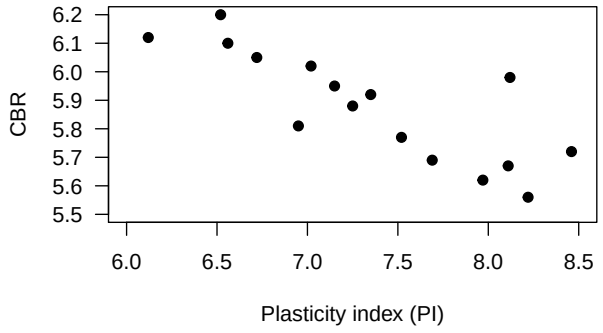


FIGURE 34.16: *The relationship between CBR and PI in sixteen soil samples*

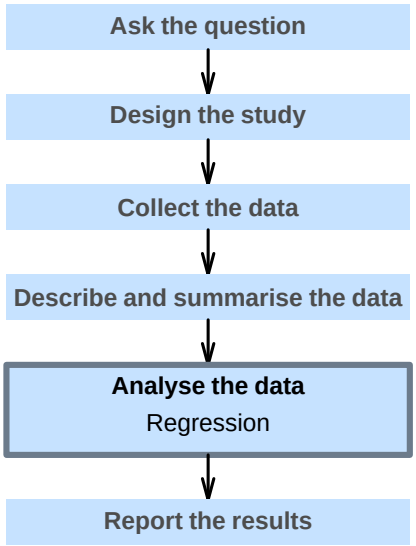
35

Regression

So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, understand the tools of inference, to form *confidence intervals*, and to perform *hypothesis tests*.

In this chapter, you will learn about *regression*. You will learn to:

- produce and interpret linear regression equations.
- conduct hypothesis tests for the slope in a regression line.
- produce confidence intervals for the slope in a regression line.



35.1 Introduction

In the last chapter, *correlation* was studied, which measures the *strength* of the *linear* relationship between two quantitative variables x and y. We now study **regression**, which describes *what* the linear relationship is between x and y.

The relationship is described using an *equation*, which allows us to:

1. **Predict** values of y from given values of x (Sect. 35.4); and
2. **Understand** the relationship between x and y (Sect. 35.5).

35.2 Linear equations: A review

An *example* of a regression equation is

$$\hat{y} = -4 + 2x.$$

Here, x refers to the explanatory variable, y refers to the *observed* response variable, and $\hat{y}$ refers to the *predicted* values of the response variable.

In general, the equation of a straight line is written as

$$\hat{y} = b_0 + b_1 x$$

where b_0 and b_1 are just numbers. Again, $\hat{y}$ refers to the *predicted* (not observed) values of y.



$\hat{y}$ is pronounced as 'why hat'; the 'caret' above the y is called a 'hat,' and designates a predicted value (of y).

Example 35.1 (Regression equations)**.** In the regression equation $\hat{y} = 15 - 102x$, we have $b_0 = 15$ and $b_1 = -102$.

Think 35.1 (Regression coefficients)**.** *Consider the regression equation* $\hat{y} = -0.0047x + 2.1$. *What are the values of* b_0 *and* b_1*? (Look carefully!)*

Answer: The answer is given in the online book.

The numbers b_0 and b_1 are called *regression coefficients*, where

- b_0 is a number called the **intercept**. It is the *predicted* value of y when $x = 0$.
- b_1 is a number called the **slope**. It is, on average, how much the value of y changes when the value of x increases by 1.

We will use software to find the values of b_0 and b_1. However, we can roughly guess the values of the *intercept* by first drawing what looks like a sensible straight line through the data, and determining what that line predicts for the value of y when $x = 0$.

A rough guess of the *slope* can be made using the formula

$$\text{slope} = \frac{\text{Change in } y}{\text{Corresponding change in } x} = \frac{\text{rise}}{\text{run}}.$$

That is, a guess of the slope is the *change* in the value of y (the 'rise') divided by the corresponding *change* in the value of x (the 'run').

To demonstrate, consider the scatterplot in Fig. 35.1. I have drawn a sensible line on the graph to capture the relationship (your line may look a bit different). When $x = 0$, the regression line predicts the value of y is about to be 2, so b_0 is approximately 2.

To guess the slope, use the 'rise over run' idea; see Fig. 35.2 (the online version has an animation). When the value of x increases from 1 to 5 (a change of $5 - 1 = 4$), the corresponding values of y change from 5 to 17 (a change of $17 - 5 = 12$). Then, use the formula:

$$\begin{aligned}\frac{\text{rise}}{\text{run}} &= \frac{17-5}{5-1} \\ &= \frac{12}{4} = 3.\end{aligned}$$

The value of b_1 is about 3. The regression line is approximately $\hat{y} = 2 + (3 \times x)$, usually written as

$$\hat{y} = 2 + 3x.$$

The *intercept* has the same measurement units as the response variable. For example, with the red-deer data the intercept is measured in 'grams,' the measurement units of the molar weight.

The measurement unit for the *slope* is the 'measurement units of the response variable,' per 'measurement units of the explanatory variable.' For example, with the red-deer data the slope has the units of 'grams per year.'

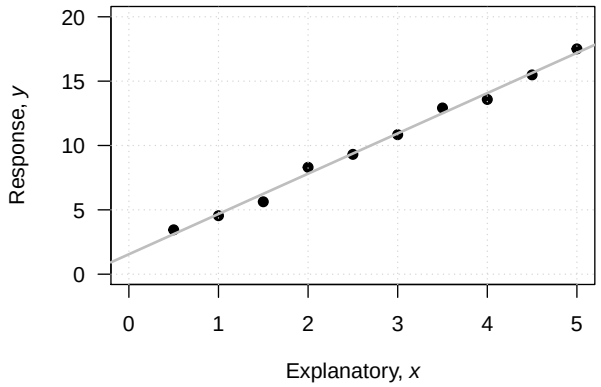


FIGURE 35.1: *An example scatterplot*

Example 35.2 (Estimating regression parameters)**.** A study (Dunn and Smyth 2018) examined the number of cyclones y in the Australian region each year from 1969 to 2005, and the relationship with a climatological index called the *Ocean Nino Index* (ONI, x)[1]; see (Fig. 35.3),

When the value of x is zero, the predicted value of y is about 12; b_0 is about 12. (You may get something slightly different.) Notice that the intercept is the predicted value of y when $x = 0$, which is *not* at the left of the graph.

To guess the value of b_1, use the 'rise over run' idea. When x is about -2, the predicted value of y is about 17. When x is about 2, the predicted value of y is about 8. So when the value of x changes by $2 - (-2) = 4$, the value of y changes by $8 - 17 = -9$ (a *decrease* of about 9). Hence, the value of b_1 is approximately $-9/4 = -2.25$. (You may get something slightly different.) Notice that the relationship has a *negative* direction, so the slope must be negative.

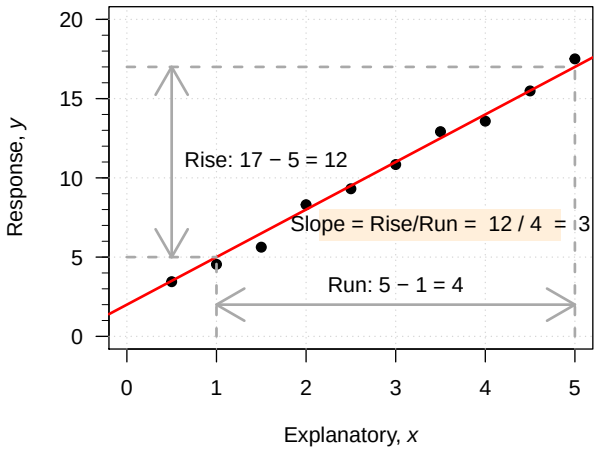


FIGURE 35.2: *Making a guess of the slope, using rise-over-run*

Using these guesses of $b_0 = 12$ and $b_1 = -2.25$, the regression line is approximately

$$\hat{y} = 12 - 2.25x.$$

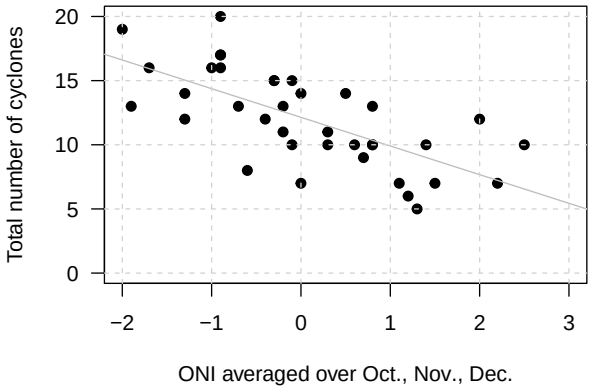


FIGURE 35.3: *The number of cyclones in the Australian region each year from 1969 to 2005, and the ONI for October, November, December*

In this section, we have seen how to understand a linear regression equation, and how an equation can be used to describe a fitted line. The above method gives a very crude guess of the values of the intercept b_0 and the slope b_1. In practice, *many* reasonable lines could be drawn through a scatterplot of data. However, one of those lines is the 'best fitting line' in some sense[2]. Software calculates this 'line of best fit' for us.

[2]For those who want to know: The 'line of best fit' is the line such that the sum of the *squared* vertical distances between the observations and the line is as small as possible.

35.3 Regression using software

In the population, the intercept is denoted by β_0 and the slope is denoted by β_1. These population values are unknown, and are estimated by the statistics b_0 and b_1 respectively.



The symbol β is the Greek letter 'beta,' pronounced 'beater' (as in 'egg beater').

The formulas for computing b_0 and b_1 are ugly, so we will use software to do the calculations. As usual, the values of these population parameters are unknown, and the values of the sample statistics will change from sample to sample (so they have *sampling variation*).

For the red deer data again (Fig 33.1), part of the relevant output is shown in Fig. 35.4 (using jamovi) and Fig. 35.5 (using SPSS). From the output, the *slope* b_1 in the sample is $b_1 = -0.181$, and the y-intercept b_0 in the sample is $b_0 = 4.398$. That is, the values of b_0 and b_1 are in the column labelled `Estimate` in jamovi, or the column labelled `B` in SPSS. These are the values of the two *regression coefficients*; then

$$\hat{y} = 4.398 + (-0.181 \times x),$$

which is usually written more simply as

$$\hat{y} = 4.398 - 0.181x.$$



The *sign* of the slope b_1 and the correlation coefficient r are always the same. For example, if the slope is negative, the correlation coefficient will also be negative.

Model Coefficients - Weight

Predictor	Estimate	SE	t	p
Intercept	4.398	0.2312	19.02	< .001
Age	−0.181	0.0289	−6.27	< .001

FIGURE 35.4: *jamovi output for the red-deer data*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	4.398	.231		19.020	.000
	Age (in years)	-.181	.029	-.584	-6.275	.000

a. Dependent Variable: Molar weight (in grams)

FIGURE 35.5: *SPSS output for the red-deer data*

Example 35.3 (Regression coefficients)**.** The regression equation for the cyclone data (Fig. 35.3) can be found from the jamovi output (Fig. 35.6):

$$\hat{y} = 12.14 - 2.23x,$$

where x is the ONI (averaged over October, November, December) and y is the number of cyclones. These values are close the guesses made in Example 35.2.

Model Coefficients - Total

Predictor	Estimate	SE	t	p
Intercept	12.14	0.452	26.85	< .001
OND	−2.23	0.404	−5.52	< .001

FIGURE 35.6: *jamovi output for the cyclone data*

35.4 Regression for predictions

The regression equation for the red deer data

$$\hat{y} = 4.398 - 0.181x$$

can be used to make *predictions*. For example, we could predict the *average* molar weight for deer 10 years old. Since x represents the age, use $x = 10$ in the regression equation:

$$\begin{aligned} \hat{y} &= 4.398 - (0.181 \times 10) \\ &= 4.398 - 1.81 \\ &= 2.588. \end{aligned}$$

Male red deer aged 10 years old are predicted to have a *mean* molar weight of 2.588 grams. Some individual male red deer aged 10 will have molars weighing *more* than this, and some will have molars weighing *less* than this. The model predicts that the *mean* molar weight for male red deer aged 10 will be about 2.588 grams.

Think 35.2 (Predicting)**.** *For male red deer* 12 *years of age, what is the predicted mean molar weight?*

Think 35.3 (Predicting)**.** *For male red deer* 20 *years of age, what is the predicted mean molar weight?*

Answer: Prediction: $4.398 - (0.181 \times 20) = 0.778$, or about 0.78 grams.

This last prediction *may* be a useful prediction... but it also may be rubbish. The oldest deer in the data is aged 14.4 years, so the regression line may not even apply for deer aged over 14.4 years of age (red deer may not even live to 20 years of age). The prediction *may* be sensible... but it *may not* be either. We don't know whether the prediction is sensible or not, because we have no data for deer aged over 14.4 years to inform us. Making prediction outside the range of the available data is called *extrapolation*, and *extrapolation* beyond the data can lead to nonsense predictions.

Definition 35.1 (Extrapolation)**.** *Extrapolation* refers to making predictions outside the range of the available data. Extrapolation beyond the data can lead to nonsense predictions.



Extrapolating can lead to nonsense predictions.

35.5 Regression for understanding

The regression equation can be used to *understand* the relationship between the two variables. Consider again the red deer regression equation:

$$\hat{y} = 4.398 - 0.181x. \tag{35.1}$$

What does it tell us about the relationship between x and y?

35.5.1 The meaning of b_0

b_0 is the *predicted* value of y when $x = 0$. Equation (35.1) predicts a molar weight of 4.398 for a deer zero years of age, which is likely to be nonsense: it is *extrapolating* beyond the data (the youngest deer in the sample is aged 4.4 years).



The value of the intercept b_0 is sometimes meaningful, but is often meaningless. The value of the slope b_1 is usually of greater interest, as it explains the *relationship* between the two variables.

35.5.2 The meaning of b_1

b_1 tells us how much the value of y changes (on average) when the value of x *increase* by one. For the red-deer data, b_1 tells us how much the molar weight changes (on average) when age increases by one year.

Each extra year of age is associated with a change of -0.181 grams in molar weight; that is, a *decrease* in molar weight by a mean of 0.181. The molars of some individual deer will lose

more weight than this in some years, and some will lose less weight than this in some years... the value is a *mean* weight loss per year.

To demonstrate, for $x = 10$, y is predicted to be $\hat{y} = 2.588$. For deer one year older than this (i.e. $x = 11$) we predict y to be $b_1 = -0.181$ higher (or, equivalently, 0.181 *lower*). That is, we would predict $\hat{y} = 2.588 - 0.181 = 2.407$. (This is the same prediction made by using $x = 11$ in Eq. (35.1).)



If the value of b_1 is *positive*, then the predicted values of y *increase* as the values of x *increase*.

If the value of b_1 is *negative*, then the predicted values of y *decrease* as the values of x *increase*.

This interpretation of b_1 explains the relationship: Each extra year of age reduces the weight of the molars by 0.181 grams, on average, in male red deer. The units of the slope are the units of the response variable divided by the units of the explanatory variable (so in the deer example, the slope is -0.181 grams per year).

Observe what happens if the slope is *zero*. Since b_1 is the change in y (on average) when x increase by one, $b_1 = 0$ means that the value of y changes by *zero* if the value of x changes by one. In other words, if the value of x changes, the predicted value of y doesn't change. This is equivalent to saying that there is *no relationship* between the variables. (We would also find $r = 0$.)



If the value of the slope is zero, there is *no linear relationship* between x and y. In this case, the correlation coefficient is also zero.

35.6 Hypothesis testing

35.6.1 Introduction

The regression line is computed from the *sample*, assuming a linear relationship actually exists in the *population*. The (unknown) regression line in the *population* is

$$\hat{y} = \beta_0 + \beta_1 x.$$

From the *sample*, the *estimate* of the *population* regression line (Appendix C) is

$$\hat{y} = b_0 + b_1 x.$$

That is, the intercept in the population is β_0 (estimated by b_0), and the slope in the population is β_1 (estimated by b_1). The *sample* can be used to ask questions about the *population* regression coefficients. As usual, the sample values can vary from sample to sample (and so have a *sampling distribution*).

Usually questions are asked about the *slope*, because the slope explains the *relationship* between the two variables (Sect. 35.5).

35.6.2 Hypotheses: Assumption

The null hypothesis is the usual 'no relationship' hypothesis. In this context, 'no relationship' means that the slope is zero (Sect. 35.5.2). Hence, the null hypotheses (about the *population*) is:

- H_0: $\beta_1 = 0$.

This hypothesis proposes that b_1 is not zero because of sampling variation. As part of the decision-making process, the null hypothesis is initially **assumed** to be true.

For the red deer data (Sect. 33.2), determining if a relationship exists between the age of the deer, and the weight of their molars, would test these hypotheses:

- H_0: $\beta_1 = 0$;
- H_1: $\beta_1 \neq 0$

The parameter is β, the population slope for the regression equation predicting molar weight from age.

The alternative hypothesis is two-tailed, based on the RQ.

35.6.3 Sampling distribution: Expectation

Assuming the null hypothesis is true (that $\beta_1 = 0$), we can describe what values the sample slope b_1 are **expected** to take, through *sampling variation*. The variation in the sample slope from sample to sample can be described (Fig. 35.7) using:

- an approximate normal distribution,
- with a mean of $\beta_1 = 0$ (from H_0), and
- a standard deviation, called the *standard error of the slope*, of s.e.(b_1).

The standard error is found using software (jamovi: Fig. 35.8; SPSS: Fig. 35.9).

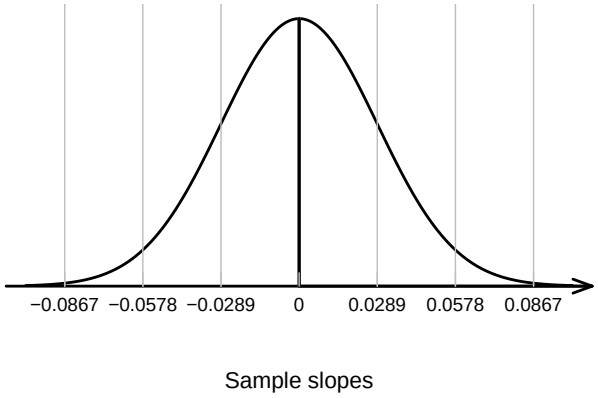


FIGURE 35.7: *The distribution of sample slope for the red deer data, if the population slope is 0*

35.6.4 The test statistic: Observation

The **observed** sample slope was $b_1 = -0.181$. The *test statistic* would be found using the usual approach:

$$t = \frac{b_1 - \beta_1}{\text{s.e.}(b_1)} = \frac{-0.181 - 0}{0.0289} = -6.27,$$

where the values of b_1 and s.e.(b_1) are taken from the software output. The t-score is also reported by the software.

35.6.5 P-value: Consistency with assumption

To determine if the statistic is **consistent** with the null hypothesis, the P-value can be approximated using the 68–95–99.7 rule, or taken from software output (jamovi: Fig. 35.8; SPSS: Fig. 35.9). Using software, the P-value is $P < 0.001$.

We write:

> The sample presents very strong evidence ($t = -6.27$; one-tailed $P < 0.001$) that the slope in the population between age of the deer and molar weight is not zero (slope: -0.181).

Model Coefficients - Weight

Predictor	Estimate	SE	t	p
Intercept	4.398	0.2312	19.02	< .001
Age	−0.181	0.0289	−6.27	< .001

FIGURE 35.8: *jamovi output for the red-deer data*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	4.398	.231		19.020	.000
	Age (in years)	-.181	.029	-.584	-6.275	.000

a. Dependent Variable: Molar weight (in grams)

FIGURE 35.9: *SPSS output for the red-deer data*

> **Example 35.4** (Emergency department patients)**.** A study examined the relationship between the number of emergency department (ED) patients and the number of days following the distribution of monthly welfare monies (Brunette et al. 1991) from 1986 to 1988 in Minneapolis, USA.

The data (extracted from Fig. 2 of Brunette et al. (1991)) is plotted in Fig. 35.10, and the jamovi analysis shown in Fig. 35.11.

The regression line is estimated as

$$\hat{y} = 150.19 - 0.348x,$$

where y represents the mean number of ED patients, and x the number of days since welfare distribution.

This regression equation suggests that each extra day after welfare distribution is associated with a *decrease* in the number of ED patients of about 0.35. (It may be easier to understand this way: 'each 10 extra days after welfare distribution is associated with a *decrease* in the number of ED patients of about $10 \times 0.35 = 3.5$.')

The scatterplot and the regression equation suggests a negative relationship between the number of ED patients and the days after distribution. However, we know that every sample is likely to be different, so the relationship may not actually be present in the population. So we test these hypotheses:

- H_0: $\beta_1 = 0$ where β_1 is the *population* slope
- H_1: $\beta_1 \neq 0$ (i.e., two-tailed, based on the authors' aim)

The output shows that the test statistic is $t = -7.45$, which is very large; unsurprisingly, the two-tailed P-value is very small: $P < 0.001$.

We write:

> There is very strong evidence ($t = -7.45$; two-tailed $P < 0.001$) of a relationship between the mean number of ED patients and the number of days after welfare distribution (slope: -0.348).

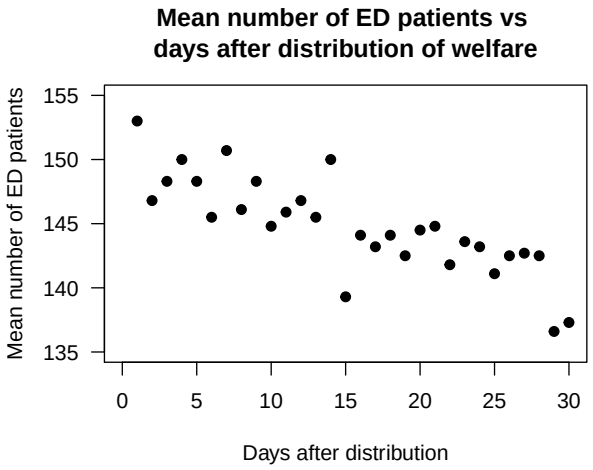


FIGURE 35.10: *Scatterplot of the number of emergency department patients, and the number of days since distribution of welfare*

Linear Regression

Model Fit Measures

Model	R	R^2
1	0.81543	0.66493

Model Coefficients - Mean ED patients

Predictor	Estimate	SE	t	p
Intercept	150.18575	0.828561	181.2610	< .00001
Days since welfare	−0.34790	0.046672	−7.4541	< .00001

FIGURE 35.11: *jamovi output for the emergency department data*

35.7 Confidence intervals

Reporting the CI for the slope is also useful, which can be obtained from software or computed manually.

Think 35.4 (Approximate 95% CI)**.** *Using the output (jamovi: Fig. 35.8; SPSS: Fig. 35.9), what is the* approximate *95% CI for* β_1*?*

CIs have the form

$$\text{statistic} \pm (\text{multiplier} \times \text{standard error}),$$

The multiplier is two for an approximate 95% CI, so (using the standard error reported by the software), we obtain $-0.181 \pm (2 \times 0.029)$, or -0.181 ± 0.058, or from -0.239 to -0.123.

Software can be asked to produce *exact* CIs too (jamovi: Fig. 35.12; SPSS: Fig. 35.13). The *approximate* and *exact* 95% CIs are very similar.

Model Coefficients - Weight

			95% Confidence Interval			
Predictor	Estimate	SE	Lower	Upper	t	p
Intercept	4.398	0.2312	3.937	4.858	19.02	< .001
Age	−0.181	0.0289	−0.239	−0.124	−6.27	< .001

FIGURE 35.12: *jamovi output for the red-deer data, including the CIs*

We write:

The sample presents very strong evidence ($P < 0.001$; $t = -6.275$) of a relationship between age and the weight of molars in male red deer (slope: -0.181; $n = 78$; 95% CI from -0.239 to -0.124) in the population.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	4.398	.231		19.02	.000	3.937	4.858
	Age (in years)	-.181	.029	-.584	-6.275	.000	-.239	-.124

a. Dependent Variable: Molar weight (in grams)

FIGURE 35.13: *SPSS output for the red-deer data, including the CIs*

Example 35.5 (Emergency department patients)**.** In Example 35.4, the jamovi output does not give the 95% CI for the slope.

However, since CIs have the form

$$\text{statistic} \pm (\text{multiplier} \times \text{standard error}),$$

the CI is easily computed:

$$-0.34790 \pm (2 \times 0.046672),$$

or -0.34790 ± 0.093344. This is equivalent to -0.441 to -0.255.

Combining with information from Example 35.4, we write:

> The sample presents very strong evidence ($P < 0.001$; $t = -7.45$) of a relationship between the mean number of ED patients and the numbers of days after welfare distribution (slope: -0.348; $n = 30$; 95% CI from -0.441 to -0.255) in the population.

35.8 Statistical validity conditions

As usual, these results hold under certain conditions to be met. The conditions for which the test is statistically valid are the same as for correlation (Sect. 34.4.3):

1. The relationship is approximately linear.
2. The variation in the response variable is approximately constant for all values of the explanatory variable.
3. The sample size is at least 25.

The sample size of 25 is a rough figure here, and some books give other values.

In addition to the statistical validity condition, the test will be

- **internally valid** if the study was well designed; and
- **externally valid** if the the sample is a simple random sample from the population.

Example 35.6 (Statistical validity)**.** For the red deer data, the relationship is approximately linear, and the variation in molar weight appears to be somewhat constant for various ages (Fig. 33.1), so regression is appropriate. The sample size is $n = 78$. The results from the hypothesis test should be statistically valid.

Think 35.5 (Statistical validity)**.** *Are the conditions for statistically validity met for the cyclones data (Fig. 35.3)?*

Example 35.7 (Emergency department patients)**.** In Example 35.4, the scatterplot (Fig. 35.10) shows that the relationship is approximately linear, that the variation in ED patients seems reasonably constant for different numbers of days after distribution, and the sample size is larger than 30.

The test (Example. 35.4) and the CI (Example. 35.5) should be statistically valid.

35.9 Example: Obstructive sleep apnoea

In a study of obstructive sleep apnoea (OSA) in adults with Down Syndrome (de Carvalho et al. 2020), $n = 60$ adults underwent a sleep study and had various data recorded. The main response variable of interest was OSA severity (measured using the Respiratory Event Index, REI). REI means the average number of episodes of sleep disruption (according to specific criteria) per hour of sleep.

One research question is

> Among Down Syndrome adults, is there a relationship between the REI and neck size?

Here, x is the neck size (in cm), and y is the REI value.

Part of the data are shown in Table 35.1.

Using the jamovi output (Fig. 35.15) the value of the slope and y-intercept in the *sample* are $b_0 = -0.193$ and $b_1 = 0.047$.

The values of the slope and y-intercept in the *sample* are $b_0 = -24.47$ and $b_1 = 1.36$. The regression equation is

$$\hat{y} = -24.47 + 1.36x.$$

The slope means that for each one centimetre increase in neck circumference, the number of sleep disruptions per hour increase (on average) by about 1.36.

Each sample will produce slightly different *sample* slopes, so we can test to see if the slope in the *population* is non-zero due to sampling variation, using a hypothesis test:

TABLE 35.1: *Part of the Obstructive sleep apnoea data set*

Age	Gender	BMI	Neck circumference (cm)	REI
21	Male	20.3	37.0	46.0
24	Male	24.1	40.5	19.3
26	Male	25.2	38.0	12.4
39	Female	40.8	41.0	58.6
21	Female	35.0	37.0	12.7
29	Male	29.2	41.0	38.8
20	Male	25.8	42.0	24.4
21	Male	20.9	37.0	7.0
19	Female	20.5	32.0	37.6
27	Male	22.4	39.0	21.7

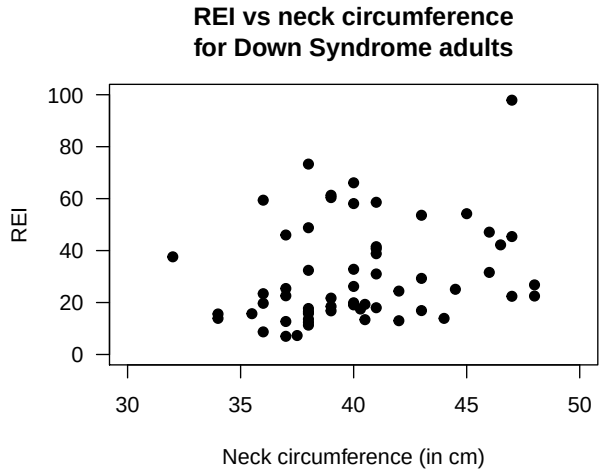


FIGURE 35.14: *Scatterplot of the neck circumference vs REI for Down Syndrome adults*

Linear Regression

Model Fit Measures

Model	R	R^2
1	0.26440	0.069907

Model Coefficients - REI

Predictor	Estimate	SE	t	p
Intercept	−24.4728	26.37586	−0.92785	0.35733
Neck	1.3663	0.65441	2.08790	0.04121

FIGURE 35.15: *jamovi output for the Obstructive sleep apnoea data*

- H_0: $\beta_1 = 0$;
- H_1: $\beta_1 \neq 0$ (that is, two-tailed).

The parameter is β_1. From the software output, $t = 2.09$ and the two-tailed P-value is $P = 0.041$. This means there moderate evidence that the neck circumference is associated with greater REI.

The *approximate* 95% CI for the population slope β_1 is

$$1.3663 \pm (2 \times 0.65441),$$

or from 0.057 to 2.68.

From the scatterplot (Fig. 35.14) the results appear statistically valid. We write:

> The sample presents moderate evidence ($t = 2.09$; two-tailed $P = 0.041$) of a relationship between neck circumference and REI (slope: 1.36; $n = 60$; 95% CI from 0.057 to 2.68) in the population.

35.10 Example: Food digestibility

A study evaluated various food for sheep (Moir 1961). One combination of variables assessed is shown in Fig. 33.6.

The RQ is:

> Does the digestible energy requirement of feed *increase* with dry matter digestibility percentage (and if so, what is the relationship)?

In this study, x is the dry matter weight digestibility percentage, and y is the digestible energy. Part of the data are shown in Table 35.2. Using the software output (Fig. 35.16 (jamovi); Fig. 35.17 (SPSS)), the values of the slope and y-intercept in the *sample* are $b_0 = -0.193$ and $b_1 = 0.047$. The regression equation is

$$\hat{y} = -0.193 + 0.047x.$$

The slope means that when the dry matter weight digestibility increases by 1 percentage point, the digestible energy increases, on average, by 0.047 Cal/gram.

