Statistical Techniques for Biological and Environmental Sciences

Brad Duthie

2023-03-19

Contents

Preface **11**
Why this module is important . . . 11
Intended learning outcomes (ILOs) . . . 12
Accessibility . . . 12
Teaching overview . . . 12
Canvas . . . 14
Assessment overview . . . 15
Practicals . . . 19
Optional help hours . . . 19
Jamovi statistical software . . . 20
Timetable . . . 21
License . . . 22

I. Background mathematics and data organisation **23**

Week 1 Overview **25**

1. Background mathematics **27**
1.1. Numbers and operations . . . 28
1.2. Logarithms . . . 32
1.3. Order of operations . . . 35

2. Data organisation **39**
2.1. Tidy data . . . 40
2.2. Data files . . . 42
2.3. Managing data files . . . 43

3. Practical: Preparing data **47**
3.1. Exercise 1: Transferring data to a spreadsheet . . . 47
3.2. Exercise 2: Making spreadsheet data tidy . . . 52
3.3. Exercise 3: Making data tidy again . . . 53
3.4. Exercise 4: Tidy data and spreadsheet calculations . . . 54
3.5. Summary . . . 58

II. Statistical concepts 59

Week 2 Overview 61

4. Populations and samples 63

5. Types of variables 65

6. Accuracy, precision, and units 67
6.1. Accuracy . . . 67
6.2. Precision . . . 67
6.3. Systems of units . . . 68
6.4. Other examples of units . . . 70

7. Uncertainty propogation 71
7.1. Adding or subtracting errors . . . 71
7.2. Multiplying or dividing errors . . . 72
7.3. Applying formulas for combining errors . . . 73

8. Practical. Introduction to Jamovi 75
8.1. Exercise for summary statistics . . . 75
8.2. Exercise on transforming variables . . . 80
8.3. Exercise on computing variables . . . 83
8.4. Summary . . . 86

III. Summary statistics 87

Week 3 Overview 89

9. Decimal places, significant figures, and rounding 91
9.1. Decimal places and significant figures . . . 91
9.2. Rounding . . . 92

10. Graphs 95
10.1. Histograms . . . 95
10.2. Barplots and pie charts . . . 96
10.3. Box-whisker plots . . . 98

11. Measures of central tendency 103
11.1. The mean . . . 103
11.2. The mode . . . 106
11.3. The median and quantiles . . . 107

12. Measures of spread 109
12.1. The range . . . 109

12.2. The inter-quartile range . . . 110
12.3. The variance . . . 111
12.4. The standard deviation . . . 115
12.5. The coefficient of variation . . . 116
12.6. The standard error . . . 116

13. *Practical*. Plotting and statistical summaries in Jamovi 121
13.1. Reorganise the dataset into a tidy format . . . 121
13.2. Histograms and box-whisker plots . . . 124
13.3. Calculate summary statistics . . . 127
13.4. Reporting decimals and significant figures . . . 129
13.5. Comparing across sites . . . 130

IV. Probability models and the Central Limit Theorem 133

Week 4 Overview 135

14. Introduction to probability models 137
14.1. An instructive example . . . 138
14.2. Biological applications . . . 141
14.3. Sampling with and without replacement . . . 142
14.4. Probability distributions . . . 143
14.5. Summary . . . 153

15. The Central Limit Theorem (CLT) 155
15.1. The distribution of means is normal . . . 155
15.2. Probability and z-scores . . . 159

16. *Practical*. Probability and simulation 165
16.1. Probabilities from a dataset . . . 166
16.2. Probabilities from a normal distribution . . . 169
16.3. Central limit theorem . . . 174

V. Statistical inference 181

Week 5 Overview 183

17. Confidence intervals (CIs) 185
17.1. Normal distribution CIs . . . 186
17.2. Binomial distribution CIs . . . 189

18. The t-interval 193

19. *Practical.* z- and t- intervals **199**
19.1. Confidence intervals with distrACTION 200
19.2. Confidence intervals from z- and t-scores 204
19.3. Confidence intervals for different sample sizes (t- and z-) 205
19.4. Proportion confidence intervals 207
19.5. Another proportion confidence interval 208

VI. Hypothesis testing **211**

Week 6 Overview **213**

20. What is hypothesis testing? **215**
20.1. How ridiculous is our hypothesis? 215
20.2. Statistical hypothesis testing 217
20.3. P-values, false positives, and power 218

21. The t-test **221**
21.1. One sample t-test 221
21.2. Independent samples t-test 224
21.3. Paired sample t-test 228
21.4. Assumptions of t-tests 230
21.5. Non-parametric alternatives 232
21.6. Summary 237

22. *Practical.* Hypothesis testing and t-tests **239**
22.1. Exercise on a simple one sample t-test 240
22.2. Exercise on a paired t-test 244
22.3. Wilcoxon test 245
22.4. Independent samples t-test 246
22.5. Mann-Whitney U Test 250

VII. Review of parts I-V **253**

Week 7 Overview (Reading week) **255**

VIII Analysis of Variance (ANOVA) **257**

Week 8 Overview **259**

23. Analysis of variance **261**
23.1. The F-distribution 263
23.2. One-way ANOVA 265

23.3. Assumptions of ANOVA 270

24. Multiple comparisons **273**

25. Kruskall-Wallis H test **275**

26. Two-way ANOVA **279**

27. *Practical*. ANOVA and associated tests **285**
27.1. One-way ANOVA (site) 285
27.2. One-way ANOVA (profile) 289
27.3. Multiple comparisons 291
27.4. Kruskall-Wallis H test 293
27.5. Two-way ANOVA 295

IX. Counts and Correlation **299**

Week 9 Overview **301**

28. Frequency and count data **303**
28.1. The Chi-square distribution 304
28.2. Chi-squared goodness of fit 307
28.3. Chi-squared test of association 310

29. Correlation **315**
29.1. Scatterplots 315
29.2. The correlation coefficient 316
29.3. Correlation hypothesis testing 322

30. *Practical*. Analysis of counts and correlations **325**
30.1. Survival goodness of fit 326
30.2. Colony goodness of fit 328
30.3. Chi-Square test of association 329
30.4. Pearson product moment correlation test 330
30.5. Spearman rank correlation test 333
30.6. Untidy goodness of fit 334

X. Linear Regression **337**

Week 10 Overview **339**

31. Regression key concepts **341**

32. Regression validity **343**

33. Introduction to multiple regression **345**

34. Model selection (maybe remove this?) **347**

35. *Practical.* Using regression **349**
35.1. Regression Exercise 1 349
35.2. Regression Exercise 2 349
35.3. Regression Exercise 3 349
35.4. Regression Exercise 4 349

XI. Randomisation approaches **351**

Week 11 Overview **353**

36. Introduction to randomisation **355**

37. Assumptions of randomisation **357**

38. Bootstrapping **359**

39. Monte Carlo **361**

40. *Practical.* Using R **363**
40.1. R Exercise 1 363
40.2. R Exercise 2 363
40.3. R Exercise 3 363

XII. Statistical Reporting **365**

Week 12 Overview **367**

41. Reporting statistics **369**

42. More introduction to R **371**

43. More getting started with R **373**

44. *Practical.* Using R **375**
44.1. R Exercise 1 375
44.2. R Exercise 2 375
44.3. R Exercise 3 375

XIIIReview of parts (VII-XII) **377**

Module summary **379**

A. Common Marking Scheme **381**

B. Uncertainty derivation **383**

C. Statistical tables **387**
C.1. Wilcoxon signed rank critical values 388
C.2. Mann-Whitney U critical values 389

Preface

Welcome to SCIU4T4, Statistical Techniques! Statistical techniques are tools that allow us to make inferences about the world using data. These tools are indispensable in the sciences, and their importance in research continues to grow. Developing a statistical understanding will improve your ability to critically evaluate the scientific literature and conduct your own scientific research. In this module, you will learn important skills for working with biological and environmental data sets. Many of these skills will be directly applied in subsequent modules.

This preface introduces how the module will be structured. We recognise that this can be a daunting module, especially for students who are not confident with mathematics or computer skills. Our hope this semester is to build your confidence, knowledge, interest, and appreciation for statistics. We will do this by presenting the learning material in an interesting and accessible way, and by assessing your learning fairly and transparently, and with plenty of detailed feedback. If you stick with it, then by the end of this semester you should have all of the tools that you need to conduct your own statistical analysis, or to interpret the analysis of other researchers. We are looking forward to helping you learn!

Why this module is important

Nearly all research in the biological and environmental sciences relies on data analysis of some kind. Statistical literacy is therefore important, not just for *doing* research, but also for *understanding* and *evaluating* the research of other scientists. Throughout this book, we will illustrate the importance of statistics using examples inspired by, or directly sourced from, real-world projects in the biological and environmental sciences. Many of these examples, including those used in lab practicals, will draw from research projects conducted at the University of Stirling. Several examples will focus on research that is important for addressing major global challenges in sustainability, food security, conservation, or the spread of disease. Other examples will focus on research that addresses fundamental scientific questions about ecology and evolution. We hope that you will find topics that interest you, and that this module will inspire you to learn more about statistics.

Intended learning outcomes (ILOs)

Modules at the University of Stirling all include a set of Intended Learning Outcomes (ILOs). As the name implies, these ILOs define the core learning outcomes for a module. In statistical techniques, there are 4 ILOs around which the rest of the module is based.

1. Manipulate datasets and characterise their statistical properties.
2. Demonstrate an understanding of null hypothesis testing.
3. Choose and apply the correct statistical test to unseen data using statistical software.
4. Interpret the results of statistical tests in order to generate conclusive statements on scientific problems.

The methods you learn in this module will be applied in future modules, including your fourth year honours dissertation project. It will also provide you with key data analysis skills that might be useful to you in the future, both in your personal and professional life. During the semester, we will highlight the relevance of these skills using real examples from biological and environmental sciences and use readings and software that are free and open access (and therefore accessible to you even after the module and your degree have completed).

Accessibility

We are committed to making this module accessible, which is why the material in this book is available in multiple formats (online, PDF, audio). Figures in this book should have accessible contrast and colour, and informative alt images. The font size, type, and colour scheme should also be adjustable in the online version. Links should make it easy to navigate to different parts of the book. Lectures are provided in manageable chunks and with the option for captions. We have tried to be as clear as possible about the timetable, learning content, and assessments. If some aspect of the module is inaccessible or unclear to you, then please let us know, and we will address it!

Teaching overview

The learning content of this module will be delivered both online and face-to-face. Each week, there will be new chapters to read in this book, new lectures to watch on Canvas, and a new practice quiz. These can all be completed online at any time during the week. If possible, it is probably best to read and watch lectures before the weekly practical, then take the quiz after the practical. Weekly practicals are on Wednesday afternoon (Group A) or Thursday morning (Group B), depending on which group you sign up for

(you only need to attend one). Weekly face-to-face optional help sessions are on Friday afternoons.

- **Practical Group A**: WED 13:05-15:55 in Cottrell 2A17
- **Practical Group B**: THU 09:05-11:55 in Cottrell 2A17
- **Optional Help**: FRI 15:05-17:55 in Cottrell 1A13

These are all of the scheduled face-to-face sessions, with the exception of one help session at the very end of the semester. See the full schedule for all of the specific dates, times, and locations of sessions.

Table 1.: Summary of online and face-to-face teaching.

Online material	Face-to-face sessions
Book chapters	Weekly practicals
Lecture videos	Option help sessions
Practice quizzes	

For each weak, links to all chapters, lectures, and quizzes are available in this book (e.g., the Week 1 Overview), and on the Learning and Teaching page on Canvas.

Book chapters

The book chapters that you need to read for each week are all listed at the start of each section in this book (e.g., Week 1), and in each week's Learning and teaching content on Canvas. On Canvas, this includes links to the book chapters online, PDF copies of the chapters, and audio recordings of the module coordinator reading them. Reading the book online gives you the option of adjusting the text size (small or large), font (serif or sans serif), and background (white, sepia, or night). It should be quite readable on a mobile phone, or on a computer screen. The PDF of the book is identical, but organised more in the form of a traditional textbook. An eBook is in development, but the equations do not render well.

Book chapters are generally quite short, and longer chapters are broken down into manageable subsections. Wherever relevant, links are provided to other chapters and references, and to interactive applications that make concepts easier to visualise. These applications will also be embedded into Canvas.

> Click here for an example interactive application.

Information that is interesting but not critical to know is generally relegated to footnotes.

Additional readings

This book is the only one that you **need** to read to do well in this module, but each week also includes some readings that are recommend, suggested, or advanced. Recommended readings provide similar information to what is in this book, but more in-depth or from a slightly different perspective. All recommended readings will be free to view or download, so you will never need to pay for this material. Suggested readings provide a bit more context for the taught material, but might not always be directly relevant to the learning material of the module. Advanced readings go beyond the taught material and are sometimes quite technical or mathematically dense.

Table 2.: How to interpret different reading recommendations for each week.

Reading	Category	Purpose
Required	**Required**	Important to read (material will be assessed)
Recommended	**Optional**	Useful to better understand required reading
Suggested	**Optional**	Provides helpful context; not critical reading
Advanced	**Optional**	Additional concepts and primary literature

Wherever possible, readings are free and open access (always the case for required and recommended readings), or are from inexpensive books. References cited in this book are also useful sources of information, but a lot of these references are from expensive statistics textbooks. Some of these textbooks are available in the library, and all of them are owned by the module coordinator (happily shared upon request).

Canvas

This module will be taught using Canvas. You should be enrolled in the University of Stirling Canvas module "SCIU4T4 - Statistical Techniques (2022/3)". If for some reason you cannot access the module on Canvas, then please email the module coordinator as soon as possible (alexander.duthie@stir.ac.uk).

In Canvas, you will find links to all learning content, including chapters to this book, video lectures, practice quizzes, tests, and exams. For your benefit, there is some redundancy between the information on Canvas and the information in this book. All of the information in this preface is also available on Canvas. Similarly, weekly links to readings, lectures, practicals, and assessments are posted at the start of the sections of this book and in the Learning and Teaching content on Canvas. You will also receive weekly announcements at 08:00 on Mondays summarising what needs to be done each week (again, with relevant links). Our objective here is to provide a clear structure for the module and regular reminders so that you always know what is happening and when.

Some aspects of the module must be completed on Canvas. All of the links below are also on the main Canvas page.

- **Discussions** are a place where you can ask questions about the module. Each week, a new topic will be introduced so that you can ask questions pertaining to that week (we are not strict about which week you post in; this is more just to help keep everything organised). You can also ask questions anonymously on the **SCIU4T4 Padlet**. It is completely fine to ask for practice problems in the discussion boards, but please try to be as specific as possible about what you want to practice (this helps us come up with good questions).
- **Quizzes** are where all practice quizzes and assessments will be located. You must take these on Canvas, and you will need a laptop or desktop computer to complete them.
- **People** is where you can sign up for a practical group (A or B).

In general, it is probably also easiest to watch lecture videos from within Canvas, although this should also be possible outside of Canvas using links in this book.

Assessment overview

This module includes 2 formative (i.e., ungraded) assessments and 3 summative (i.e., graded) assessments.

Table 3.: Summary of module assessments, whether they are for practice (formative) or for a grade (summative), how much they count for the final grade, the weeks of material that the test includes, and the time of the test.

Assessment	Type	Weighting	Weeks	Time
Test 1F	Formative	0%	1-4	22 FEB at 10:00-12:00
Test 1S	Summative	25%	1-6	15 MAR at 10:00-12:00
Test 2S	Summative	25%	7-10	05 APR at 10:00-12:00
Mock Exam	Formative	0%	All	To be determined
Exam	Summative	50%	All	To be determined

All tests and exams are taken online on Canvas. You can access the assessments in the table above using the 'Quizzes' link on canvas, or access an assessment directly with the above links.

Tests and exams are completed online. Tests and exams are open book and open note. You are free to use any learning materials from the module, but you **must complete your test independently**. In other words, you cannot confer with anyone else about the test or exam material during the test or exam (except for module instructors if you

have question). During all tests and exams, the module coordinator will be available by email and MS Teams chat.

It is important to have access to a computer for all tests and exams. To complete tests and exams, you will need to run statistical analyses on a laptop or desktop computer (a mobile phone will not work, or will at least be extremely inconvenient to the extent that you might not be able to complete the assessment on time). If you do not have access to a laptop, you can borrow one from the library. If you believe that you might not have access to a computer during a test or exam, for whatever reason, please let the module coordinator know as soon as possible. We will work something out!

The goal of all assessments is to evaluate your learning. Assessment questions will never be written to deliberately trick you, nor will they be intentionally written to be confusing. We want to evaluate your understanding of the learning content, not your ability to decipher test questions. Most questions will therefore be asked in the same way as questions in lab practicals and practice quizzes, or explained in this book. If you practice the skills introduced in these practicals, do the required readings, and complete weekly practice quizzes, then you should be well-prepared for tests and exams.

University policy requires all module assessments to align with the University of Stirling Common Marking Scheme. This means that the numeric mark (0-100) awarded needs to match up with the descriptor attainment of learning outcomes. If you attain most of the intended learning outcomes (ILOs) of the module, then you should receive passing grade (40+):

> "Acceptable attainment of most intended learning outcomes, displaying a qualified familiarity with a minimally sufficient range of relevant materials, and a grasp of the analytical issues and concepts which are generally reasonable, albeit insecure" (University of Stirling Common Marking Scheme).

However, extremely high grades (80+) are intended to indicate truly exceptional work that goes above and beyond what is required to pass the module:

> "[D]emonstrates outstanding quality evidenced by an ability to engage critically and analytically with source material, exhibits independent lines of argument, is highly original and uses an extremely wide range of relevant sources where appropriate" (University of Stirling Common Marking Scheme).

The full University of Stirling Common Marking Scheme is replicated in Appendix A of this book. To align SCIU4T4 assessments to this scheme, all tests and exams will be set with 3 different types of questions:

1. **Fundamental questions: (at least 50%).** These questions will assess the most fundamental skills of the module. Questions will focus on specific protocols that have been introduced in lectures and this book, and they will test your ability to

complete these protocols and correctly report or interpret what you have done. For example, you might be asked to answer a specific question about a dataset, correctly calculate a statistic from a dataset in Jamovi (e.g., the mean, or confidence intervals), or interpret statistical output (e.g., output from a particular statistical test). These are the kinds of questions that you can prepare for with practice and repetition. And we are happy to provide practice questions in the Canvas Discussion upon request.

2. **Conceptual questions: (at least 30%).** These questions will assess more advanced learning content in the lectures and this book, and they might therefore require a greater depth of understanding of the relevant statistical concepts that you have learned. Instead of simply repeating a statistical calculation or procedure, these questions might require you to demonstrate an appreciable understanding of the underlying statistical concepts. For example, instead of simply asking you to report the answer for a specific calculation, a question might leave it up to you to use your understanding of statistics to decide what calculation needs to be performed. Or, rather than asking you to run a particular statistical test and provide a result, a conceptual question might require you to choose and then correctly apply the appropriate test based on properties of a dataset.

3. **Advanced questions: (no more than 20%).** These questions are intended to provide you with the opportunity to demonstrate an exceptional degree of understanding, problem-solving, and analytical skills. Advanced questions might require you to apply knowledge gained from independent learning outside the lectures and this book. The topics of these questions will be relevant for the range of weeks that the assessment includes (e.g., for Test 1S, you will only be asked material that is relevant to topics from weeks 1-6), but they will not be questions that have been previously introduced or explored. Again, the point of these questions is to allow you to demonstrate exceptional understanding of the material and independence in your statistical learning (not to shake your confidence).

Question types will be clearly separated into different sections on all assessments, and will be presented in order 1-3. We strongly recommend that you attempt these questions in this order, and that you do not attempt advanced (3) questions until you are first feeling confident about your answers to fundamental (1) and conceptual questions (2). Separating questions into these categories is intended to ensure alignment with the university Common Marking Scheme, and to give you a clearer idea of what to expect.

Tests

Tests are written so that they can be completed in 1 hour, but you will have 2 hours to complete them. At the time of writing, this is the university recommended approach for accommodating any technical difficulties that might arise during online assessments,

and for accommodating students with an Agreed Record of University Access Adjustments (ARUUA). If you have any concerns, then please get in touch with the module coordinator. We will listen, and do our best to help!

All tests will consist of 25 questions in total. Questions will be a combination of multiple choice, multiple answer, and fill in the blank. Practice quizzes, and especially the first formative test (Test 1F), will match the format of summative tests (Test 1S and Test 2S) as closely as possible.

About 1 hour before each test, you will receive an email at your University of Stirling email address that includes any datasets needed to complete the test. The email will contain 'SCIU4T4' in the subject line. If you have not received this email 30 minutes before the test starts, then please let the module coordinator know as soon as possible. It is entirely fine to have a look at the datasets before the test starts. Be sure to get comfortable in your test environment!

Exams

Exams are basically long tests. Exams are written so that they can be completed in 2 hours, but you will have 4 hours to complete them (see the Tests section above for an explanation of the timing). Exams will consist of 50 questions in total. Questions will be multiple choice, multiple answer, and fill in the blank. As with tests, you will receive an email at your University of Stirling email address about 1 hour before an exam that includes any datasets needed to complete the exam.

The formative mock exam will match the format and content of the summative exam as closely as possible.

Feedback

After all tests and the mock exam, you will receive a detailed feedback report. This report will walk you through the assessment questions, explain how to answer them correctly, and, where possible, explain why you might have answered questions incorrectly. This report will be made available to you on Canvas within 21 working days of the assessment (in practice, we will try to get assessments graded as quickly as possible).

Extenuating circumstances

If you encounter personal difficulties that make it impossible to sit a test or exam, then you will need to file for Extenuating Circumstances. For more information, or to submit a request, you can go to the Extenuating Circumstances webpage. If you need help, or want some guidance, then please feel free to get in touch with the module coordinator. Discussions surrounding extenuating circumstances will be kept strictly confidential.

Note that because all assessments are tests or exams, it is not possible to apply for an extension request as you would for an assignment in other modules (i.e., tests and exams cannot be taken late except under extenuating circumstances).