Each sample will produce slightly different *sample* slopes, so we can test to see if the slope in the *population* is non-zero due to sampling variation, using a hypothesis test:

- H_0: $\beta_1 = 0$;
- H_1: $\beta_1 > 0$.

The parameter is β_1, the population slope for the regression equation predicting digestible energy from dry matter weight.

TABLE 35.2: *Part of the digestibility data set*

Dry matter digestibility	Digestible energy	Energy
30.5	27.8	1.243
63.0	61.5	2.750
62.8	60.4	2.701
50.0	49.5	2.213
60.3	58.7	2.681
64.1	63.0	2.877
63.7	62.8	2.895
63.4	62.8	2.895
65.4	64.2	2.952
68.1	66.5	3.059

The alternative hypothesis is *one*-tailed, based on the RQ.

From the software output, $t = 39.322$, which is huge; the **two-tailed** P-value is $P < 0.001$. Since we have a **one-tailed** alternative hypothesis, the P-value is less than $0.001/2 = 0.0005$. There is *very* strong evidence that the digestible energy increases as the dry matter weight digestibility increases.

The *approximate* 95% CI for the population slope β_1 is

$$0.047 \pm (2 \times 0.001),$$

or from 0.045 to 0.049.

Model Coefficients - Energy

Predictor	Estimate	SE	t	p
Intercept	−0.1927	0.07646	−2.52	0.017
DryMatterDigest	0.0467	0.00119	39.32	< .001

FIGURE 35.16: *jamovi output for the sheep-feed data*

The results are statistically valid. We write:

> The sample presents very strong evidence ($t = 39.322$; one-tailed $P < 0.0005$) of a relationship between dry matter weight digestibility and the digestible energy (slope: 0.047; $n = 36$; 95% CI from -0.045 to -0.049) in the population.

35.11 Summary

In this chapter, we have learnt about **regression**, which mathematically describes the relationship between two *quantitative* variables. The response variable is denoted by y, and the explanatory variable by x. The linear relationship between them (the **regression equation**), in the sample, is

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.989[a]	.978	.978	.062926

a. Predictors: (Constant), Dry matter digestibility (in %)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.123	1	6.123	1546.251	.000[b]
	Residual	.135	34	.004		
	Total	6.257	35			

a. Dependent Variable: Digestibility energy content (in caolries/gram)

b. Predictors: (Constant), Dry matter digestibility (in %)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-.193	.076		-2.520	.017
	Dry matter digestibility (in %)	.047	.001	.989	39.322	.000

a. Dependent Variable: Digestibility energy content (in caolries/gram)

FIGURE 35.17: *SPSS output for the sheep-feed data*

$$\hat{y} = b_0 + b_1 x,$$

where b_0 is a number (the **intercept**), b_1 is a number (the **slope**), and the 'hat' above the y indicates that the equation gives an *predicted* mean value of y for the given x value.

The *intercept* is the predicted mean value of y when the value of x is zero. The *slope* is how much the predicted mean value of y changes, on average, when the value of x *increases* by 1.

The regression equation can be used to make *predictions* or to *understand* the relationship between the two variables. Predictions made with values of x outside the values of x used to create the regression equation (called *extrapolation*) may not be reliable.

In the population, the regression equation is

$$\hat{y} = \beta_0 + \beta_1 x.$$

To test a hypothesis about a population slope β_1, based on the value of the sample slope b_1, **assume** the value of β_1 in the null hypothesis (usually zero) to be true. Then, the sample slope varies from sample to sample and, under certain statistical validity conditions, varies with an approximate normal distribution centered around the hypothesised value of β_1, with a standard deviation of s.e.(b_1). This distribution describes what values of the sample slope could be **expected** in the sample if the value of β_1 in the null hypothesis was true. The *test statistic* is

$$t = \frac{b_1 - \beta_1}{\text{s.e.}(b_1)},$$

where β_1 is the hypothesised value given in the null hypothesis (usually zero). The t-value is like a z-score, and so an approximate P**-value** can be estimated using the 68–95–99.7 rule.

35.12 Quick review questions

A study of athletes (Telford and Cunningham 1991) examined the relationship between the height and weight of $n = 37$ rowers at the Australian Institute of Sport (AIS), as shown in Fig. 35.18.

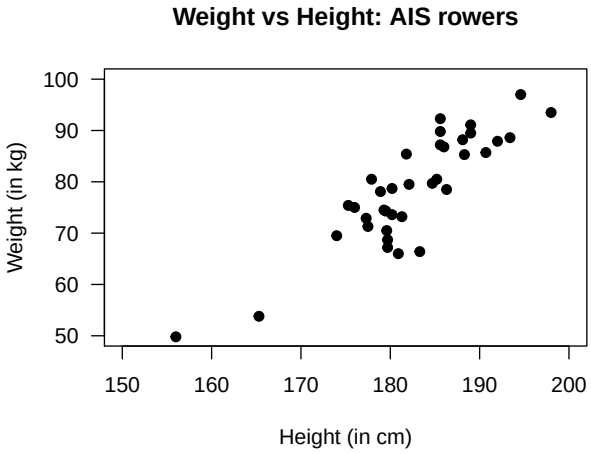


FIGURE 35.18: *Scatterplot of Weight against Height rowers at the AIS*

1. Using the 'rise over run' idea, the *slope* is approximately
2. The y-variable is
3. The regression equation is $\hat{y} = -138 + 1.2x$. What does x represent?
4. What does the 'hat' above the y mean?
5. To two decimal places, what weight would be predicted for a rower who is 180cm tall?
6. The standard error of the slope is 0.112. What is the value of the *test statistic* (to one decimal place) to test if the population slope is zero?
7. True or false? The P-value for this test will be *very small*
8. True or false? The *units* of the slope are kg/cm
9. True or false? Making a prediction for the weight of a rower weighing 220 kg would be an example of *extrapolation*

35.13 Exercises

Selected answers are available in Sect. D.32.

Exercise 35.1. In wastewater treatment facilities, air from biofiltration is passed through a membrane and dissolved in water, and is transformed into harmless byproducts (Chitwood and Devinny 2001; Devore and Berk 2007). The removal efficiency y (in %) may depend on the inlet temperature (in °C; x). The RQ is

> In treating biofiltation wastewater, how does the removal efficiency depend on the inlet temperature?

Using the scatterplot (Fig. 35.19) and software output (Fig. 35.20 (jamovi); Fig. 35.21 (SPSS)), answer these questions.

1. Write down the value of the slope (b_1) and y-intercept (b_0).
2. Write down the regression equation.
3. Interpret the slope (b_1).
4. Test for a relationship in the population, by first writing the hypotheses.
5. Write down the t-score and P-value from the software output.
6. Determine an *approximate* 95% CI for the population slope β_1.
7. Write a conclusion.

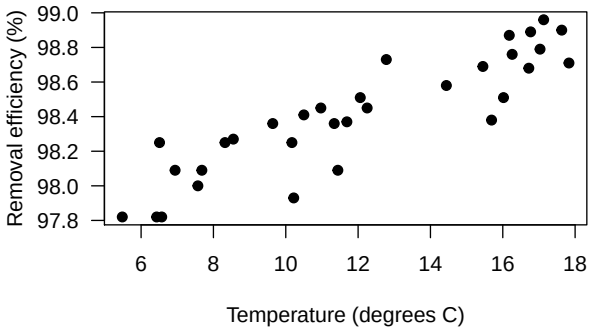


FIGURE 35.19: *The relationship between removal efficiency and inlet temperature*

Model Coefficients - Removal

Predictor	Estimate	SE	t	p
Intercept	97.4986	0.08894	1096.2	< .001
Temp	0.0757	0.00705	10.7	< .001

FIGURE 35.20: *jamovi regression output for the removal-efficiency data*

Exercise 35.2. A study (Myers (1990), p. 75) of American footballers measured the right-leg strengths x of 13 players (using a weight lifting test), and the distance y they punt a football (with their right leg).

1. Use the plot (Fig. 35.22) to guess of the values of the intercept and slope.
2. Using the jamovi output (Fig. 35.23), write down the value of the slope (b_1) and y-intercept (b_0).
3. Hence write down the regression equation.
4. Interpret the slope (b_1).

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Inlet temperature (in deg C)[b]	.	Enter

a. Dependent Variable: Removal efficiency (in %)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.891[a]	.794	.787	.15517

a. Predictors: (Constant), Inlet temperature (in deg C)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.779	1	2.779	115.400	.000[b]
	Residual	.722	30	.024		
	Total	3.501	31			

a. Dependent Variable: Removal efficiency (in %)

b. Predictors: (Constant), Inlet temperature (in deg C)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	97.499	.089		1096.169	.000
	Inlet temperature (in deg C)	.076	.007	.891	10.742	.000

a. Dependent Variable: Removal efficiency (in %)

FIGURE 35.21: *SPSS regression output for the removal-efficiency data*

5. Write the hypotheses for testing for a relationship in the population
6. Write down the t-score and P-value from the output.
7. Determine an *approximate* 95% CI for the population slope β_1.
8. Write a conclusion.

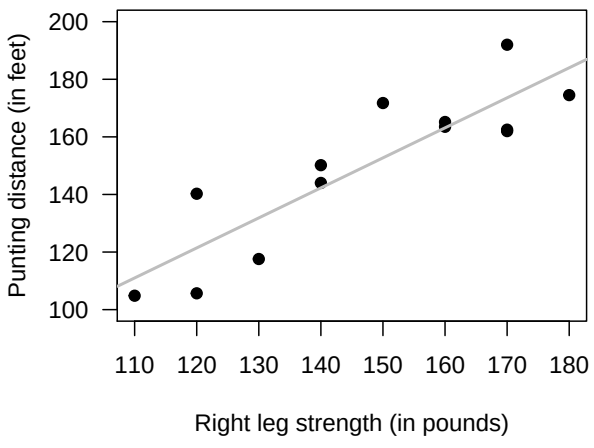


FIGURE 35.22: *Punting distance and right leg strength*

Model Coefficients - Punt

Predictor	Estimate	SE	t	p
Intercept	−3.69	25.265	−0.146	0.886
Right	1.04	0.169	6.162	< .001

FIGURE 35.23: *jamovi regression output for the punting data*

Exercise 35.3. A study (Amin and Mahmood-ul-Hasan 2019) of gas engines measured the throttle angle (x) and the manifold air pressure (y) as a fraction of the maximum value.

1. The value of r is given in the paper as 0.972986604. Comment on this value, and what it means.
2. Comment on the use of a regression model, based on the scatterplot (Fig. 35.24).
3. The authors fitted the following regression model: $y = 0.009 + 0.458x$. Identify errors that the researchers have made when giving this regression equation (there are more than one).
4. Critique the researchers' approach.

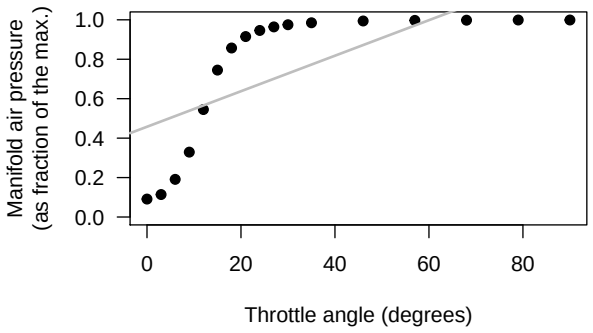


FIGURE 35.24: *Manifold air pressure plotted against throttle angle for an internal-combustion gas engine*

Exercise 35.4. A study of hot mix asphalt (Panda et al. 2018) created $n = 42$ samples of asphalt and measured the volume of air voids and the bitumen content by weight (Fig. 35.25). Use the software output (Fig. 35.26) to answer these questions.

1. Write down the regression equation.
2. Interpret what the regression equation means.
3. Perform a test to determine is there is a relationship between the variables.
4. Predict the mean percentage of air voids by volume when the percentage bitumen is 5.0%. Do you expect this to be a good prediction? Why or why not?
5. Predict the mean percentage of air voids by volume when the percentage bitumen is 6.0%. Do you expect this to be a good prediction? Why or why not?

Exercise 35.5. A study of $n = 31$ people on the use of sunscreen (Heerfordt et al. 2018) explored the relationship between the time (in minutes) spent on sunscreen application x and the amount (in grams) of sunscreen applied (y). The fitted regression equation was $\hat{y} = 0.27 + 2.21x$.

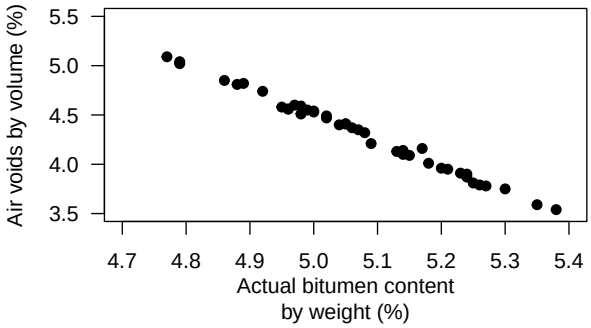


FIGURE 35.25: *Air voids in bitumen samples*

Model Coefficients - AirVoids

Predictor	Estimate	SE	t	p
Intercept	17.47	0.1757	99.4	< .001
Bitumen	−2.59	0.0346	−74.9	< .001

FIGURE 35.26: *jamovi regression output for the bitumen data*

1. Interpret the meaning of b_0 and b_1. Do they seems sensible?
2. According to the article, a hypothesis for testing β_0 produced a P-value much larger than 0.05. What does this mean?
3. If someone spent 8 minutes applying sunscreen, how much sunscreen would you predict that they used?
4. The article reports that $R^2 = 0.64$. Interpret this value.
5. What is the value of the correlation coefficient?

Exercise 35.6. One study (Bhargava et al. 1985) stated:

> In developing countries [...] logistic problems prevent the weighing of every newborn child. A study was performed to see whether other simpler measurements could be substituted for weight to identify neonates of low birth weight and those at risk.
> — Bhargava et al. (1985), p. 1617

The relationship between infant chest circumference (in cm) x and birth weight (in grams) y was given as:

$$\hat{y} = -3440.2403 + 199.2987x.$$

The correlation coefficient was $r = 0.8696$ with $P < 0.001$.

1. Based on the regression equation only, could chest circumference be used as a useful predictor of birth weight? Explain.
2. Based on the correlation information only, could chest circumference be used as a useful predictor of birth weight? Explain.
3. Interpret the intercept and the slope of the regression equation.
4. What units of measurement are the intercept and slope measured in?
5. Predict the birth weight of an infant with a chest circumference of 30cm.

6. Critique the way in which the regression equation and correlation coefficient are reported.

Part IX

Reporting, writing and reading research

36

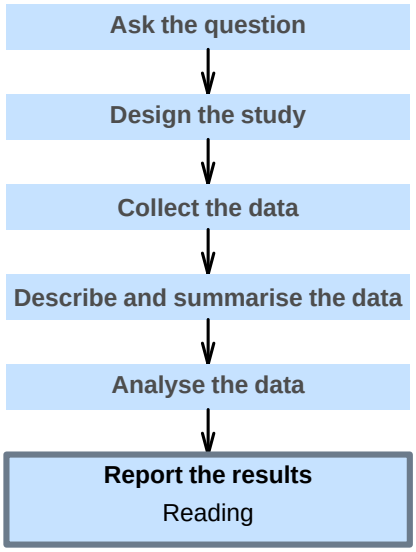
Reading research



So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, understand the tools of inference, to form *confidence intervals*, and to perform *hypothesis tests*.

In this chapter, you will learn to read about the research of others. You will learn to:

- read and understand research.



36.1 Introduction

Science requires reading the research of others. Research is usually communicated in *journal articles* (also called *papers*), or sometimes in *presentations* (conferences; seminars). Milllions of journal articles are available (many online), and this book references many articles.

At some time during your university studies, you will need to read articles: so you know *why* your discipline does things as it does, the evidence for doing so, and open questions in your discipline. Understanding the language of research is important for understanding these articles.

However, reading a research article can be hard work... A good place to start is to read the *Abstract* (sometimes called a *Summary*, or *Overview*): a useful overview of the whole paper (without details).

To understand a paper, the six steps of the research process can be used as a guide::

1. **Ask the question**: What *research question* is the paper answering? Are inclusion and/or exclusion criteria given?
2. **Design the study**: How did the authors *design the study*? Is the study designed to maximize internal and external validity? What are the design limitations?
3. **Collect the data**: How did the authors *collect the data*? Could the study be approximately repeated if needed?
4. **Describe and summarise the data**: Is the data *summary* appropriate, complete and clear?
5. **Analyse the data**: Is the *analysis* appropriate, accurate, valid and clear?
6. **Report the results**: Are the results accurately, appropriately and well *reported*? What is the answer to the RQ? What other questions have emerged?

In the examples that follow, some extracts from articles will be studied.

36.2 Example 1: Reading research

A study (Fritts et al. 2018) explored the impact of adding herbs and spices to the consumption of vegetables by adolescent school children. Part of the Abstract states (slightly edited for brevity):

> *Purpose*: We evaluated whether new vegetable recipes using herbs and spices would increase preference for vegetables served to adolescents at this school.
> *Methods*: To evaluate recipe acceptance, we assessed liking (100 mm visual analog scales) among students (n = 96–110; aged 14–18 years) for 8 plain (oil and salt) and 8 seasoned vegetables. Liking ratings between plain and seasoned vegetables were compared with paired t-tests...
> *Results*: Students reported higher liking for several seasoned recipes compared to plain: broccoli ($P = 0.02$), vegetable dip ($P < 0.0001$), black beans and corn ($P < 0.001$) and cauliflower ($P < 0.0001$).
> *Conclusions*: Common herbs and spices improved liking for several school lunch vegetables compared to plain varieties among rural high school students...
> — Fritts et al. (2018), p. 125

Later we read this (again, slightly edited):

> This is a cross-sectional study assessing preference for plain and seasoned vegetables in a population of middle/high school students (aged 14–18 years) attending a rural Pennsylvania public school.
> — Fritts et al. (2018), p. 126

Even using this (small amount) of information, much can be learnt about the study. For example:

1. **Ask the question**: The POCI elements are:

- *Population*: 'middle/high school students (aged 14–18 years) attending a rural Pennsylvania public school'
- *Outcome*: The *mean difference* in taste ratings between plain and seasoned vegetables. The taste ratings are given using a '100 mm visual analog scale.'
- *Comparison*: There is **no** comparison: Every member of the population is treated the same way. A comparison exists if different subsets of the population are treated differently (for example, one group of students is given plain vegetables, and a different group is given seasoned vegetables).
- *Intervention*: No; there is no comparison, so there is no comparison to be allocated.

2. **Design the study**: Since this RQ is *descriptive*, the study is * descriptive. *The participants were probably not* blinded*, since the presence of seasoning was probably obvious.
3. **Collect the data**: No details are given about the data collection.
4. **Describe and summarise the data**: The Abstract gives no summary data (since eight vegetables were studied, this would have consumed too much space I guess).
5. **Analyse the data**: The data were analysed using *paired* t-tests, one for each different vegetable used. (Each subject gave two ratings for each vegetable: one for *plain* vegetables and one for *seasoned* vegetables),
6. **Report the results**: Evidence exists of a mean difference (that students preferred the seasoned vegetables) in many cases, but not all (the Abstract states that eight vegetables were used, with statistically significant differences for five).

From this information, the RQ is something like:

> For middle/high school students (aged 14–18 years) attending a rural Pennsylvania public school, is there a mean difference in taste ratings (measured on a '100 mm visual analog scale') between plain and seasoned vegetables?

For more details, the whole paper could be read.

36.3 Example 2: Reading research

Consider this Abstract (Groves 2010):

> **Objective** To determine whether the author's 20.9 lb (9.5 kg) carbon frame bicycle reduced commuting time compared with his 29.75 lb (13.5 kg) steel frame bicycle.
> **Design** Randomised trial.
> **Setting** Sheffield and Chesterfield, United Kingdom, between mid-January 2010 and mid-July 2010.
> **Participants** One consultant in anaesthesia and intensive care.
> **Main outcome measure** Total time to complete the 27 mile (43.5 kilometre) journey from Sheffield to Chesterfield Royal Hospital and back.
> **Results** The total distance travelled on the steel frame bicycle during the study period was 809 miles (1302 km) and on the carbon frame bicycle was 711 miles (1144 km). The difference in

> the mean journey time between the steel and carbon bicycles was 00:00:32 (hr:min:sec; 95% CI -00:03:34 to 00:02:30; $P = 0.72$).
> **Conclusions** A lighter bicycle did not lead to a detectable difference in commuting time. Cyclists may find it more cost effective to reduce their own weight rather than to purchase a lighter bicycle.
> — Groves (2010), p. 341

Based on this Abstract, again we can learn many things about the study.

1. **Ask the question**: The POCI elements are:
   - *Population*: The trips by *this* rider, on *his* bikes, on *his* route to work. This is not easy to identify, but notice that there are many examples of this rider, on his bikes, on his route. For example, there are not many examples of different bikes, different riders, or different routes.
   - *Outcome*: 'Total time to complete the 27 mile (43.5 kilometre) journey.'
   - *Comparison*: Between the steel-frame and carbon-frame bicycles.
   - *Intervention*: Yes, because the elements of the population (the different commutes) can be randomly allocated to be taken with the steel- or carbon-frame bikes.
2. **Design the study**: The study is 'randomised controlled trial,' a type of experimental study. Random allocation has been used.
3. **Collect the data**: The Abstract gives no information.
4. **Describe and summarise the data**: The Abstract gives no summary data for each bike, but summarises the *difference* between the means: 32 seconds (95% CI between -3:34 and 2:30 minutes, but *which* bike produces the faster mean time is not stated).
5. **Analyse the data**: Though not stated, probably a two-sample t-test.
6. **Report the results**: 'A lighter bicycle did not lead to a detectable difference in commuting time': There is no evidence that the carbon-frame bicycle reduced the commmuting time (for this rider, on his route to work, with his bikes...). In any case, the difference between the two mean commuting times is 32 seconds... over a 43.5 kilometre journey: Hardly of any *practical* importance (Sect. 28.8)!

The RQ may be:

> For trips made by one cyclist (on his bikes, on his route to work), is the mean time to complete the 43.5 kilometre the same for the steel-frame and carbon-frame bicycles?

This is a poor RQ: it is not relevant or interesting (Sect. 2.6) to anyone except this single rider: The results are relevant to one person in the entire world...

Another thing to observe: The RQ is *one*-tailed (does the carbon frame bicycle *reduce* commuting time), but the conclusion gives a *two*-tailed P-value. (This may not be obvious, but a one-tailed P-value cannot be larger than 0.5.)

This is a strange study... However, it appeared in a Christmas edition of *BMJ*, which contains more 'light-hearted' articles:

> While the subject matter may be more light-hearted, research papers in the Christmas issue adhere to the same high standards of novelty, methodological rigour, reporting transparency, and readability as apply in the regular issue.
> — From https://www.bmj.com/about-bmj/resources-authors/article-types/christmas-issue

36.4 Exercises

Selected answers are available in Sect. D.33.

Exercise 36.1. A research article (Duncan et al. 2018) examined the accuracy of step counts recorded on iPhones. The paper records this information about the selection of participants:

> Participants were recruited through word of mouth and posters displayed around the [researcher's] university. Participants were eligible if they were ambulatory, ≥ 18 years of age, and owned an iPhone 6 [...] or newer model.

Although 33 participants were selected, the authors note some parts of the study used a smaller sample size because:

> ... one [subject] lost their phone during the observation period, [and] the other opted out of the [...] test due to personal circumstances.

The paper notes that previous studies have been able to:

> [...] demonstrate the accuracy of the iPhone pedometer function in laboratory test conditions. However, no studies have attempted to evaluate evidence [...] in the field.

1. What is the issue that the authors raise with previous studies?
2. Why did the authors discuss the changes in sample size for some parts of the study?
3. How would you describe the sampling method?
4. What would you call the information about given about the subjects needing to be ambulatory and 18 years of age or over?
5. Among many other things, the researchers compared the *mean diference* between the number of step counts recorded by manually counting steps and the iPhone-recorded number of steps. What type of test would be appropriate?
6. While walking at 2.5 km/h, the above test produced a P-value of 0.006. What does this mean?
7. The sample size for the part of the study mentioned above was $n = 32$. Do you think the test will be statistically valid?

Exercise 36.2. One study of hearing loss among Iranian students (Mohammadpoorasl et al. 2018) used a cross-sectional study to explore the relationship between hearing loss and headphone use. The article states that

> ... 890 students were randomly selected from five schools at QUMS (Medicine, Dentistry, Nursing and Midwifery, Public Health, and Paramedical Sciences schools) using a proportional cluster sampling method...

The participants completed a hearing test and completed a Hearing Loss Questionnaire (values are between 17 and 34: higher scores indicating more severe hearing loss).

1. What is the population?
2. Critique the sampling method: What is the implication for interpreting the results of the study?
3. Some of the results are presented in Table 36.1. What statistical test do you think was used to compare the scores for males and females?
4. What are the hypotheses being tested about 'Frequency of use?'
5. Form an approximate 95% CI for the mean hearing loss score for students who use earphones.
6. What information is needed to be able to form an approximate 95% CI for the *difference* between the hearing loss scores for females and males?

TABLE 36.1: *The Hearing Loss Questionnaire scores for various demographic variables*

Criterion	Levels	Sample size	Mean	Std. dev	P-value
Sex	Female	543	19.37	2.91	0.009
	Male	302	19.99	3.51	
Frequency of use	0, 1 times/day	194	19.2	2.87	0.001
	2 to 3 times/day	319	19.6	2.66	
	More than 3 times/day	278	20.2	3.54	
Earphone use	Yes	745	19.8	3.08	< 0.001
	No	100	19	1.71	

Exercise 36.3. The Abstract from a large study is given below:

> *OBJECTIVE*: This study aims to elucidate any existing link between energy-containing liquids, as consumed in various forms within the diet, and the effect they may have on body weight or other diseases [...]
> *METHODS*: A self-administered online survey was conducted in 2496 participants from different countries, in six languages (Spanish, English, Chinese, French, German and Portuguese). Questions referred to their soft drink and water consumption habits, physical exercise performed, presence or absence of certain diseases and medication.
> *RESULTS*: There is statistically significant difference ($p < 0.001$) in BMI and consumption of cola per week: those who consumed 0–3 cans a week have a lower BMI than those who consume >7 cans of cola a week [...] There is greater presence of obesity ($p < 0.001$), gastritis ($p < 0.001$), constipation ($p < 0.001$) and mental illness ($p = 0.003$) among people who drink cola soft drinks.
> *CONCLUSION*: Removal of energy-containing beverages from our diet may be an appropriate public health message to support those interested in preventing weight gain as well as other diseases.
> Martín et al. (2018), p. 1

Evaluate the study using the six steps of research discussed in this book.

37

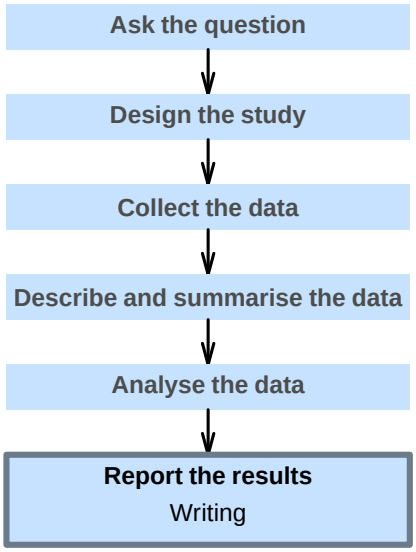
Writing research



So far, you have learnt to ask a RQ, identify different ways of obtaining data, design the study, collect the data describe the data, summarise data graphically and numerically, understand the tools of inference, to form *confidence intervals*, and to perform *hypothesis tests*.

In this chapter, you will learn to write about your own research. You will learn to:

- write research.
- appropriately structure your research writing.



37.1 Introduction

All students in scientific, engineering and health professions need to *read* the research of others; that's how to stay up-to-date with the discipline. Some students will also need to *write* about their own research or the research of others. To do so, understanding the language of research is important.

One of the most important points about writing in the scientific disciplines is to write *carefully* and *precisely*:

- Think carefully about the words you use: You do not want to just be understood, you need to make sure that you *can't* be misunderstood.

- Use the correct, technical words, and use them correctly.
- Write what you mean.
- Mean what you write.
- Be careful using words whose meaning are *lexically ambiguous* (Richardson et al. 2013; Dunn et al. 2016): Words with different meanings in science and in every-day use (Sect. 37.10).

Formal approaches to writing and reporting research exist, for experimental (CONSORT[1]) and for observational studies (STROBE[2]). We will not delve into these specifically (partly because of the wide range of disciplines adopting this book), but these websites are useful resources.

Some information in this chapter is based on Dr. Michael Lufaso's notes (http://www.unf.edu/~michael.lufaso/chem4931/lecture3.pdf) and Prof. Tony Roberts (http://www.maths.adelaide.edu.au/anthony.roberts/LaTeX/ltxwrite.php) notes.

37.2 General tips

A series of experimental studies concludes that

> … a majority of undergraduates admit to deliberately increasing the complexity of their vocabulary so as to give the impression of intelligence.
> — Oppenheimer (2006), p. 139

That is, study like to use fancy words to sounds clever. One conclusion of the study was that using 'fancy' language *does not work*: 'needless complexity leads to negative evaluations…' (Oppenheimer (2006), p. 151). One recommendation by the author is to

> … write clearly and simply if you can, and you'll be more likely to be thought of as intelligent.
> — Oppenheimer (2006), p. 153

With this in mind, a scientific paper:

- **Should** use simple, clear but technically correct language.
- **Should** present the facts in an unbiased manner.
- **Should** be clear, concise and complete.
- **Should** use facts to make statements.
- **Should** be complete enough that other professionals can repeat the study.

Likewise, a scientific paper:

- **Should not** be haphazard, jumbled or illogical.
- **Should not** be used as a personal soapbox.

[1] http://www.consort-statement.org/
[2] https://www.strobe-statement.org

- **Should not** reach conclusions not based on the reported evidence.
- **Should not** be for insiders only.
- **Should not** overstate what has been learnt from the study.

37.3 Article structure

Many scientific papers have these (or similar) sections, though it varies a lot by discipline:

- Title and authors.
- **A**bstract: A summary of the whole paper, without details.
- **I**ntroduction: *Why* was the study done, and *what* was hoping to be achieved?
- **M**aterials and method: *How* was the study done?
- **R**esults: What was found?
- **D**iscussion (or Summary, or Conclusions): What does it *mean*?
- References (or Bibliography).

Often the acronym AIMRaD or IMRaD is used to help recall these sections. These components capture certain parts of the six-step research process in this book (Fig. 37.1).