Practicals

Each week, there will be a practical session to learn and practice using statistical tools. Practical instructions are located in the last chapter of each book section (e.g., Chapter 3 for week 1, or Chapter 8 for week 2). Practical sessions are held in room Cottrell 2A17 on either Wednesdays from 13:05-15:55 (Group A) or Thursdays from 09:05-11:55 (Group B). When you get to the practical, there will be a quick introduction, then you will work through the practical in a mostly self-guided way (feel free to converse with your classmates and help one another too). Instructors will be walking around the room to answer any questions that you have and help you if you get stuck. These instructors will include the lecturer leading the week's content, the module coordinator, and module demonstrators. Module demonstrators include postgraduate researchers who have experience in data analysis from their own ongoing PhD work.

Some practicals will take the full 2 hours and 50 minutes, while others might be completed in less time. Before you complete the practical, it is a good idea to check in with one of the instructors to make sure that you have not missed anything. While practicals are not assessed, the skills that you learn during module practicals are absolutely critical for being able to correctly answer questions on the tests and exams. Practicals exist to provide a low-pressure environment for learning these skills, so we hope that you can take your time on these, relax, and focus on completing them with the support of your instructors and classmates. If you finish early, then you are welcome to continue exploring the practical topic and ask questions, but you are not required to stay the full 2 hours and 50 minutes. If for some reason you find that the practical environment is not conducive to learning for you, then please let the module coordinator know.

Optional help hours

Face-to-face help is available to you every Friday in room Cottrell 1A113 from 15:05-17:55. You are very welcome to come in during this time to ask questions about the learning content of the week, and to get some extra help with the practical if you were unable to finish it. Learning statistics requires a lot practice and repetition, so it is entirely okay to ask the same question multiple times if you think that it will help you better understand something. This is not at all an inconvenience to the teaching team; we know that this material is challenging, and we want to help you understand it!

Note that the scheduled help hours are probably the easiest way to get in touch with the teaching team for some one-on-one help. Nevertheless, you are also free to contact the

module coordinator by email, MS Teams, or some other means if you have a question. Ad hoc meetings can also be set up during the week if necessary.

Jamovi statistical software

There are a lot of different software packages for running statistical analyses. Previous versions of this module have been taught using Minitab or SPSS. This year, and for the foreseeable future, we are using Jamovi (The Jamovi Project, 2022). There are a lot of tangible benefits to using Jamovi:

- It is user-friendly, with a point and click interface that is more intuitive than other software.
- It is free and open source, which means that you will be able to use it even after you have finished this module and your degree.
- It is written by a community of statisticians and scientists with a focus on teaching.
- It works on Windows, Mac, Linux, or Chrome, and can even be operated in a browser.
- It is lightweight, and less prone to crashing than other statistical software.
- It is built on top of the R statistical programming language, which makes it easier to transition to learning R.

In short, Jamovi has a lot of advantages that make it the best option for you to learn statistics, apply your learning in future projects, and build on the skills in this module in future course work.

In the computer labs on campus, you can access Jamovi using AppsAnywhere. For your personal devices, you can download Jamovi here (recommended). You can also run Jamovi from a browser, but there are currently some time limitations for doing this (sessions are limited to 45 minutes).

This module will also introduce R, although we will not really use R by itself until Week 11. Prior to Week 11, there will be a few R commands scattered throughout lab practicals to ease you into it and accomplish some very specific tasks (e.g., generating random numbers). Because Jamovi is built on top of R, R can be run directly from within Jamovi to do these tasks. Anyone who is especially interested in learning about R and what it can do is welcome to attend fortnightly sessions of Stirling Coding Club, which is run by the module coordinator. Sessions are run online using MS Teams and do not require any active participation (i.e., you can just watch; contact the module coordinator if you want to be included in the group).

The R programming language has become by far the most popular software for doing statistics in biology and environmental sciences (we will see why in Week 11). It is highly flexible and can be used for all kinds of projects. This book was entirely written in R using software called "Rmarkdown" and "Pandoc".

Timetable

Table 4.: Dates, times, and locations of all synchronous activities in Statistical Techniques for Spring 2023.

Week	Date	Day	Time	Room	Lead	Session
1	25 JAN	WED	13:05-15:55	C2A17	BD	Preparing data (A)
1	26 JAN	THU	09:05-11:55	C2A17	BD	Preparing data (B)
1	27 JAN	FRI	15:05-17:55	C1A13	BD	Help (optional)
2	01 FEB	WED	13:05-15:55	C2A17	IJ	Stats concepts (A)
2	02 FEB	THU	09:05-11:55	C2A17	IJ	Stats concepts (B)
2	03 FEB	FRI	15:05-17:55	C1A13	IJ	Help (optional)
3	08 FEB	WED	13:05-15:55	C2A17	IJ	Summary stats (A)
3	09 FEB	THU	09:05-11:55	C2A17	IJ	Summary stats (B)
3	10 FEB	FRI	15:05-17:55	C1A13	IJ	Help (optional)
4	15 FEB	WED	13:05-15:55	C2A17	IJ	Prob models (A)
4	16 FEB	THU	09:05-11:55	C2A17	IJ	Prob models (B)
4	17 FEB	FRI	15:05-17:55	C1A13	IJ	Help (optional)
5	22 FEB	WED	10:05-11:55	Online	BD	Test 1F
5	22 FEB	WED	13:05-15:55	C2A17	IJ	Stats inference (A)
5	23 FEB	THU	09:05-11:55	C2A17	IJ	Stats inference (B)
5	24 FEB	FRI	15:05-17:55	C1A13	IJ	Help (optional)
6	01 MAR	WED	13:05-15:55	C2A17	MQ	Hypo testing (A)
6	02 MAR	THU	09:05-11:55	C2A17	MQ	Hypo testing (B)
6	03 MAR	FRI	15:05-17:55	C1A13	MQ	Help (optional)
8	15 MAR	WED	10:05-11:55	Online	BD	Test 1S
8	15 MAR	WED	13:05-15:55	C2A17	MQ	ANOVA (A)
8	16 MAR	THU	09:05-11:55	C2A17	MQ	ANOVA (B)
8	17 MAR	FRI	15:05-17:55	C1A13	MQ	Help (optional)
9	22 MAR	WED	13:05-15:55	C2A17	MQ	Counts (A)
9	23 MAR	THU	09:05-11:55	C2A17	MQ	Counts (B)
9	24 MAR	FRI	15:05-17:55	C1A13	MQ	Help (optional)
10	29 MAR	WED	13:05-15:55	C2A17	BD	Regression (A)
10	30 MAR	THU	09:05-11:55	C2A17	BD	Regression (B)
10	31 MAR	FRI	15:05-17:55	C1A13	BD	Help (optional)
11	05 APR	WED	10:05-11:55	Online	BD	Test 2S
11	05 APR	WED	13:05-15:55	C2A17	BD	Randomisation (A)
11	06 APR	THU	09:05-11:55	C2A17	BD	Randomisation (B)
11	07 APR	FRI	15:05-17:55	Tutorial	BD	Help (optional)
12	12 APR	WED	13:05-15:55	C2A17	BD	Stats reporting (A)
12	13 APR	THU	09:05-11:55	C2A17	BD	Stats reporting (B)
12	14 APR	FRI	15:05-17:55	C1A13	BD	Help (optional)

Week	Date	Day	Time	Room	Lead	Session
13	18 APR	TUE	14:05-16:55	C1A13	BD	Help (optional)

License

This book is licensed under CC BY-NC-ND 4.0. You can copy and redistribute this material however you want as long as you give credit to the authors, a copyright notice, a link to the original book, and an indication if the material has been modified. You may not use this material for commercial purposes.

Part I.

Background mathematics and data organisation

Week 1 Overview

Dates	23 January 2023 - 27 January 2023
Reading	**Required:** SCIU4T4 Workbook chapters 1-2
	Recommended: None
	Suggested: Navarro and Foxcroft (2022) Section 2.1
	Advanced: Wickham (2014) (Download)
Lectures	1.0: Numbers and operations (17:48 min; Video)
	1.1: Logarithms (6:06 min; Video)
	1.2: Order of Operations (7:00 min; Video)
	1.3: Tidy data (13:39 min; Video)
	1.4: Data files (12:15 min; Video)
Practical	Preparing data (Chapter 3)
	Room: Cottrell 2A17
	Group A: 25 JAN 2023 (WED) 13:05-15:55
	Group B: 26 JAN 2023 (THU) 09:05-11:55
Help hours	Brad Duthie
	Room: Cottrell 1A13
	27 JAN 2023 (FRI) 15:05-17:55
Assessments	Week 1 Practice quiz on Canvas

Week 1 focuses on background mathematics and data organisation.

Chapter 1 will review some background mathematics that is relevant to the statistical techniques that you will learn in this module. This information might not be new to you, but it is important to review some fundamental mathematical concepts that will be used throughout the module. Specific topics include numbers and operations, logarithms, and the order of operations.

Chapter 2 will focus on data organisation. Before actually doing any statistics, it is important to be able to organise data in a way that can be understood by other researchers and interpreted by statistical software. This chapter will focus on what to do first after data have been collected in the field or laboratory.

Chapter 3 guides you through the week 1 practical, which focuses on organising datasets and preparing them for statistical analysis. The aim of this practical is for you to learn how to take data recorded in the field, laboratory, or some other source and put it into a format that can be used in statistical programs such as Jamovi or R.

1. Background mathematics

There are at least two types of mathematical challenges that come with first learning statistics. The first challenge is simply knowing the background mathematics upon which many statistical tools rely. Fortunately, while the *theory* underlying statistical techniques does rely on some quite advanced mathematics (e.g., see Mclean et al., 1991; Rencher, 2000; Miller and Miller, 2004), the *application* of standard statistical tools to data usually does not. This module focuses on the application of statistical techniques, so all that is required is a background in some fundamental mathematical concepts such as mathematical operations (addition, subtraction, multiplication, division, and exponents), simple algebra, and probability. This chapter will review these operations and the mathematical symbols used to communicate them.

The second mathematical challenge that students face when learning statistics for the first time is a bit more subtle. Students with no statistical background sometimes have an expectation that statistics will be similar to previously learned mathematical topics such as algebra, geometry, or trigonometry. In some ways, this is true, but in a lot of ways statistics is a much different way of thinking than any of these topics. A lot of mathematical subjects focus on questions that have very clear right or wrong answers (or, at least, this is how they are often taught). If, for example, we are given the lengths of two sides of a right triangle, then we might be asked to calculate the hypotenuse of the triangle using Pythagorean theorem ($a^2 + b^2 = c^2$, where c is the hypotenuse). If we know the length of the two sides, then the length of the hypotenuse has a clear correct answer (at least, on a Euclidean plane). In statistics, answers are not always so clear cut. Statistics, by its very nature, deals with uncertainty. While all of the standard rules of mathematics still apply, statistical questions such as, "Can I use this statistical test on my data?", "Do I have a large enough sample size?", or even "Is my hypothesis well-supported?" often do not have unequivocal 'correct' answers. Being a good statistician often means making well-informed, but ultimately at least somewhat subjective, judgements about how to make inferences from data.

For the purpose of assessments in this module (tests and exams), please note that we will only ask questions that **do** have clear and correct answers. This is to keep the module assessment fair and transparent. For example, we will not ask you questions like, "Can I use this statistical test on my data" unless the answer is a very clear yes or no. And we will not ask you questions like, "Is my hypothesis well-supported", but specify what we mean instead by asking questions such as, "should you reject the null hypothesis at the $\alpha = 0.05$ level of Type I error" (we will worry about what this means later). We

will give practice questions, a practice test, and a practice exam, so that the nature of assessment questions is clear before you are actually assessed for a grade.

For now, we will move on to looking at numbers and operations, logarithms, and order of operations. These topics will be relevant throughout the semester, so it is important to understand them and be able to apply them when doing calculations.

1.1. Numbers and operations

Calculating statistics and reading statistical output requires some knowledge of numbers and basic mathematical operations. This section is a summary of the basic mathematical tools that will be used in introductory statistics. Much of this section is inspired by Courant et al. (1996) and chapter 2 of Pastor (2008). This section will be abridged to focus on only the numbers and mathematical operations relevant to this book. The objective here is to present some very well-known ideas in an interesting way, and to intermix them with bits of information that might be new and interesting. For doing statistics, what you really *need* to know here are the operations and the notation; that is, how operations such as addition, multiplication, and exponents are calculated and represented mathematically.

We can start with the *natural* numbers, which are the kinds of numbers that can be counted using fingers, toothpicks, pebbles, or any discrete sets of objects.

$$1, 2, 3, 4, 5, 6, 7, 8, ...$$

There are an infinite number of natural numbers (we can represent the set of all of them using the symbol $\mathbb{N}$). For any given natural number, we can always find a higher natural number using the operation of addition. For example, a number higher than 5 can be obtained by simply adding 1 to it,

$$5 + 1 = 6.$$

This is probably not that much of a revelation, but it highlights why the natural numbers are countably infinite (for any number you can think of, N, there is always a higher number $N + 1$). It also leads to a reminder about two other important mathematical symbols for this module (in addition to $+$, which indicates addition), greater than ($>$) and less than ($<$). We know that the number 6 is greater than 5, and express this mathematically as the **inequality**, $6 > 5$. Note that the large end of the inequality faces the higher number, while the pointy end (i.e., the smaller end) faces the lower number. Inequalities are used regularly in statistics, e.g., to indicate when a probability of something is less than a given value (e.g., $P < 0.05$, which can be read 'P is less than 0.05'). We might also use the symbols $\geq$ or $\leq$ to indicate when something is greater

than or equal to ($\geq$) or less than or equal to ($\leq$) a particular value. For example, $x \geq 10$ indicates that some number x has a value of 10 or higher.

Whenever we add one natural number to another natural number, the result is another natural number, a sum (e.g., $5 + 1 = 6$). If we want to go back from the sum to one of the values being summed (i.e., get from 6 to 5), then we need to subtract,

$$6 - 1 = 5.$$

This operation is elementary mathematics, but a subtle point that is often missed is that the introduction of subtraction creates the need for a broader set of numbers than the natural numbers. We call this broader set of numbers the *integers* (we can represent these using the symbol $\mathbb{Z}$). If, for example, we want to subtract 5, from 1, we get a number that cannot be represented on our fingers,

$$1 - 5 = -4.$$

The value -4 is an integer (but *not* a natural number). Integers include 0 and all negative whole numbers,

$$..., -4, -3, -2, -1, 0, 1, 2, 3, 4, ...$$

Whenever we add or subtract integers, the result is always another integer.

Now, suppose we wanted to add the same value up multiple times. For example,

$$2 + 2 + 2 + 2 + 2 + 2 = 12.$$

The number 2 is being added 6 times in the equation above to get a value of 12. But we can represent this sum more easily using the operation of multiplication,

$$2 \times 6 = 12.$$

The 6 in the equation just represents the number of times that 2 is being added up. The equation can also be written as $2(6) = 12$, or sometimes, `2*6 = 12` (i.e., the asterisk is sometimes used to indicate multiplication). Parentheses indicate multiplication when no other symbol separates them from a number. This rule also applies to numbers that come immediately before variables. For example, $2x$ can be interpreted as *two times x*. When multiplying integers, we always get another integer. Multiplying two positive numbers always equals another positive number (e.g., $2 \times 6 = 12$). Multiplying a positive and a negative number equals a negative number (e.g., $-2 \times 6 = -12$). And multiplying two negative numbers equals a positive number (e.g., $-2 \times -6 = 12$). There

are multiple ways of thinking about why this last one is true (see, e.g., Askey, 1999, for one explanation), but for now we can take it as a given.

As with addition and subtraction, we need an operation that can go back from multiplied values (the product) to the numbers being multiplied. In other words, if we multiply to get $2 \times 6 = 12$ (where 12 is the product), then we need something that goes back from 12 to 2. Division allows us to do this, such that $12 \div 6 = 2$. In statistics, the symbol $\div$ is rarely used, and we would more often express the calculation as either $12/6 = 2$ or,

$$\frac{12}{6} = 2.$$

As with subtraction, there is a subtle point that the introduction of division requires a new set of numbers. If instead of dividing 6 into 12, we divided 12 into 6,

$$\frac{6}{12} = \frac{1}{2} = 0.5.$$

We now have a number that is not an integer. We therefore need a new broader set of numbers, the *rational* numbers (we can represent these using the symbol $\mathbb{Q}$). The rationals include all numbers that can be expressed as a *ratio* of integers. That is, p/q, where both p and q are in the set $\mathbb{Z}$.

We have one more set of operations relevant for introductory statistics. Recall that we introduced 2×6 as a way to represent $2 + 2 + 2 + 2 + 2 + 2$. We can apply the same logic to multiplying a number multiple times. For example, we might want to multiply the number 2 by itself 4 times,

$$2 \times 2 \times 2 \times 2 = 16.$$

We can represent this more compactly using an **exponent**, which is written as a superscript,

$$2^4 = 16.$$

The 4 in the equation above indicates that the 2 should be multiplied 4 times to get 16. Sometimes this is also represented by a carrot in writing or code, such that 2^4 = 16. Very occasionally, some authors will use two asterisks in a row, 2**4 = 16, probably because this is how exponents are represented in some statistical software and programming languages. One quick note that can be confusing at first is that a negative in the exponent indicates a reciprocal. For example,

$$2^{-4} = \frac{1}{16}.$$

This can sometimes be useful for representing the reciprocal of a number or unit in a more compact way than using a fraction (we will come back to this in Chapter 6).

As with addition and subtraction, and multiplication and division, we also need an operation to get back from the exponentiated value to the original number. That is, for $2^4 = 16$, there should be an operation that gets us back from 16 to 2. We can do this using the **root** of an equation,

$$\sqrt[4]{16} = 2.$$

The number under the radical symbol $\sqrt{}$ (in this case 16) is the one that we are taking the root of, and the index (in this case 4) is the root that we are calculating. When the index is absent, we assume that it is 2 (i.e., a square root),

$$\sqrt[2]{16} = \sqrt{16} = 4.$$

Note that $4^2 = 16$ (i.e., 4 squared equals 16).

Instead of using the radical symbol, we could also use a fraction in the exponent. That is, instead of writing $\sqrt[4]{16} = 2$, we could write $16^{1/4} = 2$ or $16^{1/2} = 4$. In statistics, however, the $\sqrt{}$ is more often used. Either way, this yet again creates the need for an even broader set of numbers. This is because expressions such as $\sqrt{2}$ do not equal any rational number. In other words, there are no integers p and q such that their *ratio*, $p/q = \sqrt{2}$ (the proof for why is very elegant!). Consequently, we can say that $\sqrt{2}$ is *irrational* (not in the colloquial sense of being illogical or unreasonable, but in the technical sense that it cannot be represented as a ratio of two integers). Irrational numbers cannot be represented as a ratio of integers, or with a finite or repeating decimal. Remarkably, the set of irrational numbers is larger than the set of rational numbers (i.e., rational numbers are countably infinite, while irrational numbers are uncountably infinite, and there are more irrationals; you do not need to know this or even believe it, but it is true!).

Perhaps the most famous irrational number is π, which appears throughout science and mathematics and is most commonly introduced as the ratio of a circle's circumference to its diameter. Its value is $\pi \approx 3.14159$, where the symbol $\approx$ means 'approximately'. Actually, the decimal expansion of π is infinite and non-repeating; the decimals go on forever and never repeat themselves in a predictable pattern. As of 2019, over 31 trillion (i.e., 31000000000000) decimals of π have been calculated (Yee, 2019).

The rational and irrational numbers together comprise a set of numbers called *real* numbers (we can represent these with the symbol $\mathbb{R}$), and this is where we will stop. This story of numbers and operations continues with imaginary and complex numbers (Courant et al., 1996; Pastor, 2008), but these are not necessary for introductory statistics.

1.2. Logarithms

There is one more important mathematical operation to mention that is relevant to introductory statistics. Logarithms are important functions, which will appear in multiple places (e.g., statistical transformations of variables). A logarithm tells us the exponent to which a number needs to be raised to get another number. For example,

$$10^3 = 1000.$$

Verbally, 10 raised to the power of 3 equals 1000. In other words, we need to raise 10 to the power of 3 to get a value of 1000. We can express this using a logarithm,

$$\log_{10}(1000) = 3.$$

Again, the same relationship is expressed in $10^3 = 1000$ and $\log_{10}(1000) = 3$. For the latter, we might say that the base 10 logarithm of 1000 is 3. This is actually extremely useful in mathematics and statistics. Mathematically, logarithms have the very useful property,

$$log_{10}(ab) = log_{10}(a) + log_{10}(b).$$

Historically, this has been used to make calculations easier by converting multiplication to addition (Stewart, 2008). In statistics, and across the biological and environmental sciences, we often use logarithms when we want to represent something that changes exponentially on a more convenient scale. For example, suppose that we wanted to illustrate the change in global CO_2 emissions over time (Friedlingstein et al., 2022). We could show year on the x-axis and emissions in billions of tonnes of CO_2 on the y-axis (Figure 1.1).

We can see from Figure 1.1 that global CO_2 emissions go up exponentially over time, but this exponential relationship means that the y-axis has to cover a large range of values. This makes it difficult to see what is actually happening in the first 100 years. Are CO_2 emissions increasing from 1750-1850, or do they stay about the same? If instead of plotting billions of tonnes of CO_2 on the y-axis, we plotted the logarithm of these values, then the pattern in the first 100 years becomes a bit more clear (Figure 1.2).

It appears from the logged data in Figure 1.2 that global CO_2 emissions were indeed increasing from 1750-1850. Note that Figure 1.2 presents the *natural logarithm* of CO_2 emissions on the y-axis. The natural logarithm uses Euler's number, $e \approx 2.718282$, as a base. Euler's number e is an irrational number (like π), which corresponds to the intrinsic rate of increase of a population's size in ecology (Gotelli, 2001), or, in banking, interest compounded continually (like π, e actually shows up in a lot of different places throughout science and mathematics). We probably could have just as easily used 10 as

Figure 1.1.: Global carbon dioxide emissions from 1750-2021.

Figure 1.2.: Natural logarithm of global carbon dioxide emissions from 1750-2021.

a base, but e is usually the default base to use in science (bases 10 or 2 are also often used). Note that we can convert back to the non-logged scale by raising numbers to the power of e. For example, $e^{-4} \approx 0.018$, $e^{-2} \approx 0.135$, $e^0 = 1$, and $e^2 = 7.390$.