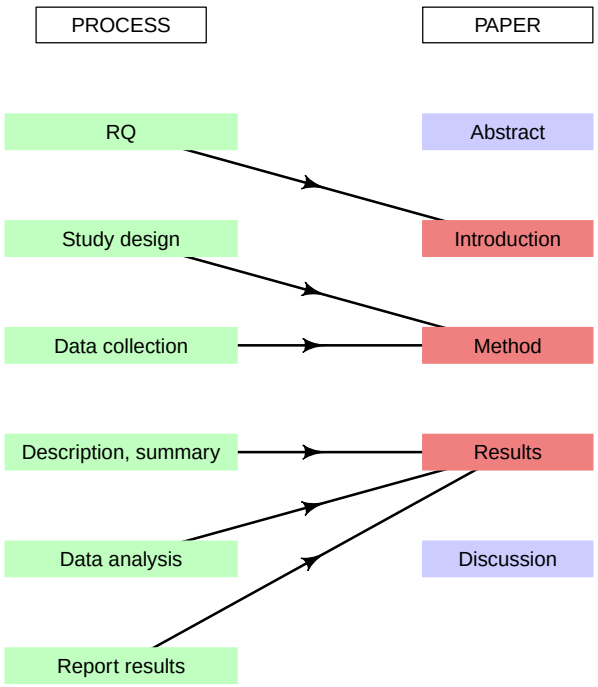


FIGURE 37.1: *The connection between the paper and the steps we have studied. The Abstract briefly covers all aspects of the study, and the Discussion combines elements from all areas also.*

37.4 Writing scientifically: Title

Titles are important: they can easily discourage a reader from engaging with an article. A title should be a clear description of the main purpose of the article, and be:

> ... accurate, specific, concise, and informative, must not contain abbreviations, and must never be dull.
> Peat et al. (2002), p. 93.

Titles sometimes pose questions ('Does asthma reduce linear growth?') or answer questions ('Linear growth deficit in asthmatic children'; Peat et al. (2002), p. 98).

Example 37.1 (Article title)**.** A good example of a title is:

> Beauty sleep: experimental study on the perceived health and attractiveness of sleep deprived people
> — Axelsson et al. (2010)

Example 37.2 (Article title)**.** A *poor* example of an article title is:

> The nucleotide sequence of a 3.2 kb segment of mitochondrial maxicircle DNA from Crithidia fasciculata containing the gene for cytochrome oxidase subunit III, the N-terminal part of the apocytochrome b gene and a possible frameshift gene; further evidence for the use of unusual initiator triplets in trypanosome mitochondria
> — Sloof et al. (1987)

Extra example: A *very* poor example of an article title is:

> Reaction of a bidentate ligands (4,4'-dimethyl 2,2'-bipyridine) with planar-chiral chloro-bridged ruthenium: Synthesis of cis-dicarbonyl[4,4'-dimethyl-2,2'-bipyridine- κ O1, κ O2] { 2-[tricarbonyl(η 6-phenylene- κ C1)chromium]pyridine-κN } ruthenium hexafluorophosphate
> — Hijazi et al. (2013a)

This article was also retracted due to unethical conduct of one author (Hijazi et al. 2013b).

37.5 Writing scientifically: Abstract

The Abstract is a short section at the start of an article which summarises the *whole* paper; it is *not* an introduction! An Abstract includes the most important and interesting parts of the

research. The Abstract is often the most important part of any article, as it is the only part that many people will read.

Writing the Abstract after the paper is fully written is often sensible. Some (but not all) journals require a *structured abstract*, where the Abstract contains sections to be briefly completed (see Sect. 36.2). These abstracts are usually much easier for a reader to follow.

The Standards for Reporting Diagnostic Accuracy (STARD) statement (Cohen et al. 2017) list *essential* items for Abstracts; these are (slightly adapted):

- *Background and Objectives*: List the study objectives (the RQ).
- *Methods*: Describe:
  - The process of *data collection*;
  - The *type* of study;
  - The *inclusion and exclusion criteria* for individuals;
  - The *settings* in which the data were collected;
  - The *sampling* method (e.g., random or convenience sample);
  - The tools or methods used to *collect the data*.
- *Results*: Provide
  - The number of individuals in all groups included in the analysis;
  - Estimates of precision of estimates (e.g., confidence intervals);
  - Results of analysis (e.g., hypothesis tests).
- *Discussion*: Provide
  - A general *interpretation* of the results;
  - *Implications* for practice, including the intended use of the index test;
  - Limitations of the study.

These loosely align with the six steps of research used in this book.

Example 37.3 (Structured abstract)**.** A research study examined the long-term effects of mortality after amptutation (Singh and Prasad 2016). The (structured) Abstract (slightly edited for brevity) is repeated below:

> *Background*: Mortality after amputation is known to be extremely high and is associated with a number of patient features. We wished to calculate this mortality after first-time lower-limb amputation and investigate whether any population or treatment factors are associated with worse mortality.
> *Objective*: To follow up individuals after lower limb amputation and ascertain the mortality rate as well as population or treatment features associated with mortality.
> *Study design*: A prospective cohort study.
> *Methods*: Prospective lower-limb amputations over 1 year ($N = 105$) at a Regional Rehabilitation Centre were followed up for 3 years.
> *Results*: After 3 years, 35 individuals in the cohort had died, representing a mortality of 33%. On initial univariate analysis, those who died were more likely to have diabetes mellitus ($\chi^2 = 7.16$, df $= 1$, $p = 0.007$) and less likely to have been fitted with a prosthesis ($\chi^2 = 5.84$, df $= 1$, $p = 0.016$) [...] Diabetes (odds ratio $= 3.04$, confidence intervals $= 1.25 - 7.40$, $p = 0.014$) and absence of prosthesis-fitting (odds ratio $= 2.60$, confidence interval $= 1.16 - 6.25$, $p = 0.028$) were independent predictors of mortality.
> *Conclusion*: Mortality after amputation is extremely high and is increased in individuals with diabetes or in those who are not fitted with a prosthesis after amputation.
> — Singh and Prasad (2016), p. 545

37.6 Writing scientifically: Introduction

The introduction has many purposes:

1. To gain the interest of readers, and encourage them to read more of the article.
2. To set up the context and background for the paper.
3. To define the language and definitions used in the study.
4. To allow the reader to become familiar with the theoretical groundwork of the subject.
5. To state the purpose of the paper: Why it was written, and what the authors hope to learn.
6. To show how the research fills a gap in existing knowledge.

The introduction provides a clear statement of the study's RQ (sometimes stated as the Purpose, Aim, Objective, etc.). The introduction often includes a literature review too, though sometimes a literature review is a separate section.

37.7 Writing scientifically: Materials and methods

The Materials and Methods section explains how the data were obtained:

- How the *sample* was obtained.
- How the data were *collected* (the data collection *protocol*).

- How the data were *analysed* (including the software used, and the statistical methods used).
- What specialized equipment was used (don't list pencils, rulers, paper, etc.!).

37.8 Writing scientifically: Results

The Results summarise what was found from the data: The Results section:

- Shows all the relevant findings from the research.
- Presents a summary of the data set: the number of cases, the number of missing values, and a verbal description of all variables. Unless the data set is small, the data itself is usually not given. (Sometimes, the data may be presented in an Appendix, or in an online supplement.)
- Presents tabular, numerical and/or graphical summary of the data and relationships of importance.
- Gives a brief verbal interpretation of these summaries.
- Gives the *results* from any hypothesis tests and confidence intervals.
- Identifies trends, consistencies, anomalies etc.
- Does *not* contain an interpretation or explanation of the results of the tests (that is the purpose of the Discussion).

37.9 Writing scientifically: Discussion and conclusion

Sometimes, articles have *separate* Discussion and Conclusion sections; sometimes they are combined. The Discussion section:

- Summarises the results.
- Gives a short evaluation of the results: Be concise.
- Answers the stated RQ.
- Discusses limitations (Sect. 7.9), strengths, weaknesses, problems, challenges.
- Tries to anticipate and respond to potential questions about the research.

Readers should reach the conclusions based on the *evidence* presented.

37.10 Writing carefully: Lexically ambiguous words

As noted in Sect. 37.1, writing *carefully* and *precisely* is incredibly important in science.

One aspect of writing well is using *lexically ambiguous words* carefully. *Lexically ambiguous words* have a different meaning in other scientific disciplines, or in their usual every-day use (Richardson et al. 2013; Dunn et al. 2016). If you are unsure of the definitions used in this book, make use of the Glossary (Appendix G).

Here are some lexically ambiguous words to be wary of:

- **Average**: In statistics and research, 'average' can refers to *any* way of measuring the typical value (Sect. 13.2), including the mean and the median, but also other measures too. Use the specific word 'mean' or 'median' if that is what you actually intend!
- **Confidence**: In statistics and research, the word 'confidence' is usually used in the phrase 'confidence interval,' where it has a specific meaning (Sect. 21.2).
- **Comparison**: In statistics and research, a 'comparison' (Sect. 2.3.3) is when the sample and population can be separated into two or more groups that are either treated differently (e.g., one group is given a placebo, and one a treatment) or are fundamentally different (e.g., aged under 40, or aged 40 or over).
- **Control**: In statistics and research, a 'control' refers to a specific situation, and is helpful for maximising internal vaidity (Def. 7.6).
- **Correlation**: In statistics and research, correlation describes the relationship between two *quantitative* variables (Sect. 34.1).
- **Estimate**: In statistics and research, 'estimating' usually means to find a sample value (i.e., to make a calculation) to estimate an unknown population parameter, rather than the colloquial use where it often means to take a guess (Sect. 2.5).
- **Experiment**: In statistics and research, an experiment is a specific type of research study (Sect. 3.4). Use the word 'study' to talk about experimental and observational studies more generally.
- **Graph**: In statistics and research, a 'graph' is used to summarise data (Chap. 12).
- **Independent**: This words has *many* uses in statistics and research, in science, and in general use. We use the word 'independent' in this book to refer to events that do not impact each other in a probabilistic sense (Sect. 16.5).

- **Intervention**: In statistics and research, an 'intervention' (Sect. 2.3.4) is how the researchers manipulate the comparison or connection.
- **Normal**: In statistics and research, 'normal' usually refers to the 'normal distribution' (Chap. 17.3). If this is *not* the meaning you intend to convey, consider using the word 'usual.'
- **Odds**: In statistics and research, 'odds' has a specific meaning (Sect. 14.2) and is *different* that probability, whereas 'probability' and 'odds' are often used interchangeably in general usage.
- **Population**: In statistics and research, the 'population' refers to a larger group of interest (Sect. 2.3.1), whereas in general use 'population' usualy refers to groups of people.
- **Random**: In statistics and research, 'random' has a specific meaning, but in general usage it often means 'haphazard.'
- **Regression**: In statistics and research, 'regression' refers to the mathematical relationship between two quantitative variables (Sect. 35).
- **Sample**: In statistics and research, we say (for example) that we 'have taken one sample of 30 fungi' (Sect. 5.1); in some disciplines, this could be described as 'taking 30 samples.'
- **Significant**: This is perhaps the most mis-used word in scientific writing. In statistics and research, 'significance' is usually understood to refer to 'statistical significance' (Sect. 28.6). If this is *not* the meaning you intend to convey, consider using the word 'substantial.'

- **Variable**: In statistics and research, a 'variable' is something that can vary from individual to individual (Def. 2.11).

Furthermore, some *symbols* may have different meanings in research and statistics than in some other scientific disiplines; again, care is needed when using these symbols:

- β: In this book, β refers to the regression parameters (Sect. 35.3).
- ρ: In this book, ρ refers to the population correlation coefficient (Sect. 34.1).
- $\pm$: In this book, the symbol $\pm$ is used for confidence intervals to describe a *range* of values in which the population parameter probably lies (Sect. 21).

37.11 Constructing tables

Good tables can be time-consuming to construct. In general, tables:

- **Should** be discussed (not just simply presented) in the text.
- **Should** be clear and uncluttered (consider using multiple tables if necessary).
- **Should** typically have captions *above*.
- **Should** have very few horizontal lines, and probably *no* vertical lines. (Tables produced for online reading may have different rules.)
- **Should** include units of measurement (such as kg) where appropriate.
- **Should** be able to be understood without reference to the paper, as far as possible.
- **Should** make the most important comparisons easy to make.
- **Should** align the data for easy comparisons, if possible (for example, decimal places all beneath each other in columns).

37.12 Constructing figures and graphs

Good figures can be hard to produce, but their purpose should always be kept in mind: *To display the data in the simplest, clearest possible way.* In general, figures:

- **Should** be discussed (not just simply presented) in the text.
- **Should** be clear and uncluttered.
- **Should** typically have captions *below*.
- **Should** includes units of measurement (such as kg) where appropriate.
- **Should** be able to be understood without reference to the paper, as far as possible.
- **Should** have any colours or different line types explained.
- **Should** use easy-to-read fonts and colours: Make sure the writing is sufficiently large when the figure is placed in the article.
- **Should not** include *chart junk* such as artificial third dimensions (Su 2008).

37.13 Other elements

Many research articles have other sections too:

- **References**
  - Give the full citations of any work referenced, in the required format (such as APA, Harvard, etc.).
  - Organise and format references correctly: Many disciplines or journals have very strict guidelines for how references must be listed and formatted.
- **Acknowledgements**: Thank people who legitimately contributed to the report, and research funding bodies. One attempt (Allen et al. 2019) has identified many different ways in which people can contribute: see, for example, https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement
- **Appendices**
  - In the appendices, place important material that would break the flow of the document's narrative.
  - Appendices may include large tables, images, detailed discussions of technical details, ...
  - Sometimes, appendices may be placed online.

37.14 Style

Different disciplines have their own styles. You may need to read articles from your discipline or target journal to see how to write in that style. Here are some general style recommendations:

- Use short sentences.
- Many disciplines prefer the passive voice ('The blood pressure was measured using...') rather than active voice ('We measured the blood pressure using...'), although this is not universal[3].
- Most disciplines prefer the past tense ('When the concrete cylinder was...') rather than the present tense ('When the concrete cylinder is...').
- Use inclusive language ('firefighter' rather than 'fireman').
- Use, but do not rely upon, the spell checker.
- Don't talk down to the reader.
- Omit any words, phrases, sentences that add nothing to the paper.
- Ensure correct content, grammar, spelling, punctuation, format.
- Check for commonly misused words: there/their; your/you're; affect/effect; chose/choose; etc.
- Ensure capitalisation is correct.

[3] https://theconversation.com/we-should-use-i-more-in-academic-writing-there-is-benefit-to-first-person-perspective-131898

- Use apostrophes (**not** apostrophe's!) correctly.
- Crucially, be unambiguous: Sentences should say what they mean, and should mean what they say:

> Don't write so that you can be understood; write so that you can't be misunderstood.
> — Attributed to William Howard Taft

Example 37.4 (Short sentences)**.** The first sentence should be accessible and engaging. Here is a very poor first sentence:

> Until recently, atypical hemolytic uremic syndrome (aHUS), conventionally defined in the pediatric literature as a syndrome of the triad of renal failure, microangiopathic hemolytic anemia, and thrombocytopenia without a prodrome of hemorrhagic diarrhea, has received little attention in adult practice because the patients are commonly given the diagnosis of thrombotic thrombocytopenic purpura (TTP) or TTP/HUS and treated as TTP with plasma exchange, augmented in refractory cases with rituximab and sometimes even splenectomy.
> — Tsai (2014), p. 187

Extra example: This sentence appeared in a published article (Salimirad and Srimathi 2016):

> 600 teachers, from both Government and Private Schools, have been *drowned* by random sampling.
> — Salimirad and Srimathi (2016), p. 14; emphasis added

This sentence is poor: No-one has ever been *drowned* by random sampling. Possibly, the authors mean that teachers were 'overwhelmed by participation in many research studies'...

However, later the article states:

> Using random sampling a total number of 600 teachers were selected from...
> — Salimirad and Srimathi (2016), p. 17

So the initial wording is *wrong*, and I suspect the sample probably wasn't *random* either!

37.15 Plagiarism

Research involves using other people's ideas and research to develop new conclusions, or confirm existing conclusion. All sources used when writing research should be acknowledged, otherwise you are committing plagiarism. The Macquarie Dictionary[4] defines plagiarism as

> the appropriation or imitation of another's ideas and manner of expressing them, as in art, literature, etc., to be passed off as one's own.

[4] https://www.macquariedictionary.com.au/

— *The Macquarie Dictionary*: https://www.macquariedictionary.com.au/

Plagiarism is a serious offence: theft of intellectual property. **Do not plagiarise**:

- Do not take parts of sentences or complete sentences directly from papers.
- Use quotes if necessary and cite work (sparingly).
- Plagiarism does not just apply to words and text. It also applies to images, ideas, etc.

Plagiarism can destroy people's careers and affect the reputation and status of the University. Those caught plagiarising will be penalised. Penalties can range from having to resubmit assignments and being marked down, to failing that assignment, failing the course, or (in very serious cases) expulsion from university.

Example 37.5 (Plagiarism)**.** The *Indian Journal of Dermatology* published an article discussing plagiarism, in an attempt to discourage it (Shamim 2014). Unfortunately, the article was retracted[5] because parts of the article were plagiarised:

> This article is being retracted as the manuscript has been found to be copied from [...] the dissertation entitled 'Developing a comprehensive guideline for overcoming and preventing plagiarism at the international level based on expert opinion with the Delphi method' by Dr. Mehdi Mokhtari.

Example 37.6 (Plagiarism)**.** Many examples of plagiarism in academia are given in Shahabuddin (2009).

37.16 Final comments

Finally, excellent advice, with a humorous slant, appears on the offical *plain language* website of the US government[6], including these gems:

- Avoid Alliteration. Always.
- Prepositions are not words to end sentences with.
- Avoid cliches like the plague. (They're old hat.)
- Eschew ampersands & abbreviations, etc.
- Contractions aren't necessary.
- One should never generalize.
- Be more or less specific.
- Exaggeration is a billion times worse than understatement.
- Don't repeat yourself, or say again what you have said before.
- Don't use commas, that, are not, necessary.
- Never use a big word when a diminutive alternative would suffice.
- Use youre spell chekker to avoid mispeling and to catch typograhpical errers.

[6] https://www.plainlanguage.gov/resources/humor/how-to-write-good/

- Use the apostrophe in it's proper place and omit it when its not needed.
- If you reread your work, you can find on rereading a great deal of repetition can be avoided by rereading and editing.

37.17 Quick review questions

1. What is the correct word to complete this sentence? 'The subject were told to eat [?????] snacks at about 8am.'
2. What is the correct word to complete this sentence? 'Seedlings were transplanted [?????] pots containing one of three different soils.'
3. What is the correct word to complete this sentence? 'Each kangaroos was observed for signs that [?????] tracking device caused discomfort.'
4. What is the biggest problem with this sentence? 'We took 50 samples of students; the mean age was 26.2 years.'
5. What is the biggest problem with this text? 'Subjects are not blinded. Because the subjects would clearly know they were in a study.'
6. What is the biggest problem with this text? 'The sample of pedestrians were all taken on a Thursday.'

37.18 Exercises

Selected answers are available in Sect. D.34. 73

Exercise 37.1. Consider the NHANES data again. In preparing a paper about this study, suppose Fig. 37.2 and Tables 37.1 were produced. Critique these.

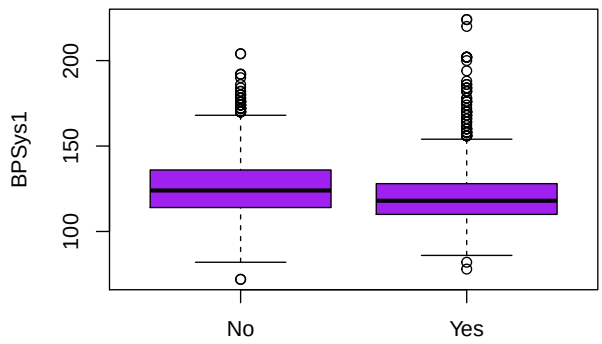


FIGURE 37.2: *A boxplot*

Exercise 37.2. In a student project at the university where I work, the students recorded the

TABLE 37.1: *A table of results*

	Mean	Std dev
Current smoker	206.6	46
Current non-smoker	214.64	48.79
Difference	8.03	
95% CI	1.25	14.8

reading speed for students reading a portion of text, and compared the reading speed for two different fonts. Their RQ was:

> Which font allows [. . .] students to read a pangram the fastest, between a default and what is considered to be a 'easy to read' font.

In their Abstract, the conclusion was:

> The Georgia font was the fastest to be read and is therefore the faster of the two.

1. Explain why this is a poorly-worded RQ. Rewrite the RQ.
2. Explain what is wrong with the conclusion given in the Abstract. Rewrite the statement.

Exercise 37.3. In a student project at the university where I work, the students compared the heights that students could jump vertically, starting from a crouch (squat) or standing (counter movement jump; CMJ) position. Every student in the study performed both jumps. Critique their numerical summary (Table 37.2).

TABLE 37.2: *The information showing how much higher the (standing) jump height is compared to the squat jump*

Sample size	Mean	Standard deviation	Standard error	Confidence interval (95%)	t value	P value
50	7.48	4.674	0.661	6.152 to 8.808	11.316	0

Exercise 37.4. In a student project, the aim was 'to determine if the proportion of males and females that use disposable cups on [the university] Campus is the same.' The two variables observed on each person in the study were:

- Whether or not the person used a disposable cup;
- The sex of the person.

In reporting the results in their Abstract, the students state:

> Based on the sample results, the 95% confidence interval for the population mean number of disposable cups used by males and females is between 0.690 and 1.625. Meaning that the population mean is likely to fall between those two intervals.

Critique this statement.

Exercise 37.5. In a student project, the aim was 'to determine if the average hang time is different between two types of paper plane designs.' The two variables in the study were:

- The plane design (Basic Dart; Hunting Flight);
- The hang time of the flight of the plane (in seconds).

In reporting the results in their Abstract, the students state:

> Very strong evidence proving a difference ($P = .000$) between the Basic Dart mean hang time (881.84 ± 140.73ms) and the Hunting Flight mean hang time (1504.19 ± 699.86ms). 95% CI for the means of The Basic Dart ($829.29 - 934.39$) and the Hunting Flight ($1242.86 - 1765.52$).

Critique this statement.

Exercise 37.6. An article (Baur et al. 2012) includes this in the Abstract:

> Cardiovascular disease (CVD) accounts for 45% of on-duty fatalities among firefighters, occurring primarily in firefighters with excess CVD risk factors in patterns resembling the metabolic syndrome (MetSyn). Additionally, firefighters have a high prevalence of obesity and sedentary behavior suggesting that MetSyn is also common. Therefore we assessed the prevalence of MetSyn in firefighters and its association with cardiorespiratory fitness (CRF) in a cross-sectional study of 957 male career firefighters.
> — Baur et al. (2012), p. 2331

1. Critique Table 37.3.
2. Critique Fig. 37.3. What would be a better graph to use?

TABLE 37.3: *The OR and 95% CI of MetSyn as a function of increasing METS and age (continuous) model 1: unadjusted, model 2 adjusted for age or cardiorespiratory fitness (CRF) (METS)*

	OR (95% CI)	p-value
Model 1		
CRF	0.691 (0.634–0.752)	<0.0001
Age	1.037 (1.020–1.055)	<0.0001
Model 2		
CRF	0.693 (0.630–0.762)	<0.0001
Age	1.002 (0.982–1.021)	0.8713

Exercise 37.7. A study (Baughman et al. 2007) gave this information:

> The aim of our study was to determine the range of 6MWD [6-minute walk distance] in an unselected group of sarcoidosis patients. We performed a prospective study of sarcoidosis patients followed up in one tertiary sarcoidosis clinic.
> — Baughman et al. (2007), p. 208

Critique the graph in Fig. 37.4 which appears in the paper.

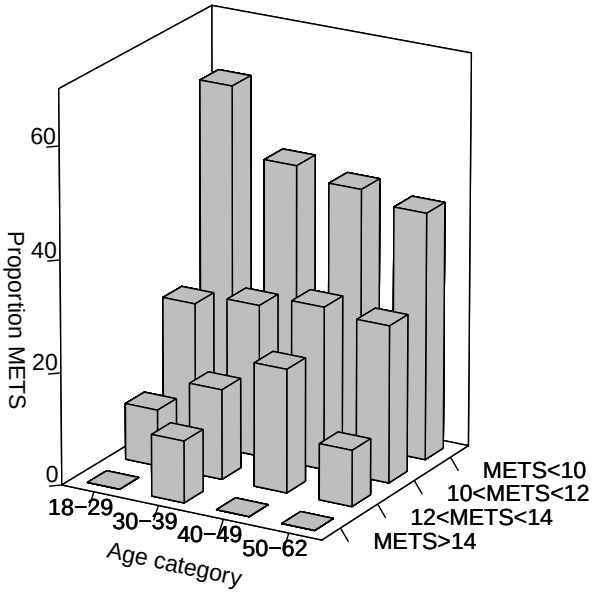


FIGURE 37.3: *A graph like that in the Baur et al. paper*

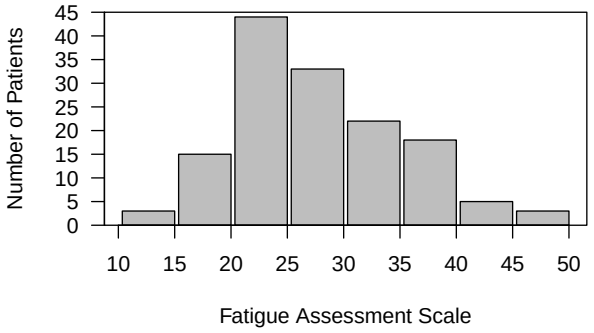


FIGURE 37.4: *A graph like that from the Baughman et al. paper*

Part X

Appendices

A

Appendix: Data sets

The online version of this book allows you to download some of the data sets used in this book.

B

Appendix: Tables

This Appendix contains tables that may be useful:

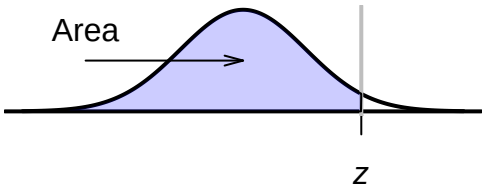
- Random numbers (Appendix B.1).
- z-tables: Find the area associated with a normal distribution *given* the z-score (Appendices B.2 and B.3).

B.1 Random numbers

TABLE B.1: *Some random numbers*

	Column I	**Column II**	**Column III**	**Column IV**	**Column V**
Row A	11842	47296	07131	54741	51478
Row B	81350	86719	04218	05743	79254
Row C	10830	29873	12359	84663	97382
Row D	06347	75554	54131	32967	28765
Row E	22802	55128	53481	70240	67505
Row F	42190	57424	07551	41161	40386
Row G	76083	58417	87965	71185	04743
Row H	96497	36007	17651	41714	61513
Row I	10860	06252	06849	98413	47644
Row J	78402	33807	76769	32977	16998
Row K	18034	38543	24810	19493	72874
Row L	31163	37163	21520	70253	52571
Row M	16265	36253	95759	62948	95499
Row N	29709	09419	15589	41775	96269
Row O	38297	97531	90757	81822	44920
Row P	14144	87467	28876	66399	06587
Row Q	31104	86099	31383	65177	82801
Row R	90954	60377	76574	11588	03820
Row S	57157	46771	09983	88908	57154
Row T	64421	05736	99976	64764	62738
Row U	24214	06705	65012	22088	28699
Row V	37532	29499	25979	61953	73180
Row W	68097	39123	74160	04245	33575
Row X	87575	49322	64087	88218	99540
Row Y	75810	71842	06844	16728	41771
Row Z	42773	21977	62587	33452	93245

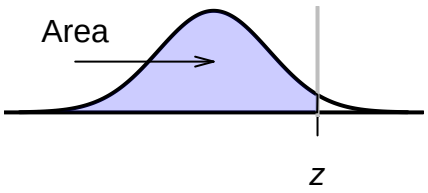
B.2 Normal distribution: negative z-values probabilities



The table gives the probability (area) that a z-score is **less** than the z-score looked up. For example: Look up $z = -1.87$; the area *less than* $z = -1.87$ is about 0.0307, or about 3.07%.

TABLE B.2: *The probability that the area is less than the value of z that is looked up*

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641



B.3 Normal distribution: positive z-values probabilities

The table gives the probability (area) that a z-score is **less** than the z-score looked up. For example: Look up $z = 1.87$; the area *less than* $z = 1.87$ is about 0.9693, or about 96.9%.

TABLE B.3: *The probability that the area is less than the value of z that is looked up*

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

C

Appendix: Symbols, formulas, statistics and parameters

Symbols used

TABLE C.1: *Some symbols used*

Symbol	Meaning	Reference
H_0	Null hypothesis	Sect. 28.2
H_1	Alternative hypothesis	Sect. 28.2
df	Degrees of freedom	Sect. 31.4
CI	Confidence interval	Chap. 21
s.e.	Standard error	Def. 18.3
n	Sample size	
χ^2	The chi-squared test statistic	Sect. 31.4

Confidence intervals

Almost all **confidence intervals** have the form

$$\text{statistic} \pm (\text{multiplier} \times \text{s.e.}(\text{statistic})).$$

Notes:

- The multiplier is *approximately* 2 for an *approximate* 95% CI (based on the 68–95–99.7 rule).
- multiplier $\times$ s.e.(statistic) is called the *margin of error*.
- Confidence intervals for *odds ratios* are slightly different, so **this formula does not apply for odds ratios**. For the same reason, a standard error for ORs is not given.

Hypothesis testing

For many **hypothesis tests**, the *test statistic* is a t-scores, which has the form:

$$t = \frac{\text{statistic} - \text{parameter}}{\text{s.e.}(\text{statistic})}$$

Notes:

- Since t-scores are a little like z-scores, the 68–95–99.7 rule can be used to *approximate* P-values.
- Tests involving *odds ratios* do not use t-scores, so **this formula does not apply for tests involving odds ratios**.
- For tests involving odds ratios, the *test statistic* is a χ^2 score and not t-score. For the same reason, a standard error for ORs is not given.
- The χ^2 statistic is approximately like a z-score with a value of (where df is the 'degrees of freedom' given in the software output):

$$\sqrt{\frac{\chi^2}{\text{df}}}.$$

	Parameter	Statistic	Standard error	S.E. formula reference
Proportion	p	$\hat{p}$	$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}}$	Def. 20.2
Mean	μ	$\bar{x}$	$\text{s.e.}(\bar{x}) = \frac{s}{\sqrt{n}}$	Def. 22.1
Standard deviation	σ	s		
Mean difference	μ_d	$\bar{d}$	$\text{s.e.}(\bar{d}) = \frac{s_d}{\sqrt{n}}$	Def. 23.2
Diff. between mean	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\text{s.e.}(\bar{x}_1 - \bar{x}_2)$	*
Odds ratio	Pop. OR	Sample OR	s.e.(sample OR)	*
Correlation	ρ	r		
Slope of regression line	β_1	b_1	$\text{s.e.}(b_1)$	*
Intercept of regression line	β_0	b_0	$\text{s.e.}(b_0)$	*
R-squared		R^2		

D

Appendix: Answers to end-of-chapter exercises

This Appendix contains answers to *most* (not all) exercises. Some are fully worked, and some are only brief solutions.