1.3. Order of operations

Every once in a while, a maths problem like the one below seems to go viral online,

$$x = 8 \div 2\,(2+2)\,.$$

Depending on the order in which calculations are made, some people will conclude that $x = 16$, while others conclude that $x = 1$ (Chernoff and Zazkis, 2022). The confusion is not caused by the above calculation being difficult, but by peoples' differences in interpreting the rules for what order calculations should be carried out. If we first divide $8/2$ to get 4, then multiply by $(2 + 2)$, we get 16. If we first multiply 2 by $(2 + 2)$ to get 8, then divide, we get 1. The truth is that even if there is a 'right' answer here (Chernoff and Zazkis, 2022), the equation could be written more clearly. We might, for example, rewrite the above to more clearly express the intended order of operations,

$$x = \frac{8}{2}\,(2+2) = 16.$$

We could write it a different way to express a different intended order of operations,

$$x = \frac{8}{2(2+2)} = 1.$$

The key point is that the order in which operations are calculated matters, so it is important to write equations clearly, and to know the order of operations to calculate an answer correctly. By convention, there are some rules for the order in which calculations should proceed.

1. Anything within parentheses should always be calculated first.
2. Exponents and radicals should be applied second
3. Multiplication and division should be applied third
4. Addition and subtraction should be done last

These conventions are not really rooted in anything fundamental about numbers or operations (i.e., we made these rules up), but there is a logic to them. First, parentheses are a useful tool for being unequivocal about the order of operations. We could, for example, always be completely clear about the order to calculate by writing something like $(8/2) \times (2+2)$ or $8/(2(2+2))$, although this can get a bit messy. Second, rules 2-4 are ordered by the magnitude of operation effects; for example, exponents have a bigger

effect than multiplication, which has a bigger effect than addition. In general, however, these are just standard conventions that need to be known for reading and writing mathematical expressions. In this module, you will not see something ambiguous like $x = 8 \div 2\,(2+2)$, but you should be able to correctly calculate something like this,

$$x = 3^2 + 2\,(1+3)^2 - 6 \times 0.$$

First, remember that parentheses come first, so we can rewrite the above,

$$x = 3^2 + 2\,(4)^2 - 6 \times 0.$$

Exponents come next, so we can calculate those,

$$x = 9 + 2\,(16) - 6 \times 0.$$

Next comes multiplication and division,

$$x = 9 + 32 - 0.$$

Lastly, we calculate addition and subtraction,

$$x = 41.$$

In this module, you will very rarely need to calculate something with this many different steps. But you will often need to calculate equations like the one below,

$$x = 20 + 1.96 \times 2.1.$$

It is important to remember to multiply 1.96×2.1 *before* adding 20. Getting the order of operations wrong will usually result in the calculation being completely off.

One last note is that when operations are above or below a fraction, or below a radical, then parentheses are implied. For example, we might have something like the fraction below,

$$x = \frac{2^2 + 1}{3^2 + 2}.$$

Although rules 2-4 still apply, it is implied that there are parentheses around both the top (numerator) and bottom (denominator), so you can always read the above equation like this,

$$x = \frac{(2^2+1)}{(3^2+2)} = \frac{(4+1)}{(9+2)} = \frac{5}{11}.$$

Similarly, anything under the $\sqrt{\ }$ can be interpreted as being within parentheses. For example,

$$x = \sqrt{3+4^2} = \sqrt{(3+4^2)} \approx 4.47.$$

This can take some getting used to, but with practice, it will become second nature to read equations with the correct order of operations.

2. Data organisation

In the field or the lab, data collection can be messy. Often data need to be recorded with a pencil and paper, and in a format that is easiest for writing in adverse weather or a tightly controlled laboratory. Sometimes data from a particular sample, such as a bird nest (Figure 2.1), cannot all be collected in one place.

Figure 2.1.: Dr Becky Boulton collects data from nest boxes in the field (A), then processes nest material in the lab (B).

Data are sometimes missing due to circumstances outwith the researcher's control, and data are usually not collected in a format that is immediately ready for statistical analysis (e.g., Figure 2.2). Consequently, we often need to reorganise data from a lab or field book to a spreadsheet on the computer.

Fortunately, there are some generally agreed upon guidelines for formatting data for statistical analysis. This chapter introduces the tidy format (Wickham, 2014), which can be used for structuring data files for statistical software. This chapter will provide an example of how to put data into a tidy format, and how to save a dataset into a

DATE (m	SPECIES	SITE NO.	TREE NO	FRUIT NO	FRT LENG	FRT WID	FRT
5/9/10	F-pet	70	70	1	15	18	14
5/10/10	F-pet	70	70	2	17	19	15
5/10/10	F-pet	70	70	3	21	21	16
5/11/10	F-pet	70	70	4			
5/11/10	F-pet	70	70	5	15	16	14
5/11/10	F-pet	70	70	6	16	16	15

Figure 2.2.: A portion of a lab notebook used to record measurements of fig fruits from different trees in Baja, Mexico, in 2010.

file that can be read and used in statistical software such as Jamovi or R (The Jamovi Project, 2022; R Core Team, 2022).

2.1. Tidy data

After data are collected, they need to be stored digitally (i.e., in a computer file, such as a spreadsheet). This should happen as soon as possible so that back up copies of the data can be made. Nevertheless, retaining field and lab notes as a record of the originally collected data is also a good idea. Sometimes it is necessary to return to these notes, even years after data collection. Often we will want to double-check to make sure that we copied a value or observation correctly from handwritten notes to a spreadsheet. Note that sometimes data can be input directly into a spreadsheet or mobile application, bypassing handwritten notes altogether, but it is usually helpful to have a physical copy of collected data.

Most biological and environmental scientists store data digitally in the form of a spreadsheet. Spreadsheets enable data input, manipulation, and calculation in a highly flexible way. Most spreadsheet programs even have some capacity for data visualisation and statistical analysis. For the purposes of statistical analysis, spreadsheets are probably most often used for inputting data in a way that can be used by more powerful statistical software. Commonly used spreadsheet programs are MS Excel, Google Sheets, LibreOffice Calc. The interface and functions of these programs are very similar, nearly identical for most purposes. They can all open and save the same file types (e.g., XLSX, ODS, CSV), and they all have the same overall look, feel, and functionality for data input, so the program used is mostly a matter of personal preference. In this text, we

will use LibreOffice because it is completely free and open source, and easily available to download at http://libreoffice.org. Excel and Google Sheets are also completely fine to use.

Site	Tree	Fruit	Tree_Lat	Tree_Lon	Foundress_Pollinators	Fruit_Volume_mm	Poll	SO1	SO2
S70	T70	F1	23.73629	-109.83987	4	1978.2	138	0	0
S70	T70	F2	23.73629	-109.83987	3	2535.55	97	3	0
S70	T70	F3	23.73629	-109.83987	1	3692.64	58	3	0
S70	T70	F4	23.73629	-109.83987	1	NA	39	0	0
S70	T70	F5	23.73629	-109.83987	1	1758.4	129	0	0
S70	T70	F6	23.73629	-109.83987	1	2009.6	77	0	0
S70	T70	F7	23.73629	-109.83987	1	1648.5	74	0	0

Figure 2.3.: A LibreOffice spreadsheet showing data from fig fruits collected in 2010. Each row is a unique sample (fruit), and columns record properties of the fruit.

Spreadsheets are separated into individual rectangular cells, which are identified by a specific column and row (Figure 2.3). Columns are indicated by letters, and rows are indicated by numbers. We can refer to a specific cell by its letter and number combination. For example, the active cell in Figure 2.3 is F3, which has a value of '3' indicating the value recorded in that specific measurement (in this case, foundress pollinators in the fig fruit). We will look more at how to interact with the spreadsheet in the Chapter 3 lab practical, but for now we will focus on how the data are organised.

There are a lot of potential ways that data could be organised in a spreadsheet. For good statistical analysis, there are a few principles that are helpful to follow. Whenever we collect data, we record observations about different units. For example, we might make one or more measurements on a tree, a patch of land, or a sample of soil. In this case, trees, land patches, or soil samples are our units of **observation**. Each attribute of a unit that we are measuring is a **variable**. These variables might include tree heights and leaf lengths, forest cover in a patch of land, or carbon and nitrogen content of a soil sample. Tidy datasets that can be used in statistical analysis programs are defined by three characteristics (Wickham, 2014).

1. Each variable gets its own column.
2. Each observation gets its own row.
3. Different units of observation require different data files.

If, for example, we were to measure the heights and leaf lengths for 4 trees, we might organise the data as in Table 2.1.

Table 2.1.: Hypothetical tidy dataset in which each column of data is a variable and each row of data is an observational unit (tree).

Tree	Species	Height (m)	Leaf length (cm)
1	Oak	20.3	8.1
2	Oak	25.4	9.4
3	Maple	18.2	12.5
4	Maple	16.7	11.3

By convention (Wickham, 2014), variables tend to be in the left-most columns if they are known in advance or fixed in some way by the data collection or experiment (e.g., tree number or species in Table 2.1). In contrast, variables that are actually measured tend to be in the right-most columns (e.g., tree height or leaf length). This is more for readability of the data; statistical software such as Jamovi will not care about the order of data columns.

2.2. Data files

Data can be saved using many different file types. File type is typically indicated by an extension following the name of a file and a full stop. For example, "photo.png" would indicate a PNG image file named "photo". A peer-reviewed journal article might be saved as a PDF, e.g., "Wickham2014.pdf". A file's type affects what programs can be used to open it. One relevant distinction to make is between text files and binary files.

Text files are generally very simple. They only allow information to be stored as plain text; no colour, bold, italic, or anything else is encoded. All of the information is just made up of characters on one or more lines. This sounds so simple as to be almost obsolete; what is the point of not allowing anything besides plain text? The point is that text files are generally much more secure for long-term storage. The plain text format makes data easier to recover if a file is corrupted, readable by a wider range of software, and more amenable to version control (version control is a tool that essentially saves the whole history of folder, and potentially different versions of it in parallel; it is not necessary for introductory statistics, but is often critical for big collaborative projects). There are many types of text files with extensions such as TXT, CSV, HTML, R, CPP, or MD. For data storage, we will use comma separated value (CSV) files. As the name implies, CSV files include plain text separated by commas. Each line of the CSV file is a new row, and commas separate information into columns. These CSV files can be opened in any text editor, but are also recognised by nearly all spreadsheet programs and statistical software. The data shown in Figure 2.3 are from a CSV file called "wasp_data.csv". Figure 2.4 shows the same data when opened with a text editor.

Figure 2.4.: A plain text comma-separated value (CSV) file showing data from fig fruits collected in 2010. Each line is a unique row and sample (fruit), and commas separate the data into columns in which the properties of fruit are recorded. The file has been opened in a program called 'Mousepad', but it could also be opened in any text editor such as gedit, Notepad, vim, or emacs. It could also be opened in spreadsheet programs such as LibreOffice Calc, MS Excel, or Google Sheets, or in any number of statistical programs.

The data shown in Figure 2.4 are not easy to read or work with, but the format is highly effective for storage because all of the information is in plain text. The information will therefore always look *exactly* the same, and can be easily recovered by any text editor, even after years pass and old software inevitably becomes obsolete.

Binary files are different from text files and contain information besides just plain text. This information could include formatted text (e.g., bold, italic), images, sound, or video (basically, anything that can be stored in a file). The advantages of being able to store this kind of information are obvious, but the downside is that the information needs to be interpreted in a specific way, usually using a specific program. Examples of binary files include those with extensions such as DOC, XLS, PNG, GIF, MP3, or PPT. Some file types such as DOCX are not technically binary files, but a collection of zipped files (which, in the case of DOCX, include plain text files). Overall, the important point is that saving data in a text file format such as CSV is generally more secure.

2.3. Managing data files

Managing data files (or any files) effectively requires some understanding of how files are organised on a computer or cloud storage. In mobile phone applications, file organisation is often hidden, so it is not obvious where a file actually goes when it is saved on a device. Many people find files in these applications using a search function. The ability to search for files like this, or at least the tendency to do so regularly, is actually a relatively new phenomenon. And it is an approach to file organisation that does not work quite as well on non-mobile devices (i.e., anything that is not a phone or tablet), especially for big projects. On laptop and desktop computers, it is really important to know *where* files

are being saved, and to ideally have an organisational system that makes it easy to find specific files without having to use a search tool.

On a computer, files are stored in a series of nested folders. You can think of the storage space on a computer, cloud, or network drive, as a big box. The big box can contain other smaller boxes (folders, in this analogy), or it can contain items that you need (files, in this analogy). Figure 2.5 shows the general idea. On this computer, there is a folder called 'brad', which has inside it 5 other folders (Figure 2.5A). Each of the 5 inner folders is used to store more folders and files for a specific module from 2006. Clicking on the 'Biostatistics' folder leads to the sub-folders inside it, and to files saved specifically for a biostatistics module (e.g., homework assignments, lecture notes, and an exam review document). Files on a computer therefore have a location that we can find using a particular **path**. We can write the path name using slashes to indicate nested folders. For example, the file 'HW9.scx' in Figure 2.5B would have the path name '/home/brad/Spring_2006/Biostatistics/HW9.scx'. Each folder is contained within slashes, and the file name itself is after the last slash.

Figure 2.5.: File directory of a computer showing (A) the file organisation of modules taken during spring 2006. Within one folder (B), there are multiple sub-folders and files associated with a biostatistics module.

These path names might look slightly different depending on the computer operating system that you are using. But the general idea of files nested within folders is the same. Figure 2.6 shows the same folder 'Spring_2006' saved in a different location, on OneDrive.

Windows has the same general file organisation (Figure 2.7). Path names for storing files on the hard drive of a Windows computer look something like "C:\Users\MyName\Desktop\Spring_2006\Biostatistics\HW9.scx". The 'C:\' is the root directory of the hard drive; it is called 'C' for historical reasons ('A:\' and 'B:\'

Figure 2.6.: OneDrive file directory showing the file organisation of modules taken during spring 2006.

used to be for floppy disks; the 'A:\' floppy disks had about 1.44 MB of storage, and 'B:\' had even less, so these are basically obsolete).

The details are not as important as the idea of organising files in a logical way that allows you to know roughly where to find important files on a computer or cloud drive. It is usually a good idea to give every unique project or subject (e.g., a university module, a student group, holiday plans, health records) its own folder. This makes it much easier to find related files such as datasets, lecture notes, or assignments when necessary. It is usually possible to right click somewhere in a directory to create a new folder. In Figure 2.7, there is even a 'New folder' button in the toolbar with a yellow folder icon above it. It takes some time to organise files this way, and to get used to saving files in specific locations, but it is well worth it in the long-term.

Figure 2.7.: Windows file directory showing the file organisation of modules taken during spring 2006. In this case, the 'Spring_2006' folder is located on the desktop; the path to the folder is visible in the toolbar above the folders.

3. Practical: Preparing data

In this practical, we will use a spreadsheet to organise datasets following the tidy approach explained in Chapter 2, then save these datasets as CSV files to be opened in Jamovi statistical software (The Jamovi Project, 2022). The data organisation in this lab can be completed using LibreOffice Calc, MS Excel, or Google Sheets. In the computer lab, MS Excel is probably the easiest program to use, either through AppsAnywhere or within a browser. The screenshots below will mostly be of LibreOffice Calc, but the instructions provided will work on any of the three aforementioned spreadsheet programs. You can download a PDF of this practical here, or a DOCX of just the questions here.

There are 4 data exercises in this practical. All of these exercises will focus on organising data into a tidy format. Being able to do this will be essential for later practicals and assessments, and for future modules (especially fourth year dissertation work). Exercise 1 uses handwritten field data that need to be entered into a spreadsheet in a tidy format. These data include information shown in Figure 2.2, plus tallies of seed counts. The goal is to get all of this information into a tidy format and save it as a CSV file. Exercise 2 presents some data on the number of eggs produced by five different fig wasp species (more on these in Chapter 8). The data are in an untidy format, so the goal is to reorganise them and save them as a tidy CSV file. Exercise 3 presents counts of the same five fig wasp species as in Exercise 2, which need to be reorganised in a tidy format. Exercise 4 presents data that are even more messy. These are morphological measurements of the same five species of wasps, including lengths and widths of wasp heads, thoraxes, and abdomens. The goal in this exercise is to tidy the data, then estimate total wasp volume from the morphological measurements using mathematical formulas, keeping in mind the order of operations from Chapter 1.

3.1. Exercise 1: Transferring data to a spreadsheet

Exercise 1 focuses on data collected from the fruits of fig trees collected from Baja, Mexico in 2010 (Duthie et al., 2015; Duthie and Nason, 2016). Due to the nature of the work, the data needed to be recorded in notebooks and collected in two different locations. The first location was the field, where data were collected identifying tree locations and fruit dimensions. Baja is hot and sunny; fruit measurements were made with a ruler and recorded in a field notebook. These measurements are shown in Figure 2.2, which is reproduced again in Figure 3.1.

Figure 3.1.: A fully grown Sonoran Desert Rock Fig in the desert of Baja, Mexico.

The second location was in a lab in Iowa, USA. Fruits were dried and shipped to Iowa State University so that seeds could be counted under a microscope. Counts were originally recorded as tallies in a lab notebook (Figure 3.2). The goal of Exercise 1 is to get all of this information into a single tidy spreadsheet.

The best place to start is with an empty spreadsheet, so open a new one in LibreOffice Calc, MS Excel, or Google Sheets. Remember that each row will be a unique observation; in this case, a unique fig fruit from which measurements were recorded. Each column will be a variable of that observation. Fortunately, the data in Figure 3.2 are already looking quite tidy. The information here can be put into the spreadsheet mostly as written in the notebook. But there are a few points to keep in mind:

1. It is important to start in column A and row 1; do not leave any empty rows or columns because when we get to the statistical analysis in Jamovi, Jamovi will assume that these empty rows and columns signify missing data.
2. There is no need to include any formatting (e.g., bold, underline, colour) because it will not be saved in the CSV or recognised by Jamovi.
3. Missing information, such as the empty boxes for the fruit dimensions in row 4 in the notebook (Figure 3.2) should be indicated with an '`NA`' (capital letters, but without the quotes). This will let Jamovi know that these data are missing.
4. The date is written in an American style of month-day-year, which might get confusing. It might be better to have separate columns for year, month, and day,

DATE (m	SPECIES	SITE NO.	TREE NO	FRUIT NO	FRT LENG	FRT WID	FRT
5/9/10	F-pet	70	70	1	15	18	14
5/10/10	F-pet	70	70	2	17	19	15
5/10/10	F-pet	70	70	3	21	21	16
5/11/10	F-pet	70	70	4			
5/11/10	F-pet	70	70	5	15	16	14
5/11/10	F-pet	70	70	6	16	16	15

Figure 3.2.: A portion of a lab notebook used to record measurements of fig fruits from different trees in 2010.

and to write out the full year (2010).

The column names in Figure 3.2 are (1) Date, (2) Species, (3) Site number, (4) Tree number, (5) Fruit length in mm, (6) Fruit width in mm, and (7) Fruit height in mm. All of the species are *Ficus petiolaris*, which is abbreviated to "F-pet" in the field notebook. How you choose to write some of this information down is up to you (e.g., the date format, capitalisation of column names), but when finished, the spreadsheet should be organised like the one in Figure 3.3.

	A	B	C	D	E	F	G	H	I	J
1	Year	Month	Day	Species	Site number	Tree number	Fruit number	Fruit length (mm)	Fruit width (mm)	Fruit height (mm)
2	2010	5	9	F_petiolaris	70	70	1	15	18	14
3	2010	5	10	F_petiolaris	70	70	2	17	19	15
4	2010	5	10	F_petiolaris	70	70	3	21	21	16
5	2010	5	11	F_petiolaris	70	70	4	NA	NA	NA
6	2010	5	11	F_petiolaris	70	70	5	15	16	14
7	2010	5	11	F_petiolaris	70	70	6	16	16	15

Figure 3.3.: A spreadsheet with data organised in a tidy format and nearly ready for analysis.

This leaves us with the data that had to be collected later in the lab. Small seeds needed to be meticulously separated from other material in the fig fruit, then tallied under a microscope. Tallies from this notebook are shown in Figures 3.4 and 3.5.

Fortunately, the summed tallies have been written and circled in the right margin of the notebook, which makes inputting them into a spreadsheet easier. But it is important to also recognise this step as a potential source of human error in data collection. It is possible that the tallies were counted inaccurately, meaning that the tallies on the left do not sum to the numbers in the right margins. It is always good to be able to go back and check. There are at least two other potential sources of human error in counting

Figure 3.4.: Tallies of seed counts collected from 4 fig fruits in Baja, Mexico in 2010.

Figure 3.5.: Tallies of seed counts collected from 2 fig fruits in Baja, Mexico in 2010.

seeds and inputting them into the spreadsheet, one before, and one after counting the tallies. Fill in 1 and 3 below with potential causes of error.

1.
2. Tallies are not counted correctly in the lab notebook
3.

Next, create a new column in the spreadsheet and call it "Seeds" (use column K). Fill in the seed counts for each of the six rows. The end result will be a tidy dataset that is ready to be saved as a CSV.

What you do next depends on the spreadsheet program that you are using and how you are using it. If you are using LibreOffice Calc or MS Excel on a your computer, then you should be able to simply save your file as something like "Fig_fruits.csv", and the program will recognise that you intend to save as a CSV file (in MS Excel, you might need to find the pulldown box for 'Save as type:' under the 'File name:' box and choose 'CSV'). If you are using Google Sheets, you can navigate in the toolbar to `File > Download > Comma-separated values (.csv)`, which will start a download of your spreadsheet in CSV format. If you are using MS Excel in a browser online, then it is a bit more tedious. At the time of writing, the online version of MS Excel does not allow users to save or export to a CSV. It will therefore be necessary to save as an XLSX, then convert to CSV later in another spreadsheet program (either a local version of MS Excel, LibreOffice Calc, or Google Sheets).

Save your file in a location where you know that you can find it again. It might be a good idea to create a new folder on your computer or your cloud storage online for files in Statistical Techniques. This will ensure that you always know where your data files are located and can access them easily.

3.2. Exercise 2: Making spreadsheet data tidy

Exercise 2 is more self-guided than Exercise 1. After reading Chapter 2 and completing Exercise 1, you should have a bit more confidence in organising data in a tidy format. Here we will work with a dataset that includes counts of the number of eggs collected from fig wasps, which are small species of insects that lay their eggs into the ovules of fig flowers (Weiblen, 2002). You can **download the dataset here**, or recreate it from Figure 3.6.