- Answers to Chap. 1 (Introduction): Sect. D.1.
- Answers to Chap. 2 (Research questions): Sect. D.2.
- Answers to Chap. 3 (Research design): Sect. D.3.
- Answers to Chap. 4 (Ethics): Sect. D.4.
- Answers to Chap. 5 (Sampling): Sect. D.5.
- Answers to Chap. 6 (Factors that influence the response variable): Sect. D.6.
- Answers to Chap. 7 (Designing experiments): Sect. D.7.
- Answers to Chap. 8 (Designing observational studies): Sect. D.8.
- Answers to Chap. 9 (Interpretation): Sect. D.9.
- Answers to Chap. 10 (Collecting data): Sect. D.10.
- Answers to Chap. 11 (Describing variables): Sect. D.11.
- Answers to Chap. 12 (Graphs): Sect. D.12.
- Answers to Chap. 13 (Numerical summaries for quantitative data): Sect. D.13.
- Answers to Chap. 14 (Numerical summaries for qualitative data): Sect. D.14.
- Answers to Chap. 15 (Making decisions): Sect. D.15
- Answers to Chap. 16 (Probability): Sect. D.16.
- Answers to Chap. 17 (Sampling distributions): Sect. D.17.
- Answers to Chap. 18 (Sampling variation): Sect. D.18.
- Answers to Chap. 20 (CIs for one proportion): Sect. D.19.
- Answers to Chap. 21 (More about forming CIs): Sect. D.20.
- Answers to Chap. 22 (CIs for one mean): Sect. D.21.
- Answers to Chap. 23 (CIs for paired data): Sect. D.22.
- Answers to Chap. 24 (CIs for two independent means): Sect. D.23.
- Answers to Chap. 25 (CIs for odds ratios): Sect. D.24.
- Answers to Chap. 27 (Tests for one mean): Sect. D.25.
- Answers to Chap. 28 (More about hypothesis tests): Sect. D.26.
- Answers to Chap. 29 (Tests for paired mean): Sect. D.27.
- Answers to Chap. 30 (Tests for two independent mean): Sect. D.28.
- Answers to Chap. 31 (Tests for odds ratios): Sect. D.29.
- Answers to Chap. 33 (Relationships between two quantitative variables): Sect. D.30.
- Answers to Chap. 34 (Correlation): Sect. D.31.
- Answers to Chap. 35 (Regression): Sect. D.32.
- Answers to Chap. 36 (Reading research): Sect. D.33.
- Answers to Chap. 37 (Writing research): Sect. D.34.

D.1 Answers: Introduction

Answers to exercises in Sect. 1.9.

Answer to Exercise 1.1: The RQ requires numerical information to be answered, such as the *average time* taken to apply the tourniquets. This RQ would be answered using a **quant**itative RQ.

Answer to Exercise 1.2: The RQ does not require numerical information to be answered. This RQ would be answered using a **qual**itative RQ.

D.2 Answers: RQs

Answers to exercises in Sect. 2.13.

Answer to Exercise 2.1: See Table D.1.

TABLE D.1: *Terms matched with their operational definitions*

Term	Definition
Rainwater	Rainwater from a rainwater collection tank on your property
Bottled	Water sold in bottles by food companies that is widely available to the public for purchase and consumption
Tap	Water you presently use throughout your dwelling (home)
Recycled	Highly purified wastewater deemed by scientists as safe for human consumption
Desalinated	Highly purified seawater deemed by scientists and public health officials as safe for human consumption

Answer to Exercise 2.2: **1.** P: University students. **2.** O: Resting diastolic blood pressure. **3.** C: between students who regularly drive to USC and those who regularly ride their bicycles. **4.** No intervention. **5.** Relational. **6.** What is meant by 'regularly'; 'university student' (on-camus and online? undergraduate *and* postgraduate? full-time and part-time?). **7.** Resting diastolic blood pressure; whether they regularly drive to university or regularly ride their bicycles.

Answer to Exercise 2.3: **1.** P: Children aged under 3 in a Peruvian peri-urban community; O: diarrohea status; C: nutritional status; No intervention. **2.** Hard to be sure; perhaps something like: 'In children aged under 3 in a Peruvian peri-urban community, is there a a relationship between diarrohea status and nutritional status?' **3.** Relational. **4.** How is 'diarrohea status' measured? Likewise, how is 'nutritional status' measured? There are probably others. **5.** Response: diarrohea status; explanatory: nutritional status.

Answer to Exercise 2.4: Recall that the outcome is used to describe a *group* (the population), not the individuals.

1. The *percentage* of vehicles that crash. **2.** The *average* jump height. **3.** The *average* number of tomatoes per plant. **4.** The *percentage* of people who own a car.

Answer to Exercise 2.5: Recall that the explanatory variable is what is actually measured on the individuals in the population.

1. The type of car fuel. **2.** The type of coffee. **3.** The dose of iron supplement. **4.** The diet.

Answer to Exercise 2.6: **1.** *Does* have a comparison (between a group of people in winter, and a different group of people in summer). The outcome is 'the percentage of people wearing hats.' **2.** *Does not* have a comparison. Two subsets of the population are not being compared: instead, each person is measured twice. So an Outcome may be 'average *change* in cholesterol levels.' **3.** *Does not* have a comparison. Two subsets of the population are not being compared: instead, each person gets two measurements. So an Outcome may be 'average *difference* between right- and left-leg balance times.' **4.** *Does* have a comparison: The three subsets of the population are being compared: the three groups of tomato plants. The Outcome is 'average yield' (which could be measured in kg/plant, tomatos/plant, kg/hectare, etc).

Answer to Exercise 2.7: The *unit of observation* is the *animal*: The animals, for example, are weighed.

The *unit of analysis* is the *pen*, as the food is allocated to the animals in the whole pen. In addition, the animals in the same pen are not independent: they compete for the same space, food, resources, and would all have similar environments that they share.

Answer to Exercise 2.8: The *population* surely is not 10 adults; that sounds like the sample. It does not make clear how many fonts are being compared (or which fonts are being used).

Perhaps try this:

> Among Australian adults, is the time taken to read a passage of text different when Arial font is used compared to Times Roman font?

Answer to Exercise 2.9: The RQ is about comparing *groups*, so it should talk about the *average* lung capacity of males and females. Perhaps:

> Of students that study at USC, Sippy Downs, do males have a larger *average* lung capacity than females?

D.3 Answers: Research designs

Answers to exercises in Sect. 3.10.

Answer to Exercise 3.1: The researchers could decide which beams go into Group A and into Group B. Researchers could also allocate treatments to the groups: they could select what treatments is applied to each group of beams. This is a *true experiment*.

Answer to Exercise 3.2: The researchers had no say in who was in hospital at the time: they could not allocate the patients to the two groups (overlay; matress). this is a *quasi-experiment*.

Answer to Exercise 3.3: **1.** *P*: Perhaps people in a suburb of the Sunshine Coast; *O*: number of doctor's visits in the next six months; *C*: between people owning a pet for those six months, and those who do not own a pet for those six months. **2.** For an experiment, we would need to *intervene* to *give* subjects a pet, or *not* give them a pet. **3.** For an observational study, we would *not* intervene: We would find the subjects who *already* owned a pet, or who did not *already* own a pet.

Answer to Exercise 3.4: **1.** *P*: A bit vague from this small extract: people of some kind; *O*: the *average change* in body weight over two years; *C*: Between the four diets; *I*: The diets seems to be have been imposed. **2.** Experimental: The diets have been *imposed* by the researchers, with the intent of changing the outcome (the weight change). **3.** Probably a true experiment. **4.** The individuals: those from whom the weight change is taken. **5.** The individuals: the diets are allocated to each individual. **6.** The *change* in body weight over two years. **7.** The type of diet.

D.4 Answers: Ethics

Answers to exercises in Sect. 4.6.

Answer to Exercise 4.1: Answers will vary.

Answer to Exercise 4.2: Answers will vary.

Answer to Exercise 4.3: Answers will vary.

D.5 Answers: Sampling

Answers to exercises in Sect. 5.14.

Answer to Exercise 5.1: A tricky thing here is that some books are not physically in the library, as they have been borrowed.

1. Simple random sample: A list of all the books held by the USC library is needed. This may be possible for a librarian (it may not be, and would be really huge), it certainly is not possible for a student or non-library stafff member. In principle though, number each book, and randomly select a sample from that list. **2.** Stratified: Use locations (Sippy Downs; Fraser Coast; Caboolture; Gympie; Southbank; SCHI) as strata, and then a random sample of all the book in each locations. **3.** Cluster: Consider each set of shelves as a cluster, and randomly select some shelves, and determine the number of pages in each book on the selected shelves. **4.** Multistage: Consider taking a random of campuses, then a random sample of the sets of shelves in the selected libraries, then selecting a random shelf from

each one, then a small number of random book from each shelf. **5.** Convenience: Finding books in the libraries within reach and easily accessible and on the shelves, **6.** Multistage perhaps.

Answer to Exercise 5.2: **1.** Multi-stage. **2.** It's a bit like stratified... but not quite. **3.** Convenience. **4.** The last is poor. The second is bit odd but is probably OK. The first might be the best.

Answer to Exercise 5.3: **1.** Convenience, but by approaching every 10th person they are trying to make it a little more representative... but they can do a lot better. **2.** Convenience, but by approaching every 5th person and going every day for a week they are trying to make it a little more representative... but they can do a lot better. **3.** Self-selecting. **4.** Convenience. At least the researcher is trying to get a more representative sample, by going every day for two weeks, and at different times and locations each week, and approaching someone every 15 minutes. **5.** The fourth is the best, but it is still far from 'random.' **6.** None.

Answer to Exercise 5.4: A bit like *cluster sampling* (randomly taking a small sample from many groups, and taking everyone (or everything) in those selected groups)... but not every person in the selected schools would respond (they would decide if they responded). A *combination* of *cluster* and *voluntary response* sampling.

D.6 Answers: Overview of internal validity

Answers to exercises in Sect. 6.8.

Answer to Exercise 6.1: Presumably *all* are extraneous variables, as all are possibly related to the response variable (incidence of depression): That is why the researchers obtained this information. *None* can be lurking variables, as the researchers measure or observe all of them.

To be a confounding variable, the extraneous variable should be related to *both* the response variable (incidence of depression) *and* the explanatory variable (diet quality). As a result, all of the extraneous variables could potentially be confounding variables.

Answer to Exercise 6.2: Response variable: something like 'risk of developing a cancer of the digestive system.' Explanatory variable: 'whether or not the participants drank green tea at least three times a week.'

Lurking variable: 'health consciousness of the participants,' because the researchers don't seem to have measured or observed this.

Answer to Exercise 6.3: Older children would probably be more likely to be smokers, and would be larger in general: age would be a confounding variable. Age is easy to record, and usually *is* recorded in these types of studies, so probably *not* a lurking variable. (The age, height and gender of each child *is* recorded.)

D.7 Answers: Designing experimental studies

Answers to exercises in Sect. 7.12.

Answer to Exercise 7.2: The observer effect. The researcher is directly contacting the subjects, so may unintentionally influence their responses.

Answer to Exercise 7.3: **1.** Randomly allocate the type of water to the subject (or the *order* in which the subjects taste-test each drink.) **2.** The subjects do not know which type of water they are drinking. **3.** The person providing the water and receiving the ratings does not know which type of water they are drinking. **4.** Hard to find a control. **5.** Any random sampling is good, if possible.

Answer to Exercise 7.4: **1.** *Response*: The amount of sunscreen used; *Explanatory*: The time spent on sunscreen application. **2.** They were looking at potential confounding variables. **3.** If the mean of both the response and explanatory variables was different for females and males, then the sex of the participant would be a *confounding* variable, and this would need to be factored into the analysis of the data. **4.** The participants are blinded to what is happening in the study.

D.8 Answers: Designing observational studies

Answers to exercises in Sect. 8.9.

Answer to Exercise 8.1: **1.** Since this is an *observational study*, we *cannot* allocate stduents to receive bottled or tap water (because then the study would be an *experimental study*). In an *experiment* we could randomly allocate students to receive *either* bottled or tap water and have them rate the taste (or even randomly allocate students to receive bottled or tap water *first*, then swap to the other type of water, and each student would then provide *two* ratings). **2.** The students would not be aware of which water they would be drinking. **3.** Neither the students nor the researchers who give the students the water would know which type of water the students are drinking. **4.** We can't really set up a control here. **5.** Any of the random sampling methods are possible, and are preferred. In practice, perhaps use a convenience sample, but try to get a sample as representative as possible (Sect. 5.9).

Answer to Exercise 8.2: Yes. Consider a study of the effect of smoking: non-smokers are the control. However, in an observational study, cases cannot be allocated to be controls.

Answer to Exercise 8.3: No. People can know they are being observed.

Answer to Exercise 8.4: The descriptions indicates that patients probably knew they were involved, so the Hawthorne effect should be considered when interpreting the results.

D.9 Answers: Interpretation

Answers to exercises in Sect. 9.7.

Answer to Exercise 9.1: Population: 'USC students on-campus.' External validity refers to whether the results apply to other members of *this* population, not to people outside this population (such as members of the general public).

Answer to Exercise 9.2: **1.** P: Aircraft passengers aged 18 and over. O: Unclear; something about 'composite of death or major traumatic injury.' C: Between wearing a parachute and wearing a backpack. I: Yes: Having participants wear the parachute or backpack. **2.** Experimental: The researchers decide if the participants use a parachute or backpack. **3.** *Explanatory*: 'whether or not a parachute is worn.' *Response*: harder to understand; is it 'whether or not the participant dies or sustains a major injury?' **4.** These results won't apply in the real world; not ecologically valid. In the real world, parachutes are used at high altitude, for example. **5.** The study is not very useful! **6.** Speaking loosely: That jumping from a small plane that is on the ground, parachutes are equally effective as backpacks in keeping people safe.

Answer to Exercise 9.3: Because the sample is not a random sample, the researchers are (rightly) noting that the results may not *generalise* to all hospitals. Because the data was only collected at night, perhaps the data is not *ecologically valid*.

D.10 Answers: Data collection

Answers to exercises in Sect. 10.5.

Answer to Exercise 10.1: People aged 18 do not have a category.

Answer to Exercise 10.2: The second. The first is *leading*: Should *concerned* dog owners...

Answer to Exercise 10.3: **1.** The phrase 'Do you agree' is leading. Placing *RIGHT DIRECTION* in capitals is leading. Besides, everyone wants their country to head in the *right* direction... but what that means varies from person to person. **2.** The phrase 'Do you agree' is leading. Phrases like *unwavering commitment*, *respect* and *incredible veterans and TROOPS* are all leading and undefined. **3.** The word *revitalize* is leading.

D.11 Answers: Describing variables

Answers to exercises in Sect. 11.5.

Answer to Exercise11.1: *Foliage biomass*: quantitative continuous. *Tree diameter* (in cm): quantitative continuous. *Age of the tree* (in years): quantitative continuous. *Origin of the tree*: Qualitative nominal.

Answer to Exercise 11.2: **1.** Systolic blood pressure: quantitative continuous. **2.** Program of enrolment: qualitative nominal. **3.** Academic grade: qualitative ordinal. **4.** Number of times people visited the doctor last year: quantitative discrete.

Answer to Exercise 11.3: **1.** Age: qualitative ordinal. **2.** Gender: qualitative nominal. **3.** Location: qualitative nominal. **4.** Social media use: qualitative ordinal. **5.** BMI: quantitative continuous. **6.** Total sitting time, in minutes per day: quantitative continuous.

Answer to Exercise 11.4: *Gender*: Qualitative nominal. *Age*: Quantitative continuous. *Height*: Quantitative continuous. *Weight*: Quantitative continuous. *GMFCS*: Qualitative ordinal.

Answer to Exercise 11.5: *Fertilizer dose*: Quantitative continuous. *Soil nitrogen*: Quantitative continuous. *Fertilizer source*: Qualitative nominal.

Answer to Exercise 11.6: *Response of kangaroos*: Qualitative ordinal. (Or perhaps nominal?) *Height of drone*: 'Height' is quantitative, but with just four values used it would probably be treated as qualitative ordinal. *Mob size*s: Quantitative discrete. *Sex*: Qualitative nominal.

Answer to Exercise 11.7: *Location* is the only variable (something observed from the *individuals*). The *number of people* and the *percentage of people* who died at each location is a *summary* of the data collected from the individuals. 'Location' is a *nominal*, *qualitative* variable, with seven *levels*.

D.12 Answers: Graphs

Answers to exercises in Sect. 12.13.

Answer to Exercise 12.1: None of them are *bad* graphs. I'd prefer the bar chart, but any are OK.

Answer to Exercise 12.2: A graph of the individual variables is always useful as a starting point: so a *bar chart* for the origin, and a *histogram* for the others.

But *relationships* are the main focus. Relationships between foliage biomass and tree origin: *boxplot*. Relationships between foliage biomass and the other variables: *scatterplot*. On the scatterplot, the different origins of the trees could be *encoded* by using different colours or plotting symbols.

Answer to Exercise 12.3: `Gender` and `GFMCS`: both qualitative; the others are quantitative. Relationships between two *quantitative* variables: use a scatterplot. Relationships between two *qualitative* variables: (say) a side-by-side bar chart. With one of each: boxplot. See Fig. D.1 for some examples.

Answer to Exercise 12.4: *Fertilizer* (quantitative): histogram (response variable). *Soil nitrogen* (quantitative): Histogram (explanatory variable). *Source* (qualitative nominal): Barchart (explanatory variable). *Relationships*: Between fertilizer dose and soil nitrogen: scatterplot. *Source* could be encoded using different *coloured* points.

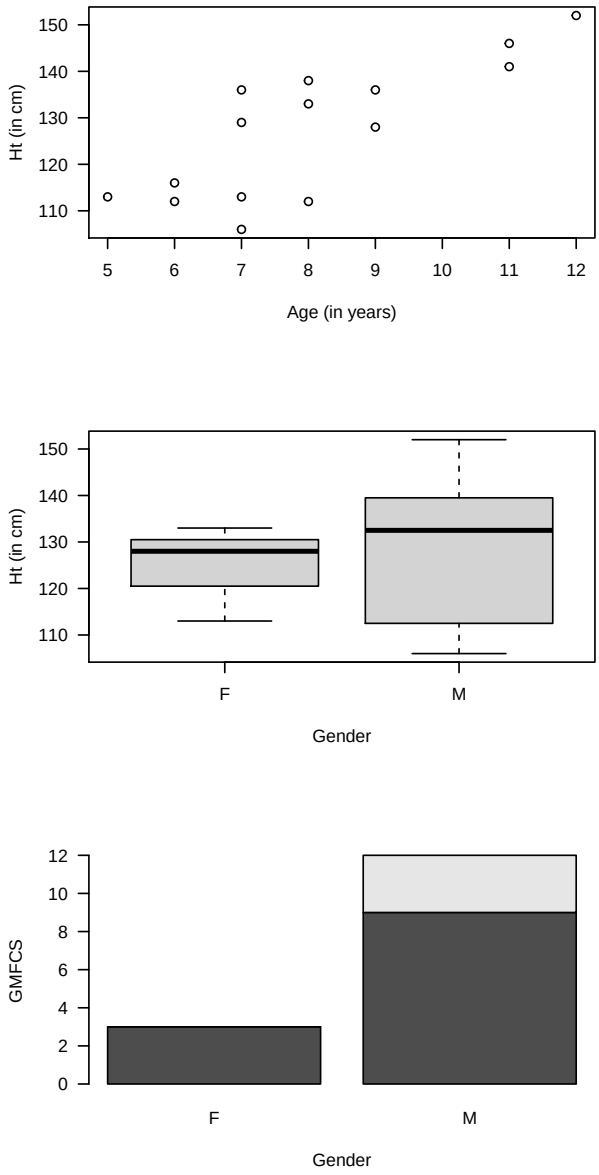


FIGURE D.1: *Some graphs from the cerebral palsy data*

Answer to Exercise 12.5: A bar chart (or dot chart). A pie chart would *not* be appropriate, as respondents could select more than one option.

Answer to Exercise 12.6: In general, female basketball players are taller than female netballer players (the first, second and third quartiles are all greater for basketball players). For the second and third quartiles, the differences look quite substantial. The minimum heights are similar.

Answer to Exercise 12.7: What do the different plotting symbols mean? The labels on the axes are not helpful. The vertical axis goes up to 35, but could easily stop at 20. See Fig. D.2.

Answer to Exercise 12.8: The graph is *inappropriate*! Both variables are qualitative, but the graph is a scatterplot (used for two *quantitative* variables). What does that plot even tell you?

A stacked or side-by-side barchart should be used (Fig. D.3).

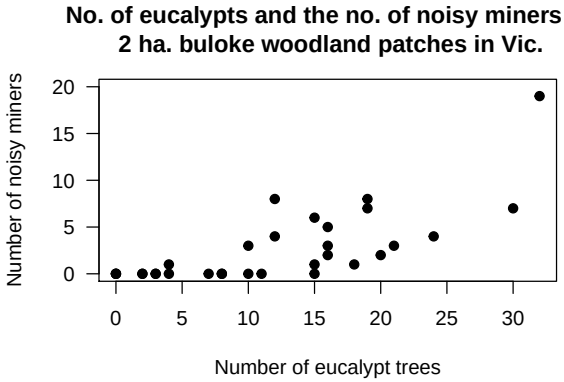


FIGURE D.2: *The number of noisy miners and the number of eucalyptus trees*

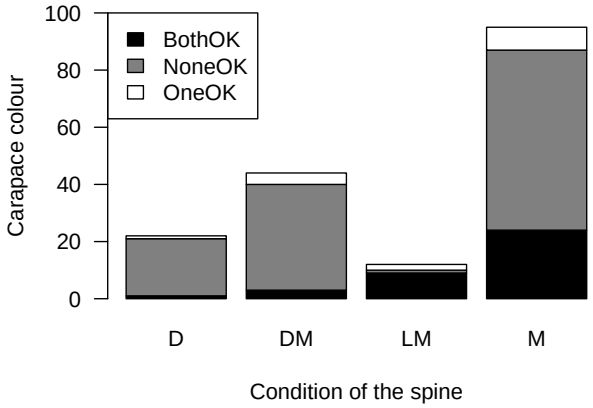


FIGURE D.3: *The colour of female horseshoe crabs and the condition of their spines. There are no missing values.*

Answer to Exercise 12.9: **1.** Response variable: *Change* in MADRS (quantitative continuous). **2.** Explanatory variable: treatment group (qualitative nominal with three levels). **3.** Response variable: Histogram. Explanatory: bar chart. Relationship: boxplot.

Answer to Exercise 12.10: See Fig. D.4.

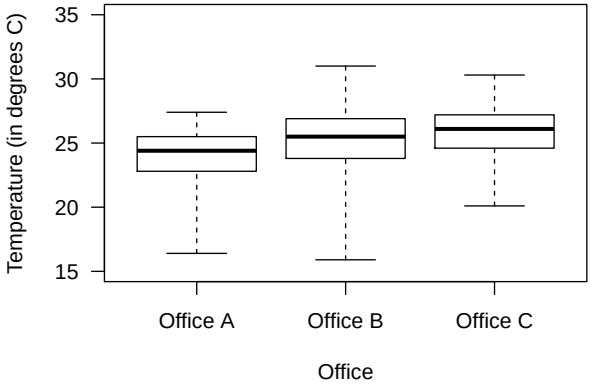


FIGURE D.4: *Boxplot of the office temperatures*

Answer to Exercise 12.11: Variable is the 'Sport' (*qualitative*). The bars can be ordered any way. *Skewness makes no sense*: It only makes sense to talk about skewness for *quantitative* variables.

D.13 Answers: Numerical summaries for quantitative data

Answers to exercises in Sect. 13.10.

Answer to Exercise 13.1: Probably the median as slightly skewed right, with some outliers. *Both* the mean and median *can* be quoted...

Answer to Exercise 13.2: **1.** Sample mean: 0.467. **2.** Sample median: 3.35. **3.** Range: 29.6 (from -19.8 to 9.8). **4.** Sample standard deviation: 10.40263. (SOI has no units of measurement.)

Answer to Exercise 13.3: **A**: II (median; IQR). **B**: I (mean; standard deviation). **C**: III (median; IQR).

Answer to Exercise 13.4: See Fig D.5. Worker 2 is faster in general (more panels installed per minute), including one *fast* outlier. Workers 1 and 3 have similar medians, but Worker 3 is more consistent (smaller IQR).

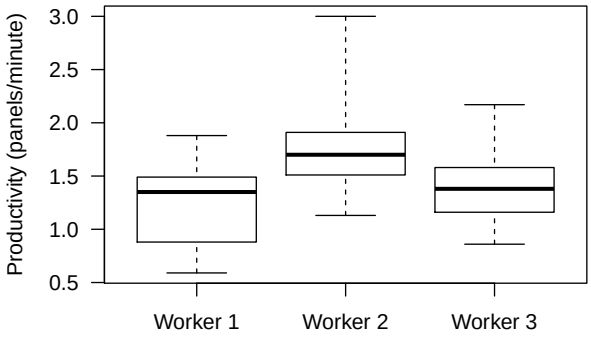


FIGURE D.5: *The boxplot for the productivity data*

D.14 Answers: Numerical summaries for qualitative data

Answers to exercises in Sect. 14.9.

Answer to Exercise 14.1: **1.** *Vomited*: 0.50 had beer then wine; 0.50 had wine only.
Didn't vomit: 0.738 had beer then wine, 0.262 had wine only. They tell us the proportion that drank various things, among those who did and didn't vomit. **2.** *Beer then wine*: 8.8% vomited and 91.2% didn't; *Wine only*: 21.4% vomited and 78.6% didn't. They tell us the percentage that vomited, for each drinking type. **3.** $(6+6)/(6+6+62+22) = 0.125$. **4.** $6/22 = 0.2727$. **5.** $6/62 = 0.096774$. **6.** $0.27272/0.096774 = 2.82$. **7.** $0.096774/0.27272 = 0.354$.

Answer to Exercise 14.2: **1.** $91/(91+188) = 0.32616$. **2.** $188/91 = 2.0659$, or about 2.07. **3.** $22/13 = 1.6923$, or about 1.69. **4.** $13/(13+22) = 0.37142$, or about 37.1%. **5.** $2.0659/1.6923 = 1.22$.

Answer to Exercise 14.3: **1.** 21/114, or about 18.4%. **2.** 14/54, or about 25.9%. **3.** 7/60, or about 11.7%. **4.** 21/93, or about 0.226. **5.** 14/40, or 0.35. **6.** 7/53, or about 0.132. **7.** 0.35/0.132, or about 2.7. **8.** The odds of no August rainfall in Emerald is 2.7 times higher in months with non-positive SOI.

Answer to Exercise 14.4: **1.** 45.9%. **2.** 61.4%. **3.** 0.848. **4.** 1.59. **5.** 1.15. **6.** 0.533. **7.** The odds of reporting back pain from carrying school bags, comparing boys to girls.

D.15 Answers: Making decisions

Answers to exercises in Sect. 15.8.

Answer to Exercise 15.1: **1.** Yes! Seems likely there is a problem (we can't be certain). **2. Assuming** the die was fair, I would not **expect** to get a **6** ten times in a row; sounds highly unusual.

Answer to Exercise 15.2: **1.** That the population mean is 12 inches, as claimed. We have no evidence to refute this claim. **2.** *First*: the *population* mean diameter is $\mu = 12$ inches; the *sample* mean is not 12 inches due to sampling variation. *Second*: the *population* mean diamter isn't 12 inches, reflected in the sample. **3.** 11.48 is 0.52 inches from the target of 12; seems unlikely that the sample mean would be that far from 12 inches through sampling variation alone. **4.** $\bar{x} = 11.25$ inches isfurther from $\mu = 12$ that $\bar{x} = 11.48$: claim probably not supported. **5.** Smaller sample sizes: sample mean would vary more (in general, larger samples give more precise estimates).

D.16 Answers: Probability

Answers to exercises in Sect. 16.8.

Answer to Exercise 16.1: **1.** *Probability* draw a King: $4/52 = 0.07692$. **2.** *Odds* draw a King: $4/48 = 0.08333$. **3.** *Probability* draw a picture card: $16/52 = 0.3077$. **4.** *Odds* draw a picture card: $16/(52 - 16) = 0.4444$. **5.** *Not* independent. This is like Example 16.10. **6.** *Are* independent. What happens on the die does not change what happens with the cards.

Answer to Exercise 16.2: Only a 50–50 chance if the events were equally likely... they clearly are not.

Answer to Exercise 16.3: **1.** $9/16$; about 56.3%. **2.** $6/57$; about 0.105. (Or, 10.5% if expressed as a percentage.) **3.** The *number* of pilots in each age group.

Answer to Exercise 16.4: **1.** *Not independent* events: If it rains, less likely to walk to work than if it doesn't rain. **2.** *Not independent* events: A smoker is far more likely to suffer from lung cancer than a non-smoker. **3.** *Independent* events: My rubbish is collected, rain or not.

Answer to Exercise 16.5: **1.** Expect $100 \times 0.99 = 99$ people to return a positive test result. **2.** Expect $100 \times (1 - 0.98) = 2$ people to return a positive test result.

A positive test result may or may not mean the person has the disease.

Answer to Exercise 16.6: The reasoning assumes that the three outcomes (HH, TT, HT) are *equally likely*, which is not true. For example, consider tossing a 20-cent coin (shown in lower-case, normal font) and a 1-dollar coin (shown in capitals, **bold** font). The *four* outcomes are: h**H**, h**T**, t**H** t**T**.

D.17 Answers: Sampling distributions

Answers to exercises in Sect. 17.12.

Answer to Exercise 17.1: **1.** $z = (8 - 8.8)/2.7 = 0.2962$, or $z = -0.30$. From tables, the probability is 0.3821, or about 38.2%. **2.** $z = 0.07$; probability is $1 - 0.52379 = 0.4721$, or about 47.2%. **3.** The z-scores are $z_1 = -0.67$ and $z_2 = 0.44$; the probability is $0.6700 - 0.2514 = 0.4186$, or about 41.9%. (Draw a diagram!) **4.** Using the tables 'backwards': z-score is about 1.04; corresponding tree diameter is $x = 8.8 + (1.04 \times 2.7) = 11.608$, or about 11.6 inches. About 15% of tress will have diameters larger than about 11.6 inches.

Answer to Exercise 17.2: **1.** $z = (39 - 40)/1.64 = -0.6097561$, or $z = -0.61$. Using tables: probability *less than* this value of z is 0.2709, so the answer is $1 - 0.2709 = 0.7291$, or about 72.9%. **2.** $z = (37 - 40)/1.64 = -1.83$; probability is 0.0336, about 3.4%. **3.** The two z-scores: $z_1 = -4.878$ and $z_2 = -1.83$. Drawing a diagram, probability is $0.0336 - 0 = 0.0336$, or about 3.4%. **4.** The z-score: 1.64 (or 1.65). Gestation length: $x = 40 + (1.64 \times 1.64) = 42.7$ (same answer to one decimal place using $z = 1.65$). 5% of gestation lengths *longer* than about 42.7 weeks. **5.** z-score is -1.64 (or -1.65). Gestation length: $x = 40 + (-1.64 \times 1.64) = 37.3$ (same answer to one decimal place using $z = -1.65$). 5% of gestation lengths *shorter* than about 37.3 weeks.

Answer to Exercise 17.3: z-score: about $z = 2.05$. Corresponding IQ: $x = 100 + (2.05 \times 15) = 130.75$. An IQ greater than about 130 is required to join Mensa.

Answer to Exercise 17.4: An IQ score lower than about 80.8 leads to a rejection by the US military.

Answer to Exercise 17.5: **1:** C; **2:** A; **3:** B; **4:** D.

Answer to Exercise 17.6: **1:** A; **2:** C; **3:** B; **4:** D.

Answer to Exercise 17.7: Be *very* careful: work with the *number of minutes from the mean, or from 5:30pm*. The standard deviation already is in decimal, but converted to minutes, standard deviation is 120 minutes, plus $0.28 \times 60 = 16.8$ minutes. The standard deviation is 136.8 minutes.

1. 9pm is 3 hours and 30 minutes from 5:30pm: 210 minutes. z-score: $z = (210 - 0)/136.8 = 1.54$; probability: $1 - 0.9382 = 0.0618$, or about 6.2%. **2.** $z = (5 - 5.5)/2.28 = -0.22$; probability: 0.4129$, or about 41.3%. **3.** z-scores are $z_1 = -0.22$ and $z_2 = 0.22$; probability: $0.5871 - 0.4129 = 0.1742$, or about 17.4%. **4.** z-score is 0.52; time is $x = 0 + (0.52 \times 136.8) = 71.136$ minutes after 5pm; about one hour and 11 minutes after 5:30pm, or 6:41pm. **5.** z-score: -1.04; time is $x = 0 + (-1.04 \times 136.8) = -141.272$, or 141.272 minutes *before* 5pm; about two hours and 21 minutes before 5:30pm, or 3:09pm.

D.18 Answers: Sampling variation

Answers to exercises in Sect. 18.8

Answer to Exercise 18.1: **1.** Standard deviation. **2.** Standard error. More specifically, the standard error of the mean. **3.** Standard deviation. **4.** Standard error. More specifically, the standard error of the proportion.

Answer to Exercise 18.2: **1.** No: Population proportions don't vary from sample to sample. **2.** Yes: varies from sample to sample. **3.** Yes: varies from sample to sample. **4.** Yes: varies from sample to sample. **5.** No: Population odds don't vary from sample to sample.

Answer to Exercise 18.3: The *standard error of the mean* is used to describe how much the sample mean is likely to vary from sample to sample. Alternatively, it describes how precisely the sample mean is estimating the (unknown) population mean.

D.19 Answers: CIs for one proportion

Answers to exercises in Sect. 20.11.

Answer to Exercise 20.1: $\hat{p} = 2182/6882 = 0.317059$ and $n = 6882$. So:

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{0.317059 \times (1 - 0.317059)}{6882}} = 0.005609244.$$

The CI is $0.317059 \pm (2 \times 0.005609244)$, or 0.317059 ± 0.01121849.

Rounding sensibly: 0.317 ± 0.011 (notice we keep lots of decimal places in the *working*, but round the final answer).

Answer to Exercise 20.2: $\hat{p} = 8/154 = 0.05194805$; $\text{s.e.}(\hat{p}) = 0.0017833$; approximate 95% CI is $0.05194 \pm (2 \times 0.0017833)$, or 0.0519 ± 0.0358, equivalent to 0.016 to 0.088. The CI is statistically valid.

Answer to Exercise 20.3: Use $\hat{p} = 708/864 = 0.8194444$ and $n = 864$. Standard error: $\text{s.e.}(\hat{p}) = 0.01308604$; approximate 95% CI is $0.8194444 \pm (2 \times 0.01308604)$. The CI is statistically valid.

Answer to Exercise 20.4: **1.** Approximately $n = 1/(0.05^2) = 400$. **2.** Approximately $n = 1/(0.025^2) = 1600$. **3.** To *halve* the width of the interval, *four* times as many people are needed.

Answer to Exercise 20.5: After 3000 hours: $\hat{p} = 0.2143$; $\text{s.e.}(\hat{p}) = 0.06331$. The CI is from 0.088 to 0.341. The statistical validity conditions are satisfied.

After 400 hours: $\hat{p} = 0$; $\text{s.e.}(\hat{p}) = 0$. The CI is from 0 to 0: clearly silly (implies no sampling variation). This is because the statistical validity conditions are **not** satisfied.

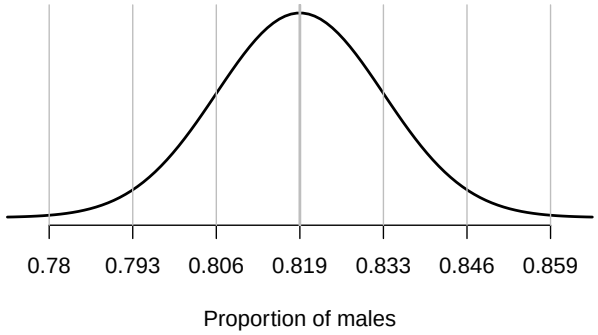


FIGURE D.6: *The sampling distribution of the proportion of males in samples of 864 people with hiccups*

D.20 Answers: More about formings CIs

Answers to exercises in Sect. 21.5.

Answer to Exercise 21.1: The conclusion states that the interval is one in which they are reasonably sure (i.e., 95% sure) that the *sample* proportion will lie. But the researcher knows *exactly* what the sample proportion is: it is $\hat{p} = 0.314$.

CIs give intervals in which we are reasonably certain that the *population* value is within, because the population proportion is unknown.

(In addition, the CI is a 68% anyway, not a 95% CI as claimed.)

Answer to Exercise 21.2: The CI is not about *individual* trees; it is about a sample statistic. Presumably, it should read something like 'This CI means that between 22.3% and 40.5% of trees are infected with apple scab.'

D.21 Answers: CIs for one mean

Answers to exercises in Sect. 22.9.

Answer to Exercise 22.1: Standard error: s.e. $= s/\sqrt{n} = 0.43/\sqrt{45} = 0.06410062$ (keeping lots of decimal places in the working). *Approximate* 95% CI: $2.85 \pm (2 \times 0.06410062)$, or 2.85 ± 0.1282012, or from 2.72 litres to 2.98 litres.

Answer to Exercise 22.2: Standard error: s.e. $= s/\sqrt{n} = 7571.74/\sqrt{58} = 994.2182$ (keeping lots of decimal places in the working). *Approximate* 95% CI is: 4967.984 micrograms to 8944.86 micrograms.

Answer to Exercise 22.3: Approximate 95% CI for the mean brushing time: 29.9 seconds to 36.1 seconds.

Answer to Exercise 22.4: **1.** Standard error: s.e.$(\bar{x}) = 651.1/\sqrt{199} = 46.15526$; approximate 95% CI: 754.1ml to 938.7ml. **2.** They don't seem very good at estimating (the article reports that the guesses ranged from 50ml to 3000ml). **3.** The sample size is much larger than 25; the CI should be statistically valid. **4.** Using the margin-of-error as 50, and $s = 651.1$:

$$\left(\frac{2 \times 651.1}{50}\right)^2 = 678.2899.$$

We would need about 679 participants (remembering to round *up*).

5. Using margin-of-error as 25, and $s = 651.1$:

$$\left(\frac{2 \times 651.1}{25}\right)^2 = 2713.16.$$

Need about 2714 participants (remembering to round *up*). **6.** To *halve* the width of the margin of error, *four* times as many subjects are needed.

Answer to Exercise 22.5: *None* of these interpretations are acceptable. **1.** CIs are not about how individual observations vary; they are about how a *statistic* varies (in this case, the sample mean). In addition, CIs are about populations and not samples. **2.** CIs are not about how individual observations vary; they are about how a *statistic* varies (in this case, the sample mean. **3.** This doesn't make sense: *samples* can't vary between two values. Sample *statistics* vary. In addition, CIs are about populations, not samples. **4.** This doesn't make sense: *populations* can't vary between two values. Even population *parameters* don't vary. **5.** The population parameter does not vary. It is a fixed (but unknown) value to be estimated. (If the value of the population mean was constantly changing, it would be very hard to estimate...) **6.** We know *exactly* what the sample mean is ($\bar{x} = 1.3649$mmol/L: We don't need a interval for the sample mean. **7.** We know *exactly* what the sample mean is ($\bar{x} = 1.3649$mmol/L: We don't need a interval for the sample mean.

Answer to Exercise 22.6: *Neither* is correct. To learn about the variation in *individuals* trees, use the *standard deviation* rather than the standard error. The standard error tells us about the sample mean diameter, not about individual trees.

D.22 Answers: CIs for paired data

Answers to exercises in Sect. 23.12.

Answer to Exercise 23.1: Mean of the *differences*: 5.2; standard error 3.6. *Approximate* 95% CI: $5.2 \pm (2 \times 3.06)$, or 5.2 ± 6.12, from -0.92 to 11.22. Mean taste preference between preferring it better *with* dip by up to 11.2mm on the 100mm visual analog scale, or preferring it *without* dip by a little (up to -0.9mm on the 100mm visual analog scale. (Understanding how the *differences* are defined is needed to understand where this came from.)

A useful summary might be like Table D.2.

Answer to Exercise 23.2: **1.** Computing differences as `Before` minus the `After` measurements seems sensible: the average blood pressure *decrease*, the purpose of the drug. **2.** The differences (when

TABLE D.2: *A numerical summary for the brocilli data*

	Mean	Standard deviation	Standard error
Raw	56	26.6	2.64679892595857
With dip	61.2	28.7	2.85575673590267
Differences	5.2		3.06

defined as*reductions*): 9, 4, 21, 3, 20, 31, 17, 26, and so on. **3.** Mean difference: 18.933; standard deviation: 9.027; standard error: $9.027/\sqrt{15} = 2.331$. *Approximate* 95% CI: 14.271 to 23.56 mm Hg. **4.** Exact 95% CI: 13.934 to 23.93 mm Hg from output. **5.** The first uses *approximate* multipiers. The second uses exact multipliers.

Answer to Exercise 23.3: **1.** *Approximate* 95% CI for *reduction*: $0.66 \pm (2 \times 0.37)$, or -0.08 to 1.4: average could be an *increase* of up to 0.08 to a *reduction* of up to 1.4 on the given scale for women. **2.** Sample size is not larger than 25, but close: probably reasonably statistically valid.

D.23 Answers: CIs for two means

Answers to exercises in Sect. 24.13.

Answer to Exercise 24.1: **1.** Table D.3. **2.** From SPSS, *exact* 95% CI: 0.05438 to 0.11501 (bottom row). Exact 95% CI for the difference between the mean direct HDL cholesterol concentrations: 0.05438 to 0.11501 mm Hg higher for non-smokers.

TABLE D.3: *A summary table for the NHANES data; statistics in mm Hg*

	Mean	Standard deviation	Standard error	Sample size
Non-smokers	1.3924	0.42792	0.01048	1668
Smokers	1.3077	0.42353	0.01137	1388
Differences	0.0847		0.01546	

Answer to Exercise 24.2: **1.** Placebo group: $3.62/\sqrt{176} = 0.2728678$ days; echinacea group $3.31/\sqrt{183} = 0.2446822$ days. **2.** $0.53 \pm (2 \times 0.367)$, or -0.204 to 1.264 days. **3.** Placebo *minus* echinacea: the difference between the means show how much *longer* symptoms last with placebo, compared to echinacea. **4.** $6.34 \pm (2 \times 0.2446822)$, or 5.85 to 6.83 days. **5.** Sample sizes are large, so the CIs statistically valid.

The difference between the means is an average of 0.53 days; about half a day (quicker on echinacea). Probably not that important when a cold last for almost seven days.

Answer to Exercise 24.3: **1.** Exercise group: $1.4/\sqrt{10} = 0.4427189$; splinting group: $1.1/\sqrt{10} = 0.3478505$. **2.** Splinting *minus* exercise: the difference are how much greater the pain is with splinting. **3.** $0.3 \pm (2 \times 0.563$, or from -0.826 to 1.426: 0.826 greater pain with exercise to 1.426 greater pain with splinting. **4.** $1.1 \pm (2 \times 0.3478505)$: from 0.404 to 1.796. **5.** Sample sizes are small; CIs may not be statistically valid, roughly correct only.

D.24 Answers: CIs for odds ratios

Answers to exercises in Sect. 25.9.

Answer to Exercise 25.1: **1.** $99/62 = 1.596774$; about 1.60. **2.** $216/115 = 1.878261$; about 1.88. **3.** $1.596774/1.878261 = 0.850$, as in the output. **4.** A few ways; for example: For every 100 men with a smooth scar, about 85 women with a smooth scar. **5.** (Graph not shown, but use a stacked or side-by-side bar chart.) **6.** Table D.4. **7.** Exact 95% CI for the OR, from the output: 0.576 to 1.255. **8.** If study repeated study many times (with the same numbers of men and women), about 95% of the CIs would contain population OR. In practice: population OR is probably between 0.576 and 1.255.

TABLE D.4: *The odds and percentage of having smooth scars, for women and men*

	Odds with smooth scars	Percentage with smooth scars	Sample size
Women:	1.60	61.5%	161
Men:	1.88	65.3%	331
Odds ratio:	0.850		

Answer to 25.2: The output can be interpreted in one of two ways (Sect. 25.2):

- Odds are the odds of swimming at the beach; OR compares these odds between those without an ear infection, to those with an ear infection.
- Odds are the odds of *not* having an ear infection; OR compares these odds for beach swimmers to non-beach swimmers.

Answer to 25.3: **1.** Table D.5. **2.** Table D.6. **3.** OR: Odds of a 1800-hr turbine getting a fissure is 0.389 times the odds of a 3000-hr turbine getting a fissure. **4.** CI from 0.133 to 1.14. Plausible values for the population OR that may have produced the sample OR likely to be between these values.

TABLE D.5: *The number of fissures for two sets of turbines, run for different numbers of hours*

	Fissures	No fissures	Total
About 1800 hours	7	66	73
About 3000 hours	9	33	42
Total	16	106	122

TABLE D.6: *The numerical summary for the fissures data*

	Odds with fissures	Percentage with fissures	Sample size
About 1800 hours	0.1061	9.59%	73
About 3000 hours	0.2727	21.43%	42
Odds ratio	0.389		

Answer to Exercise 25.4: Odds of no rainfall (non-positive SOI): $14/40 = 0.35$. Odds of no rainfall (negative SOI): $7/53 = 0.1320755$. Required OR is $0.35/0.1320755 = 2.65$, as in output. 95% CI from 0.979 to 7.174.

Answer to Exercise 25.5: The 95% CI is from 0.151 to 0.408. The OR of not wearing a hat, comparing males to females (males*less* likely to be *not* wearing a hat; rewording, males *more* likely to be wearing a hat).

D.25 Answers: Tests for one mean

Answers to exercises in Sect. 27.13.

Answer to Exercise 27.1: **1.** H_0: $\mu = 7725$; H_1: $\mu \neq 7725$ (two tailed). **2.** $\bar{x} = 6753.64$ and s.e.$(\bar{x}) = s/\sqrt{n} = 1142.123/\sqrt{11} = 344.363$. **3.** $t = (6753.64 - 7725)/344.363 = -2.821$, as in output. This 'large'; expect small P-value; software confirms this: two-tailed $P = 0.018$. **4.** Moderate evidence ($P = 0.018$) that the *mean* energy intake is not meeting the recommended daily energy intake (mean: 6753.6kJ; std. dev.: 1142.1kJ).

Answer to Exercise 27.2: H_0: $\mu = 120$ and H_1: $\mu \neq 120$ (two-tailed), where μ is the mean time *in seconds*. Standard error: s.e.$(\bar{x}) = 23.8/\sqrt{85} = 2.581472$. t-score: $(60.3 - 120)/2.581472 = -23.13$, which is *huge*; P-value will be *really* small. *Very* strong evidence ($P < 0.001$) that children do not spend 2 minutes (on average) brushing their teeth (mean: 60.3s; std. dev.: 23.8s).

Answer to Exercise 27.3: H_0: $\mu = 50$ and H_1: $\mu > 50$ (one-tailed), where μ is the mean mental demand. Standard error: s.e.$(\bar{x}) = 22.05/\sqrt{22} = 4.701076$. t-score: $(84 - 50)/4.701076 = 7.23$, which is very large; P-value will be very small. *Very* strong evidence ($P < 0.001$) that the mean mental demand is *greater* than 50. (Notice we say *greater* than, because of the RQ and the alternative hypothesis.)

Answer to Exercise 27.4: Physical:* $t = -1.28$; Mental:* $t = 1.80$. The P-values both larger than 5%. No evidence that the mean score for patients is different than the general population score.

Answer to Exercise 27.5: H_0: $\mu = 12$ and H_1: $\mu \neq 12$ (two-tailed), where μ is the mean weight in grams. Standard error: s.e.$(\bar{x}) = 0.60652/\sqrt{43} = 0.09249343$. t-score: $(14.9577 - 12)/0.09249343 = 31.98$, which is **huge**; P-value will be very small. *Very* strong evidence ($P < 0.001$) that the mean weight of a Fun Size Cherry Ripe bar is not 12 grams (mean: 14.9577; std. dev.: 0.067g), and they may be larger.

Answer to Exercise 27.6: H_0: $\mu = 1000$ and H_1: $\mu \neq 1000$, where μ is the population mean guess of the spill volume. Standard error: 46.15526. t-score: $(846.4 - 1000)/46.15526 = -3.33$, which is very large (and negative), so the P-value will be very small. Very strong evidence that the mean guess of blood volume is not 1000,ml, the actual value. The sample is much larger than 25: the test is statistically valid.

Answer to Exercise 27.7: Hypotheses have the form H_0: $\mu =$ pre-determined target, and H_1: $\mu \neq$ pre-determined target. t-scores: $t_1 = 0.318$, $t_2 = 2.347$, $t_3 = -0.466$, $t_4 = -0.726$. P-values will be large, except for second test. No evidence that the instruments are dodgy, except perhaps for the first instrument for mid-level LH concentrations. Should be statistically valid.

While assessing the means is useful, how *variable* the measurements are is also useful (but beyond us).

D.26 Answers: More about hypothesis tests

Answers to exercises in Sect. 28.12.

Answer to Exercise 28.1: Using the 68–95–99.7 rule: **1.** Very small; certainly less than 0.003 (99.7% between -3 and 3). **2.** Very small; bit bigger than 0.003 (99.7% between -3 and 3). **3.** Large-ish: Between 0.32 (68% between 1 and -1) and 5% (95% between -1 and 1), but closer to 0.32. **4.** Bit smaller than 0.32 (68% between -1 and 1). **5.** *Very* large! Almost 0.50. **6.** *Very* small; *much* smaller than 0.003.

Answer to Exercise 28.2: The answers are *half* the values given to Exercise 28.1. Using the 68–95–99.7 rule: **1.** Very small; certainly less than 0.0015 (99.7% between -3 and 3). **2.** Very small; bit bigger than 0.0015 (99.7% between -3 and 3). **3.** Large-ish: Between 0.16 (68% between 1 and -1) and 2.5% (95% between -1 and 1), but closer to 0.16. **4.** Bit smaller than 0.16 (68% between -1 and 1). **5.** *Very* large! Almost 0.25. **6.** *Very* small; *much* smaller than 0.0015.

Answer to Exercise 28.3: Using the 68–95–99.7 rule, the P-value is *just under* 0.05, and hence 'small' If the t-score was 0.0499, the P-value would be *just larger* than 0.05 and hence 'big.'

The difference between 0.0501 and 0.0499 is trivial though... it is silly to jump from 'evidence supports the alternative hypothesis!' to the complete opposite conclusion 'evidence doesn't support the alternative hypothesis!' over such a minor difference.

Answer to Exercise 28.4: **1.** Hypotheses are about *population* parameters like μ, not *sample* statistics like $\bar{x}$. **2.** Hypotheses are about parameters like μ, not statistics like $\bar{x}$. The value of 36.8051 is a sample mean, but hypothesis are meant to be written *before* the data are collected. In any case, these hypotheses are asking to test if the *sample* mean is 36.8051... which we *know* it is. **3.** 36.8051 is a sample mean, but hypothesis are meant to be written down *before* the data are collected. **4.** 36.8051 looks like a sample mean, but hypothesis are meant to be written down *before* the data are collected. **5.** Hypotheses are about parameters like μ, not statistics like $\bar{x}$. **6.** This would be fine, if the RQ was one-tailed... but it is two-tailed.

Answer to Exercise 28.5: **1.** The conclusion is about the **mean** energy intake (*population* mean energy intake specifically).
2. Conclusions are *never* about sample statistics. We want to know what the *statistic* that tells us about the *population* parameter **3.** The conclusion is about the population **mean** energy intake.

Answer to Exercise 28.6: **1.** Slight evidence of a difference in lifetime between the two brands. **2.** No. The difference is 0.29 hours, or about 17 minutes. A difference of 17 minutes in over 5 hours of use is trivial. **3.** Conclusion: little evidence of a difference between the mean lifetimes. That's cumbersome for advertising. A common advertising trick: "No other battery lasts longer!"... meaning there is no evidence of a difference in means. **4.** Price!

D.27 Answers: Tests for paired means

Answers to exercises in Sect. 29.12.

Answer to Exercise 29.1: H_0: $\mu_d = 0$ and H_1: $\mu_d > 0$: *differences* are positive when the dip rating is better than the raw rating. $t = (5.2 - 0)/3.06 = 1.699$; the *approximate* one-tailed P-value, from using the 68–95–99.7 rule, is somewhere between 16% and 2.5%. So we cannot be sure if the P-value is larger than 0.05... but it is likely that it is (since the calculated t-score is quite a long distance from $z = 1$). the evidence *probably* doesn't support the alternative hypothesis.

Answer to Exercise 29.2: **1.** Because it is the bloood pressure *reduction*, and a reduction is what the drug is meant to produce, so expect the reductions to be positive nunmbers. **2.** Differences shown below. **3.** Histogram of *differences*: Fig. D.7. **4.** H_0: $\mu_d = 0$ and H_1: $\mu_d > 0$ (because the differences are *reductions*). **5.** $t = 8.12$. **6.** $P = 0.001 \div 2 = 0.0005$ (one-tailed test). **7.** Very strong evidence ($P = 0.0005$) that the drug reduces the average systolic blood pressure (mean reduction: 8.6 mm Hg) in the population.

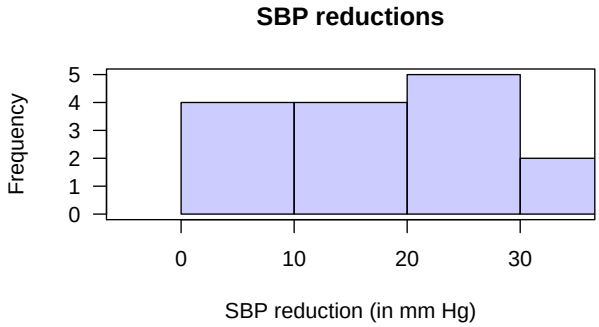


FIGURE D.7: *A histogram of the systolic blood pressure reductions (in mm Hg)*

Answer to Exercise 29.3: H_0: $\mu_d = 0$ and H_1: $\mu_d > 0$, where *differences* are positive when the intention to smoke is reduced after exercise. $t = (0.66 - 0)/0.37 = 1.78$; P-value larger than 0.05: the evidence doesn't support the alternative hypothesis. No evidence ($P > 0.05$) that the mean 'intention to smoke' reduced after exercise in women (mean change in intention to smoke: -0.66; std. error: 0.37).

Answer to Exercise 29.4: H_0: $\mu_d = 0$ and H_1: $\mu_d > 0$, where *differences* refer to the *reduction* in ferritin. $\bar{d} = -424.25$ and $s = 2092.693$ and $n = 20$, so $t = -0.90663$. t is 'small'; $P > 0.05$ (actually $P = 0.376$): the evidence doesn't support the alternative hypothesis. Since $n < 25$, the test may not be statistically valid (the histogram of data (Fig. D.8) suggests that the population *might* have a normal distribution), though the P-value is very large so it probably makes little difference.

D.28 Answers: Tests for two means

Answers to exercises in Sect. 30.12.

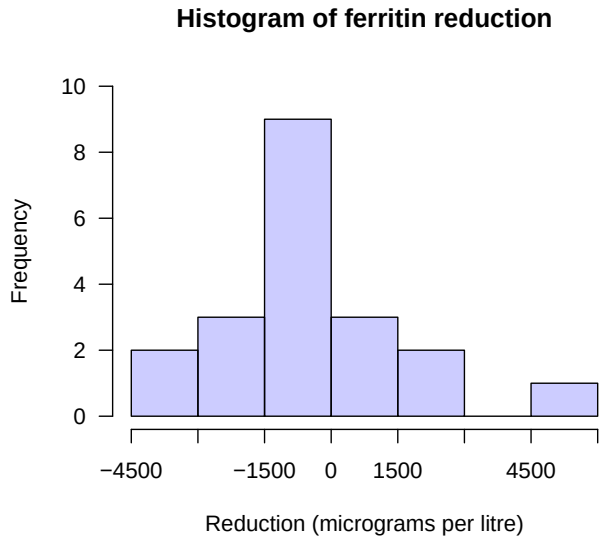


FIGURE D.8: *A histogram of the change in ferritin concentration*

Answer to Exercise 30.1: H_0: $\mu_S - \mu_{NS} = 0$ and H_1: $\mu_S - \mu_{NS} \neq 0$. From output: $t = 5.478$ and $P < 0.001$. Very strong evidence to support H_1.

Answer to Exercise 30.2: **1.** Table D.7. **2.** H_0: $\mu_C - \mu_{SO} = 0$ and H_1: $\mu_C - \mu_{SO} \neq 0$. Then $t = ((51 - 56) - 0)/3.3044 = -1.513$; P-value larger than 5%. Sample size are small; test may not be statistically valid. **3.** H_0: $\mu_C - \mu_{SO} = 0$ and H_1: $\mu_C - \mu_{SO} \neq 0$. Then $t = ((36 - 47) - 0)/4.0689 = -2.70$; P-value smaller than 5%. Sample size are small; the test may not be statistically valid.

TABLE D.7: *The physical profile of conventional and special operation paramedics in Western Australia*

	Conventional	**Special Operations**
Sample size	11	11
Grip strength (in kg)		
Mean	51	56
Standard deviation	8	9
Standard error	1.86	2.71
Push-ups (per minute)		
Mean	36	47
Standard deviation	10	11
Standard error	2.36	3.3

Answer to Exercise 30.3: H_0: $\mu_M - \mu_F = 0$ and H_1: $\mu_M - \mu_F \neq 0$. From output, $t = -2.285$; (two-tailed) P-value is 0.024 Moderate evidence to support H_1: Moderate evidence ($P = 0.024$) that the mean internal body temperature is different for females (mean: 36.886°C) and males (mean: 36.725°C).

The difference between the means, of 0.16 of a degree, is hardly of any *practical* importance in everyday use.

Answer to Exercise 30.4: **1.** H_0: The means are equal: $\mu_I = \mu_{NI}$ or $\mu_I - \mu_{NI} = 0$. H_1: The means are not equal: $\mu_I \neq \mu_{NI}$ or $\mu_I - \mu_{NI} \neq 0$. **2.** CI from -22.54 to -11.95: the mean sugar consumption between 11.95 and 22.54 kg/person/year *greater* in industrialised countries. **3.** Very strong evidence in the

sample ($P < 0.001$) that the mean annual sugar consumption per person is different for industrialised (mean: 41.8 kg/person/year) and non-industrialised (mean: 24.6 kg/person/year) countries (95% CI for the difference 11.95 to 22.54).

D.29 Answers: Tests for odds ratios

Answers to exercises in Sect. 31.13.

Answer to Exercise 31.1: The missing entries: Odds: 1.15; Percentage: 58.1%.

$\chi^2 = 4.593$; approximately $z = \sqrt{4.593/1} = 2.14$; expect small P-value. Software gives $P = 0.032$. Evidence that the difference between the sample proportions is unlikely to be due to sampling variation. The test is statistically valid.

> The sample provides *moderate evidence* (chi-square $= 4.593$; two-tailed $P = 0.032$) that the *population* odds of finding a male sandfly in eastern Panama is different at 3 feet above ground (odds: 1.15) compared to 35 feet above ground (odds: 1.71; OR: 0.67; 95% CI from 0.47 to 0.97).

Answer to Exercise 31.2: One option: H_0: The population OR is one; H_1: The population OR is not one. From software, $\chi^2 = 0.667$; $P = 0.414$, which is large. No evidence ($P = 0.414$) that the odds of having a smooth scar is different for women and men (chi-square: 0.667). The test is statistically valid.

Answer to Exercise 31.4: One option: H_0: The population OR is one; H_1: The population OR is not one. From software, $\chi^2 = 3.845$; $P = 0.050$. Moderate evidence ($P = 0.05$) that the odds of having no rainfall is different for non-positive SOI Augusts and negative-SOI Augusts (chi-square: 3.845). The test is statistically valid.

Answer to Exercise 31.5: **1.** $22/366 \times 100 = 6.0\%$. **2.** $79/386 \times 100 = 20.5\%$. **3.** $22/344 = 0.06395349$, or about 0.0640. **4.** $79/307 = 0.257329$, or about 0.257. **5.** $0.257/0.0640 = 4.02$. **6.** $0.0640/0.257 = 0.249$. **7.** From 0.151 to 0.408. **8.** $\chi^2 = 33.763\%$ (approximately $z = 5.81$) and $P < 0.001$. **9.** Strong evidence ($P < 0.001$; $\chi^2 = 33.763$; $n = 752$) that the odds of wearing hat is different for males (odds: 0.257) and females (odds: 0.0640; OR: 0.249, 95% CI from 0.151 to 0.408). **10.** Yes.