Using what you have learned in Chapter 2 and Exercise 1, create a tidy version of the wasp egg loads dataset. For a helpful hint, it might be most efficient to open a new spreadsheet and copy and paste information from the old to the new.

How many columns did you need to create the new dataset? _____

Are there any missing data in this dataset? _____

Figure 3.6.: An untidy dataset of egg loads from fig wasps of five different species, including two unnamed species of the genus *Heterandrium* (Het1 and Het2) and three unnamed species of the genus *Idarnes* (LO1, SO1, and SO2).

Save the tidy dataset to a CSV file. It might be a good idea to check with classmates and an instructor to confirm that the dataset is in the correct format.

3.3. Exercise 3: Making data tidy again

Exercise 3, like Exercise 2, is self-guided. The data are presented in a fairly common, but untidy, format, and the challenge is to reorganise them into a tidy dataset that is ready for statistical analysis. Table 3.1 shows the number of different species of wasps counted in 5 different fig fruits. Rows list all of the species and columns list the fruits, with the counts in the middle This is an efficient way to present the data so that they are all easy to see, but this will not work for running statistical analysis.

Table 3.1.: An efficient but untidy way to present count data. Counts of different species of fig wasps (rows) are from 5 different fig fruits (columns). Data were originally collected from Baja, Mexico in 2010.

Species	Fruit_1	Fruit_2	Fruit_3	Fruit_4	Fruit_5
Het1	0	0	0	1	0
Het2	0	2	3	0	0
LO1	4	37	0	0	3
SO1	0	1	0	3	2
SO2	1	12	2	0	0

This exercise might be a bit more challenging than Exercise 2. The goal is to use the above information to create a tidy dataset. Remember that each observation (wasp counts, in this case) should get its own row, and each variable should get its own column. Try creating a tidy dataset from the information in Table 3.1, then save the dataset to a CSV file. As with Exercise 2, it might be good to confer with classmates and an instructor to confirm that the dataset is in the correct format and will work for statistical analysis.

3.4. Exercise 4: Tidy data and spreadsheet calculations

Exercise 4 requires some restructuring and calculations. The dataset that will be used in this exercise includes morphological measurements from five species of fig wasps, the same species used in Exercise 2. **Download this dataset from the file wasp_morphology_untidy.xlsx (XLSX file) or wasp_morphology_untidy.ods (ODS open-source file).** Both files contain identical information, so which one you use is a matter of personal preference. This dataset is about as untidy as it gets. First note that there are multiple sheets in the spreadsheet, which is not allowed in a tidy CSV file. You can see these sheets by looking at the very bottom of the spreadsheet, which will have separating tabs called Het1, Het2, LO1, SO1, and SO2 (Figure 3.7).

Figure 3.7.: Spreadsheets can include multiple sheets. This image shows that the spreadsheet containing information for fig wasp morphology includes five separate sheets, one for each species.

You can click on all of the different tabs to see the measurements of head length, head width, thorax length, thorax width, abdomen length, and abdomen width for wasps of each of the 5 species. All of the measurements are collected in millimeters. Note that the individual sheets contain text formatting (titles highlighted, and in bold), and there is a picture of each wasp in its respective sheet. The formatting and pictures are a nice touch for providing some context, but they cannot be used in statistical analysis. The first task is to create a tidy version of this dataset. Probably the best way to do this is to create a new spreadsheet entirely and copy-paste information from the old. It is good idea to think about how the tidy dataset will look before getting started. What columns should this new dataset include? Write your answer below.

How many rows are needed? _____

When you are ready, create the new dataset. Your dataset should have all of the relevant information about wasp head, thorax, and abdomen measurements. It should look something like Figure 3.8.

	A	B	C	D	E	F	G	H
1	Species	Head_Length_mm	Head_Width_mm	Thorax_Length_mm	Thorax_Width_mm	Abdomen_Length_mm	Abdomen_Width_mm	
2	Het1	0.566	0.698	0.767	0.494	1.288	0.504	
3	Het1	0.505	0.607	0.784	0.527	1.059	0.43	
4	Het1	0.511	0.622	0.769	0.511	1.107	0.504	
5	Het1	0.479	0.601	0.766	0.407	1.242	0.446	
6	Het1	0.545	0.707	0.828	0.561	1.367	0.553	
7	Het1	0.525	0.651	0.852	0.59	1.408	0.618	
8	Het2	0.497	0.607	0.781	0.487	1.248	0.601	
9	Het2	0.45	0.565	0.696	0.432	1.092	0.504	
10	Het2	0.557	0.637	0.792	0.445	1.24	0.469	
11	Het2	0.519	0.563	0.814	0.443	1.221	0.623	
12	Het2	0.43	0.53	0.621	0.372	1.034	0.546	
13	LO1	0.43	0.517	0.897	0.394	1.176	0.71	
14	LO1	0.357	0.469	0.722	0.326	0.875	0.435	
15	LO1	0.383	0.488	0.678	0.468	1.097	0.609	
16	LO1	0.433	0.562	0.858	0.456	1.061	0.521	
17	LO1	0.402	0.527	0.823	0.438	1.266	0.777	
18	LO1	0.426	0.508	0.723	0.377	1.097	0.654	
19	SO1	0.365	0.513	0.67	0.4	1.124	0.575	
20	SO1	0.361	0.483	0.624	0.385	1.095	0.55	
21	SO1	0.377	0.508	0.725	0.391	0.973	0.389	
22	SO1	0.302	0.379	0.498	0.279	0.682	0.358	
23	SO2	0.394	0.538	0.712	0.406	1.006	0.655	
24	SO2	0.353	0.423	0.64	0.35	0.963	0.541	
25	SO2	0.363	0.513	0.686	0.457	1.025	0.523	
26	SO2	0.329	0.432	0.648	0.388	0.975	0.414	
27	SO2	0.364	0.511	0.684	0.367	0.972	0.505	

Figure 3.8.: A tidy dataset of wasp morphological measurements from 5 species of fig wasps collected from Baja, Mexico in 2010.

Next comes a slightly more challenging part, which will make use of some of the background mathematics reviewed in Chapter 1. Suppose that we wanted our new dataset to include information about the volumes of each of the three wasp body segments, and wasp total volume. To do this, let us assume that the wasp head is a sphere (it is not, exactly, but this is probably the best estimate that we can get under the circumstances). Calculate the head volume of each wasp using the following formula,

$$V_{head} = \frac{4}{3}\pi \left(\frac{Head_L + Head_W}{4} \right)^3 .$$

In the equation above, $Head_L$ is head length (mm) and $Head_W$ is head width (note, $(Head_L + Head_W)/4$ estimates the radius of the head). You can replace π with the approximation $\pi \approx 3.14$. To make this calculation in your spreadsheet, find the cell in which you want to put the head volume. By typing in the `=` sign, the spreadsheet will know to start a new calculation or function in that cell. Try this with an empty cell by typing "`= 5 + 4`" in it (without quotes). When you hit 'Enter', the spreadsheet will make the calculation for you and the number in the new cell will be 9. To see the equation again, you just need to double-click on the cell.

To get an estimate of head volume into the dataset, we can create a new column of data. To calculate V_{head} for the first wasp in row 2 of Figure 3.8, we could select the spreadsheet cell H2 and type the code, `=(4/3)*(3.14)*((B2+C2)/4)^3`. Notice that the code recognises `B2` and `C2` as spreadsheet cells, and takes the values from these cells when doing these calculations. If the values of `B2` or `C2` were to change, then so would the calculated value in H2. Also notice that we are using parentheses to make sure that the order of operations is correct. We want to add head length and width before dividing by 4, so we type `((B2+C2)/4)` to ensure with the innermost parentheses that head length and width are added before dividing. Once all of this is completed, we raise everything in parentheses to the third power using the `^3`, so `((B2+C2)/4)^3`. Different mathematical operations can be carried out using the the symbols in Table 3.2.

Table 3.2.: List of mathematical operations available in a spreadsheet.

Symbol	Operation
`+`	Addition
`-`	Subtraction
`*`	Multiplication
`/`	Division
`^`	Exponent
`sqrt()`	Square-root

The last operation in Table 3.2 is a function that takes the square-root of anything within the parentheses. Other functions are also available that can make calculations across cells (e.g., `=SUM` or `=AVERAGE`), but we will ignore these for now.

Once head volume is calculated for the first wasp in cell H2, it is very easy to do the rest. One nice feature of a spreadsheet is that it can usually recognise when the cells need to change (B2 and C2, in this case). To get the rest of the head volumes, we just need to select the bottom right of the H2 cell. There will be a very small square in this bottom right (see Figure 3.9), and if we drag it down, the spreadsheet will do the same calculation for each row (e.g., in H3, it will use B3 and C3 in the formula rather than B2 and C2).

Another way to achieve the same result is to copy (Ctrl + C) the contents of cell H2, highlight cells H3-H27, then paste (Ctrl + V). However you do it, you should now have a new column of calculated head volume.

Next, suppose that we want to calculate thorax and abdomen volumes for all wasps. Unlike wasp heads, wasp thoraxes and abdomens are clearly not spheres. But it is perhaps not entirely unreasonable to model them as ellipses. To calculate wasp thorax and abdomen volumes assuming an ellipse, we can use the formula,

fx Σ = =(4/3)*(3.14)*((B2+C2)/4)^3

	E	F	G	H
n	Thorax_Width_mm	Abdomen_Length_mm	Abdomen_Width_mm	Head_vol
'67	0.494	1.288	0.504	0.132108157
'84	0.527	1.059	0.43	
'69	0.511	1.107	0.504	
'66	0.407	1.242	0.446	

Figure 3.9.: A dataset of wasp morphological measurements from 5 species of fig wasps collected from Baja, Mexico in 2010. Head volume (column H) has been calculated for row 2, and to calculate it for the remaining rows, the small black square in the bottom right of the highlighted cell H2 can be clicked and dragged down to H27.

$$V_{thorax} = \frac{4}{3}\pi \left(\frac{Thorax_L}{2}\right)\left(\frac{Thorax_W}{2}\right)^2.$$

In the equation above, $Thorax_L$ is thorax length (mm) and $Thorax_W$ is thorax width. Substitute $Abdomen_L$ and $Abdomen_W$ to instead calculate abdomen volume ($V_{abdomen}$). What formula will you type into your empty spreadsheet cell to calculate V_{thorax}? Keep in mind the order of operations indicated in the equation above.

Now fill in the columns for thorax volume and abdomen volume. You should now have 3 new columns of data from calculations of the volumes of the head, thorax, and abdomen of each wasp. Lastly, add 1 final column of data for total volume, which is the sum of the 3 segments.

There are a lot of potential sources of error and uncertainty in these final volumes. What are some reasons that we might want to be cautious about our calculated wasp volumes? Explain in 2-3 sentences.

Save your wasp morphology file as a CSV. This was the last exercise of the practical. You should now be comfortable formatting tidy datasets for use in statistical software. Next week, we will begin using Jamovi to do some descriptive statistics and plotting.

3.5. Summary

Completing this practical should give you the skills that you need to prepare datasets for statistical analysis. There are many additional features of spreadsheets that were not introduced (mainly because we will do them in Jamovi), but could be useful to learn. For example, if we wanted to calculate the sum of all head lengths, we could use the function `=sum(B2:B27)` in any spreadsheet cell (where B2 is the head length of the first wasp, and B27 is the head length of the last wasp). Other functions such as `=count()`, `=min()`, `=max()`, or `=average()` can be similarly used for calculations. If you have time at the end of the lab, we recommend exploring the spreadsheet interface and seeing what you can do.

Part II.

Statistical concepts

Week 2 Overview

Dates	30 January 2023 - 3 February 2023
Reading	**Required:** SCIU4T4 Workbook chapters 4-7
	Recommended: Navarro and Foxcroft (2022) Section 3.3-3.9
	Suggested: Rowntree (2018) Chapter 2
	Advanced: None
Lectures	2.0: Why study statistics? (18:13 min; Video)
	2.1: Populations and samples (6:47 min; Video)
	2.2: Types of variables (11:00 min; Video)
	2.3: Units, precision, and accuracy (9:06 min; Video)
	2.4: Uncertainty propagation (11:44 min; Video)
Practical	Introduction to Jamovi (Chapter 8)
	Room: Cottrell 2A17
	Group A: 01 FEB 2023 (WED) 13:05-15:55
	Group B: 02 FEB 2023 (THU) 09:05-11:55
Help hours	Ian Jones
	Room: Cottrell 1A13
	03 FEB 2023 (FRI) 15:05-17:55
Assessments	Week 2 Practice quiz on Canvas

Week 2 focuses on general statistical concepts, data, and measurement.

Chapter 4 focuses on key concepts that will be used throughout this module. In particular, it is important to understand the difference between a **population** and a **sample**, and to recognise that there are many types of variables in statistics.

Chapter 5 introduces different variable types. Different types of variables have different characteristics, which will affect how these variables are best visualised in figures and analysed with statistical hypothesis tests introduced later in the semester. A variable's type will rarely be stated explicitly when doing scientific research, and will not always be provided in assessments for this module. Being able to infer variable type is therefore an important skill.

Chapter 6 focuses on units of measurement, and how these units are communicated in text. Units are essential in scientific measurement, and we will use them throughout the

module to indicate the type and scale of data measurement. We are not expecting you to memorise all scientific units, so a table on units is provided.

Chapter 7 will introduce the propogation of measurement errors. This is important to understand because no measurement is perfectly accurate, and predicting how errors in measurement combine is fundamental to understanding measurement accuracy.

Chapter 8 guides you through the Week 2 practical, which is an introduction to Jamovi. This aim of this practical is to become familiar with the Jamovi interface and comfortable importing data into Jamovi to collect some descriptive statistics.

4. Populations and samples

When we collect data, we are recording some kind of observation or measurement. If we are working in a forest, for example, we might want to measure the heights of different trees, or measure the concentration of carbon in the soil. The idea might be to use these measurements to make some kind of inference about the forest. But as scientists, we are almost always limited in the amount of data that we can collect. We cannot measure everything, so we need to collect a *sample* of data and use it to make inferences about the *population* of interest. For example, while we probably cannot measure the height of every tree in a forest, nor can we measure the concentration of carbon at every possible location in the forest's soil, we can collect a smaller number of measurements and still make useful conclusions about overall forest tree height and carbon concentration.

Statistics thereby allows us to approximate properties of entire populations from a limited number of samples. This needs to be done with caution, but before getting into the details of how, it is important to fully understand the difference between a **population** and a **sample** to avoid confusing these two concepts. A **population** is the entire set of possible observations that could be collected. Some examples will make it easier to understand:

- All of the genes in the human genome
- All individuals of voting age in Scotland
- All common pipistrelle bats in the United Kingdom

These populations might be important for a particular research question. For example, we might want to know something about the feeding behaviours of pipistrelle bats in the UK. But there is no way that we can find and observe the behaviour of every single bat, so we need to take a subset of the population (a sample) instead. Examples of samples include the following:

- A selection of 20 human genes
- A pub full of Scottish voters
- 40 caught common pipistrelle bats

It is important to recognise that the word "population" means something slightly different in statistics than it does in biology. A biological population, for example, could be defined as all of the individuals of the same species in the same general location. A statistical population, in contrast, refers to a set of observations (i.e., things that we can measure). Sokal and Rohlf (1995) provide a more technical definition for "population",

> In statistics, population always means the *totality of individual observations about which inferences are to be made, existing anywhere in the world or at least within a definitely specified sampling area limited in space and time* [p. 9, emphasis theirs].

They define a sample to be "a collection of individual observations selected by a specified procedure" (Sokal and Rohlf, 1995). For our purposes, it is not necessary to be able to recite the technical definitions, but it is important to understand the relationship between a population and a sample. When we collect data, we are almost always taking a small sample of observations from a much larger number of possible observations in a population.

Figure 4.1.: A conceptual figure illustrating how a statistical population relates to a statistical sample. The population is represented by 35 smiling faces enclosed within a dashed box. The sample is represented by a solid box within the dashed box, within which there are 3 smiling faces. Hence, we have a sample of 3 measurements from the total population.

5. Types of variables

A variable is any property that is measured in an observation (Sokal and Rohlf, 1995); i.e., anything that varies among things that we can measure (Dytham, 2011). We can summarise how these measurements vary with summary statistics, or visually with figures. Often, we will want to predict one variable from a second variable. In this case, the variable that we want to predict is called the **response variable**, also known as the **dependent variable** or **Y** variable ('dependent' because it *depends* on other variables, and 'Y' because this is the letter we often use to represent it). The variable that we use to predict our response variable is the **explanatory variable**, also known as the **independent variable** or **X** variable ('independent' because it does not depend on other variables, and 'X' because this is the letter most often used to represent it). There are several different types of variables:

- **Categorical** variables take on a fixed number of discrete values (Spiegelhalter, 2019). In other words, the measurement that we record will assign our data to a specific category. Examples of categorical variables include species (e.g., "Robin", "Nightingale", "Wren") or life history stage (e.g., "egg", "juvenile", "adult"). Categorical variables can be either nominal or ordinal.

  - **Nominal** variables do not have any inherent order (e.g., classifying land as "forest", "grassland", or "urban").
  - **Ordinal** variables do have an inherent order (e.g., "low", "medium", and "high" elevation).

- **Quantitative** variables are variables represented by numbers that reflect a magnitude. That is, unlike categorical variables, we are collecting numbers that really mean something tangible (in contrast, while we might represent low, medium, and high elevations with the numbers 1, 2, and 3, respectively, this is just for convenience; a value of '2' does not always mean 'medium' in other contexts). Categorical variables can be either discrete or continuous.

  - **Discrete** variables can take only certain values. For example, if we want to record the number of species in a forest, then our variable can only take discrete counts (i.e., integer values). There could conceivably be any natural number of species (1, 2, 3, etc.), but there could not be 2.51 different species in a forest; that does not make sense.

- **Continuous** variables can take any real value within some range of values (i.e., any number that can be represented by a decimal). For example, we could measure height to as many decimals as our measuring device will allow, with a range of values from zero to the maximum possible height of whatever it is we are measuring. Similarly, we could measure temperature to any number of decimals, at least in theory, so temperature is a continuous variable.

The reason for organising variables into all of these different types is that different types of variables need to be handled in different ways. For example, it would not make sense to visualise a nominal variable in the same way as a continuous variable. Similarly, the choice of statistical test to apply to answer a statistical question will almost always depend on the types of variables involved. If presented with a new data set, it is therefore very important to be able to interpret the different variables and apply the correct statistical techniques (this will be part of the assessment for this module).

6. Accuracy, precision, and units

The science of measurements is called "metrology", which, among other topics, focuses on measurement accuracy, precision, and units (Rabinovich, 2013). We will not discuss these topics in depth, but they are important for statistical techniques because measurement, in the broadest sense of the word, is the foundation of data collection. When collecting data, we want measurements to be accurate, precise, and clearly defined.

6.1. Accuracy

When we collect data, we are trying to obtain information about the world. We might, for example, want to know the number of seedlings in an area of forest, the temperature of the soil at some location, or the mass of a particular animal in the field. To get this information, we need to make measurements. Some measurements can be collected by simple observation (e.g., counting seedlings), while others will require measuring devices such as a thermometer (for measuring temperature) or scale (for measuring mass). All of these measurements are subject to error. The *true* value of whatever it is that we are trying to measure (called the "measurand") can differ from what we record when collecting data. This is true even for simple observations (e.g., we might miscount seedlings), so it is important to recognise that the data we collect comes with some uncertainty. The **accuracy** of a measurement is defined by how close the measurement is to the *true* value of what we are trying to measure (Rabinovich, 2013).

6.2. Precision

The **precision** of a measurement is how consistent it will be if measurement is replicated multiple times. In other words, precision describes how similar measurements are expected to be. If, for example, a scale measures an object to be the exact same mass every time it is weighed (regardless of whether the mass is accurate), then the measurement is highly precise. If, however, the scale measures a different mass each time the object is weighed (for this hypothetical, assume that the true mass of the object does not change), then the measurement is not as precise.

One way to visualise the difference between accuracy and precision is to imagine a set of targets, with the centre of the target representing the true value of what we are trying to measure (Figure 6.1)[1].

Figure 6.1.: A conceptual figure illustrating the difference between accuracy and precision. Points in (A) are both accurate and precise, points in (B) are accurate, but not precise, points in (C) are precise but not accurate, and points in (D) are neither accurate nor precise.

Note again that accuracy and precision are not necessarily the same. Measurement can be accurate but not precise (Figure 6.1B) or precise but not accurate (Figure 6.1C).

6.3. Systems of units

Scientific units are standardised with the Système International D'Unités (SI). Having standardised units of measurement is highly important to ensure measurement accuracy (Quinn, 1995). Originally, these units were often defined in terms of physical artefacts. For example, the kilogram (kg) was once defined by a physical cylinder of metal housed in the Bureau International des Poids et Mesures (BIPM). In other words, the mass of a metal sitting at the BIPM *defined* what a kg was, with the mass of every other measurement being based on this physical object (Quinn, 1995). This can potentially present a problem if the mass of that one object changes over time, thereby causing a change in how a kg is defined. Where possible, it is therefore preferable to define units in terms of fundamental constants of nature. In 2019, for example, the kg was redefined in terms of the Planck constant, a specific atomic transition frequency, and the speed of light (Stock et al., 2019). This ensures that measurements of mass remain accurate over time because what a kg represents in terms of mass cannot change.

We can separate units into base units and derived units. Table 6.1 below lists some common base units for convenience (Quinn, 1995). You do not need to memorise these units, but it is good to be familiar with them. We will use these units throughout the module.

[1]This figure was released into the public domain by Egon Willighagen on 8 March 2014.

Table 6.1.: Base units of SI measurements. For details see Quinn (1995).

Measured Quantity	**Name of SI unit**	Symbol
Mass	kilogram	kg
Length	metre	m
Time	second	s
Electric current	ampere	A
Temperature	kelvin	K
Amount of a substance	mole	mol
Luminous intensity	candela	cd

We can also define derived SI units from the base units of Table 6.1; examples of these derived SI units are provided in Table 6.2. Again, you do not need to memorise these units, but it is good to be aware of them.

Table 6.2.: Examples of derived SI units.