Answer to Exercise 31.6: **1.** Low exposure (in order): 73.7%, 72.5%, 85.6%. High exposure (in order): 26.3%, 27.5%, 14.4%. **2.** In order: 2.80, 2.64, 5.92. **3.** Various ways; probably the easiest: H_0: No association between level of exposure and type of interaction (in the population). **4.** Table D.8. **5.** Approximately $z = \sqrt{20.923/2} = 3.23$: expect small P-value. **6.** Very strong evidence in the sample of an association between level of exposure and type of interaction in the population ($\chi^2 = 20.923$; $P < 0.001$).

TABLE D.8: *The expected counts for the phone-use data*

	Answer call	Respond to text	Reply to email
Low exposure	275.7	275.7	272.61
High exposure	81.3	81.3	80.39

D.30 Answers: Relationships between two quantitative variables

Answers to exercises in Sect. 33.6.

Answer to Exercise 33.1: Linear, positive, very little variation (i.e., strong relationship). Of note: The observation in the bottom left is very different from the rest of the data, but still maintains the linear relationship.

Answer to Exercise 33.2: Non-linear; higher wind speed related to higher DC output (in general); a small to moderate amount of variation. The DC output increases as wind speed increases, but not linearly.

Answer to Exercise 33.3: The relationship probably linear... but a few observations at the top right look a bit different. Variation seems to increase a little as the Age increases. Of note:: A few observations in the top right of the scatterplot seem to not follow the linear relationship.

Answer to Exercise 33.4: Approximately linear; positive relationship; variation seems to get larger for a larger number of cases.

D.31 Answers: Correlation

Answers to exercises in Sect. 34.8.

Answer to Exercise 34.1: Many correct answers.

Answer to Exercise 34.2: **1.** $R^2 = 0.881^2 = 77.6\%$. About 77.6% of the variation in punting distance can be explained by the variation in right-leg strength. **2.** H_0: $\rho = 0$ and H_1: $\rho \neq 0$. P-value *very* small. Very strong evidence of a correlation in the population.

Answer to Exercise 34.3: The plot looks linear; $n = 25$; variation doesn't seem constant.

Answer to Exercise 34.4: **1.** Very close to -1. **2.** $r = -\sqrt{0.9929} = -0.9964$. ($r$ must be negative!) **3.** Very small. This is a very large value for r on a reasonable sized sample. **4.** Yes.

Answer to Exercise 34.5: **1.** Close to -1, but not super close. **2.** $r = -\sqrt{0.6707} = -0.819$. ($r$ must be negative!) **3.** Very small. This is a large value for r on a reasonable sized sample. (The P-value turns out to be 0.000104.) **4.** Since $n < 25$, the test may not be statistically valid.

D.32 Answers: Regression

Answers to exercises in Sect. 35.13.

Answer to Exercise 35.1: **1.** $b_0 = 97.499$ (the intercept); $b_1 = 0.0764$ (the slope). **2.** $\hat{y} = 97.499 + 0.0764x$: x is the inlet temperature (in °C) and y is removal efficiency (in %). **3.** When inlet temperature increases by 1 degree C, on average the removal efficiency *increases* by 0.076 percentage points. **4.** H_0: $\beta = 0$; H_1: $\beta \neq 0$ (two-tailed;RQ implies two-tailed test). **5.** $t = 10.742$, which is huge; $P < 0.001$. **6.** $0.076 \pm (2 \times 0.007)$, or 0.076 ± 0.014, or 0.062 to 0.090.

Answer to Exercise 35.2: **1.** Intercept *not* about 110; that's where the line 'stops,' but the intercept is the predicted value of y when $x = 0$. We have to extend the line quite a bit. Using rise-over-run, guess slope is $(190 - 110)/(180 - 110) = 1.14$. **2.** $\hat{y} = -3.69 + 1.04x$, where y is punting distance (in feet), and x is right leg strength (in pounds). **3.** For each extra pound of leg strength, the punting distance increases, on average, by about 1 foot. **4.** H_0: $\beta = 0$; H_1: $\beta \neq 0$. (You could answer in terms of correlations.) The question is stated as a two-tailed question, but testing if stronger legs *increase* kicking distance seems sensible. **5.** $t = 6.16$, which is huge; $P = 0.0001$ (two-tailed). **6.** $1.0427 \pm (2 \times 0.1692)$, or 1.0427 ± 0.3384, or 0.70 to 1.4. **7.** Very strong evidence in the sample ($t = 6.16$; $P = 0.0001$ (two-tailed)) that punting distance is related to leg strength (slope: 1.0427; $n = 13$).

Answer to Exercise 35.3: **1.** *Way* too many decimal places. r is not relevant as relationship is non-linear. **2.** Regression is inappropriate: the relationship is non-linear. **3.** y should be $\hat{y}$; the slope and intercept have been *swapped* (from the plot, the intercept for their line is about 0.4, which they give as the slope). **4.** The whole thing is as dodgy-as...

Answer to Exercise 35.4: **1.** $\hat{y} = 17.47 - 2.59x$, where x is the percentage bitumen by weight, and y is the percentage air voids by volume. **2.** *Slope*: an increase in the bitumen weight by one percentage point *decreases* the average percentage air voids by volume by 2.59 percentage points. *Intercept*: dodgy (extrapolation); in principle 0% bitumen content by weight, the percentage air voids by volume is about 17.47%. **3.** $t = -74.9$: Massive! Extremely strong evidence ($P < 0.001$) of a relationship. **4.** $\hat{y} = 17.4712 - (2.5937 \times 5) = 4.5027$, or about 4.5%. Expected good prediction, as relationship is strong. **5.** $\hat{y} = 17.4712 - (2.5937 \times 6) = 1.909$, or about 1.9%. Might be a poor prediction, since this is extrapolation.

Answer to Exercise 35.5: **1.** b_0: When someone spends *no* time on sunscreen application, an average of 0.27g has been applied; nonsense. b_1: Each extra minute spent on application adds an average of 2.21g of sunscreen: sensible. **2.** The value of β_0 could be zero... which would make sense. **3.** $\hat{y} = 0.27 + (2.21 \times 8) = 17.95$; an average of about 18g. **4.** About 64% of the variation in sunscreen amount applied can be explained by the variation in the time spent on application. **5.** $r = \sqrt{0.64} = 0.8$, and need a positive value of r. A strong and positive correlation between the variables.

Answer to Exercise 35.6: **1.** No. **2.** Possibly; no idea of accuracy of predictions really. **3.** Intercept: Weight of infant with chest circumference zero; silly. Slope: average increase in birth weight (in g) for each increase in chest circumference by one cm. **4.** Intercept: cm; slope: cm/gram. **5.** $\hat{y} = 2538.7$g. **6.** Too many decimal place! Regression equation implies predicting to 0.0001 of a gram. r has too many decimal places too.

D.33 Answers: Reading research

Answers to exercises in Sect. 36.4.

Answer to Exercise 36.1: **1.** Not ecologically valid. **2.** Ethical. People understand that sometimes unexpected things happen. **3.** Convenience; self-selected. However, nothing obvious to suggest the people in the study would record different accuracies than people not in the study. **4.** Inclusion criteria. **5.** Paired t-test. **6.** Evidence in the sample that the mean difference in step-count between the two methods cannot be explained by chance: likely is a difference. **7.** From the given information: Probably valid.

Answer to Exercise 36.2: **1.** Students at that university (QUMS). **2.** A random sampling method has been used,
so results should be generalisable to the population (students at that university). **3.** t-test comparing two means. **4.** Three groups. *Null* hypothesis: the population mean hearing loss score is the same in all three groups. *Alternative* hypothesis: the mean hearing loss score is *not* the same in all three groups. **5.** Standard error: s.e.$(\bar{x}) = 3.08/\sqrt{745} = 0.1128$; CI is 19.8 ± 0.26. **6.** Need the standard error for the *difference* between two means, which is not reported.

D.34 Answers: Writing research

Answers to exercises in Sect. 37.18.

Answer to Exercise 37.1: The graph: odd colour choice; vertical axis label isn't helpful; horizontal axis isn't labelled at all; units of measurement not given.; title and/or caption would be helpful.

The table: the two limits of the CI are under the *Mean* and *Std dev* columns, but that is not what they are; units of measurement are not given; no caption, or any way to know what the table is about; number of decimal places is inconsistent. sample sizes not given; *difference*, and probably the other rows too, should report a standard error.

Answer to Exercise 37.2: RQ: P, O, C and I are not clear or explicit; the two fonts being compared should be identified. Perhaps better: 'For students, is the mean reading speed for text in the Georgia font the same as for text in Calibri font?'

Abstract: statement poorly constructed (*fonts* are not fast or slow!). Perhaps: 'The sample provided evidence that the mean reading speeds were different ($P =$???), when comparing text in Georgia font (mean: ???) and Calibri font (mean: ???; 95% CI for the difference: ??? to ???).

Answer to Exercise 37.3: No units of measurement given (they are centimetres.); jump-heights are given to 0.001 of a centimetre... which seems rather optimistic; table gives information for the *differences*, which is great but it could also provide numerical summary information for each individual jump type too; a numerical summary shouldn't include a P-value, t-score, or confidence interval.

Answer to Exercise 37.4: Variables are *qualitative*, so means inappropriate; the appropriate summary is an odds ratio, so the values almost certainly refer to the CI for the OR. Without more information, we can't really be sure what the OR means though.

Answer to Exercise 37.5: Such a study cannot *prove* anything just by itself (only a sample studied); the two CIs for the hang time for each plane design is fine... but really, the *difference* is of interest: The appropriate CI is for the difference between the mean hang times.

Answer to Exercise 37.6: **1.** Table: Reasonably good! **2.** Figure: Poor (it is 3D). Use a stacked or side-by-side bar chart.

Answer to Exercise 37.7: Reasonably good: should not be gaps between the bars of the histogram.

E

Appendix: Checklists

In this Appendix, two checklists are provided for evaluating and producing graphs (Sect. E.1) and tables (Sect. E.2). These may not be comprehensive.

E.1 A checklist for good scientific graphics

TABLE E.1: *A checklist for good scientific graphics*

	Criterion	Details
1	Caption	Does the graph have a descriptive caption under the graph: comprehensive, clear and accurate (what, where, when)?
2	Simplicity	Is the information is presented simply (e.g., not too many messages in one graph)?
3	Clarity	Is the information is presented clearly (e.g., use visible symbols; colours and symbols explained; no chart junk; no unnecessary 3-D; minimum axis tick marks; faint grid lines if needed)?
4	Accuracy	Is the information accurate (e.g., not misleading, complete)?
5	Message	Is the main message, theme, trend, or comparison of greatest importance easily seen?
6	Scaling	Are the comparisons made on linear scale (e.g., avoiding comparison of volumes or areas or angles)?
7	Organisation	Are the data are sorted meaningfully to aid interpretation?
8	Units	Are the units of measurement given?

E.2 A checklist for good scientific tables

TABLE E.2: *A checklist for good scientific tables*

	Criterion	Details
1	Caption	Does the table have a descriptive caption above the table: comprehensive, clear and accurate (what, where, when)?
2	Clarity	Are the row and columns in the table clearly labelled and explained?
3	Numbers	Are a suitable number of decimal places (or significant figures) displayed?
4	Alignment	Are the numbers aligned to allow easy comparison (e.g., on decimal points)?
5	Arrangement	Are the table columns/rows meaningfully arranged to emphasise patterns (e.g., not just alphabetically by default)?
6	Vertical lines	Are there no vertical lines except where absolutely essential?
7	Horizontal lines	Are there a minimum number of horizontal lines?
8	Units	Are the units of measurement clearly given?
9	Minimalist	Is unnecessary text avoided?

F

Appendix: Image credits

The sources of the images used in the online version of this book (in accordance with the terms of Unsplash[1], Pixabay[2] and Pexels[3]) are listed in the online book.

[1] https://unsplash.com/license
[2] https://pixabay.com/service/license/
[3] https://www.pexels.com/license/

G

Glossary

68–95–99.7 rule For *any* bell-shaped distribution, *approximately* 68% of observations lie within one standard deviation of the mean, 95% of observations lie within two standard deviations of the mean, and 99.7% of observations lie within three standard deviations of the mean. Also called the *empirical rule*.

Accuracy *Accuracy* refers to how close a *sample* estimate is to the *population* value, on average.

Alternative hypothesis The *alternative hypothesis* proposes that any difference, change or relationship observed in the sample is because a difference, change or relationship exists the *population* (that is, the difference, change or relationship cannot be explained by sampling variation).

Bell-shaped distributions See Normal distributions.

Bias *Bias* is the tendency of a sample to over- or under-estimate a population quantity.

Blinding *Blinding* when those involved in the study do not know which comparison group the study individuals are in.

A study can blind the **researcher** to knowing what comparison group the study individuals are in.

A study can blind the **participants** to knowing what comparison group they are in.

A study can blind the **analysts** to knowing what comparison group the individuals are in during analysis.

Blocking *Blocking* is when units of analysis are arranged in groups (called blocks) that are similar to one another.

Carryover effect The *carry-over effect* is when the influence of past experience(s) of the individuals carry over to influence future experience(s) of the individuals, for experimental studies (Sect. 7.4) or observational studies (Sect. 8.3).

Categorical data *Cateogorical data* is not *mathematically* numerical data: it consists of categories or labels (even if those labels are numbers). In this book, categorical data is called *qualitative* data.

Classical approach to probability In the *classical approach to probability*, the probability of an event occurring is the number of elements of the sample space included in the event, divided by the total number of elements in the sample space, *when all outcomes are equally likely*.

Cluster sampling A sample where the population is split into a large number of small groups called *clusters*, then a *simple random sample* of clusters is selected and *every* member of the chosen small groups is part of the sample.

Collusion *Collusion* occurs when people work together to produce a work, but only one gets the credit for it.

At university, collusion happens if you give or receive help in completing any form of individual assessment such as assignments and exams.

Comparison The *comparison* in the RQ identifies the small number of different, distinct subsets of the population between which the outcome is being compared. The groups being compared have either *imposed* differences, or have *existing* differences.

Conceptual definition A *conceptual definition* articulates *what* exactly is to be measured or observed in a study.

Confidence interval A *confidence interval* is an interval in which the population *parameter* is likely to be contained, if we found many samples the same way.
If we computed the 95% confidence interval (or CI) from each sample, about 95% of the CIs would contain the *statistic* of interest. This interval is called a *confidence interval*.
Alternatively, the CI can be seen as the range of plausible values for the parameter that may have produced the observed sample statistic. We studied CIs is some specific situations (there are hundreds more!):

- CIs for one proportion: Chap. 20
- CIs for one mean: Chap. 22
- CIs for a mean difference (*paired* sample mean): Chap. 23
- CIs for the difference between two means: Chap. 24
- CIs for comparing two odds: Chap. 25
- CIs for regression parameters: Sect. 35.7

Confounding *Confounding* is when a third variable influences the relationship between the response and explanatory variable.

Confounding variable A *confounding variable* (or a *confounder*) is an extraneous variable associated with the response and explanatory variables.

Conditions The *conditions* of interest that those in the observational study can be exposed to.

Connection The *connection* in the RQ identifies another quantity of interest that varies, that may be related to the outcome.

Continuous data *Continuous* quantitative data has (at least in theory) an infinite number of possible values between any two given values.

Control A *control* is a unit of analysis without the treatment applied (but as similar as possible in every other way to other units of analysis).

Convenience sample A sample where individuals are selected because they are convenient for the researcher.

Data *Data* refers items of information obtained from a study (such as height of seedlings, or the type of medication given).

Data set A *data set* refers to a collection of data from a study.

Descriptive study A *descriptive study* is one where the researchers only focus on collecting, measuring, assessing or describing an outcome in the population.

Discrete data *Discrete* quantitative data has a countable number of possible values between any two given values of the variable.

Ecological validity A study is *ecologically valid* if the study methods, materials and context approximate the real situation being studied.

Event An *event* is any combination of the elements in the sample space.

Exclusion criteria *Exclusion criteria* are characteristics that disqualify potential individuals from being included in the study.

Empirical rule For *any* bell-shaped distribution, *approximately* 68% of observations lie within one standard deviation of the mean, 95% of observations lie within two standard

deviations of the mean, and 99.7% of observations lie within three standard deviations of the mean. Also called the *68–95–99.7 rule*.

Experiment In an *experimental study* (or an *experiment*), the researchers intervene to control the values of the explanatory variables (C) that are applied to the individuals. The researchers allocate treatments (i.e., apply the intervention).

Experimenter effect The *experimenter effect* is another name for the *observer effect* in experimental studies (that is, when the researchers *unintentionally* influence the behaviour of subjects).

Explanatory variable An *explanatory variable* is a variable of interest from the individuals in the study which (potentially) causes changes in, or is related to, the response variable.

External validity *Externally validity* refers to the ability to generalise the results to other groups in the population apart from the sample studied.

Extraneous variable An *extraneous* variable is any variable that is (potentially) associated with the response variable, but is not one the explanatory variable.

Extrapolation *Extrapolation* refers to making prediction outside the range of the available data. Extrapolation beyond the data can lead to nonsense predictions.

Fraud *Fraud* refers to the intent to deceive. *Fraud* can occur by:

- taking an exam for another student or letting someone take an exam for you
- falsifying or inventing research data and findings
- altering or fabricating information
- forging a document
- falsifying past academic records or employment details in order to gain entrance into the university

Hawthorne effect The *Hawthorne effect* is the tendency of people (or animals, or...) to behave differently if they know (or think) they are being observed, in experimental studies (Sect. 7.5) or observational studies (Sect. 8.4).

Hypothesis A *hypothesis* is a possible answer to a (research) question. More specifically, see *null hypothesis* or *alternative hypothesis*

Hypothesis test A *hypothesis test* is a way to formally answer questions about a population, based on information obtained from a sample. In this book, we have looked at some specific hypothesis tests (hundreds exist: Kanji (2006)!):

- Hypothesis tests about a single mean: Chap. 27
- Hypothesis tests about a mean difference (means of *paired* samples): Chap. 29
- Hypothesis tests comparing two means: Chap. 30
- Hypothesis tests comparing odds (or percentages): Chap. 31
- Hypothesis tests about a correlation: Sect. 34.4
- Hypothesis tests about regression parameters: Sect. 35.6

Inclusion criteria *Inclusion criteria* are characteristics that individuals must meet explicitly to be included in the study.

Independence Two events are *independent* if the probability of one event doesn't change depending on whether or not other event has happened.

Internal validity *Internally valid* refers to the strength of the association between the outcome and the comparison/connection. In a study with high internal validity, the association between the outcome and the comparison/connection can be attributed to that comparison/connection, rather than to other factors.

Intervention An *intervention* is a comparison or connection that the researchers impose upon those in the study, intending to change the outcome.

IQR The *IQR* is the range in which the middle 50% of the data lie; the difference between the third and the first quartiles.

IQR rule for identifying outliers The IQR rule can identify outliers as either:

- *mild* (observations $1.5 \times$ IQR more unusual than Q_1 or Q_3), or
- *extreme* (observations $3 \times$ IQR more unusual than Q_1 or Q_3).

Judgement sample A sample where individuals are selected, based on the researchers' judgement, depending on whether the researcher thinks they are likely to be agreeable or helpful.

Levels of a qualitative variable The *levels* (or the *values*) of a qualitative variable refer to the names of the distinct categories.

Lurking variable A *lurking variable* is an extraneous variable associated with the response and explanatory variables (that is, is a confounding variable), but whose values are not measured, assessed, described or recorded in the study.

Mean The *mean* is one way to measure the 'average' value of quantitative data. The *arithmetic mean* can be considered as the 'balance point' of the data, or the value such that the positive and negative distances from the mean add to zero.

Median The *median* is one way to measure the 'average' value of some data. The *median* is a value such that half the values are larger than the median, and half the values are smaller than the median.

Multistage sampling A sample where large groups are selected using a *simple random sample*, then smaller groups within those large groups are selected using a *simple random sample*. The simple randomly sampling can continue for as many levels as necessary.

Nominal variable A *nominal* qualitative variable is a qualitative variable where the levels *do not* have a natural order.

Normal distribution A *normal distribution* is symmetrical distribution, with most values in the centre of the distribution. The normal distribution is described by its *mean* and *standard deviation*. A picture of a normal distribution is shown in Fig. G.1. Normal distributions are also called *bell-shaped* distributions.

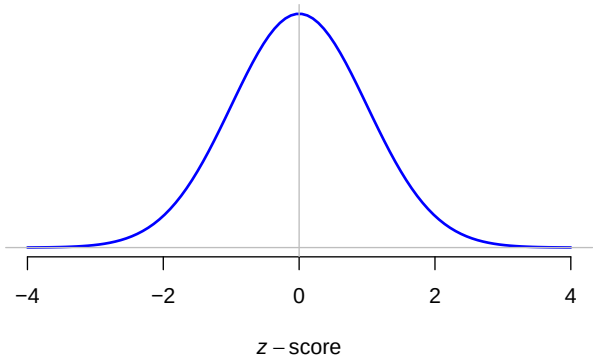


FIGURE G.1: *A normal distribution*

Null hypothesis The *null hypothesis* proposes that any difference, change or relationship observed in the sample can be explained by sampling variation (that is, no difference, change or relationship exists the *population*).

Observational study An *observational study* is one where the researchers do not impose the

comparison or connection upon those in the study to (potentially) change the response of the participants.

Observer effect The *observer effect* is when the researchers *unintentionally* influence the behaviour of subjects, in experimental studies (Sect. 7.6) or observational studies (Sect. 7.6).

Odds The *odds* of some event is the proportion (or percentage, or number) of times that an event happens, divided by the proportion (or percentage, or number) of times that the event does *not* happen.

Odds ratio The *odds ratio* is how many *times* greater the odds of an event are in one group, compared to the odds of the same event in another group.

Operational definition An *operational definition* articulates *how* to capture (identify, create, measure, assess etc.) the value.

Ordinal variable[1] An *ordinal* qualitative variable is a qualitative variable where the levels *do* have a natural order.

Outcome The *outcome* in a RQ is the result, output, consequence or effect of interest in a study, numerically summarising the entire population (or subsets of the population).

P-value A P-value is the likelihood of observing the sample results (or something even more extreme) over repeated sampling, under the assumption that the null hypothesis about the population is true.

Parameter A *parameter* is a number describing some feature of a population, and is usually estimated by a statistic.

Paired data *Paired data* is when two observations about the same variable are recorded for each unit of analysis.

Percentage A *percentage* is a proportion, multiplied by 100. Percentages are numbers between 0% and 100%.

Percentiles The pth percentile of the data is a value separating the smallest $p\%$ of the data from the rest.

Placebo A *placebo* is a treatment with no intended effect or active ingredient.

Placebo effect The *placebo effect* is when individuals report perceived or actual effects without having received the treatment or condition, in experimental studies (Sect. 8.6) or observational studies (Sect. 8.6).

Plagiarism *Plagiarism* is using other people's ideas and research to develop new conclusions, or confirm existing conclusion. All sources used when writing research should be acknowledged, otherwise you are committing plagiarism.

Plagiarism can be deliberate or accidental:

- Deliberate—for instance, if a student intentionally copies the work of others and pretends it is their own work.
- Accidental—for instance, if a student has poor notetaking skills or doesn't know how to reference correctly, and they inadvertently present someone else's ideas and words as their own.

Population The *population* is the group of individuals (or cases, or subjects if the individuals are people) from which the total set of observations of interest could be made, and to which the results will (hopefully) generalise.

Precision *Precision* refers to how likely it is that the sample values will be similar or close together, and not vary much from sample to sample.

Proportion A *proportion* is a fraction out of a total. Proportions are numbers between 0 and 1.

Protocol A *protocol* is a predefined procedure detailing the design and implementation of studies, and for data collection.

Qualitative data *Qualitative data* is not *mathematically* numerical data: it consists of categories or labels (even if those labels are numbers). Also called *categorical* data.

Quantitative data *Quantitative data* is *mathematically* numerical data: the numbers themselves have numerical meaning, and it makes sense to be able to perform mathematical operation on them. Most data that are counted or measured will be quantitative. Also called *scale data*.

Quantitative research *Quantitative research* summarises and analyses data using numerical methods, such as producing averages and percentages.

Quartiles *Quartiles* describe the variation and shape of data:

- The first quartile Q_1: A value that separates the smallest 25% of observations from the largest 75%. The Q_1 is like the median of the *smaller* half of the data, halfway between the minimum value and the median.
- The second quartile Q_2: A value that separates the smallest 50% of observations from the largest 50%. (This is the *median*.)
- The third quartile Q_3: The value that separates the smallest 75% of observations from the largest 25%. The Q_3 is like the median of the *larger* half of the data, halfway between the median and the maximum value.

Quasi-experiment In a *quasi-experiment*, the researchers (a) allocate treatments to groups of individuals (i.e., do not decide the values of the Comparison/Connection used), but (b) do not determine who or what is in those groups.

Random In research and statistics, *random* means "determined completely by chance."

Range The *range* is the maximum value minus the minimum value.

Relative frequency approach to probability In the *relative frequency approach to probability*, the probability of an event is (approximately) the number of times the outcomes of interest has appeared in the past, divided by the number of 'attempts' in the past.

Representative samples A representative sample is one where the individuals *in* the sample are not likely to be different the individuals *not in* the sample, at least for the variables of interest.

Response variable A *response variable* is the variable used to measure, assess or describe the outcome on each individual in the population.

Sample A *sample* is a subset of the population of interest which is actually studied, and from which information is collected.

Sample space The *sample space* is a list of all possible and distinct results after administering a procedure whose result is unknown beforehand. is a list of the results after administering a procedure whose result is unknown beforehand.

Sampling distribution A *sampling distribution* is the distribution of some sample statistic, showing how its value varies from one sample to sample.

Sampling frame The *sampling frame* is a list of all the members of the population (the individuals, or cases, or subjects).

Sampling variation *Sampling variation* refers to how much a sample estimate (a *statistic*) is likely to vary from sample to sample, because each sample is different.

Scale data *Scale data* is *mathematically* numerical data: the numbers themselves have numer-

ical meaning, and it makes sense to be able to perform mathematical operation on them. Most data that are counted or measured will be quantitative. In this book, scale data is called *quantitative data*.

Simple random sample A sample where *every* possible sample of the same size has *same* chance of being selected.

Standard deviation The *standard deviation* is, approximately, the average distance that observations are away from the mean.

Standard deviation rule for identifying outliers For approximately symmetric distributions, any observation more than three standard deviations from the mean can be considered an outlier.

Standard error A *standard error* is the standard deviation of all possible values of the sample estimate (from samples of a certain size). Any quantity estimated from a sample has a standard error.

Stratified sampling A sample where the population is split into a small number of large (usually homogeneous) groups called *strata*, then cases are selected using a *simple random sample* from *each* stratum.

Statistic A *statistic* is a number describing some feature of a sample (to estimate a population parameter).

Statistical validity A result is *statistically valid* if the conditions for the underlying mathematical calculations and assumptions to be approximately correct are met. Every confidence interval and hypothesis test has statistical validity conditions.

Subjective approach to probability In the *subjective approach to probability*, various factors are incorporated, perhaps subjectively, to determine the probability of an event.

Systematic sampling A sample where the first case is *randomly* selected; then, every nth individual is selected.

Treatments *Treatments* are the conditions of interest that those in the study can be exposed to (in the comparison/connection). In experiments, treatments are imposed by researchers.

True experiment In a true experiment, the researchers (a) allocate treatments to groups of individuals (i.e., decide the values of the Comparison/Connection used), and (b) determine who or what is in those groups.

Unit of observation The *unit of observation* is the 'who' or 'what' which are observed, from which measurements are taken and data collected.

Unit of analysis The *unit of analysis* is the 'who' or 'what' about which generalizations and conclusions are made; the smallest independent 'who' or 'what' for which information is analysed. Units of analysis should not typically share a common underlying source.

Unstandardizing formula When the z-score is known, the *unstandardizing formula* determines the corresponding value of the observation x.

Values of a qualitative variable The *levels* (or the *values*) of a qualitative variable refer to the names of the distinct categories.

Variable A *variable* is a single aspect or characteristic associated with each of a group of individuals under consideration, that can vary from individual to individual.

Voluntary* response (self-selecting) sample A sample where individuals participate if they wish to.

z-score A *z-score* measure how many standard deviations a value is from the mean. In symbols:

$$z = \frac{x - \mu}{\sigma},$$

where x is the value, μ is the mean of the distribution, and σ is the standard deviation of the distribution.

References

Aedo-Ortiz DM, Olsen ED, Kellogg LD. Simulating a harvester-forwarder softwood thinning: A software evaluation. Forest Products Journal. 1997;47(5):36–41.

Affonso CM, Kezunovic M. Probabilistic assessment of electric vehicle charging demand impact on residential distribution transformer aging. 2018 IEEE international conference on probabilistic methods applied to power systems (PMAPS). IEEE; 2018. p. 1–6.

Agbayani S, Fortune SME, Trites AW. Growth and development of North Pacific gray whales (*Eschrichtius robustus*). Journal of Mammalogy. 2020;101(3):742–54.

Agresti A, Franklin CA. Statistics: The art and science of learning from data. 3rd edition. Pearson Education Limited; 2007;

Allen AM, Abdelwahab NM, Carlson S, Bosch TA, Eberly LE, Okuyemie K. Effect of brief exercise on urges to smoke in men and women smokers. Addictive Behaviors. 2018;77:34–7.

Allen L, O'Connell A, Kiermer V. How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. Learned Publishing. Wiley Online Library; 2019;32(1):71–4.

Alley S, Wellens P, Schoeppe S, Vries H de, Rebar AL, Short CE, et al. Impact of increasing social media use on sitting time and body mass index. Health Promotion Journal of Australia. 2017;28:91–5.

Aloy AB, Vallejo Jr. BM, Juinio-Meñez MA. Increased plastic litter cover affects the foraging activity of the sandy intertidal gastropod *Nassarius pullus*. Marine Pollution Bulletin. 2011;62:1772–9.

AlTarawneh G, Thorne S. A pilot study exploring spreadsheet risk in scientific research. arXiv preprint arXiv:170309785. 2017;

Altman DG. Practical statistics for medical research. Chapman & Hall; 1991.

Amin AA, Mahmood-ul-Hasan K. Robust active fault-tolerant control for internal combustion gas engine for air–fuel ratio control with statistical regression-based observer model. Measurement and Control. 2019;0020294018823031.