Measured Quantity	**Name of unit**	**Symbol**	**Definition in SI units**	**Alternative in derived units**
Energy	Joule	J	$m^2\ kg\ s^{-2}$	$N\ m$
Force	Newton	N	$m\ kg\ s^{-2}$	$J\ m^{-1}$
Pressure	Pascal	Pa	$kg\ m^{-1}\ s^{-2}$	$N\ m^{-2}$
Power	Watt	W	$m^{-2}\ kg\ s^{-3}$	$J\ s^{-1}$
Frequency	Hertz	Hz	s^{-1}	
Radioactivity	Becquerel	Bq	s^{-1}	

When numbers are associated with units, it is important to recognise that the units must be carried through and combined when calculating an equation. As a very simple example, if want to know the speed at which an object is moving, and we find that it has moved 10 metres in 20 seconds, then we calculate the speed and report the correct units as below,

$$speed = \frac{10\,m}{20\,s} = 0.5\,m/s = 0.5\,m\,s^{-1}.$$

Notice that the final units are in metres per second, which can be written as m/s or $m\,s^{-1}$ (remember that raising s to the -1 power is the same as $1/s$; see Chapter 1 for a quick reminder about superscripts). It is a common error to calculate just the numeric components of a calculation and ignore the associated units. Often on assessments, we will ask you not to include units in your answer (this is just for convenience on the tests and exam), but recognising that units are also part of calculations is important.

6.4. Other examples of units

Remember that an exponent (indicated by a superscript, e.g., the 3 in m^3) indicates the number of times to multiply a base by itself, so $m^3 = m \times m \times m$.

6.4.1. Units of density

Density (ρ) is calculated by,

$$\rho = \frac{mass}{volume} = \frac{kg}{m^3}.$$

The units of density are therefore mass per unit volume, $kg\,m^{-3}$.

6.4.2. Mass of metal discharged from a catchment

The mass of metal carried by a stream per unit time (M) is given by multiplying the concentration of metal per unit volume (C) of water by the volume of water discharged per unit time (V),

$$M = C \times V.$$

This equation is useful in showing how units can cancel each other out. If we calculate the above with just the units (ignoring numbers for C and V),

$$M = \frac{mg}{l} \times \frac{l}{s} = \frac{mg}{s}.$$

Notice above how the l units on the top and bottom of the equation cancel each other out, so we are left with just mg/s.

6.4.3. Soil carbon inventories

For one final example, the inventory of carbon I within a soil is given by the specific carbon concentration C (g of carbon per kg of soil), multiplied by the depth of soil analysed (D, measured in m), and by the density (ρ, measured in $kg\,m^{-3}$),

$$I = C \times D \times \rho = \frac{g \times m \times kg}{kg \times m^3} = \frac{g}{m^2} = g\,m^{-2}.$$

Notice above how the kg on the top and bottom of the fraction cancel each other out, and how one m on the top cancels out one m on the bottom, so that what we are left with is grams per metre squared ($g\,m^{-2}$).

7. Uncertainty propogation

Nothing can be measured with perfect accuracy, meaning that every measurement has some associated error. The measurement error might be caused by random noise in the measuring environment, or by mistakes made by the person doing the measuring. The measurement error might also be caused by limitations or imperfections associated with a measuring device. The device might be limited in its measurement precision, or perhaps it is biased in its measurements due to improper calibration, manufacture, or damage from previous use. All of this generates uncertainty with respect to individual measurements.

Recall from Chapter 6 the difference between precision and accuracy. We can evaluate the precision and accuracy of measurements in different ways. Measurement precision can be estimated by replicating a measurement (i.e., taking the same measurement over and over again). The more replicate measurements made, the more precisely a value can be estimated. For example, if we wanted to evaluate the precision with which the mass of an object is measured, then we might repeat the measurement with the same scale multiple times and see how much mass changes across different measurements. To evaluate measurement accuracy, we might need to measure a value in multiple different ways (e.g., with different measuring devices). For example, we might repeat the measurement of an object's mass with a different scale (i.e., a different physical scale used for measuring the mass of objects).

Sometimes it is necessary to combine different measured values. For example, we might measure the mass of 2 different bird eggs in a nest separately, then calculate the total mass of both the 2 eggs combined. The measurement of each egg will have its own error, and these errors will propagate to determine the error of the total egg mass for the nest. How this error propagates differs depending on if they are being added or subtracted, or if they are being multiplied or divided.

7.1. Adding or subtracting errors

In the case of our egg masses, we can assign the mass of the first egg to the variable X and the mass of the second egg to the variable Y. We can assign the total mass to the variable Z, where $Z = X + Y$. The errors associated with the variables X, Y, and Z can be indicated by E_X, E_Y, and E_Z, respectively. In general, if the variable Z is

calculated by adding (or subtracting) 2 or more values together, then this is the formula for calculating E_Z,

$$E_Z = \sqrt{E_X^2 + E_Y^2}.$$

Hence, for the egg masses, the error of the combined masses (E_Z^2) equals the square root of the error associated with the mass of egg 1 squared (E_X^2) plus the error associated with the mass of egg 2 squared (E_Y^2). It often helps to provide a concrete example. If the error associated with the measurement of egg 1 is $E_X^2 = 2$, and the error associated with the measurement of egg 2 is $E_Y^2 = 3$, then we can calculate,

$$E_Z = \sqrt{2^2 + 3^2} \approx 3.61.$$

Note that the units of E_Z are the same as Z (e.g., grams).

7.2. Multiplying or dividing errors

Multiplying or dividing errors works a bit differently. As an example, suppose that we need to measure the total area of a rectangular field. If we measure the length (L) and width (W) of the field, then the total area is the product of these measurements, $A = L \times W$. Again, there is going to be error associated with the measurement of both length (E_L) and width (E_W). How the error of the total area (E_A) is propagated by E_L and E_W is determined by the formula,

$$E_A = A\sqrt{\left(\frac{E_L}{L}\right)^2 + \left(\frac{E_W}{W}\right)^2}.$$

Notice that just knowing the error of each measurement (E_L and E_W) is no longer sufficient to calculate the error associated with the measurement of the total area. We also need to know L, W, and A. If our field has a length of $L = 20$ m and width of $W = 10$ m, then $A = 20 \times 10 = 200\, m^2$. If length and width measurements have associated errors of $E_L = 2$ m and $E_W = 1$ m, then,

$$E_A = 200\sqrt{\left(\frac{2}{20}\right)^2 + \left(\frac{1}{10}\right)^2} \approx 28.3\, m^2.$$

Of course, not every set of measurements with errors to be multiplied will be lengths and widths (note, however, that the units of E_A are the same as A, m^2). To avoid confusion, the general formula for multiplying or dividing errors is below, with the variables L,

W, and A replaced with X, Y, and Z, respectively, to match the case of addition and subtraction explained above,

$$E_Z = Z\sqrt{\left(\frac{E_X}{X}\right)^2 + \left(\frac{E_Y}{Y}\right)^2}.$$

Note that the structure of the equation is the exact same, just with different letters used as variables. It is necessary to be able to apply these equations correctly to estimate combined error.

7.3. Applying formulas for combining errors

It is not necessary to understand why the equations for propagating different types of errors are different, but a derivation is provided in Appendix B for the curious.

8. Practical. Introduction to Jamovi

This practical focuses on learning how to work with datasets in Jamovi (The Jamovi Project, 2022). Jamovi is available in the university laboratory computers through AppsAnywhere. You can also download it on your own computer for free or run it directly from a browser. For an introduction to what Jamovi is and why we are using it in this module, see the introduction of this workbook and Sections 3.3-3.9 of Navarro and Foxcroft (2022). In this practical, we will work with two datasets, both of which are based on real biological and environmental studies conducted by researchers at the University of Stirling.

The first dataset includes measurements of soil organic carbon (grams of Carbon per kg of soil) from the topsoil and subsoil collected in a national park in Gabon. These data were collected by Dr Carmen Rosa Medina-Carmona in an effort to understand how pyrogenic carbon (i.e., carbon produced by the charring of biomass during a fire) is stored in different landscape areas (Santín et al., 2016; Preston and Schmidt, 2006). **Download these data here: soil_organic_carbon.csv** (right click on the links and choose 'Save Link As…', then save it with the extension '.csv').

The second dataset includes measurements of figs from trees of the Sonoran Desert Rock Fig (*Ficus petiolaris*) in Baja, Mexico. These data were collected by Dr Brad Duthie in an effort to understand coexistence in a fig wasp community (Duthie et al., 2015; Duthie and Nason, 2016). Measurements include fig lengths, widths, and heights in centimeters from 4 different fig trees, and the number of seeds in each fruit. **Download these data here: fig_fruits.csv** (right click and "Save Link As…").

This lab will use the soil organic carbon dataset in Exercise 8.1 for summary statistics. The fig fruits will be used for Exercise 8.2 on transforming variables and Exercise 8.3 on computing a variable. Some of these exercises will be similar to what we did in the week 1 practical from Chapter 3, but in Jamovi rather than a separate spreadsheet.

8.1. Exercise for summary statistics

Download the soil organic carbon dataset if you have not already done so (right click and "Save Link As…"), and save it in a location where you know you will be able to find it again, then open Jamovi. Once Jamovi is open, you can import the dataset by clicking on the three horizontal lines in the upper left corner of the tool bar, then selecting 'Open' (Figure 8.2).

Figure 8.1.: Three images showing the process of collecting data for the dimensions of figs from trees of the Sonoran Desert Rock Fig in Baja, Mexico. (A) Processing fig fruits, which included measuring the diameter of figs along three diferent axes of length, width, and height, (B) a fig still attached to a tree with a fig wasp on top of it, and (C) a sliced open fig with seeds along the inside of it.

Figure 8.2.: The Jamovi toolbar including tabs for opening files, Variables, Data, Analyses, and Edit. To open a file, select the three horizontal lines in the upper right

You might need to click 'Browse' in the upper right of Jamovi to find the file. Figure 8.3 below shows how this will look when you browse for a data file.

Figure 8.3.: The Jamovi interface for opening a file with the 'Import' tab selected. Options for browsing to a file on the computer are available in the upper right, which opens the window in the foreground.

Once the data are imported, you should see two separate columns. The first column will show soil organic carbon values for topsoil samples, and the second column will show soil organic carbon values for subsoil samples. These data are not formatted in a tidy way. We need to fix this so that each row is a unique observation and each column is a variable (see Chapter 2). It might be easiest to reorganise the data in a spreadsheet such as LibreOffice Calc or Microsoft Excel. But you can also edit the data directly in Jamovi by clicking on the 'Data' tab in the toolbar (see Figures 8.2 and 8.4). The best way to reorganise the data in Jamovi is to double-click on the third column of data next to 'subsoil' (see Figure 8.4).

Figure 8.4.: The Jamovi toolbar is shown with the soil organic carbon dataset. In Jamovi, double-clicking above column three where it says 'CLICK HERE' will allow you to input a new variable.

After double-clicking on the location shown in Figure 8.4, there will be three buttons

visible. You can click the 'New Data Variable' to insert a new variable named 'soil_type' in place of the default name 'C'. Keep the 'Measure type' as 'Nominal', but change the 'Data type' to 'text'. When you are done, click the > character to the right so that the variable is fixed (Figure 8.5).

Figure 8.5.: The Jamovi toolbar is shown with the input for creating a new data variable. The new variable added is to indicate the soil type (topsoil or subsoil), so it needs to be a nominal variable with a data type of text.

After typing in the new variable 'soil_type', add another variable called 'organic_carbon'. The organic_carbon variable should have a measure type of 'Continuous' and a data type of 'Decimal'. After both soil_type and organic_carbon variables have been set, you can click the up arrow with the upper right circle (Figure 8.5) to get the new variable window out of the way.

With the two new variables created, we can now rearrange the data in a tidy way. The first 19 rows of soil_type should be 'topsoil', and the remaining 15 rows should be 'subsoil'. To do this quickly, you can write 'topsoil' in the first row of soil_type and copy-paste into the remaining rows. You can do the same to write 'subsoil' in the remaining rows 20-34. Next, copy all of the topsoil values in column 1 into the first 19 rows of column 4, and copy all of the subsoil values in column 2 into the next 15 rows. After doing all of this, your column 3 (soil_type) should have the word 'topsoil' in rows 1-19 and 'subsoil' in rows 20-34. The values from columns 1 and 2 should now fill rows 1-34 of column 4. You can now delete the first column of data by right clicking on the column name 'topsoil' and selecting 'Delete Variable'. Do the same for the second column 'subsoil'. Now you should have a tidy data set with two columns of data, one called 'soil_type' and one called 'organic_carbon'. You are now ready to calculate some descriptive statistics from the data.

First, we can calculate the minimum, maximum, and mean of all of the organic carbon values (i.e., the 'grand' mean, which includes both soil types). To do this, select the 'Analyses' tab, then click on the left-most button called 'Exploration' in the toolbar (Figure 8.6).

Figure 8.6.: The Jamovi toolbar where the tab 'Analyses' can be selected at the very top. Below this tab, the button 'Exploration' can be clicked to calculate descriptive statistics.

After clicking on 'Exploration', a pull-down box will appear with an option for 'Descriptives'. Select this option, and you will see a new window with our two columns of data in the left-most box. Click once on the 'organic_carbon' variable and use the right arrow to move it into the 'Variables' box. In the right-most panel of Jamovi, a table called 'Descriptives' will appear, which will include values for the organic carbon mean, minimum, and maximum. Write these values on the lines below, and remember to include units.

Grand Mean: _____

Grand Minimum: _____

Grand Maximum: _____

These values might be useful, but recall that there are two different soil types that need to be considered, topsoil and subsoil. The mean, minimum, and maximum above pools both of these soil types together, but we might instead want to know the mean, minimum, and maximum values for topsoil and subsoil separately. Splitting organic carbon by soil types is straightforward in Jamovi. To do it, go back to the Exploration → Descriptives option and again put 'organic_carbon' in the Variables box. This time, however, notice the 'Split by' box below the Variables box. Now, click on 'soil_type' in the descriptives and click on the lower right arrow to move soil type into the 'Split by' box. The table of descriptives in the right window will now break down all of the summary statistics by soil type. First, write the mean, minimum, and maximum topsoil values below.

Topsoil Mean: _____

Topsoil Minimum: _____

Topsoil Maximum: _____

Next, do the same for the mean, minimum, and maximum subsoil values.

Subsoil Mean: _____

Subsoil Minimum: _____

Subsoil Maximum: _____

From the values above, the mean of organic carbon sampled from the topsoil appears to be greater than the mean of organic carbon sampled from the subsoil. Assuming that Jamovi has calculated the means correctly, we can be confident that the topsoil *sample* mean is higher. But what about the *population* means? Think back to concepts of populations versus samples from Chapter 4. Based on these samples in the dataset, can we really say for certain that the population mean of topsoil is higher than the population mean of subsoil? Think about this, then write a sentence below about how confident we can be about concluding that topsoil organic carbon is greater than subsoil organic carbon.

What would make you more (or less) confident that topsoil and subsoil population means are different? Think about this, then write another sentence below that answers the question.

Note that there is no right or wrong answer for the above two questions. The entire point of the questions is to help you reflect on your own learning and better link the concepts of populations and samples to the real dataset in this practical. Doing this will make the statistical hypothesis testing that comes later in the module more clear.

8.2. Exercise on transforming variables

In this next exercise, we will work with the fig fruits dataset (right click and "Save Link As…"). Open this dataset into Jamovi. Note that there are 5 columns of data, and all of the data appear to be in a tidy format. Each row represents a separate fig fruit, while each column represents a measured variable associated with the fruit. The first several rows should look like the below.

```
##   Tree Length_cm Width_cm Height_cm Seeds
## 1    A       1.5      1.8       1.4   238
## 2    A       1.7      1.9       1.5   198
## 3    A       2.1      2.1       1.6   220
## 4    A       1.5      1.6       1.4   188
## 5    A       1.6      1.6       1.5   139
## 6    A       1.5      1.4       1.5   173
```

The dataset includes the tree from which the fig was sampled in column 1 (A, B, C, and D), then the length, width, and heights of the fig in cm. Finally, the last column shows how many seeds were counted within the fig. Use the Descriptives option in Jamovi to find the grand (i.e., not split by Tree) mean length, width, and height of figs in the dataset. Write these means down below (remember the units).

Grand Mean length: _____

Grand Mean height: _____

Grand Mean width: _____

Now look at the different rows in the Descriptives table of Jamovi. Note that there is a row for 'Missing', and there appears to be one missing value for fig width and fig height. This is very common in real datasets. Sometimes practical limitations in the field prevent data from being collected, or something happens that causes data to be lost. We therefore need to be able to work with datasets that have missing data. For now, we will just note the missing data and find them in the actual data set. Go back to the 'Data' tab in Jamovi and find the figs with a missing width and height value. Report the rows of these missing values below.

Missing width row: _____

Missing height row: _____

Next, we will go back to working with the actual data. Note that the length, width, and height variables are all recorded in cm to a single decimal place. Suppose we want to transform these variables so that they are represented in mm instead of cm. We will start by creating a new column 'Length_mm' by transforming the existing 'Length_cm' column. To do this, click on the 'Data' tab at the top of the toolbar again, then click on the 'Length_cm' column name to highlight the entire column. Your screen should look like the image in Figure 8.7.

With the 'Length_cm' column highlighted, click on the 'Transform' button in the toolbar. Two things happen next. First, a new column appears in the dataset that looks identical to 'Length_cm'; ignore this for now. Second, a box appears below the toolbar allowing us to type in a new name for the transformed variable. We can call this variable 'Length_mm'. Below, note the first pulldown menu 'Source variable'. The source value should be 'Length_cm', so we can leave this alone. The second pulldown menu 'using

Figure 8.7.: The Jamovi toolbar where the tab 'Data' is selected. The length (cm) column is highlighted and will be transformed by clicking on the Transform button in the toolbar above

transform' will need to change. To change the transform from 'None', click the arrow and select 'Create New Transform' from the pulldown. A new box will pop up allowing us to name the transformation. It does not matter what we call it (e.g., 'cm_to_mm' is fine). Note that there are 10 mm in 1 cm, so to convert from cm to mm, we need to multiply the values of 'Length_cm' by 10. We can do this by appending a `* 10` to the lower box of the transform window, so that it reads `= $source * 10` (Figure 8.8).

Figure 8.8.: The Jamovi toolbar where the tab 'Data' is selected. The box below shows the transform, which has been named 'cm to mm'. The transformation occurs by multiplying the source (Length mm) by 10. The dataset underneath shows the first few rows with the transformed column highlighted (note that the new 'Length mm' column is 10 times the length column.

When we are finished, we can click the down arrow inside the circle in the upper right to get rid of the transform window, then the up arrow inside the circle in the up-

per right to get rid of the transformed variable window. Now we have a new column called 'Length_mm', in which values are 10 times greater than they are in the adjacent 'Length_cm' column, and therefore represent fig length in mm. If we want to, we can always change the transformation by double-clicking the 'Length_mm' column. For now, apply the same transformation to fig width and height, so we have three new columns of length, width, and height all measured in mm (note, if you want to, you can use the saved transformation 'cm_to_mm' that you used to transform length, saving some time). At the end of this, you should have eight columns of data, including three new columns that you just created by transforming the existing columns of Length_cm, Width_cm, and Height_cm into the new columns Length_mm, Width_mm, and Height_mm. Find the means of these three new columns and write them below.

Grand Mean length (mm): _____

Grand Mean height (mm): _____

Grand Mean width (mm): _____

Compare these means to the means calculated above in cm. Do the differences between means in cm and the means in mm make sense?

8.3. Exercise on computing variables

In this last exercise, we will compute a new variable 'fig_volume'. Because of the way that the dimensions of the fig were measured in the field, we need to make some simplifying assumptions when calculating volume. We will assume that fig fruits are perfect spheres, and that the radius of each fig is half of its measured width (i.e., 'Width_mm / 2'). This is obviously not ideal, but sometimes practical limitations in the field make it necessary to make these kinds of simplifying assumptions. In this case, how might assuming that figs are perfectly spherical affect the accuracy of our estimated fig volume? Write a sentence of reflection on this question below, drawing from what you have learned this week about accuracy and precision of measurements.

Now we are ready to make our calculation of fig volume. The formula for the volume of a sphere (V) given its radius r is,

$$V = \frac{4}{3}\pi r^3.$$

In words, sphere volume equals four thirds times π, times r cubed (i.e., r to the third power). If this equation is confusing, remember that π is approximately 3.14, and taking r to the third power means that we are multiplying r by itself 3 times. We could therefore rewrite the equation above,

$$V = \frac{4}{3} \times 3.14 \times r \times r \times r.$$

This is the formula that we can use to create our new column of data for fig volume. To do this, double-click on the first empty column of the dataset, just to the right of the 'Seeds' column header. You will see a pull down option in Jamovi with 3 options, one of which is 'NEW COMPUTED VARIABLE'. This is the option that we want. We need to name this new variable, so we can call it 'fig_volume'. Next, we need to type in the formula for calculating volume. First, in the small box next to the f_x, type in the (4/3) multiplied by 3.14 as below.

```
= (4/3) * 3.14 *
```

Next, we need to multiply by the variable 'Width_mm' divided by 2 (to get the radius), three times. We can do this by clicking on the f_x box to the left. Two new boxes will appear; the first is named 'Functions', and the second is named 'Variables'. Ignore the functions box for now, and find 'Width_mm' in the list of variables. Double click on this to put it into the formula, then divide it by 2. You can repeat this two more times to complete the computed variable as shown in Figure 8.9.

Figure 8.9.: The Jamovi toolbar where the tab 'Data' is selected. The box below shows the new computed variable 'fig volume', which has been created by calculating the product of 4/3, 3.14, and Width mm three times.

Note that we can get the cube of 'Width_mm' more concisely by using the carrot character (^). That is, we would get the same answer shown in Figure 8.9 if we instead typed the below in the function box.

```
= (4/3) * 3.14 * (Width_mm/2)^3
```

Note that the order of operations is important here, which is why there are parentheses around `Width_mm/2`. This calculation needs to be done before taking the value to the power of 3. If we instead had written, `Width_mm/2^3`, then Jamovi would first take the cube of 2 ($2 \times 2 \times 2 = 8$), then divided `Width_mm` by this value giving a different and incorrect answer. When in doubt, it is always useful to use parentheses to specify what calculations should be done first.

You now have the new column of data 'fig_volume'. Remember that the calculations underlying apply to the units too. The width of the fig was calculated in mm, but we have taken width to the power of 3 when calculating the volume. In the spaces below, find the mean, minimum, and maximum volumes of all figs and report them in the correct units.

Mean: _____

Minimum: _____

Maximum: _____

Finally, it would be good to plot these newly calculated fig volume data. These data are continuous, so we can use a histogram to visualise the fig volume distribution. To make a histogram, go to the Exploration → Descriptives window in Jamovi (the same place where you found the mean, minimum, and maximum). Now, look on the lower left-hand side of the window and find the pulldown menu for 'Plots'. Click 'Plots', and you should see several different plotting options. Check the option for 'Histogram' and see the new histogram plotted in the window to the right. Draw a rough sketch of the histogram in the area below.

Finally, we should save the file that we have been working on. There are two ways to save a file in Jamovi, and it is a good idea to save both ways. The first way is to use Jamovi's own (binary) file type, which has the extension OMV. This will not only save the data (including the calculated variables created within Jamovi), but also any

analyses that we have done (e.g., calculation of minimums, maximums, and means) or graphs that we have made (e.g., the histogram). To do this, click on the three horizontal lines in the upper left of the Jamovi toolbar, then select 'Save As'. Choose an appropriate name (e.g., 'SCIU4T4_Week2_practical.omv'), then save the file in a location where you know that you will be able to find it again. Like, all binary files, an OMV file cannot be opened as plain text. Hence, it might be a good idea to save the dataset as a CSV file (note, this will not save any of the analyses or graphs). To do this, click on the three horizontal lines in the upper left of the toolbar again, but this time click 'Export'. Give the file an appropriate name (e.g., 'SCIU4T4_Week2_data'), then choose 'CSV' from the pulldown menu below (Figure 8.10).

Figure 8.10.: The Jamovi Export menu in which data are be saved as a CSV using the pulldown menu below the filename

Make sure to choose a save location that you know you will be able to find again (to navigate through file directories, click 'Browse' in the upper right). To save, click on 'Export' in the upper right (Figure 8.10).

8.4. Summary

You should now know some of the basic tools for working with data, calculating some simple descriptive statistics, plotting a histogram, and saving output and data in Jamovi. These skills will be used throughout the module, so it is important to be comfortable with them as the analyses become more complex. If you still have time at the end of the lab practical, it might be a good idea to explore other features in Jamovi.

Part III.

Summary statistics

Week 3 Overview

Dates	6 February 2023 - 10 February 2023
Reading	**Required:** SCIU4T4 Workbook chapters 9-12 **Recommended:** Navarro and Foxcroft (2022) Chapter 5 and Chapter 4.1 **Suggested:** Rowntree (2018) Chapter 3 **Advanced:** None
Lectures	3.0: Decimal places and significant figures part 1 (7:52 min; Video) 3.1: Decimal places and significant figures part 2 (7:08 min; Video) 3.2: Graphs (10:29 min; Video) 3.3: Box-whisker plots (8:07 min; Video) 3.4: The mean (16:52 min; Video) 3.5: The mode (6:54 min; Video) 3.6: The median and quantiles (8:04 min; Video) 3.7: Mean, mode, median, and resistance (8:35 min; Video) 3.8: The variance (9:40 min; Video) 3.9: The standard deviation (6:17 min; Video) 3.10: The standard deviation (7:46 min; Video) 3.11: The standard deviation (13:23 min; Video)
Practical	Plotting and statistical summaries (Chapter 13) Room: Cottrell 2A17 Group A: 08 FEB 2023 (WED) 13:05-15:55 Group B: 09 FEB 2023 (THU) 09:05-11:55
Help hours	Ian Jones Room: Cottrell 1A13 10 FEB 2023 (FRI) 15:05-17:55
Assessments	Week 3 Practice quiz on Canvas

Week 3 focuses on descriptive statistics, how to report them, interpret them, and communicate them with graphs.

Chapter 9 focuses on how to report numbers with accuracy and precision. In practice, this means reporting values with the correct number of digits (decimal places and significant figures), and rounding appropriately.

Chapter 10 introduces different types of graphs for communicating data visually. The chapter focuses specifically on histograms, pie charts, barplots, and box-whisker plots.

Chapter 11 introduces measures of central tendency. These are measures that describe the centre of the data using a single number. Measures of central tendency in this chapter include the mean, the mode, the median, and quantiles.

Chapter 12 introduces on measures of spread. In contrast to measures of central tendency, which focus on the centre of a dataset, measures of spread focus on how much the data are spread out. Measures of spread in this chapter in clude the range, the inter-quartile range, the variance, the standard deviation, the coefficient of variation, and the standard error.

Chapter 13 guides you through the week 3 practical. The aim of this practical is to learn how to use Jamovi to generate plots introduced in Chapter 10, and to find measures of central tendency and spread introduced in Chapter 11 and Chapter 12, respectively, and report them accurately using the knowledge from Chapter 9.

9. Decimal places, significant figures, and rounding

When making calculations, it is important that any numbers reported are communicated with accuracy and precision. This means reporting numbers with the correct number of digits. This chapter focuses on correctly interpreting the decimal places and significant figures of a number, and correctly rounding. In your assessments, you will frequently be asked to report an answer to a specific number of decimal places or significant figures, and you will be expected to round numbers correctly.

9.1. Decimal places and significant figures

A higher number of digits communicates a greater level of accuracy. For example, the number 2.718 expresses a higher precision than 2.7 does. Reporting 2.718 implies that we know the value is somewhere between 2.7175 and 2.1785, but reporting 2.7 only implies that we know the value is somewhere between 2.65 and 2.75 (Sokal and Rohlf, 1995). These numbers therefore have a different number of *decimal places* and a different number of *significant figures*. Decimal places and significant figures are related, but not the same.

Decimal places are conceptually easier to understand. These are just the number of digits to the right of the decimal point. For example, 2.718 has 3 decimal places, and 2.7 has 1 decimal place.

Significant figures are a bit more challenging. These are the number of digits that you need to infer the accuracy of a value. For example, the number 2.718 has 4 significant figures and 2.7 has 2 significant figures. This sounds straightforward, but it can get confusing when numbers start or end with zeros. For example, the number 0.045 has only 2 significant figures because the first two zeros only serve as placeholders (note that if this were a measurement of 0.045 m, then we could express the exact same value as 45 mm, so the zeros are not really necessary to indicate measurement accuracy). In contrast, the measurement 0.045000 has 5 significant figures because the last 3 zeros indicate a higher degree of accuracy than just 0.045 would (i.e., we know the value is somewhere between 0.44995 and 0.45005, not just 0.0445 and 0.0455). Lastly, the measurement 4500 has only 2 significant figures because the last 2 zeros are only serving as a placeholder to indicate magnitude, not accuracy (if we wanted to represent 4500 with 4 significant figures, we could use scientific notation and express it as 4.500×10^3).

Here is a table with some examples of numbers, their decimal places, and their significant figures.

Table 9.1.: Numbers are presented in rows of the first column. Decimal places and significant figures for each row number are presented in the second and third column, respectively.

Number	Decimal places	Significant figures
3.14159	5	6
0.0333	4	3
1250	0	3
50000.0	1	6
0.12	2	2
1000000	0	1

It is a good idea to double-check that the values in these tables make sense. For assessments, make sure that you are confident that you can report your answer to a given number of decimal places or significant figures.

9.2. Rounding

Often if you are asked to report a number to a specific number of decimals or significant figures, you will need to round the number. Rounding reduces the number of significant digits in a number, which might be necessary if a number that we calculate has more significant digits than we are justified in expressing. There are different rules for rounding numbers, but in this module, we will follow Sokal and Rohlf (1995). When rounding to the nearest decimal, the last decimal written should not be changed if the number that immediately follows is 0, 1, 2, 3, or 4. If the number that immediately follows is 5, 6, 7, 8, or 9, then the last decimal written should be increased by 1.

For example, if we wanted to round the number 3.141593 to 2 significant digits, then we would write it as 3.1 because the digit that immediately follows (i.e., the third digit) is 4. If we wanted to round the number to 5 significant digits, then we would write it as 3.1416 because the digit that immediately follows is 9. And if we wanted to round 3.141593 to 4 significant digits, then we would write it as 3.142 because the digit that immediately follows is 5. Note that this does not just apply for decimals. If we wanted to round 1253 to 3 significant figures, then we would round by writing it as 1250.

Here is a table with some examples of numbers rounded to a given significant figure.

Table 9.2.: Numbers to be rounded are presented in rows of the first column. The significant figures to which rounding is desired is in the second column, and the third column shows the correctly rounded number.

Original number	Significant figures	Rounded number
23.2439	4	23.24
10.235	4	10.24
102.39	2	100
5.3955	3	5.40
37.449	3	37.4
0.00345	2	0.0035

In this module, it will be necessary to round calculated values to a specified decimal or significant figure. It is therefore important to understand the rules for rounding and why the values in the table above are rounded correctly.

10. Graphs

Graphs are useful tools for visualising and communicating data. Graphs come in many different types, and different types of graphs are effective for different types of data. This chapter focuses on 4 types of graphs: (1) histograms, (2) pie charts, (3) barplots, and (4) box-whisker plots.

After collecting or obtaining a new dataset, it is almost always a good idea to plot the data in some way. Visualisation can often highlight important and obvious properties of a dataset more efficiently that inspecting raw data, calculating summary statistics, or running statistical tests. When making graphs to communicate data visually, it is important to ensure that the person reading the graph has a clear understanding of what is being presented. In practice, this means clearly labelling axes with meaningful descriptions and appropriate units, including a descriptive caption, and indicating what any graph symbols mean. In general, it is also best to make the simplest graph possible for visualising the data, which means avoiding unnecessary colour, three-dimensional display, or unnecessary distractions from the information being conveyed (Dytham, 2011; Kelleher and Wagener, 2011). It is also important to ensure that graphs are as accessible as possible, e.g., by providing strong colour contrast and appropriate colour combinations (Elavsky et al., 2022), and alternative text for images where possible. As a guide, the histogram, pie chart, barplot, and box-whisker plot below illustrate good practice when making graphs.

10.1. Histograms

Histograms illustrate the distribution of continuous data. They are especially useful visualisation tools because it is often important to assess data at a glance and make a decision about how to proceed with a statistical analysis. The histogram shown in Figure 10.1 provides an example using the fig fruits data set from the practical in Chapter 8 (for a step-by-step demonstration of how a histogram is built, see this interactive application[1]).

The histogram in Figure 10.1 shows how many fruits there are for different intervals of width, i.e., the frequency with which fruits within some width interval occur in the data. For example, there are 6 fruits with a width between 1.0 and 1.2, so for this interval on the x-axis, the bar is 6 units in height on the y-axis. In contrast, there is only 1 fig fruit

[1] Here is the full URL: https://bradduthie.shinyapps.io/build_histogram/

Figure 10.1.: Example histogram fig fruit width (cm) using data from 78 fig fruits collected in 2010 from Baja, Mexico.

that has a width greater than 2.0 cm (the biggest is 2.1 cm), so we see that the height of the bar for the interval between 2.0 and 2.2 is only 1 unit in frequency. The bars of the histogram touch each other, which reinforces the idea that the data are continuous (Dytham, 2011; Sokal and Rohlf, 1995).

Click here for an interactive application showing how histograms are built.

It is especially important to be able to read and understand information from a histogram because it is often necessary to determine if the data are consistent with the assumptions of a statistical test. For example, the *shape* of the distribution of fig fruit widths might be important for performing a particular test. For the purposes of this chapter, the *shape* of the distribution just means what the data look like when plotted like this in a histogram. In this case, there is a peak toward the centre of the distribution, with fewer low and high values (this kind of distribution is quite common). Different distribution shapes will be discussed more in Week 4.

10.2. Barplots and pie charts

While histograms are an effective way of visualising continuous data, barplots (also known as 'bar charts' or 'bar graphs') and pie charts can be used to visualise categorical data. For example, in the fig fruits data set from Chapter 8, 78 fig fruits were collected

from 4 different trees (A, B, C, and D). A barplot could be used to show how many samples were collected from each tree (see Figure 10.2).

Figure 10.2.: Example bar plot showing how many fruits were collected from each of 4 trees (78 collected in total) in 2010 from Baja, Mexico.

In Figure 10.2, each tree is represented by a separate bar on the x-axis. Unlike a histogram, the bars do not touch each other, which reinforces the idea that different categories of data are being shown (in this case, different trees). The height of a bar indicates how many fruits were sampled for each tree. For example, 14 fruits were sampled from tree A, and 22 fruits were sampled from tree B. At a glance, it is therefore possible to compare different trees and make inferences about how they differ in sampled fruits.

Pie charts are similar to barplots in that both present categorical data, but pie charts are more effective for visualising the relative quantity for each category. That is, pie charts illustrate the percentage of measurements for each category. For example, in the case of the fig fruits, it might be useful to visualise what percentage of fruits were sampled from each tree. A pie chart could be used to evaluate this, with pie slices corresponding to different trees and the size of each slice reflecting the percentage of the total sampled fruits that came from each tree (Figure 10.3).

Pie charts can be useful in some situations, but in the biological and environmental sciences, they are not used as often as barplots. In contrast to pie charts, barplots present the absolute quantities (in Figure 10.2, e.g., the actual number of fruits sampled per tree), and it is still possible with barplots to infer the percentage each category contributes to

Figure 10.3.: Example pie plot showing the percentage of fruits that were collected from each of 4 trees (78 collected in total) in 2010 from Baja, Mexico.

the total from the relative sizes of the bars. Pie charts, in contrast, only illustrate relative percentages unless numbers are used to indicate absolute quantities. Unless percentage alone is important, barplots are often the preferred way to communicate count data.

10.3. Box-whisker plots

Box-whisker plots (also called boxplots) can be used to visualise distributions in a different way than histograms. Instead of presenting the full distribution, as in a histogram, a box-whisker plot shows where summary statistics are located (summary statistics are explained in Chapter 11 and Chapter 12). This allows the distribution of data to be represented in a more compact way, but does not show the full shape of a distribution. Figure 10.4 compares a box-whisker plot of fig fruit widths (10.4a) with a histogram of fig fruit widths (10.4b). In other words, both of the panels ('a' and 'b') in Figure 10.4 show the same information in two different ways (note that these are the same data as presented in Figure 10.1).

To show how the panels of Figure 10.4 correspond to one another more clearly, Figure 10.5 shows them again, but with points indicating where the summary statistics shown in the boxplot (Figure 10.5a) are located in the histogram (Figure 10.5b). These summary statistics include the median (black circles of Figure 10.5), quartiles (red squares of Figure 10.5), and the limits of the distribution (i.e., the minimum and maximum values; blue triangles of Figure 10.5). Note that in boxplots, if outliers exist, they are presented as separate points.

One benefit of a boxplot is that it is possible to show the distribution of multiple variables simultaneously. For example, the distribution of fig fruit width can be shown for each of the four trees side by side on the same x-axis of a boxplot (Figure 10.6). While it is possible to show histograms side by side, it will quickly take up a lot of space.

Figure 10.4.: Boxplot (a) of fig fruit widths (cm) for 78 fig fruits collected in 2010 in Baja, Mexico. Panel (b) presents the same data as a histogram.

The boxplot in Figure 10.6 can be used to quickly compare the distribution of Trees A-D. The point at the bottom of the distribution of Tree A shows an outlier. This outlier is an especially low value of fig fruit width compared to the other fruits of Tree A.

Figure 10.5.: Boxplot (a) of fig fruit widths (cm) for 78 fig fruits collected in 2010 in Baja, Mexico. Panel (b) presents the same data as a histogram. Points in the boxplot indicate the median (black circle), first and third quartiles (red squares), and the limits of the distribution (blue triangles). Corresponding locations are shown on the histogram in panel (b).

Figure 10.6.: Boxplot of fig fruit widths (cm) collected from 4 separate trees sampled in 2010 from Baja, Mexico.

11. Measures of central tendency

Summary statistics describe properties of data in a single number (e.g., the mean), or a set of numbers (e.g., quartiles). This chapter focuses on summary statistics that describe the centre of a distribution. It also introduces quantiles, which divide a distribution into different percentages of the data (e.g., the lowest 50% or highest 75%). Throughout this section, verbal and mathematical explanations of summary statistics will be presented alongside histograms or boxplots that convey the same information. The point of doing this is to help connect the two ways of summarising the data. All of the summary statistics that follow describe calculations for a *sample* and are therefore estimates of the true values in a *population*. Recall from Chapter 4 the difference between a population and a sample. This module focuses on statistical techniques, not statistical theory, so summary statistics will just focus on how to estimate statistics from sampled data instead of how statistics are defined mathematically[1].

11.1. The mean

The arithmetic mean (hereafter just *the mean*[2]) of a sample is one of the most commonly reported statistics when communicating information about a dataset. The mean is a measure of central tendency, so it is located somewhere in the middle of a distribution. Figure 11.1 shows the same histogram of fig fruit widths shown in Figure 10.1, but with an arrow indicating where the mean of the distribution is located

The mean is calculated by adding up the values of all of the data and dividing this sum by the total number of data (Sokal and Rohlf, 1995). This is a fairly straightforward calculation, so we can use the mean as an example to demonstrate some new mathematical notation that will be used throughout the module. We will start with a concrete example with actual numbers, then end with a more abstract equation describing how any sample mean is calculated. The notation might be a bit confusing at first, but learning it will make understanding statistical concepts easier later in the module. There are a lot of equations in what follows, but this is because we want to explain what is happening as clearly as possible, step by step. We start with the following 8 values.

[1] If interested, a good textbook for learning about theoretical statistics and the mathematics underlying what we do in this module is Miller and Miller (2004).

[2] There are other types of means, such as the geometric mean or the harmonic mean, but we will not use these at all in this module.

Figure 11.1.: Example histogram fig fruit width (cm) using data from 78 fig fruits collected in 2010 from Baja, Mexico. The mean of the distribution is indicated with an arrow.

```
4.2, 5.0, 3.1, 4.2, 3.8, 4.6, 4.0, 3.5
```

To calculate the mean of a sample, we just need to add up all of the values and divide by 8 (the total number of values),

$$\bar{x} = \frac{4.2 + 5.0 + 3.1 + 4.2 + 3.8 + 4.6 + 4.0 + 3.5}{8}.$$

Note that we have used the symbol $\bar{x}$ to represent the mean of x, which is a common notation (Sokal and Rohlf, 1995). In the example above, $\bar{x} = 4.05$.

Writing the full calculation above is not a problem because we only have 8 points of data. But sample sizes are often much larger than 8. If we had a sample size of 80 or 800, then there is no way that we could write down every number to show how the mean is calculated. One way to get around this is to use ellipses and just show the first and last couple of numbers,

$$\bar{x} = \frac{4.2 + 5.0 + ... + 4.0 + 3.5}{8}.$$

This is a more compact, and perfectly acceptable, way to write the sample mean. But it is often necessary to have an even more compact way of indicating the sum over a set

of values (i.e., the top of the fraction above). To do this, each value can be symbolised by an x, with a unique subscript i, so that x_i corresponds to a specific value in the list above. The usefulness of this notation, x_i, will become clear soon. It takes some getting used to, but the table below shows each symbol with its corresponding value to make it more intuitive.

Table 11.1.: A sample dataset that includes eight values.

Symbol	Value
x_1	4.2
x_2	5.0
x_3	3.1
x_4	4.2
x_5	3.8
x_6	4.6
x_7	4.0
x_8	3.5

Note that we can first replace the actual values with their corresponding x_i, so the mean can be written as,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8}{8}.$$

Next, we can rewrite the top of the equation in a different form using a summation sign,

$$\sum_{i=1}^{8} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8.$$

Like the use of x_i, the summation sign $\sum$ takes some getting used to, but here it just means "sum up all of the x_i values". You can think of it as a big 'S' that just says "sum up". The bottom of the S tells you the starting point, and the top of it tells you the ending point, for adding numbers. Verbally, we can read this as saying, "starting with $i = 1$, add up all of the x_i values until $i = 8$". We can then replace the long list of x values with a summation,

$$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{8}.$$

This looks a bit messy, so we can rewrite the above equation. Instead of dividing the summation by 8, we can multiply it by 1/8, which gives us the same answer,

$$\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i.$$

There is one more step. We have started with 8 actual values and ended with a compact and abstract equation for calculating the mean. But if we want a general description for calculating *any* mean, then we need to account for sample sizes not equal to 8. To do this, we can use N to represent the sample size. In our example, $N = 8$, but it is possible to have a sample size be any finite value above zero. We can therefore replace 8 with N in the equation for the sample mean,

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

There we have it. Verbally, the above equation tells us to multiply $1/N$ by the sum of all x_i values from 1 to N. This describes the mean for any sample that we might collect.

11.2. The mode

The mode of a dataset is simply the value that appears most often. As a simple example, we can again consider the sample dataset of 8 values.

```
4.2, 5.0, 3.1, 4.2, 3.8, 4.6, 4.0, 3.5
```

In this dataset, the values 5.0, 3.1, 3.8, 4.6, 4.0, and 3.5 are all represented once. But the value 4.2 appears twice, once in the first position and once in the fourth position. Because 4.2 appears most frequently in the dataset, it is the mode of the dataset.

Note that it is possible for a dataset to have more than one mode. Also, somewhat confusingly, distributions that have more than one peak are often described as multimodal, even if the peaks are not of the same height (Sokal and Rohlf, 1995). For example, the histogram in Figure 11.2 might be described as bimodal because it has two distinct peaks (one around 10 and the other around 14), even though these peaks are not the same size.

In very rare cases, data might have a U-shape. The lowest point of the U would then be described as the antimode (Sokal and Rohlf, 1995).

Figure 11.2.: Example histogram of a hypothetical dataset that has a bimodal distribution.