Anastasiadis E, Rajan P, Winchester CL. Framing a research question: The first and most vital step in planning research. Journal of Clinical Urology. 2015;8(6):409–11.

Andersen EB. Multiplicative Poisson models with unequal cell rates. Scandinavian Journal of Statistics. 1977;4:153–8.

Anonymous. Green tea cuts risk of cancer. The Sunday Mail. 2012;19.

Axelsson J, Sundelin T, Ingre M, van Someren EJW, Olsson A, Lekander M. Beauty sleep:

Experimental study on the perceived health and attractiveness of sleep deprived people. British Medical Journal. 2010;341.

Azwari NFSM, Hamsa AAK. Evaluating actual speed against the permissible speed of vehicles during free-flow traffic conditions. Jurnal Kejuruteraan. 2021;33(2):183–91.

Bacho Z, Lajangang FJE, Khin NY, Shah SS, Chia YK, Jalil E, et al. The effects of comprehensive core body resistance exercise on lower extremity motor function among stroke survivors. Journal of physics: Conference series. IOP Publishing; 2019. p. 012025.

Badiou A, Marsh C, Gauchet M. Exploring data: An introduction to data analysis for social scientists. Cambridge, UK: Polity Press; 1988.

Bambauer J. Defending the dog. Oregon Law Review. HeinOnline; 2012;91:1203.

Banerjee A, Chaudhury S. Statistics without tears: Populations and samples. Industrial Psychiatry Journal. Wolters Kluwer–Medknow Publications; 2010;19(1):60.

Barr D, DeBruine L. webex: Create interactive web exercises in 'R Markdown' [Internet]. 2019. Available from: https://CRAN.R-project.org/package%20=%20webex.

Barr N, Holmes M, Roiko A, Dunn P, Lord B. Self-reported behaviors and perceptions of Australian paramedics in relation to hand hygiene and gloving practices in paramedic-led health care. American Journal of Infection Control. 2017;45(7):771–8.

Barrett B, Brown R, Rakel D, Mundt M, Bone K, Barlow S, et al. Echinacea for treating the common cold: A randomized trial. Annals of Internal Medicine. 2010;153(12):769–77.

Bartareau TM. Estimating the live body weight of American black bears in Florida. Journal of Fish and Wildlife Management. 2017;8(1):234–9.

Baughman RP, Sparkman BK, Lower EE. Six-minute walk test and health status assessment in sarcoidosis. Chest. 2007;132(1):207–13.

Baur DM, Christophi CA, Kales SN. Metabolic syndrome is inversely related to cardiorespiratory fitness in male career firefighters. Journal of Strength and Conditioning Research. 2012;26(9):2331–7.

Beaman L, Karlan D, Thuysbaert B, Udry C. Profitability of fertilizer: Experimental evidence from female rice farmers in Mali. American Economic Review. 2013;103(3):381–6.

Becker H, Stuifbergen AK, Sands D. Development of a scale to measure barriers to health promotion activities among persons with disabilities. American Journal of Health Promotion. 1991;5(6):449–54.

Berger RL. Nonstandard operator precedence in Excel. Computational Statistics & Data Analysis. Elsevier; 2007;51(6):2788–91.

Bhargava SK, Ramji S, Kumar A, Mohan MAN, Marwah J, Sachdev HP. Mid-arm and chest circumferences at birth as predictors of low birth weight and neonatal mortality in the community. BMJ. British Medical Journal Publishing Group; 1985;291(6509):1617–9.

Bingham CR, Simons-Morton BG, Pradhan AK, Li K, Almani F, Falk EB, et al. Peer passenger norms and pressure: Experimental effects on simulated driving among teenage males.

Transportation research part F: Traffic Psychology and Behaviour. Elsevier; 2016;41:124–37.

Bird AR, Vuaran MS, King RA, Noakes M, Keogh J, Morell MK, et al. Wholegrain foods made from a novel high-amylose barley variety (*Himalaya 292*) improve indices of bowel health in human subjects. British Journal of Nutrition. 2008;99:1032–40.

Blair BF, Lamb MC. Evaluating concentrations of pesticides and heavy metals in the U.S. Peanut crop in the presence of detection limits. Peanut Science. 2017;44:124–33.

Bock DE, Velleman PF, De Veaux RD. Stats: Modeling the world [Internet]. Addison-Wesley; 2010. Available from: https://books.google.com.au/books?id%20=%20zHEJPwAACAAJ.

Botelho AM, de Camargo AM, Dean M, Fiates GMR. Effect of a health reminder on consumers' selection of ultra-processed foods in a supermarket. Food Quality and Preference. 2019;71:431–7.

Brockmann HJ. Satellite male groups in horseshoe crabs, *Limulus polyphemus*. Ethology. 1996;102:1–21.

Brown L, Whitney CL, Hunt RC, Addario M, Hogue T. Do warning lights and sirens reduce ambulance response times? Prehospital Emergency Care. 2000;4(1):70–4.

Brunette DD, Kominsky J, Ruiz E. Correlation of emergency health care use, 911 volume, and jail activity with welfare check distribution. Annals of Emergency Medicine. Elsevier; 1991;20(7):739–42.

Brunton E, Bolin J, Leon J, Burnett S. Fright or flight? Behavioural responses of kangaroos to drone-based monitoring. Drones. 2019;3(2):41.

Bryson MC. The Literary Digest poll: Making of a statistical myth. The American Statistician. 1976;30(4):184–5.

Budgett S, Pfannkuch M, Regan M, Wild CJ. Dynamic visualizations and the randomization test. Technology Innovation in Statistics Education. 2013;7(2).

Bulte E, Beekman G, Di Falco S, Hella J, Lei P. Behavioral responses and the impact of new agricultural technologies: Evidence from a double-blind field experiment in Tanzania. American Journal of Agricultural Economics. Wiley Online Library; 2014;96(3):813–30.

Burch PRJ. Smoking and lung cancer: The problem of inferring cause. Journal of the Royal Statistical Society, Series A. 1978;141(4):437–77.

Center for Disease Control and Prevention. National Center for Health Statistics. Third National Health and Nutrition Examination Survey, 1988–1994, NHANES III Laboratory Data File [Internet]. Hyattsville, MD: Public Use Data File Documentation Number 76200; U.S. Department of Health; Human Services, Centers for Disease Control; Prevention; 1996. Available from: https://wwwn.cdc.gov/nchs/data/nhanes3/1a/readme.txt.

Center for Disease Control and Prevention (CDC). National Center for Health Statistics. National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health; Human Services, Centers for Disease Control; Prevention; 1988--1994.

Chan E, Taylor S, Marriott J, Barger B. Exploration of attitudes and barriers to bringing patient's own medications to the Emergency Department: A survey of paramedics. Australasian Journal of Paramedicine. 2008;6(4).

Chang W. webshot: Take screenshots of web pages [Internet]. 2018. Available from: https://CRAN.R-project.org/package%20=%20webshot.

Chapman CA. Association patterns of spider monkeys: The influence of ecology and sex on social organisation. Behavioral Ecology and Sociobiology. 1990;26:409–14.

Chapman D, Peiffer J, Abbiss CR, Laursen PB. A descriptive physical profile of Western Australian male paramedics. Journal of Emergency Primary Health Care. 2007;5(1).

Charig CR, Webb DR, Payne SR, Wickham JEA. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. British Medical Journal. 1986;292:879–82.

Chatterjee S, Sarkar K. Retraction note to: Surface-functionalized gold nanoparticles mediate bacterial transformation: A nanobiotechnological approach. Biotechnology Letters [Internet]. 2015;37(7):1527–8. Available from: https://doi.org/10.1007/s10529-015-1826-0.

Checkley W, Gilman RH, Black RE, Lescano AG, Cabrera L, Taylor DN, et al. Effects of nutritional status on diarrhea in Peruvian children. Journal of Pediatrics [Internet]. 2002;140(2):210–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11865273.

Chitwood DE, Devinny JS. Treatment of mixed hydrogen sulfide and organic vapors in a rock medium biofilter. Water Environment Research. Water Environment Federation; 2001;73(4):426–35.

Choi HK, Curhan G. Soft drinks, fructose consumption, and the risk of gout in men: Prospective cohort study. British Medical Journal. 2008;336(7639):309–12.

Christensen HA, Herrer A, Telford SR. Enzootic cutaneous leishmaniasis in eastern Panama: II: Entomological investigations. Annals of Tropical Medicine & Parasitology. 1972;66(1):55–66.

Clark RL, Famodu OA, Holásková I, Infante AM, Murray PJ, Olfert IM, et al. Educational intervention improves fruit and vegetable intake in young adults with metabolic syndrome components. Nutrition Research. 2019;62:89–100.

Cohen JF, Korevaar DA, Gatsonis CA, Glasziou PP, Hooft L, Moher D, et al. STARD for abstracts: Essential items for reporting diagnostic accuracy studies in journal or conference abstracts. BMJ. British Medical Journal Publishing Group; 2017;358:j3751.

Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. Journal of health and social behavior. 1983;385–96.

Collett P, O'Shea G. Pointing the way to a fictional place: A study of direction giving in Iran and England. European Journal of Social Psychology. Wiley Online Library; 1976;6(4):447–58.

Comerford E, Durante J, Goldsworthy R, Hall V, Gooding J, Quinn B. Motivations for kerbside dumping: Evidence from Brisbane, Australia. Waste Management. 2018;78:490–6.

Cook HB, Burt MJ, Collett JA, Whitehead MR, Frampton CMA, Chapman BA. Adult coeliac disease: Prevalence and clinical significance. Journal of Gastroenterology and Hepatology. 2000;15(9):1032–6.

Corbie-Smith G. The continuing legacy of the Tuskegee Syphilis Study: Considerations for clinical investigation. The American Journal of the Medical Sciences. Elsevier; 1999;317(1):5–8.

Coudert F-X. Correcting the scientific record: Retraction practices in chemistry and materials science. Chemistry of Materials. 2019;31.

Cox LS, Tiffany ST, Christen AG. Evaluation of the brief questionnaire of smoking urges (QSU-brief) in laboratory and clinical settings. Nicotine & Tobacco Research. 2001;3(1):7–16.

Cressie NAC, Sheffield LJ, Whitford HJ. Use of the one sample t-test in the real world. Journal of Chronic Diseases. 1984;37(2):107–14.

Cross J, Lam T, Arndell J, Quach J, Reed B, Thyer L, et al. Impact of hand dominance on effectiveness of chest compressions in a simulated setting: A randomised, crossover trial. Australasian Journal of Paramedicine. 2019;16.

Crozier GKD, Schulte-Hostedde AI. Towards improving the ethics of ecological research. Science and engineering ethics. Springer; 2015;21(3):577–94.

Curfman GD. Is exercise beneficial—or hazardous—to your heart? NEJM. 1993;329:1730–1.

Dala SR, Fowlkes EB, Hoadley B. Risk analysis of the space shuttle: Pre-Challenger prediction of failure. Journal of the American Statistical Association. 1989;84(408):945–57.

Danielsson L, Papoulias I, Petersson E-L, Carlsson J, Waern M. Exercise or basic body awareness therapy as add-on treatment for major depression: A controlled study. Journal of Affective Disorders. 2014;168:98–106.

Davidson J. A survey of the satisfaction of upper limb amputees with their prostheses, their lifestyles, and their abilities. Journal of Hand Therapy. 2002;15(1):62–70.

Dawson P, Han I, Lynn D, Lackey J, Baker J, Martinez-Dawson R. Bacterial transfer associated with blowing out candles on a birthday cake. Journal of Food Research. 2017;6(4):1–5.

de Carvalho AA, Amorim FF, Santana LA, de Almeida KJQ, Santana ANC, Neves F de AR. STOP-Bang questionnaire should be used in all adults with Down Syndrome to screen for moderate to severe obstructive sleep apnea. PloS ONE. 2020;15(5):e0232596.

Delaney LJ, Currie MJ, Huang H-CC, Lopez V, Van Haren F. "They can rest at home": An observational study of patients' quality of sleep in an Australian hospital. BMC Health Services Research. BioMed Central; 2018;18(1):524.

Devore JL, Berk KN. Modern mathematical statistics with applications. Thomson Higher Education; 2007.

Dexter B, King R, Harrison SL, Parisi AV, Downs NJ. A pilot observational study of environmental summertime health risk behavior in central Brisbane, Queensland: Opportunities to raise sun protection awareness in Australia's sunshine state. Photochemistry and Photobiology. 2019;95(2):650–5.

Dexter CE, Appleby RG, Scott J, Edgar JP, Jones DN. Individuals matter: Predicting koala road crossing behaviour in south-east Queensland. Australian Mammalogy. 2018;40(1):67–75.

Dianat I, Sorkhi N, Pourhossein A, Alipour A, Asghari-Jafarabadi M. Neck, shoulder and low back pain in secondary schoolchildren in relation to schoolbag carriage: Should the recommended weight limits be gender-specific? Applied Ergonomics. 2014;45:437–42.

Dillon MP, Fortington LV, Akram M, Erbas B, Kohler F. Geographic variation of the incidence rate of lower limb amputation in Australia from 2007–12. PLoS ONE. 2017;12(1).

Dokur M, Petekkaya E, Karadağ M. Media-based clinical research on selfie-related injuries and deaths. Ulus Travma Acil Cerrahi Derg. 2018;24(2):129–35.

Doll R, Hill AB. Smoking and carcinoma of the lung. British Medical Journal. 1950;221(ii):739–48.

Doll R, Hill AB. The mortality of doctors in relation to their smoking habits. British Medical Journal. 1954;1(4877):1451.

Doosti-Irani O, Golzarian MR, Aghkhani MH, Sadrnia H, Doosti-Irani M. Development of multiple regression model to estimate the apple's bruise depth using thermal maps. Postharvest Biology and Technology. Elsevier; 2016;116:75–9.

Duncan MJ, Wunderlich K, Zhao Y, Faulkner G. Walk this way: Validity evidence of iPhone health application step count in laboratory and free-living conditions. Journal of Sports Sciences. 2018;36(15):1695–704.

Dunn PK. A simple dataset for demonstrating common distributions. Journal of Statistics Education. 1999;7(3).

Dunn PK. Bootstrap confidence intervals for predicted rainfall quantiles. International Journal of Climatology. 2001;21(1):89–94.

Dunn PK. Assessing claims made by a pizza chain. Journal of Statistical Education [Internet]. 2012;20(1). Available from: www.amstat.org/publications/jse/v20n1/dunn.pdf[2].

Dunn PK. Comparing the lifetimes of two brands of batteries. Journal of Statistical Education. 2013;21(1).

Dunn PK, Carey MD, Richardson AM, McDonald C. Learning the language of statistics: Challenges and teaching approaches. Statistics Education Research Journal. 2016;15(1).

Dunn PK, Smyth GK. GLMsData: Generalized linear model data sets [Internet]. 2017. Available from: https://CRAN.R-project.org/package%20=%20GLMsData.

Dunn PK, Smyth GK. Generalized linear models with examples in R. Springer; 2018.

Edwards NM, Myer GD, Kalkwarf HJ, Woo JG, Khoury PR, Hewett TE, et al. Outdoor temperature, precipitation, and wind speed affect physical activity levels in children: A longitudinal cohort study. Journal of Physical Activity and Health. 2015;12(8):1074–81.

Egbue O, Long S. Barriers to widespread adoption of electric vehicles: An analysis of consumer attitudes and perceptions. Energy Policy. Elsevier; 2012;48:717–29.

[2] **https://www.amstat.org/publications/jse/v20n1/dunn.pdf**

Egbue O, Long S, Samaranayake VA. Mass deployment of sustainable transportation: Evaluation of factors that influence electric vehicle adoption. Clean Technologies and Environmental Policy. Springer; 2017;19(7):1927–39.

Ehlers N. On corneal thickness and intraocular pressure. II: A clinical study on the thickness of the corneal stroma in glaucomatous eyes. Acta ophthalmologica. Wiley Online Library; 1970;48(6):1107–12.

Ejtahed HS, Mohtadi-Nia J, Homayouni-Rad A, Niafar M, Asghari-Jafarabadi M, Mofid V. Probiotic yogurt improves antioxidant status in type 2 diabetic patients. Nutrition. 2012;28:539–43.

Elton C, Nicholson M. The ten-year cycle in numbers of the lynx in canada. The Journal of Animal Ecology. JSTOR; 1942;215–44.

Farrar MB, Wallace HM, Xu C-Y, Nguyen TTN, Tavakkoli E, Joseph S, et al. Short-term effects of organo-mineral enriched biochar fertiliser on ginger yield and nutrient cycling. Journal of Soils and Sediments. Springer; 2018;1–5.

Fayet-Moore F, Peters V, McConnell A, Petocz P, Eldridge AL. Weekday snacking prevalence, frequency, and energy contribution have increased while foods consumed during snacking have shifted among Australian children and adolescents: 1995, 2007 and 2011–12 National Nutrition Surveys. Nutrition Journal. 2017;16(65):1–4.

Feng Y, Huang Y, Ma X. The application of Student's t-test in internal quality control of clinical laboratory. Frontiers in Laboratory Medicine. 2017;1(3):125–8.

Fink A. The survey handbook. The survey kit. SAGE Publications, Incorporated; 1995.

Flanagan-Hyde P. Confound it! I can't keep these variables straight. STATS: The Magazine for Students of Statistics. 2005;43:21–3.

Fraboni F, Puchades VM, De Angelis M, Pietrantoni L, Prati G. Red-light running behavior of cyclists in Italy: An observational study. Accident Analysis & Prevention. Elsevier; 2018;120:219–32.

Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F. Questionable research practices in ecology and evolution. PloS one. Public Library of Science; 2018;13(7):e0200303.

Friedmann E, Thomas S. Health benefits of pets for families. Marriage & Family Review. Taylor & Francis; 1985;8(3-4):191–203.

Friel SN, Curcio FR, Bright GW. Making sense of graphs: Critical factors influencing comprehension and instructional implications. Journal for Research in Mathematical Education. 2001;124–58.

Fritts JR, Fort C, Corr AQ, Liang Q, Alla L, Cravener T, et al. Herbs and spices increase liking and preference for vegetables among rural high school students. Food Quality and Preference. 2018;68:125–34.

Froud KJ, Beresford RM, Cogger NC. Impact of kiwifruit bacterial canker on productivity of cv. Hayward kiwifruit using observational data and multivariable analysis. Plant Pathology. Wiley Online Library; 2018;67(3):671–81.

Furness RW, Bryant DM. Effect of wind on field metabolic rates of breeding northern fulmars. Ecology. Wiley Online Library; 1996;77(4):1181–8.

Galletta DF, Hartzel KS, Johnson SE, Joseph JL, Rustagi S. Spreadsheet presentation and error detection: An experimental study. Journal of Management Information Systems. Taylor & Francis; 1996;13(3):45–63.

Gamble T, Walker I. Wearing a bicycle helmet can increase risk taking and sensation seeking in adults. Psychological Science. Sage Publications Sage CA: Los Angeles, CA; 2016;27(2):289–94.

Gammon CS, von Hurst PR, Coad J, Kruger R, Stonehouse W. Vegetarianism, vitamin B12, and insulin resistance in a group of predominately overweight/obese South African women. Nutrition. 2012;28:20–4.

Garnier S. viridis: Default color maps from 'matplotlib' [Internet]. 2018. Available from: https://CRAN.R-project.org/package%20=%20viridis.

Gatti UC, Migliaccio GC, Bogus SM, Schneider S. An exploratory study of the relationship between construction workforce physical strain and task level productivity. Construction Management and Economics. 2013;1–7.

George BJ, Brown AW, Allison DB. Errors in statistical analysis and questionable randomization lead to unreliable conclusions. Journal of Paramedical Sciences. 2015;6(3):153–4.

Gonzalez-Fonteboa B, Martinez-Abella F. Shear strength of recycled concrete beams. Construction and Building Materials. Elsevier; 2007;21(4):887–93.

Grabosky J, Bassuk N. Seventeen years' growth of street trees in structural soil compared with a tree lawn in New York City. Urban Forestry & Urban Greening. 2016;16:103–9.

Greenacre M. Data reporting and visualization in ecology. Polar Biology. 2016;39(2189–2205).

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. European journal of epidemiology. Springer; 2016;31(4):337–50.

Greenlee ET, DeLucia PR, Newton DC. Driver vigilance in automated vehicles: Hazard detection failures are a matter of time. Human Factors. 2018;60(4):465–76.

Groves J. Bicycle weight and commuting time: Randomised trial. British Medical Journal. 2010;341.

Guirao L, Samitier CB, Costea M, Camos JM, Majo M, Pleguezuelos E. Improvement in walking abilities in transfemoral amputees with a distal weight bearing implant. Prosthetics and Orthotics International. 2017;4(26–32).

Gunnarsson S, Mitchell J, Busch MS, Larson B, Gharacholou SM, Li Z, et al. Outcomes of physician-staffed versus non-physician-staffed helicopter transport for ST-elevation myocardial infarction. Journal of the American Heart Association. 2017;

Hald A. Statistical theory with engineering applications. New York: John Wiley; Sons; 1952.

Hale D, Shrestha PP, Gibson GE, Migliaccio GC. Empirical comparison of Design/Build

and Design/Bid/Build project delivery methods. Journal of Construction Engineering. 2009;135:579–87.

Hammond D, Reid JL, Zukowski S. Adverse effects of caffeinated energy drinks among youth and young adults in Canada: A Web-based survey. CMAJ Open. 2018;6(1):E19.

Hand DJ, Daly F, Lunn AD, McConway KY, Ostrowski E. A handbook of small data sets. London: Chapman; Hall; 1996.

Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. Sepsid *even-skipped* enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. PLoS Genetics. 2008;4(6):e1000106.

Hargreaves BR, McWilliams TP. Polynomial trendline function flaws in Microsoft Excel. Computational Statistics & Data Analysis. Elsevier; 2010;54(4):1190–6.

Harnish RJ, Nataraajan R. Attitudes toward wildlife: The impact of physical attractiveness. Psychology & Marketing. Wiley Online Library; 2020;37(12):1703–7.

Haselgrove C, Straker L, Smith A, O'Sullivan P, Perry M, Sloan N. Perceived school bag load, duration of carriage, and method of transport to school are associated with spinal pain in adolescents: An observational study. Australian Journal of Physiotherapy. Elsevier; 2008;54(3):193–200.

Härdle WK, others. Smoothing techniques: With implementation in S. Springer Science & Business Media; 1991.

Heerfordt IM, Torsnes LR, Philipsen PA, Wulf HC. Photoprotection by sunscreen depends on time spent on application. Photodermatology, Photoimmunology & Photomedicine. Wiley Online Library; 2018;34(2):117–21.

Henderson HV, Velleman PF. Building multiple regression models interactively. Biometrics. JSTOR; 1981;391–411.

Hieger J. Portrait of a homebuyer household: 2 kids and a PC. Orange County Register; 2001.

Hijazi A, Hamad H, Djukic J-P, Daher A, Reda M, Walther M, et al. Reaction of a bidentate ligands (4, 4'-dimethyl 2, 2'-bipyridine) with planar-chiral chloro-bridged ruthenium: Synthesis of cis-dicarbonyl [4, 4'-dimethyl-2, 2'-bipyridine-κO1, κO2]{2-[tricarbonyl (η6-phenylene-κC1) chromium] pyridine-κn} ruthenium hexafluorophosphate. Inorganica Chimica Acta. Elsevier; 2013a.

Hijazi A, Hamad H, Djukic J-P, Daher A, Reda M, Walther M, et al. RETRACTED: Reaction of a bidentate ligands (4, 4'-dimethyl 2, 2'-bipyridine) with planar-chiral chloro-bridged ruthenium: Synthesis of cis-dicarbonyl [4, 4'-dimethyl-2, 2'-bipyridine-κO1, κO2]{2-[tricarbonyl (η6-phenylene-κC1) chromium] pyridine-κn} ruthenium hexafluorophosphate. Inorganica Chimica Acta. Elsevier; 2013b.

Hirst JM, Stedman OJ. The epidemiology of apple scab (Venturia inaequalis (Cke.) Wint.) III. The supply of ascospores. Annals of Applied Biology. Wiley Online Library; 1962;50(3):551–67.

Holgate P. Fitting a straight line to data from a truncated population. Biometrics. 1965;21(3):715–20.

Hosseini R, Mirghotbi M, Pourvali K, Kimiagar SM, Rashidkhani B, Mirghotbi T. The effect of food service system modifications on staff body mass index in an industrial organization. Journal of Paramedical Sciences. 2015;6(1):2008–4978.

Houben AJ, D'Onofrio R, Kokelj SV, Blais JM. Factors affecting elevated arsenic and methyl mercury concentrations in small shield lakes surrounding gold mines near the Yellowknife, NT, (Canada) region. PloS one. Public Library of Science San Francisco, CA USA; 2016;11(4):e0150960.

Hurlimann A, Dolnicar S. Public acceptance and perceptions of alternative water sources: A comparative study in nine locations. International Journal of Water Resources Development. 2016;32(4):650–73.

Huskisson EC. Simple analgesics for arthritis. British Medical Journal. 1974;4:196–200.

IBM Corp. IBM SPSS statistics for Windows, version 24.0. Armonk, NY: IBM Corp; 2016.

Imtiaz S, Ali Z, Chaudhry KA, Abbas F, Anjum N. Assessment of nutritional status of anemic children. Pakistan Journal of Medical & Health Sciences. 2017;11(1):131–5.

Ingalls AG. If you smoke. Scientific American [Internet]. 1936;154(6):310–55. Available from: http://www.jstor.org/stable/26144809.

Jacka FN, Kremer PJ, Leslie ER, Berk M, Patton GC, Toumbourou JW, et al. Associations between diet quality and depressed mood in adolescents: Results from the Australian Healthy Neighbourhoods Study. Australian and New Zealand Journal of Psychiatry. 2010;44(5):435–42.

Janatuinen EK, Kemppainen TA, Julkunen RJK, Kosma VM, Mäki M, Heikkinen M, et al. No harm from five year ingestion of oats in coeliac disease. Gut. 2002;50(3):332–5.

Joglekar G, Scheunemyer JH, LaRiccia V. Lack-of-fit testing when replicates are not available. The American Statistician. 1989;43:135–43.

Julious SA, Mullee MA. Confounding and Simpson's paradox. BMJ. British Medical Journal Publishing Group; 1994;309(6967):1480–1.

Kahane CJ, Hertz E. The long-term effectiveness of center high mounted stop lamps in passenger cars and light trucks [Internet]. NHTSA; 1998. Report No.: DOT HT 8087 696. Available from: https://one.nhtsa.gov/cars/rules/regrev/evaluate/808696.html.

Kahn M. An exhalent problem for teaching statistics. Journal of Statistical Education. 2005;13(2).

Kanji GK. 100 statistical tests. Sage; 2006.

Kardish MR, Mueller UG, Amador-Vargas S, Dietrich EI, Ma R, Barrett B, et al. Blind trust in unblinded observation in ecology, evolution, and behavior. Frontiers in Ecology and Evolution. Frontiers; 2015;3:51.

Keeling KB, Pavur RJ. Numerical accuracy issues in using Excel for simulation studies. Proceedings of the 2004 winter simulation conference, 2004. IEEE; 2004. p. 1513–8.

Kelishadi R, Mozafarian N, Qorbani M, Motlagh ME, Safiri S, Ardalan G, et al. Is snack

consumption associated with meal skipping in children and adolescents? The CASPIAN-IV study. Eat Weight Disorders. 2017;22:321–8.

Kelpin SS, Moore TB, Hull LC, Dillon PM, Perry BL, Thacker LR, et al. Alcohol use and problems in daily and non-daily coffee drinking college females. Journal of Substance Use. 2018;23(6):574–8.

Kheok SW, Chong CY, McCarthy G, Lim WY, Goh KT, Razak L, et al. The efficacy of influenza vaccination in healthcare workers in a tropical setting: A prospective investigator blinded observational study. Annals, Academy of Medicine, Singapore. Academy of Medicine, Singapore; 2008;37(6):465.

Klanian MG, Diaz MD, Aranda J, Juárez CR. Integrated effect of nutrients from a recirculation aquaponic system and foliar nutrition on the yield of tomatoes *Solanum lycopersicum L.* And *Solanum pimpinellifolium*. Environmental Science and Pollution Research. Springer; 2018;1–3.

Ko W-R, Hung W-T, Chang H-C, Lin L-Y. Inappropriate use of standard error of the mean when reporting variability of study samples: A critical evaluation of four selected journals of obstetrics and gynecology. Taiwanese Journal of Obstetrics and Gynecology. Elsevier; 2014;53(1):26–9.

Kohlmeier L, Arminger G, Bartolomeycik S, Bellach B, Rehm J, Thamm M. Pet birds as an independent risk factor for lung cancer: Case-control study. British Medical Journal. 1992;305(6860):986–9.

Köchling J, Geis B, Wirth S, Hensel KO. Grape or grain but never the twain? A randomized controlled multiarm matched-triplet crossover trial of beer and wine. The American Journal of Clinical Nutrition. 2019;109(2):345–52.

Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: Moving toward research the public can really trust. Annals of Internal Medicine. Citeseer; 2007;146(6):450–3.

Lane PW. Generalized linear models in soil science. European Journal of Soil Science. 2002;53:241–51.

Larson N, Loth KA, Nanney MS. Staff training interests, barriers, and preferences in rural and urban child care programs in minnesota. Journal of Nutrition Education and Behavior. Elsevier; 2019;51(3):335–41.

Latour B. Pasteur et Pouchet: Hétérogenèse de l'historie des sciences [Pasteur and Pouchet: Heterogenesis of the history of science]. In: Serres M, editor. Eléments d'histoire des sciences [elements of history of science]. Paris: Bordas; 1989. p. 423–45.

Laws N. Characterisation and strategic treatment of dystrophic muscle [PhD thesis]. Department of Biology; Physical Sciences, University of Southern Queensland; 2005.

LeBlanc VR, MacDonald RD, McArthur B, King K, Lepine T. Paramedic performance in calculating drug dosages following stressful scenarios in a human patient simulator. Prehospital Emergency Care. 2005;9(4):439–44.

Lee G-W, Kim RB, Go SI, Cho HS, Lee SJ, Hui D, et al. Gender differences in hiccup patients:

Analysis of published case reports and case-control studies. Journal of Pain and Symptom Management. 2016a;51(2):278–83.

Lee Y-M, Kim S-A, Lee I-K, Kim J-G, Park K-G, Jeong J-Y, et al. Effect of a brown rice based vegan diet and conventional diabetic diet on glycemic control of patients with type 2 diabetes: A 12-week randomized clinical trial. PloS One. Public Library of Science; 2016b;11(6).