11.3. The median and quantiles

The median of a dataset is the middle value when the data are sorted. More technically, the median is defined as the value that has the same number of lower and higher values than it (Sokal and Rohlf, 1995). If there are an odd number of values in the dataset, then finding the median is often easy. For example, the median of the values {8, 5, 3, 2, 6} is 5. This is because if we sort the values from lowest to highest (2, 3, 5, 6, 8), the value 5 is exactly in the middle. It gets more complicated for an even number of values, such as the sample dataset used for explaining the mean and mode.

```
4.2, 5.0, 3.1, 4.2, 3.8, 4.6, 4.0, 3.5
```

We can order these values from lowest to highest.

```
3.1, 3.5, 3.8, 4.0, 4.2, 4.2, 4.6, 5.0
```

Again, there is no middle value here. But we can find a value that has the same number of lower and higher values. To do this, we just need to find the mean of the middle 2 numbers, in this case 4.0 and 4.2, which are in positions 4 and 5, respectively. The mean of 4.0 and 4.2 is, $(4.0 + 4.2)/2 = 4.1$, so 4.1 is the median value.

The median is a type of quantile. A quantile divides a sorted dataset into different percentages that are lower or higher than it. Hence, the median could also be called the 50% quantile because 50% of values are lower than the median and 50% of values are higher than it. Two other quantiles besides the median are also noteworthy. The first quartile (also called the "lower quartile") defines the value for which 25% of values are lower and 75% of values are higher. The third quartile (also called the "upper quartile") defines the value for which 75% of values are lower and 25% of values are higher. Sometimes this is easy to calculate. For example, if there are only five values in a dataset, then the lower quartile is the number in the second position when the data are sorted because 1 value (25%) is below it and 3 values (75%) are above it. For example, for the values {1, 3, 4, 8, 9}, the value 3 is the first quartile and 8 is the third quartile.

In some cases, it is not always this clear. We can show how quantiles get more complicated using the same 8 values as above where the first quartile is somewhere between 3.5 and 3.8.

```
3.1, 3.5, 3.8, 4.0, 4.2, 4.2, 4.6, 5.0
```

There are at least 9 different ways to calculate the first quartile in this case, and different statistical software package will sometimes use different default methods (Hyndman and Fan, 1996). One logical way is to calculate the mean between the second (3.5) and third (3.8) position as you would do for the median (Rowntree, 2018), $(3.5 + 3.8)/2 = 3.65$. Jamovi uses a slightly more complex method, which will give a value of 3.725 (The Jamovi Project, 2022).

It is important to emphasise that no one way of calculating quantiles is the one and only correct way. Statisticians have just proposed different approaches to calculating quantiles from data, and these different approaches sometimes give slightly different results. This can be unsatisfying when first learning statistics because it would be nice to have a single approach that is demonstrably correct, i.e., the *right* answer under all circumstances. Unfortunately, this is not the case here, nor is it the case for a lot of statistical techniques. Often there are different approaches to answering the same statistical question and no simple right answer. For this module, we will almost always be reporting calculations of quantiles from Jamovi, and we will clearly indicate that this is how they should be calculated for assessment questions. But it is important to recognise that different statistical tools might give different answers (Hyndman and Fan, 1996).

12. Measures of spread

It is often important to know how much a set of numbers is spread out. That is, do all of the data cluster close to the mean, or are most values distant from the mean. For example, all of the numbers below are quite close to the mean of 5.0 (3 numbers are exactly 5.0).

```
4.9, 5.3, 5.0, 4.7, 5.1, 5.0, 5.0
```

In contrast, all of the numbers that follow are relatively distant from the same mean of 5.0.

```
3.0, 5.6, 7.8, 1.2, 4.3, 8.2, 4.9
```

This chapter focuses on summary statistics that describe the spread of data. The approach in this chapter is similar to Chapter 11, which provided verbal and mathematical explanations of measures of central tendency. We will start with the most intuitive measures of spread, the range and inter-quartile range. Then, we will move on to some more conceptually challenging measures of spread, the variance, standard deviation, coefficient of variation, and standard error. These more challenging measures can be a bit confusing at first, but they are absolutely critical for doing statistics. The best approach to learning them is to see them and practice using them in different contexts, which we will do here, in the Chapter 13 practical, and throughout the semester.

12.1. The range

The range of a set of numbers is probably the most intuitive measure of spread. It is simply the difference between the highest and the lowest value of a dataset (Sokal and Rohlf, 1995). To calculate it, we just need to take the highest value minus the lowest value. If we want to be fancy, then we can write a general equation for the range of a random variable X,

$$Range(X) = \max(X) - \min(X).$$

But really, all that we need to worry about is finding the highest and lowest values, then subtracting. Consider again the two sets of numbers introduced at the beginning of the chapter. In examples, it is often helpful to imagine numbers as representing something concrete that has been measured, so suppose that these numbers are the measured masses (in grams) of leaves from two different plants. Below are the masses of plant A, in which leaf masses are very similar and close to the mean of 5.

```
4.9, 5.3, 5.0, 4.7, 5.1, 5.0, 5.0
```

Plant B masses are below, which are more spread out around the same mean of 5.

```
3.0, 5.6, 7.8, 1.2, 4.3, 8.2, 4.9
```

To get the range of plant A, we just need to find the highest (5.3 g) and lowest (4.7 g) mass, then subtract,

$$Range(Plant\ A) = 5.3 - 4.7 = 0.6$$

Plant A therefore has a range of 0.6 g. We can do the same for plant B, which has a highest value of 8.2 g and lowest value of 1.2 g,

$$Range(Plant\ B) = 8.2 - 1.2 = 7.0$$

Plant B therefore has a much higher range than plant A.

It is important to mention that the range is highly sensitive to outliers (Navarro and Foxcroft, 2022). Just adding a single number to either plant A or plant B could dramatically change the range. For example, imagine if we measured a leaf in plant A to have a mass of 19.7 g (i.e., we found a huge leaf!). The range of plant A would then be $19.7 - 4.7 = 14$ instead of 0.6! Just this one massive leaf would then make the range of plant A double the range of plant B. This lack of robustness can really limit how useful the range is as a statistical measure of spread.

12.2. The inter-quartile range

The inter-quartile range (usually abbreviated as 'IQR') is conceptually the same as the range. The only difference is that we are calculating the range between quartiles rather than the range between the highest and lowest numbers in the dataset. A general formula subtracting the first quartile (Q_1) from the third quartile (Q_3) is,

$$IQR = Q_3 - Q_1.$$

Recall from Chapter 11 how to calculate first and third quartiles. As a reminder, we can sort the leaf masses for plant A below.

```
4.7, 4.9, 5.0, 5.0, 5.0, 5.1, 5.3
```

The first quartile will be the mean between 4.9 and 5.0 (4.95). The second quartile will be the the mean between 5.0 and 5.1 (5.05). The IQR of plant A is therefore,

$$IQR_{plant\ A} = 5.05 - 4.95 = 0.1.$$

We can calculate the IQR for plant B in the same way. Here are the masses of plant B leaves sorted.

```
1.2, 3.0, 4.3, 4.9, 5.6, 7.8, 8.2
```

The first quartile of plant B is 3.65, and the third quartile is 6.70. To get the IQR of plant B,

$$IQR_{plant\ B} = 6.70 - 3.65 = 3.05.$$

An important point about the IQR is that it is more robust than the range (Dytham, 2011). Recall that if we found an outlier leaf of 19.7 g on plant A, it would change the range of plant leaf mass from 0.6 g to 14 g. The IQR is not nearly so sensitive. If we include the outlier, the first quartile for plant A changes from $Q_1 = 4.95$ to $Q_1 = 4.975$. The second quartile changes from $Q_3 = 5.05$ to $Q_3 = 5.150$. The resulting IQR is therefore $5.150 - 4.975 = 0.175$. Hence, the IQR only changes from 0.1 to 0.175, rather than from 0.6 to 14. The one outlier therefore has a huge effect on the range, but only a modest effect on the IQR.

12.3. The variance

The range and inter-quartile range were reasonably intuitive, in the sense that it is not too difficult to think about what a range of 10, e.g., actually means in terms of the data. We now move to measures of spread that are less intuitive. These measures of spread are the variance, standard deviation, coefficient of variation, and standard error. These can be confusing and unintuitive at first, but they are extremely useful. We will start with

the variance; this section is long because we want to break the variance down carefully, step by step.

The sample variance of a dataset is a measure of the expected squared distance of data from the mean. To calculate the variance of a sample, we need to know the sample size (N, i.e., how many measurements in total), and the mean of the sample ($\bar{x}$). We can calculate the variance of a sample (s^2) as follows,

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 .$$

This looks like a lot, but we can break down what the equation is doing verbally. First, we can look inside the summation ($\sum$). Here we are taking an individual measurement x_i, subtracting the mean $\bar{x}$, then squaring. We do this for each x_i, summing up all of the values from $i = 1$ to $i = N$. This part of the equation is called the **sum of squares** (SS),

$$SS = \sum_{i=1}^{N} (x_i - \bar{x})^2 .$$

That is, we need to subtract the mean from each value x_i, square the result, and add everything up. Once we have this sum, SS, then we just need to multiply by $1/(N-1)$ to get the variance.

An example of how to do the actual calculation should help make it easier to understand what is going on. We can use the same values from plant A earlier.

```
4.9, 5.3, 5.0, 4.7, 5.1, 5.0, 5.0
```

To calculate the variance of plant A leaf masses, we start with the sum of squares. That is, take 4.9, subtract the sample mean of 5.0 ($4.9 - 5.0 = -0.1$), then square the result ($(-0.1)^2 = 0.01$). We do the same for 5.3, $(5.3 - 5.0)^2 = 0.09$, and add it to the 0.01, then continue down the list of numbers finishing with 5.0. This is what the sum of squares calculation looks like all written out,

$$SS = (4.9-5)^2 + (5.3-5)^2 + (5-5)^2 + (4.7-5)^2 + (5.1-5)^2 + (5-5)^2 + (5-5)^2.$$

Remember that the calculations in parentheses need to be done first, so the next step for calculating the sum of squares would be the following,

$$SS = (-0.1)^2 + (0.3)^2 + (0)^2 + (-0.3)^2 + (0.1)^2 + (0)^2 + (0)^2.$$

Next, we need to square all of the values,

$$SS = 0.01 + 0.09 + 0 + 0.09 + 0.01 + 0 + 0.$$

If we sum the above, we get $SS = 0.2$. We now just need to multiply this by $1/(N-1)$, where $N = 7$ because this is the total number of measurements in the plant A dataset,

$$s^2 = \frac{1}{7-1}(0.2).$$

From the above, we get a variance of approximately $s^2 = 0.0333$.

Fortunately, it will almost never be necessary to calculate a variance manually in this way. Any statistical software will do all of these steps and calculate the variance for us (Chapter 13 explains how in Jamovi). The only reason that we present the step-by-step calculation here is to help explain the equation for s^2. The details can be helpful for understanding how the variance works as a measure of spread. For example, note that what we are really doing here is getting the distance of each value from the mean, $x_i - \bar{x}$. If these distances tend to be large, then it means that most data points (x_i) are far away from the mean ($\bar{x}$), and the variance (s^2) will therefore increase. The differences $x_i - \bar{x}$ are squared because we need all of the values to be positive, so that variance increases regardless of whether a value x_i is higher or lower than the mean. It does not matter if x_i is 0.1 lower than $\bar{x}$ (i.e., $x_i - \bar{x} = -0.1$), or 0.1 higher (i.e., $x_i - \bar{x} = 0.1$). In both cases, the deviation from the mean is the same. Moreover, if we did not square the values, then the sum of $x_i - \bar{x}$ values would always be 0 (you can try this yourself)[1]. Lastly, it turns out that the variance is actually a special case of a more general concept called the *covariance*, which we will look at later in Week 9 and makes the squaring of differences make a bit more sense.

We sum up all of the squared deviations to get the SS, then divide by the sample size minus 1, to get the mean squared deviation from the mean. That is, the whole process gives us the *average* squared deviation from the mean. But wait, why is it the sample size minus 1, $N - 1$? Why would we subtract 1 here? The short answer is that in calculating a *sample* variance, s^2, we are almost always trying to estimate the corresponding *population* variance (σ^2). And if we were to just use N instead of $N - 1$, then our s^2 would be a biased estimate of σ^2 (see Chapter 4 for a reminder on the difference between samples and populations). By subtracting 1, we are correcting for this bias to get a more accurate estimate of the population variance. It is not necessary to do this ourselves; statistical software like Jamovi and R will do it automatically (The

[1] If you are wondering why we square the difference $x_i - \bar{x}$ instead of just taking its absolute value, this is an excellent question! You have just invented something called the mean absolute deviation. There are some reasons why the mean absolute deviation is not as good of a measure of spread as the variance. Navarro and Foxcroft (2022) explain the mean absolute deviation, and how it relates to the variance, very well in section 4.2.3 of their textbook. We will not get into these points here, but it would be good to check out Navarro and Foxcroft (2022) for more explanation.

Jamovi Project, 2022; R Core Team, 2022). This is really all that it is necessary to know for now, but see this footnote[2] for a bit more detailed explanation to try to make this intuitive (it is actually quite cool!). Later, we will explore the broader concept of *degrees of freedom*, which explains why we need to take into account the number of parameters in a statistic that are free to vary when calculating a statistic[3].

This was a lot of information. The variance is not an intuitive concept. In addition to being a challenge to calculate, the calculation of a variance leaves us with a value in units squared. That is, for the example of plant leaf mass in grams, the variance is measured in grams squared, g^2, which is not particularly easy to interpret. For more on this, Navarro and Foxcroft (2022) have a really good section on the variance. Despite its challenges as a descriptive statistic, the variance has some mathematical properties that are very useful (Navarro and Foxcroft, 2022), especially in the biological and environmental sciences.

For example, variances are additive, meaning that if we are measuring two separate characteristics of a sample, A and B, then the variance of A+B equals the variance of A plus the variance of B; i.e., $Var(A + B) = Var(A) + Var(B)$ [4]. This is relevant to genetics when measuring heritability. Here, the total variance in the phenotype of a population (e.g., body mass of animals) can be partitioned into variance attributable to genetics plus variance attributable to the environment,

$$Var(Phenotype) = Var(Genotype) + Var(Environment).$$

This is also sometimes written as $V_P = V_G + V_E$. Applying this equation to calculate

[2]To get the true population variance σ^2, we would also need to know the true mean μ. But we can only estimate μ from the sample, $\bar{x}$. That is, what we would really want to calculate is $x_i - \mu$, but the best we can do is $x_i - \bar{x}$. The consequence of this is that there will be some error that underestimates the true distance of x_i values from the population mean, μ. Here is the really cool part; to determine the extent to which our estimate of the variance is biased by using $\bar{x}$ instead of μ, we just need to know the expected squared difference between the two values, $(\bar{x} - \mu)^2$. It turns out that this difference (i.e., the bias of our estimate s^2) is just σ^2/N; that is, the true variance of the population divided by the sample size. If we subtract this value from σ^2, so $\sigma^2 - \sigma^2/N$, then we can get the expected difference between the true variance and the estimate from the sample size. We can rearrange $\sigma^2 - \sigma^2/N$ to get $\sigma^2 \times (N-1)/N$, which means that we need to correct our sample variance by $N/(N-1)$ to get an unbiased estimate of σ^2. If all of this is confusing, that is okay! This is really only relevant for those interested in statistical theory, which is not the focus of this module.

[3]Briefly, in the case of sample variance, note that we needed to use all the values x_i in the dataset and the sample mean $\bar{x}$. But if we know what all of the x_i values are, then we also know $\bar{x}$. And if we know all but one value of x_i and $\bar{x}$, then we could figure out the last x_i. Hence, while we are using N values in the calculation of s^2, the use of $\bar{x}$ reduces the degree to which these values are free to vary. We have lost 1 degree of freedom in the calculation of $\bar{x}$, so we need to account for this in our calculation of s^2 by dividing by $N - 1$. This is another way to think about the $N - 1$ correction factor (Sokal and Rohlf, 1995) explained in the previous footnote.

[4]This has one caveat, which is not important for now. Values of A and B must be uncorrelated. That is, A and B cannot covary. If A and B covary, i.e., $Cov(A, B) \neq 0$, then $Var(A + B) = Var(A) + Var(B) + Cov(A, B)$. That is, we need to account for the covariance when calculating $Var(A + B)$.

heritability ($H^2 = V_G/V_P$) can be used to predict how a population will respond to natural selection. This is just one place where variance reveals itself to be a highly useful statistic in practice. Nevertheless, as a descriptive statistic to communicate the spread of a variable, it usually makes more sense to calculate the standard deviation of the mean.

12.4. The standard deviation

The standard deviation of the mean (s) is just the square root of the variance,

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

This is a simple step, mathematically, but it also is easier to understand conceptually as a measure of spread (Navarro and Foxcroft, 2022). By taking the square root of the variance, our units are no longer squared, so we can interpret the standard deviation in the same terms as our original data. For example, the leaf masses of plant A and plant B in the example above were measured in grams. While the variance of these masses were in g^2, the standard deviation is in g, just like the original measurements. For plant A, we calculated a leaf mass variance of $s^2 = 0.0333\,g^2$, which means that the standard deviation of leaf masses is $s = \sqrt{0.0333\,g^2} = 0.1825\,g$. Because we are reporting s in the original units, it is a very useful measure of spread to report, and it is an important one to be able to interpret. To help with the interpretation, here is an interactive tool showing how the heights of trees in a forest change across different standard deviation values[5].

> Click here for an interactive application to illustrate the standard deviation.

Here is another interactive tool showing how the shape of a histogram changes when the standard deviation of a distribution is changed[6].

> Click here for an interactive application to visualise how a histogram changes given a changing standard deviation.

The practical in Chapter 13 explains how to calculate the standard deviation in Jamovi.

[5] Here is the full URL: https://bradduthie.shinyapps.io/forest/
[6] Here is the full URL: https://bradduthie.shinyapps.io/normal_pos_neg/

12.5. The coefficient of variation

The coefficient of variation (CV) is just the standard deviation divided by the mean,

$$CV = \frac{s}{\bar{x}}.$$

Dividing by the mean seems a bit arbitrary at first, but this can often be useful for comparing variables with different means or different units. The reason for this is that the units cancel out when dividing the standard deviation by the mean. For example, for the leaf masses of plant A, we calculated a standard deviation of 0.1825 g and a mean of 5 g. We can see the units cancel below,

$$CV = \frac{0.1825\,g}{5\,g} = 0.0365.$$

The resulting CV of 0.0365 has no units; it is *dimensionless* (Lande, 1977). Because it has no units, it often used to compare measurements with much different means or with different measurement units. For example, Sokal and Rohlf (1995) suggest that biologists might want to compare tail length variation between animals with much different body sizes, such as elephants and mice. The standard deviation of tail lengths between these two species will likely be much different just because of their difference in size, so by standardising by mean tail length, it can be easier to compare relative standard deviation. This is a common application of the CV in biology, but it needs to be interpreted carefully (Pélabon et al., 2020).

Often, we will want to express the coefficient of variation as a percentage of the mean. To do this, we just need to multiply the CV above by 100%. For example, to express the CV as a percentage, we would multiply the 0.0365 above by 100%, which would give us a final answer of $CV = 3.65\%$.

12.6. The standard error

The standard error of the mean is the last measurement that we will introduce here. It is slightly different than the previous estimates in that it is a measure of the variation in the *mean* of a sample rather than the sample itself. That is, the standard error tells us how far our sample mean $\bar{x}$ is expected to deviate from the true mean μ. Technically, the standard error of the mean is the standard deviation *of sample means* rather than the standard deviation *of samples*. What does that even mean? It is easier to explain with a concrete example.

Imagine that we want to measure nitrogen levels in the water of Airthrey Loch (the loch at the centre of campus at the University of Stirling). We collect 12 water samples and

record the nitrate levels in milligrams per litre (mg/l). The measurements are reported below.

```
0.63, 0.60, 0.53, 0.72, 0.61, 0.48, 0.67, 0.59, 0.67, 0.54, 0.47, 0.87
```

We can calculate the mean of the above sample to be $\bar{x} = 0.615$, and we can calculate the standard deviation of the sample to be $s = 0.111$. We do not know what the *true* mean μ is, but our best guess is the sample mean $\bar{x}$. Suppose, however, that we then went back to the loch to collect another 12 measurements (assume that the nitrogen level of the lake has not changed in the meantime). We would expect to get values similar to our first 12 measurements, but certainly not the *exact* same measurements, right? The sample mean of these new measurements would also be a bit different. Maybe we actually go out and do this and get the following new sample.

```
0.47, 0.56, 0.72, 0.61, 0.54, 0.64, 0.68, 0.54, 0.48, 0.59, 0.62, 0.78
```

The mean of our new sample is 0.603, which is a bit different from our first. In other words, the sample means vary. We can therefore ask what is the variance and standard deviation *of the sample means*. In other words, suppose that we kept going back out to the loch, collecting 12 new samples, and recording the sample mean each time? The standard deviation of those sample means would be the standard error. **The standard error is the standard deviation of $\bar{x}$ values around the true mean μ.** But we do not actually need to go through the repetitive resampling process to estimate the standard error. We can estimate it with just the standard deviation and the sample size. To do this, we just need to take the standard deviation of the sample (s) and divide by the square root of the sample size ($\sqrt{N}$),

$$SE = \frac{s}{\sqrt{N}}.$$

In the case of the first 12 samples from the loch in the example above,

$$SE = \frac{0.111}{\sqrt{12}} = 0.032.$$

The standard error is important because it can be used to evaluate the uncertainty of the sample mean in comparison with the true mean. We can use the standard error to place confidence intervals around our sample mean to express this uncertainty. We will calculate confidence intervals in Week 5, so it is important to understand what the standard error is measuring.

If the concept of standard error is still a but unclear, we can work through one more hypothetical example. Suppose again that we want to measure the nitrogen concentration of a loch. This time, however, assume that we somehow *know* that the true mean

N concentration is $\mu = 0.7$, and that the standard deviation of water sample N concentration is $\sigma = 0.1$. Of course, we can never actually know the *true* parameter values, but we can use a computer to simulate sampling from a population in which the true parameter values are known. In Table 12.1, we simulate the process of going out and collecting 10 water samples from Airthrey Loch. This collecting of 10 water samples is repeated 20 different times. Each row is a different sampling effort, and columns report the 10 samples from each effort.

Table 12.1.: Simulated samples of nitrogen content from water samples of Airthrey Loch. Values are sampled from a normal distribution with a mean of 0.7 and a standard deviation 0.1.