Lemon J. plotrix: A package in the red light district of r. R-News. 2006;6(4):8–12.

Lennon A, Oviedo-Trespalacios O, Matthews S. Pedestrian self-reported use of smart phones: Positive attitudes and high exposure influence intentions to cross the road while distracted. Accident Analysis & Prevention. 2017;98:338–47.

Levine RV. The pace of life. American Scientist. 1990;450–9.

Lewis KP, Vander Wal E, Fifield DA. Wildlife biology, big data, and reproducible research. Wildlife Society Bulletin. Wiley Online Library; 2018;42(1):172–9.

Lipton RB, Stewart WF, Ryan Jr RE, Saper J, Silberstein S, Sheftell F. Efficacy and safety of acetaminophen, aspirin, and caffeine in alleviating migraine headache pain: Three double-blind, randomized, placebo-controlled trials. Archives of Neurology. 1998;55(2):210–7.

Loesch DZ, Stokes K, Huggins RM. Secular trend in body height and weight of Australian children and adolescents. American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists. Wiley Online Library; 2000;111(4):545–56.

London RE, Slagter HA. Statement of retraction: Effects of transcranial direct current stimulation over left dorsolateral pFC on the attentional blink depend on individual baseline performance. Journal of Cognitive Neuroscience. 2021;1.

Lord B, Cui J, Kelly A-M. The impact of patient sex on paramedic pain management in the pre-hospital setting. The American Journal of Emergency Medicine. Elsevier; 2009;27(5):525–9.

Lorenz K, Hoffmann T, Heumann C, Noack B. Effect of toothpaste containing amine fluoride and stannous chloride on the reduction of dental plaque and gingival inflammation. A randomized controlled 12-week home-use study. International Journal of Dental Hygiene. Wiley Online Library; 2019;

Lothian JB, Grey V, Lands LC. Effect of whey protein to modulate immune response in children with atopic asthma. International Journal of Food Science and Nutrition. 2006;57(3/4):204–11.

Lundin KEA, Nilsen EM, Scott HG, Løberg EM, Gjøen A, Bratlie J, et al. Oats induced villous atrophy in coeliac disease. Gut. 2003;52(11):1649–52.

MacDonald M. Is enough really enough?: Evaluation of an alcohol awareness campaign at ECU Joondalup. 2008; Available from: https://ro.ecu.edu.au/theses_hons/1032.

MacGregor GA, Markandu ND, Roulston JE, Jones JC. Essential hypertension: Effect of an oral inhibitor of angiotensin-converting enzyme. British Medical Journal. 1979;2:1106–9.

Macgregor IDM, Rugg-Gunn AJ. Survey of toothbrushing duration in 85 uninstructed English schoolchildren. Community Dentistry and Oral Epidemiology. 1979;7(5):297–8.

Macgregor IDM, Rugg-Gunn AJ. Toothbrushing duration in 60 uninstructed young adults. Community Dentistry and Oral Epidemiology. 1985;13(3):121–2.

Mackowiak PA, Wasserman SS, Levine MM. A critical appraisal of 98.6°f, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. Journal of the American Medical Association. 1992;268(12):1578–80.

Mages S, Hensel O, Zierz AM, Kraya T, Zierz S. Experimental provocation of 'ice-cream headache' by ice cubes and ice water. Cephalalgia. 2017;37(5):464–9.

Manly BFJ, Alberto JAN. Introduction to ecological sampling. CRC Press; 2014.

Mann L, Blotnicky K. Influences of physical environments on university student eating behaviors. International Journal of Health Sciences. 2017;5(2):42–52.

Manzano F, Pérez A-M, Colmenero M, Aguilar M-M, Sánchez-Cantalejo E, Reche A-M, et al. Comparison of alternating pressure mattresses and overlays for prevention of pressure ulcers in ventilated intensive care patients: A quasi-experimental study. Journal of Advanced Nursing. 2013;69(9):2099–106.

March DT, Vinette-Herrin K, Peters A, Ariel E, Blyde D, Hayward D, et al. Hematologic and biochemical characteristics of stranded green sea turtles. Journal of Veterinary Diagnostic Investigation. 2018;

Maron M. Threshold effect of eucalypt density on an aggressive avian competitor. Biological Conservation. 2007;136:100–7.

Marshman M, Dunn PK. Teaching statistics with experiential learning: A visual experience. Submitted;

Martín ISM, Vilar EG, Barrado MR, Barato VP. Soft drink consumption: Do we know what we drink and its implication on health? Mediterranean Journal of Nutrition and Metabolism. IOS Press; 2018;11(1):1–0.

Maunder A, Bessell E, Lauche R, Adams J, Sainsbury A, Fuller NR. Effectiveness of herbal medicines for weight loss: A systematic review and meta-analysis of randomized controlled trials. Diabetes, Obesity and Metabolism. Wiley Online Library; 2020;22(6):891–903.

McCarney R, Warner J, Iliffe S, Van Haselen R, Griffin M, Fisher P. The Hawthorne effect: A randomised, controlled trial. BMC medical research methodology. BioMed Central; 2007;7(1):30.

McCrory P, Meeuwisse WH, Aubry M, Cantu B, Dvořák J, Echemendia RJ, et al. Consensus statement on concussion in sport: The 4th International Conference on Concussion in Sport held in Zurich, November 2012. British Journal of Sports Medicine [Internet]. British Association of Sport; Exercise Medicine; 2013;47(5):250–8. Available from: https://bjsm.bmj.com/content/47/5/250.

McCullough BD, Wilson B. On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. Computational Statistics & Data Analysis. Elsevier; 2002;40(4):713–21.

McLinn SE, Moskal M, Goldfarb J, Bodor F, Aronovitz G, Schwartz R, et al. Comparison

of cefuroxime axetil and amoxicillin-clavulanate suspensions in treatment of acute otitis media with effusion in children. Antimicrobial Agents and Chemotherapy. 1994;38(2):315–8.

Mead R. Plant density and crop yield. Applied Statistics. 1970;19(1):64–81.

Meline T. Selecting studies for systematic review: Inclusion and exclusion criteria. Contemporary Issues in Communication Science and Disorders. 2006;33:21–7.

Mendes P. Reproducible research using biomodels. Bulletin of Mathematical Biology. Springer; 2018;80(12):3081–7.

Mélard G. On the accuracy of statistical procedures in Microsoft Excel 2010. Computational Statistics. Springer; 2014;29(5):1095–128.

Miller EE, Boyle LN. Behavioral adaptations to lane keeping systems: Effects of exposure and withdrawal. Human factors. 2019;61(1):152–64.

Mine N, Wai SH, Lim TC, Kang W. An observational study on the productivity of formwork in building construction. ISARC Proceedings of the international symposium on automation and robotics in construction. IAARC Publications; 2015. p. 1.

Mohammadpoorasl A, Hajizadeh M, Marin S, Heidari P, Ghalenoee M. Prevalence and pattern of using headphones and its relationship with hearing loss among students. Health Scope. 2018;(In Press).

Moir RJ. A note on the relationship between the digestible dry matter and the digestable energy content of ruminant diets. Australian Journal of Experimental Agriculture and Animal Husbandry. 1961;1:24–6.

Montgomery DC, Peck EA. Introduction to regression analysis. New York: Wiley; 1992.

Mullaney J, Lucke T. Practical review of pervious pavement designs. CLEAN–Soil, Air, Water. Wiley Online Library; 2014;42(2):111–24.

Myers RH. Classical and modern regression with applications. second. Duxbury; 1990.

Myers RH, Montgomery DC, Vining GG. Generalized linear models with applications in engineering and the sciences. Wiley; 2002.

Nagele P. Misuse of standard error of the mean (SEM) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. British Journal of Anaesthesia. Oxford University Press; 2003;90(4):514–6.

Nakamura N, Horibe S, Toritsuka Y, Mitsuoka T, Yoshikawa H, Shino K. Acute grade III medial collateral ligament injury of the knee associated with anterior cruciate ligament tear: The usefulness of magnetic resonance imaging in determining a treatment regimen. American Journal of Sports Medicine. 2000;31(2):261–7.

Nelson W. Applied life data. Wiley; 1982.

Norris DR. Carry-over effects and habitat quality in migratory populations. Oikos. Wiley Online Library; 2005;109(1):178–86.

O'Connor BA, Carman J, Eckert K, Tucker G, Givney R, Cameron S. Does using potting

mix make you sick? Results from a *Legionella longbeachae* case-control study in South Australia. Epidemiology and Infection. 2007;135:34–9.

Ocepek J, Roberts AEK, Vidmar G. Evaluation of treatment in the smart home IRIS in terms of functional independence and occupational performance and satisfaction. Computational and Mathematical Methods in Medicine. 2013;2013:1–0.

Ooms J. gifski: Highest quality GIF encoder [Internet]. 2018. Available from: https://CRAN.R-project.org/package%20=%20gifski.

Oppenheimer DM. Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition. 2006;20(2):139–56.

Pamphlett R. Exposure to environmental toxins and the risk of sporadic motor neuron disease: An expanded Australian case–control study. European Journal of Neurology. 2012;19:1343–8.

Panda RP, Das SS, Sahoo PK. Relation between bitumen content and percentage air voids in semi dense bituminous concrete. Journal of The Institution of Engineers (India): Series A. 2018;99(2):327–32.

Panko R. What we don't know about spreadsheet errors today: The facts, why we don't believe them, and what we need to do. arXiv preprint arXiv:160202601. 2016;

Panko RR, Sprague Jr RH. Hitting the wall: Errors in developing and code inspecting a 'simple' spreadsheet model. Decision Support Systems. Elsevier; 1998;22(4):337–53.

Paul WL, Taylor PA. A comparison of occupant comfort and satisfaction between a green building and a conventional building. Building and Environment. 2008;43:1858–70.

Peat J, Elliott E, Baur L, Keena V. Scientific writing: Easy when you know how. London: BMJ Books; 2002.

Pons PT, Haukoos JS, Bludworth W, Cribley T, Pons KA, Markovchick VJ. Paramedic response time: Does it affect patient survival? Academic Emergency Medicine. 2005;12(7):594–600.

Porter SR. Pros and cons of paper and electronic surveys. New Directions for Institutional Research. Wiley Online Library; 2004;2004(121):91–7.

Pruim R. NHANES: Data from the US National Health and Nutrition Examination Study [Internet]. 2015. Available from: https://CRAN.R-project.org/package%20=%20NHANES.

Quan Q-D, Tran H-D, Chung A-D. The relation of body score (body height/body length) and haplotype E on Phu Quoc Ridgeback dogs (*Canis familiaris*). Journal of Entomology and Zoology Studies. 2017;5:388–94.

R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: https://www.R-project.org/.

Radun I, Lajunen T. Bicycle helmets and the experimenter effect. Psychological Science. SAGE Publications Sage CA: Los Angeles, CA; 2018;29(6):1020–2.

Raftery M, Kemp S, Patricios J, Makdissi M, Decq P. It is time to give concussion an operational definition: A 3-step process to diagnose (or rule out) concussion within 48 h of injury: World Rugby guideline. British Journal of Sports Medicine [Internet]. British Association of Sport; Exercise Medicine; 2016;50(11):642–3. Available from: https://bjsm.bmj.com/content/50/11/642.

Richardson AM, Dunn PK, Hutchins R. Identification and definition of lexically ambiguous words in statistics by tutors and students. International Journal of Mathematical Education in Science and Technology. Taylor & Francis; 2013;44(7):1007–19.

Ridgewell C, Sipe N, Buchanan N. School travel modes: Factors influencing parental choice in four Brisbane schools. Urban Policy and Research. Taylor & Francis; 2009;27(1):43–57.

Robson C. Real world research. Second. Blackwell Publishing; 2002.

Romero-Blanco C, Rodríguez-Almagro J, Onieva-Zafra MD, Parra-Fernández ML, Prado-Laguna MDC, Hernández-Martínez A. Physical activity and sedentary lifestyle in university students: Changes during confinement due to the COVID-19 pandemic. International Journal of Environmental Research and Public Health. 2020;17(18):6567.

Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. Journal of the Royal Statistical Society, Series C. 1994;43(3):429–67.

Rubbo P, Helmann CL, Bilynkievycz dos Santos C, Pilatti LA. Retractions in the engineering field: A study on the web of science database. Ethics & Behavior. Taylor & Francis; 2019;29(2):141–55.

Ruscoe P, Dunn P. Fatal accident rates in private pilots (PPL) 1992–1999: An Australian review. AMSANZ annual conference & scientific meeting. 2003.

Russell J, Flood V, Yeatman H, Mitchell P. Prevalence and risk factors of food insecurity among a cohort of older Australians. Journal of Nutrition, Health and Aging. 2014;18(1):3–8.

Russell TC, Herbert CA, Kohen JL. High possum mortality on urban roads: Implications for the population viability of the common brushtail and the common ringtail possum. Australian Journal of Zoology. 2009;57:391–7.

Sacks FM, Bray GA, Carey VJ, Smith SR, Ryan DH, Anton SD, et al. Comparison of weight-loss diets with different compositions of fat, protein, and carbohydrates. The New England Journal of Medicine. 2009;360(9):859–73.

Salimirad F, Srimathi NL. The relationship between, psychological well-being and occupation self-efficacy among teachers in the city of Msore, India. International Journal of Indian Psychology. 2016;3:14–21.

Schepaschenko D, Shvidenko A, Usoltsev VA, Lakyda P, Luo Y, Vasylyshyn R, et al. A dataset of forest biomass structure for Eurasia. Scientific Data. 2017;4:1–1.

Schmid AB, Elliott JM, Stridwick MW, Little M, Coppeiters MW. Effect of splinting and exercise on intraneural edema of the median nerve in Carpal Tunnel Syndrome—an MRI study to reveal therapeutic mechanisms. Journal of Orthopaedic Research. 2012;30(8):1343–50.

Schorling JB, Roach J, Siegel M, Baturka N, Hunt DE, Guterbock TM, et al. A trial of church-

based smoking cessation interventions for rural African Americans. Preventative Medicine. 1997;26(1):92–101.

Schröder JJ, Vermeulen GD, Van der Schoot JR, Van Dijk W, Huijsmans JFM, Meuffels GJHM, et al. Maize yields benefit from injected manure positioned in bands. European Journal of Agronomy. Elsevier; 2015;64:29–36.

Schwitzgebel E. Do ethicists steal more books? Philosophical Psychology. Taylor & Francis; 2009;22(6):711–25.

Sedgwick P. Non-response bias versus response bias. BMJ. British Medical Journal Publishing Group; 2014;348:g2573.

Shahabuddin S. Plagiarism in academic. International Journal of Teaching and Learning in Higher Education. ERIC; 2009;21(3):353–9.

Shamim T. Development of a guideline to approach plagiarism in Indian scenario. Indian Journal of Dermatology. Medknow Publications & Media Pvt. Ltd.; 2014;59(5):473.

Shiffman S, West RJ, Gilbert DG. Recommendation for the assessment of tobacco craving and withdrawal in smoking cessation trials. Nicotine & Tobacco Research. 2004;6(4):599–614.

Shoemaker AL. What's normal? – temperature, gender, and heart rate. Journal of Statistics Education [Internet]. 1996;4(2). Available from: http://jse.amstat.org/v4n2/datasets.shoemaker.html.

Sidi J. slickR: Create interactive carousels with the JavaScript 'slick' library [Internet]. 2018. Available from: https://CRAN.R-project.org/package%20=%20slickR.

Siegrist M. The use or misuse of three-dimensional graphs to represent lower-dimensional data. Behaviour & Information Technology. Taylor & Francis; 1996;15(2):96–100.

Sievert C. Plotly for R [Internet]. 2018. Available from: https://plotly-r.com.

Silva EJNL, Carvalho NK, Prado MC, Zanon M, Senna PM, Souza EM, et al. Push-out bond strength of injectable pozzolan-based root canal sealer. Journal of Endodontics. 2016;42(11):1656–9.

Silverman SG, Tuncali K, Adams DF, Nawfel RD, Zou KH, Judy PF. CT fluoroscopy-guided adbominal interventions: Techniques, results, and radiation exposure. Radiology. 1999;212:673–81.

Simons JE, Holmes DT. Reproducible research and reports with R. Journal of Applied Laboratory Medicine. Oxford University Press; 2019;4(3):471–3.

Singh RK, Prasad G. Long-term mortality after lower-limb amputation. Prosthetics and Orthotics International. 2016;40(5).

Sloof P, Burg J van den, Voogd A, Benne R. The nucleotide sequence of a 3.2 kb segment of mitochondrial maxicircle DNA from crithidia fasciculata containing the gene for cytochrome oxidase subunit III, the n-terminal part of the apocytochrome b gene and a possible frameshift gene; further evidence for the use of unusual initiator triplets in trypanosome mitochondria. Nucleic Acids Research. Oxford University Press; 1987;15(1):51–65.

Smith DT, Attwell K, Evers U. Majority acceptance of vaccination and mandates across the

political spectrum in Australia. Politics. SAGE Publications Sage UK: London, England; 2020;40(2):189–206.

Smith GCS, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. BMJ. 2003;327:1459–61.

Smyth GK. OzDASL—Australasian data and story library [Internet]. 2010. Available from: http://www.statsci.org/data/index.html.

Snowden JM, Basso O. Causal inference in studies of preterm babies: A simulation study. BJOG: An International Journal of Obstetrics & Gynaecology. Wiley Online Library; 2018;125(6):686–92.

Soetaert K. Diagram: Functions for visualising simple graphs (networks), plotting flow diagrams [Internet]. 2017. Available from: https://CRAN.R-project.org/package%20=%20diagram.

Sokal RR, Rohlf FJ. Biometry: The principles and practice of statistics in biological research. Third. New York: W. H. Freeman; Company; 1995.

Solomon PR, Adams F, Silver A, Zimmer J, Veaux R De. Ginko for memory enhancement. Journal of the American Medical Association. 2002;288(7):835–40.

Stead-Richardson E, Bradshaw D, Friend T, Fletcher T. Monitoring reproduction in the critically endangered marsupial, Gilbert's potoroo (Potorous gilbertii): Preliminary analysis of faecal oestradiol-17β, cortisol and progestagens. General and comparative endocrinology. Elsevier; 2010;165(1):155–62.

Steele M, Zahr RA, Kirshbom PM, Kopf GS, Karimi M. Quality of life for historic cavopulmonary shunt survivors. World Journal for Pediatric and Congenital Heart Surgery. 2016;7(5):630–4.

Steele S. Babies by the dozen for Christmas: 24-hour baby boom. The Sunday Mail. 1997;7.

Stensballe J, Looms D, Nielsen PN, Tvede M. Hydrophilic-coated catheters for intermittent catheterisation reduce urethral microtrauma: A prospective, randomised, participant-blinded, crossover study of three different types of catheter. European Urology. 2005;48:978–83.

Sterndale SO, Miller DW, Mansfield JP, Kim JC, Pluske JR. Increasing dietary tryptophan and decreasing other large neutral amino acids increases weight gain and feed intake in weaner pigs infected with escherichia coli. Animal Production Science. CSIRO; 2017;57(12):2410–0.

Stirrat SC. Age structure, mortality and breeding in a population of agile wallabies (*Macropus agilis*). Australian Journal of Zoology. 2008;56:431–9.

Stone RC, Auliciems A. SOI phase relationships with rainfall in eastern Australia. International Journal of Climatology. 1992;12:625–36.

Stone RC, Hammer GL, Marcussen T. Prediction of global rainfall probabilities using phases of the southern oscillation index. Nature. 1996;384:252–5.

Strayer DL, Johnston WA. Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. Psychological Science. 2001;12(6):462–6.

Su Y-S. It's easy to produce chartjunk using Microsoft Excel 2007 but hard to make good graphs. Computational Statistics & Data Analysis. Elsevier; 2008;52(10):4594–601.

Sutherland J, Edwards P, Shankar B, Dangour AD. Fewer adults add salt at the table after initiation of a national salt campaign in the UK: A repeated cross-sectional analysis. British Journal of Nutrition. 2012;

Swinnen E, Baeyens J-P, Mulders B Van, Verspecht J, Degelaen M. The influence of the use of ankle-foot orthoses on thorax, spine, and pelvis kinematics during walking in children with cerebral palsy. Prosthetics and Orthotics International. 2017;2017.

Tager IB, Weiss ST, Rosner B, Speizer FE. Effect of parental cigarette smoking on the pulmonary function of children. American Journal of Epidemiology. 1979;110(1):15–26.

Talukdar DK. A study of correlation between california bearing ratio (CBR) value with other properties of soil. International Journal of Emerging Technology and Advanced Engineering. Citeseer; 2014;4(1):559–62.

Taylor MP, Camenzuli D, Kristensen LJ, Forbes M, Zahran S. Environmental lead exposure risks associated with children's outdoor playgrounds. Environmental Pollution. 2013;178:447–54.

Taylor R, Gumming R, Woodward A, Black M. Passive smoking and lung cancer: A cumulative meta-analysis. Australian and New Zealand Journal of Public Health. Wiley Online Library; 2001;25(3):203–11.

Teillet E, Urbano C, Cordelle S, Schlich P. Consumer perception and preference of bottled and tap water. Journal of sensory studies. Wiley Online Library; 2010;25(3):463–80.

Telford RD, Cunningham RB. Sex, sport, and body-size dependency of hematology in highly trained athletes. Medicine and Science in Sports and Exercise. 1991;23(7):788–94.

The jamovi Project. jamovi (version 1.0) [computer software] [Internet]. Available from: https://www.jamovi.org.

The Open University. MDST242 Statistics in Society, Unit A0: Introduction. The Open University; 1983.

Tsai H-M. A mechanistic approach to the diagnosis and management of atypical hemolytic uremic syndrome. Transfusion Medicine Reviews. Elsevier; 2014;28(4):187–97.

Tufte ER, McKay SR, Christian W, Matey JR. Visual explanations: Images and quantities, evidence and narrative. AIP; 1998.

Tully MP. Articulating questions, generating hypotheses, and choosing study designs. The Canadian Journal of Hospital Pharmacy. 2014;67(1):31.

US Department of Transportation. 2009 national household travel survey user's guide. US Department of Transportation Washington, DC; 2011.

van Helmont J-B. On the necessity of leavens in transformations. In: Conte J. L., editor. Les oevres de Jean-Baptiste van Helmont. Lyon; 1671.

van Rij J. Plotfunctions: Various functions to facilitate visualization of data and analysis

[Internet]. 2020. Available from: https://CRAN.R-project.org/package%20=%20plotfunctions.

Vanderkam D, Allaire JJ, Owen J, Gromer D, Thieurmel B. Dygraphs: Interface to 'dygraphs' interactive time series charting library [Internet]. 2018. Available from: https://CRAN.R-project.org/package%20=%20dygraphs.

Vaughn MR, Brooks E, Oorschot RAH van, Baindur-Hudson S. A comparison of macroscopic and microscopic hair color measurements and a quantification of the relationship between hair color and thickness. Microscopy and Microanalysis. Cambridge University Press; 2009;15(3):189–93.

Venables B, Hornik K. oz: Plot the Australian coastline and states [Internet]. 2016. Available from: https://CRAN.R-project.org/package%20=%20oz.

Verdecchia P, SchillacI G, Borgioni C, Ciucci A, Zampi I, Battistelli M, et al. Cigarette smoking, ambulatory blood pressure and cardiac hypertrophy in essential hypertension. Journal of Hypertension. 1995;13(10):1209–15.

Waber RL, Shiv B, Carmon Z, Ariely D. Commercial features of placebo and therapeutic. Journal of the American Medical Association. 2008;299(9):1016–7.

Wallace HJ, Fear MW, Crowe MM, Martin LJ, Wood FM. Identification of factors predicting scar outcome after burn in adults: A prospective case-control study. Burns. 2017;43:1271–83.

Wang L, Li R, Wang C, Liu Z. Driver injury severity analysis of crashes in a western China's rural mountainous county: Taking crash compatibility difference into consideration. Journal of Traffic and Transportation Engineering (English edition). Elsevier; 2020;

Weil K, Hooper L, Afzal Z, Esposito M, Worthington HV, Wijk A van, et al. Paracetamol for pain relief after surgical removal of lower wisdom teeth. Cochrane Database of Systematic Reviews. John Wiley & Sons, Ltd; 2007;(3).

Wetzel N. McDonald's french fries: Would you like small or large fries? STATS. 2005;43:12–4.

Wickham H. ggplot2: Elegant graphics for data analysis [Internet]. Springer-Verlag New York; 2016. Available from: http://ggplot2.org.

Wickham H. scales: Scale functions for visualization [Internet]. 2018. Available from: https://CRAN.R-project.org/package%20=%20scales.

Wickham H, François R, Henry L, Müller K. dplyr: A grammar of data manipulation [Internet]. 2019. Available from: https://CRAN.R-project.org/package%20=%20dplyr.

Willems JP, Saunders JT, Hunt DE, Schorling JB. Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. Southern Medical Journal. 1997;90(8):814–20.

Williams B, Boyle M. Estimation of external blood loss by paramedics: Is there any point? Prehospital and Disaster Medicine. 2007;22(6):502–6.

Williams P, Goring R, Franklin J. Mechanical chest compressions and survival in the emergency setting. Journal of Paramedic Practice. MA Healthcare London; 2021;13(2):62–8.

Wojcik LA, Thelen DG, Schultz AB, Ashton-Miller JA, Alexander NB. Age and gender differences in single-step recovery from a forward fall. Journal of Gerentology. 1999;54A(1):M44–50.

Wong P-F, Chan EY-T, Ng DK-K, Kwok K-L, Yip AY-F, Leung S-Y. Correlation between 6-min walk test and cardiopulmonary exercise test in Chinese patients. Pediatric Respirology and Critical Care Medicine. Medknow Publications; 2018;2(2):32.

Woodward M, Walker ARP. Sugar consumption and dental caries: Evidence from 90 countries. British Dental Journal. 1994;176:297–302.

Woolf K, St. Thomas MM, Hahn N, Vaughan LA, Carlson AG, Hinton P. Iron status in highly active and sedentary young women. International Journal of Sport Nutrition and Exercise Metabolism. 2009;19:519–35.

Wu K-S, Lee SS-J, Chen J-K, Chen Y-S, Tsai H-C, Chen Y-J, et al. Identifying heterogeneity in the Hawthorne effect on hand hygiene observation: A cohort study of overtly and covertly observed results. BMC Infectious Diseases. BioMed Central; 2018;18(1):369.

Wunderlich C. Das verhalten der Eiaenwarme in Krankenheitem. Leipzig, Germany: Otto Wigard; 1868.

Xie Y. animation: An R package for creating animations and demonstrating statistical methods. Journal of Statistical Software [Internet]. 2013;53(1):1–27. Available from: http://www.jstatsoft.org/v53/i01/.

Xie Y. Dynamic documents with R and knitr [Internet]. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC; 2015. Available from: https://yihui.name/knitr/.

Xie Y. Bookdown: Authoring books and technical documents with R markdown [Internet]. Boca Raton, Florida: Chapman; Hall/CRC; 2016. Available from: https://github.com/rstudio/bookdown.

Xie Y, Cheng J, Tan X. DT: A wrapper of the JavaScript library 'DataTables' [Internet]. 2018. Available from: https://CRAN.R-project.org/package%20=%20DT.

Yang P, Chan M, Carver S, Hu D. How do wombats make cubed poo? Bulletin of the American Physical Society. APS; 2018;

Yeh RW, Valsdottir LR, Yeh MW, Shen C, Kramer DB, Strom JB, et al. Parachute use to prevent death and major trauma when jumping from aircraft: Randomized controlled trial. BMJ. 2018;363:k5094.

Zagorsky JL. Are blondes really dumb? Economics Bulletin. 2016;36(1):401–10.

Zhu H. kableExtra: Construct complex table with 'kable' and pipe syntax [Internet]. 2018. Available from: https://CRAN.R-project.org/package%20=%20kableExtra.

Ziemann M, Eren Y, El-Osta A. Gene name errors are widespread in the scientific literature. Genome Biology. BioMed Central; 2016;17(1):1–3.

Zimmerman DW. A note on the interpretation of the paired-samples t-test. Journal of Educational and Behavioral Statistics. 1997;22(3):349–60.

Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. Radiology. 2003;227:617–28.

Index

P-value, 374
t-score, 373
68–95–99.7 rule, 203

bias, 76
blocking, 95
boxplot, 198

carryover effect, 98
cases, 15
Central Limit Theorem, 270
cluster sampling, 71
comparison, 18
conceptual definition, 13
conditions, 45
confidence interval
    difference between two means, 336
    mean differences, 322
    means, 308
    odds ratios, 352
    proportions, 286
confounder, 83
confounding, 83
confounding variable, 83
connection, 19
contengency table, 350
control variables, 95
convenience samples, 65

descriptive study, 44

ecological validity, 122
empirical rule, 203
encoding, 158, 175
event, 241
exclusion criteria, 16, 17, 131
experiment, 47
experimental study, 47
experimenter effect, 100
explanatory variable, 31, 32
extraneous variable, 83
extrapolation, 487

Hawthorne effect, 99
hypotheses, 370

inclusion criteria, 16
independence, 247
individuals, 15
intercept, 482
internal validity, 121
intervention, 19
IQR, 198

judgement samples, 65

levels, 350
levels of a qualitative variable, 141
lurking variable, 84, 219

multistage sampling, 72

natural variation, 371
non-random sampling, 65
non-response bias, 76
normal distribution, 253

observational study, 45
observer effect, 100, 101
one-tailed alternative hypothesis, 389
operational definition, 13
outcome, 18, 577
outliers, 164, 201
overplotting, 162

paired data, 173, 174
parameter, 188
percentage, 216
placebo, 102
placebo effect, 101
population, 6, 15, 188
practical importance, 393, 394
predictions, 486
probability, 241
proportion, 216

quasi-experiment, 49

random allocation, 97
random numbers, 67
random sapling, 97
representative samples, 73

response variable, 31
Rounding, 192
RQs, 11

sample, 6, 16, 188
sample proportion, 216
sample space, 241
sampling distribution, 270, 272, 390
    sample mean, 307
sampling frame, 67
sampling variation, 62, 390, 433
    correlation coefficient, 472
scientific process, 1
selection bias, 76
self-selecting samples, 65
sensitivity, 249
simple random samples, 67
slope, 482
specificity, 249
statistic, 188
statistical significance, 393
statistically significant, 392
stratified sampling, 70
subjects, 15
systematic sampling, 68

test statistic, 390, 433, 539
treatments, 47
true experiment, 48
two-tailed alternative hypothesis, 389

unit of analysis, 34, 349
unit of observation, 34

variable, 30
voluntary response samples, 65