Sample_1	0.74	0.63	0.67	0.82	0.74	0.62	0.54	0.60	0.60	0.62
Sample_2	0.77	0.69	0.66	0.89	0.65	0.70	0.69	0.80	0.72	0.58
Sample_3	0.66	0.70	0.68	0.67	0.66	0.70	0.57	0.76	0.61	0.58
Sample_4	0.69	0.70	0.56	0.77	0.84	0.83	0.75	0.88	0.68	0.68
Sample_5	0.74	0.70	0.78	0.62	0.73	0.81	0.90	0.55	0.68	0.71
Sample_6	0.66	0.73	0.72	0.68	0.69	0.71	0.88	0.60	0.66	0.64
Sample_7	0.65	0.73	0.61	0.75	0.74	0.69	0.67	0.63	0.63	0.59
Sample_8	0.79	0.70	0.52	0.73	0.58	0.62	0.87	0.82	0.81	0.62
Sample_9	0.99	0.72	0.72	0.66	0.66	0.77	0.64	0.74	0.63	0.64
Sample_10	0.60	0.63	0.56	0.70	0.60	0.98	0.71	0.66	0.77	0.88
Sample_11	0.85	0.88	0.68	0.68	0.60	0.40	0.61	0.65	0.87	0.66
Sample_12	0.68	0.72	0.58	0.66	0.86	0.63	0.69	0.77	0.82	0.68
Sample_13	0.82	0.72	0.85	0.72	0.47	0.71	0.64	0.59	0.64	0.85
Sample_14	0.74	0.72	0.80	0.81	0.75	0.53	0.63	0.71	0.86	0.74
Sample_15	0.57	0.68	0.80	0.57	0.58	0.75	0.65	0.64	0.90	0.51
Sample_16	0.53	0.68	0.61	0.61	0.81	0.71	0.70	0.76	0.55	0.81
Sample_17	0.60	0.83	0.67	0.56	0.65	0.70	0.76	0.80	0.74	0.62
Sample_18	0.61	0.72	0.74	0.56	0.83	0.63	0.64	0.71	0.71	0.63
Sample_19	0.60	0.68	0.67	0.85	1.03	0.74	0.79	0.75	0.77	0.65
Sample_20	0.73	0.69	0.60	0.72	0.60	0.61	0.75	0.70	0.84	0.70

We can calculate the mean of each sample by calculating the mean of each row. These 20 means are reported below.

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
## [1,] 0.658 0.715 0.659 0.738 0.722 0.697 0.669 0.706 0.717 0.709
## [2,] 0.688 0.709 0.701 0.729 0.665 0.677 0.693 0.678 0.753 0.694
```

The standard deviation of the 20 sample means reported above is 0.0263044. Now suppose that we only had Sample 1 (i.e., the top row of data). The standard deviation

of Sample 1 is $s = 0.0844327$. We can calculate the standard error from these sample values below,

$$s = \frac{0.0844327}{\sqrt{10}} = 0.0267.$$

The estimate of the standard error from calculating the standard deviation of the sample means is therefore 0.0263044, and the estimate from just using the standard error formula and data from only Sample 1 is 0.0267. These are reasonably close, and would be even closer if we had either a larger sample size in each sample (i.e., higher N) or a larger number of samples.

13. *Practical*. Plotting and statistical summaries in Jamovi

This practical focuses on applying the concepts from Chapters 9-12 in Jamovi (The Jamovi Project, 2022). The data that we will work with in this practical were collected from a research project conducted by Dr Alan Law, Prof Nils Bunnefeld, and Prof Nigel Willby at the University of Stirling (Law et al., 2014). The project focused on beaver reintroduction in Scottish habitats and its consequences for the white water lily, *Nymphaea alba*, which beavers regularly consume (Figure 13.1)[1].

Figure 13.1.: Photo of white water lillies on the water.

As an instructive example, this lab will use the data from Law et al. (2014) on the petiole diameter (mm) from *N. alba* collected from 7 different sites on the west coast of Scotland (the petiole is the structure that attaches the plant stem to the blade of the leaf). The *N. alba* dataset is available to download here (right click on the links and choose 'Save Link As…', then save it with the extension '.csv'). Note that the data are not in a tidy format, so it is important to first reorganise the data so that they can be analysed in Jamovi (13.1). Once the data are properly organised, we will use Jamovi to plot them (13.2), calculate summary statistics (13.3), apply appropriate decimals, significant figures, and rounding (13.4), and compare petiole diameters across sites (13.5).

13.1. Reorganise the dataset into a tidy format

The *N. alba* dataset is not in a tidy format. All of the numbers from this dataset are measurements of petiole diameter in mm from *N. alba*, but each row contains 7 samples because each column shows a different site. The full dataset is shown below.

[1] This figure was released into the public domain by lexej Potupin on 8 June 2018.

	Lily_Loch	Choille.Bharr	Creig.Moire	Fidhle	Buic	Linne	Beag
1	7.42	2.39	2.39	2.97	2.84	3.73	6.12
2	3.58	4.22	4.65	6.68	4.19	5.21	3.23
3	7.47	2.41	5.16	3.78	6.50	3.78	7.04
4	6.07	5.54	2.87	7.11	3.20	3.71	3.05
5	6.81	3.56	6.63	2.74	4.14	6.93	7.06
6	8.05	5.72	7.42	4.75	2.51	6.40	9.58
7	7.24	4.72	3.66	5.59	8.53	1.57	4.62
8	7.90	5.05	7.26	3.94	6.25	3.20	8.66
9	6.15	6.76	3.71	5.44	6.17	4.55	3.96
10	6.20	5.64	3.20	4.98	3.53	2.62	5.26
11	7.26	4.06	5.99	4.24	5.03	3.48	3.53
12	7.06	9.25	6.38	5.51	6.10	2.67	8.33
13	6.45	5.99	5.49	6.48	4.98	9.40	5.41
14	3.66	4.57	4.93	5.69	5.21	6.86	7.32
15	4.37	6.96	7.29	2.79	5.03	6.20	5.46
16	4.55	6.78	6.10	5.72	7.19	4.93	4.34
17	3.81	7.29	5.97	4.39	6.32	5.18	6.35
18	2.77	5.16	9.93	7.19	7.04	6.12	6.12
19	1.91	8.64	8.28	7.29	6.35	7.26	5.11
20	2.62	7.01	7.24	8.18	6.30	9.14	8.18

Remember that to make these data tidy and usable in Jamovi, we need each row to be a unique observation. What we really want then is a dataset with two columns of data. The first column should indicate the site, and the second column should indicate the petiole diameter. This can be done in two ways. First, we could use a spreadsheet programme like LibreOffice or MS Excel to create a new dataset with two columns, one column with the site information and the other column with the petiole diameters. Second, we could use the 'Data' tab in Jamovi to create two new columns of data (one for site and the other for petiole diameter). Either way, we need to copy and paste site names into the first column and petiole diameters in the second column. This is a bit tedious, and we will not ask you to do it for every dataset, but it is an important step in the process of data analysis. See Figure 13.2 for how this would look in Jamovi.

Note that to insert a new column in Jamovi, we need to right click on an existing column and select 'Add Variable' → 'Insert'. A new column will then pop up in Jamovi, and we can give this an informative name. Make sure to specify that the 'Site' column should be a nominal measure type, and the 'petiole_diameter_mm' column should be a continuous measure type. The first 6 rows of the dataset should look like the below.

	Site	petiole_diameter_mm
1	Lily_Loch	7.42
2	Lily_Loch	3.58
3	Lily_Loch	7.47

Figure 13.2.: Tidying the raw data of petiole diameters from lily pad measurements across 7 sites in Scotland. A new column of data is created by right clicking on an existing column and choosing 'Add Variable'.

```
4 Lily_Loch                6.07
5 Lily_Loch                6.81
6 Lily_Loch                8.05
```

With the reorganised dataset, we are now ready to do some analysis in Jamovi. We will start with some plotting.

13.2. Histograms and box-whisker plots

We will start by making a histogram of the full dataset of petiole diameter. To do this, we need to go to the 'Analyses' tab of the Jamovi toolbar, then select the 'Exploration' button.

Figure 13.3.: Jamovi toolbar after having selected on the Analyses tab followed by the Exploration button.

Next, select the 'Descriptives' option (Figure 13.3). This will open a new window where it is possible to create plots and calculate summary statistics. The white box on the left of the Descriptive interface lists all of the variables in the dataset. Below this box, there are options for selecting different summary statistics ('Statistics') and building different graphs ('Plots'). To get started, select the petiole diameter variable in the box to the left, then move it to the 'Variables' box (top right) using the → arrow. Next, open the Plots option at the bottom of the interface. Choose the 'Histogram' option by clicking the checkbox. A histogram will open up in the window on the right (you might need to scroll down).

Take a look at the histogram to the right (Figure 13.4). Just looking at the histogram, write down what you think the following summary statistics will be.

Mean: _____

Median: _____

Figure 13.4.: Jamovi Descriptives toolbar with petiole diameter selected and a histogram produced in the plotting window.

Standard deviation: _____

Based on the histogram, do you think that the mean and median are the same? Why or why not?

The histogram needs better labelled axes and an informative caption. To label the axes better, go back to the data tab and double click on the column heading 'petiole_diameter_mm'. Change the name of the data variable to 'Petiole diameter (mm)'. The newly named variable will then appear when a new histogram of the petiole diameter data is made. To write a caption in Jamovi, click on the 'Edit' tab at the very top of the toolbar. You will see some blue boxes above and below the histogram, and you can write your caption by clicking on the box immediately below the histogram. Write a caption for the histogram below.

If you want to save the histogram, then you can right click on it. A pop-up box will give you several options; select 'Image → Export' to save the histogram. You can save it as a PDF, PNG, SVG, or EPS (if in doubt, PNG is probably the easiest to use). You do not need to do this for this lab, but knowing how to do it will be useful for other modules, including your fourth year dissertation.

In the first example, we looked at petiole diameters across the entire dataset, but suppose that we want to see how the data are distributed for each site individually. To do this, we just need to go back to the Descriptives box (Figure 13.4) and put the 'Site' variable into the box on the lower right called 'Split by'. Do this by selecting 'Site' then using the lower → arrow to bring it to the 'Split by' box. Instead of one histogram of petiole diameters, you will now see 7 different histograms, one for each site, all stacked on top of each other. This might be useful, but all of these histograms together are a bit busy. Instead, we can use a box-whisker plot to compare the distributions of petiole diameters across different sites.

To create a box plot, simply check 'Box plot' from the Plots options (you might want to uncheck 'Histogram', but it is not necessary). You should now see all of the different sites on the x-axis of the newly created boxplot and a summary of the petiole diameters on the y-axis. Based on the boxplot, which site appears to have the highest and lowest median petiole diameter?

Highest: _____

Lowest: _____

There is one more trick with box-whisker plots in Jamovi that is useful. The current plots show a summary of each site, but it might also be useful to plot the actual data points to give some more information about the distribution of petiole diameters. You can do this by checking the option 'Data', which places the petiole diameter of each sample over the box and whiskers for each site. The y-axis shows the petiole diameter of each data point. By default, the points are jittered on the x-axis, which just means that they are placed randomly on the x-axis within a site. This is just to ensure that points will not be placed directly on top of each other if they are the same value. If you prefer, you can use the pull-down menu right below the Data checkbox to select 'Stacked' instead of 'Jittered' The stacked option will place points side by side. Think about where the points are in relation to the box and whiskers of the plot; this should help you develop an intuitive understanding of how to read box-whisker plots.

13.3. Calculate summary statistics

We can calculate the summary statistics using the 'Descriptives' option in Jamovi, just as we did with the histogram and box-whisker plots. Before doing anything else, again place the petiole diameter variable in the box of variables, but do not split the dataset by site just yet because we first want summary statistics across the entire dataset. Below the box of variables, but above the Plots options, there are options for selecting different summary statistics. Open up this new box and have a look at the different summary statistics that can be calculated. To calculate all of the variables explained in Chapter 11 and Chapter 12, check the following 11 boxes:

- N: _____
- Std. deviation: _____
- Variance: _____
- Minimum: _____
- Maximum: _____
- Range: _____
- IQR: _____
- Mean: _____
- Median: _____
- Mode: _____

- Std. error of mean: _____

When you do this, the Statistics option in Jamovi should like like it does in Figure 13.5.

Figure 13.5.: Jamovi Descriptives toolbar showing the summary statistics available to report.

Once you check these boxes, you will see a 'Descriptives' table open on the right hand side of Jamovi. This table will report all of the summary statistics that you have checked. Write down the values for the summary statistics next to the corresponding bullet points above.

Next split these summary statistics up by site. Notice the very large table that is now produced on the right hand side of Jamovi. Which of the 7 sites in the data set has the highest mean petiole diameter, and what is its mean?

Site: _____

Mean: _____

Which of the 7 sites has the lowest variance in petiole diameter, and what is its variance?

Site: _____

Variance: _____

Make sure that you are able to find and interpret these summary statistics in Jamovi. Explore different options to get more comfortable using Jamovi for building plots and reporting summary statistics. Can you find the first and third quartiles for each site? Report the third quartiles for each site below.

Beag: _____

Buic: _____

Choille-Bharr: _____

Creig-Moire: _____

Fidhle: _____

Lily_Loch: _____

Linne: _____

Next, we will look at reporting summary statistics to different significant figures.

13.4. Reporting decimals and significant figures

Using the same values that you reported above for the whole dataset (i.e., not broken down by site), report each summary statistics to two significant figures. Remember to round accurately if you need to reduce the number of significant figures from the original values to the new values below. In assessments, you will often be asked to report a particular answer to a specific number of decimal places or significant figures, so the intention here is to help you practice.

- N: _____
- Std. deviation: _____
- Variance: _____
- Minimum: _____
- Maximum: _____
- Range: _____
- IQR: _____
- Mean: _____
- Median: _____
- Mode: _____
- Std. error of mean: _____

Remember from 13.2 that you were asked to write down what you thought the mean, median, and standard deviation were just by inspecting the histogram. Compare your answers in that section with the rounded statistics listed above. Were you able to get a similar value from the histogram as calculated in Jamovi from the data? What can you learn from the histogram that you cannot from the summary statistics, and what can you learn from the summary statistics that you cannot from the histogram? Write your reflections in the space below.

Next, we will produce barplots to show the mean petiole diameter for each site.

13.5. Comparing across sites

To make a barplot that compares the mean petiole diameters across sites, we again use the Descriptives option in Jamovi. Place petiole diameter as the variable, and spit this by site. Next, go down to the plotting options and check 'Bar plot'. You will see a barplot produced in the window to the right with different sites on the x-axis. Bar heights show the mean petiole diameter for each site. Notice the intervals shown for each bar (i.e., the vertical lines in the centre of the bars that go up and down different lengths). These error bars are centred on the mean petiole diameter (bar height) and show one standard error above and below the site mean. Recall back from Chapter 12; what information do these error bars convey about the estimated mean petiole diameter?

What can you say about the mean petiole diameters across the different sites? Do these sites appear to have very different mean petiole diameters?

There were 20 total petiole diameters sampled from each site. If we were to go back out to these 7 sites and sample another 20 petiole diameters, could we **really** expect to get the exact same site means? Assuming the site means would be at least a bit different for our new sample, is it possible that the sites with the highest or lowest petiole diameters might also be different in our new sample? If so, then what does this say about our ability to make conclusions about the differences in petiole diameter among sites?

Part IV.

Probability models and the Central Limit Theorem

Week 4 Overview

Dates	13 February 2023 - 17 February 2023
Reading	**Required:** SCIU4T4 Workbook chapters 14 and 15 **Recommended:** Navarro and Foxcroft (2022) Chapter 7 **Suggested:** Rowntree (2018) Chapter 4 **Advanced:** None
Lectures	4.1: What is probability? (16:07 min; Video) 4.2: Adding and multiplying probabilities (16:18 min; Video) 4.3: Probability distributions (15:30 min; Video) 4.4: The normal distribution (15:15 min; Video) 4.5: z-scores (5:12 min; Video) 4.6: Examples using z-scores (13:22 min; Video) 4.7: More examples using z-scores (9:52 min; Video) 4.8: The binomial distribution (16:34 min; Video) 4.9: The Poisson distribution (12:57 min; Video) 4.10: The central limit theorem (5:21 min; Video) 4.11: z-score tables (14:32 min; Video)
Practical	Probability and simulation (Chapter 16) Room: Cottrell 2A17 Group A: 15 FEB 2023 (WED) 13:05-15:55 Group B: 16 FEB 2023 (THU) 09:05-11:55
Help hours	Ian Jones Room: Cottrell 3V1B 17 FEB 2023 (FRI) 15:05-17:55
Assessments	Week 4 Practice quiz on Canvas

Week 4 focuses on probability and the central limit theorem.

Chapter 14 introduces probability models and how to interpret them. The chapter also provides some examples of probability distributions that are especially relevant to biological and environmental sciences.

Chapter 15 focuses on the central limit theorem (CLT), what it is, and why it is so important in statistics.

Chapter 16 guides you through the week 4 practical. The aim of this practical is to apply the ideas from Chapter 14 and Chapter 15 in Jamovi to predict probabilities from a real dataset.

14. Introduction to probability models

Suppose that we flip a fair coin over a flat surface. There are two possibilities for how the coin lands on the surface. Either the coin lands on one side (heads) or the other side (tails), but we do not know the outcome in advance. If these two events (heads or tails) are equally likely, then we could reason that there is a 50% chance that a flipped coin will land heads up and a 50% chance that it will land heads down. What do we actually mean when we say this? For example, when we say that there is a 50% chance of the coin landing heads up, are we making a claim about our own uncertainty, how coins work, or how the world works? We might mean that we simply do not know whether or not the coin will land heads up, so a 50-50 chance just reflects our own ignorance about what will actually happen when the coin is flipped. Alternatively, we might reason that if a fair coin were to be flipped many times, all else being equal, then about half of flips should end heads up, so a 50% chance is a reasonable prediction of what will happen in any given flip. Or, perhaps we reason that events such as coin flips really are guided by chance on some deeper fundamental level, such that our 50% chance reflects some real causal metaphysical process in the world. These are questions concerning the philosophy of probability. The philosophy of probability is an interesting sub-discipline in its own right, with implications that can and do affect how researchers do statistics (Edwards, 1972; Mayo, 1996; Gelman and Shalizi, 2013; Suárez, 2020; Mayo, 2021; Navarro and Foxcroft, 2022).

In this chapter, we will not worry about the philosophy of probability[1] and instead focus on the mathematical rules of probability as applied to statistics. These rules are important for predicting real-world events in the biological and environmental sciences. For example, we might need to make predictions concerning the risk of disease spreading in a population, or the risk of extreme events such as droughts occurring given increasing global temperatures. Probability is also important for testing scientific hypotheses. For example, if we sample two different groups and calculate that they have different means (e.g., two different fields have different mean soil nitrogen concentrations), we might want to know the probability that this difference between means could have arisen by chance. Here we will introduce practical examples of probability, then introduce some common probability distributions.

[1] In the interest of transparency, this book presents a *frequentist* interpretation of probability (Mayo, 1996). While this approach does reflect the philosophical inclinations of the author, the reason for working from this interpretation has more to do with the statistical tests that are most appropriate for an introductory statistics module, which are also the tests most widely used in the biological and environmental sciences.

14.1. An instructive example

Probability focuses on the outcomes of trials, such as the **outcome** (heads or tails) of the **trial** of a coin flip. The probability of a specific outcome is the relative number of times it is expected to happen given a large number of trials,

$$P(outcome) = \frac{Number\ of\ times\ outcome\ occurs}{Total\ number\ of\ trials}.$$

For the outcome of a flipped coin landing on heads,

$$P(heads) = \frac{Flips\ landing\ on\ heads}{Total\ number\ of\ flips}.$$

As the total number of flips becomes very large, the number of flips that land on heads should get closer and closer to half the total, 1/2 or 0.5 (more on this later). The above equations use the notation $P(E)$ to define the probability (P) of some event (E) happening. Note that the number of times an outcome occurs cannot be less than 0, so $P(E) \geq 0$ must always be true. Similarly, the number of times an outcome occurs cannot be greater than the number of trials; the most frequently it can happen is in *every* trial, in which case the top and bottom of the fraction has the same value. Hence, $P(E) \leq 1$ must also always be true. Probabilities therefore range from 0 (an outcome *never* happens) to 1 (an outcome *always* happens).

It might be more familiar and intuitive at first to think in terms of percentages (i.e., from 0-100% chance of an outcome, rather than from 0-1), but there are good mathematical reasons for thinking about probability on a 0-1 scale (it makes calculations easier). For example, suppose we have two coins, and we want to calculate the probability that they will both land on heads if we flip them at the same time. That is, we want to know the probability that coin 1 lands on heads **and** coin 2 lands on heads. We can assume that the coins do not affect each other in any way, so each coin flip is **independent** of the other (i.e., the outcome of coin 1 does not affect the outcome of coin 2, and *vice versa* – this kind of assumption is often very important in statistics). Each coin, by itself, is expected to land on heads with a probability of 0.5, $P(heads) = 0.5$. When we want to know the probability that two or more independent events will happen, we *multiply* their probabilities. In the case of both coins landing on heads, the probability is therefore,

$$P(Coin_1 = heads\ \cap Coin_2 = heads) = 0.5 \times 0.5 = 0.25.$$

Note that the symbol $\cap$ is basically just a fancy way of writing 'and' (technically, the intersection between sets; see set theory for details). Verbally, all this is saying is that the probability of coin 1 landing on heads *and* the probability of coin 2 landing on heads equals 0.5 times 0.5, which is 0.25.

But why are we *multiplying* to get the joint probability of both coins landing on heads? Why not add, for example? We could just take it as a given that multiplication is the correct operation to use when calculating the probability that multiple events will occur. Or we could do a simple experiment to confirm that 0.25 really is about right (e.g., by flipping 2 coins 100 times and recording how many times both coins land on heads). But neither of these options would likely be particularly satisfying. Let us first recognise that adding the probabilities cannot be the correct answer. If the probability of each coin landing on heads is 0.5, then adding probabilities would imply that the probability of both landing on heads is $0.5 + 0.5 = 1$. This does not make any sense because we know that there are other possibilities, such as both coins landing on tails, or one coin landing on heads and the other landing on tails. Adding probabilities cannot be the answer, but why multiply?

We can think about probabilities visually, as a kind of probability space. When we have only one trial, then we can express the probability of an event along a line (Figure 14.1).

Figure 14.1.: Total probability space for flipping a single coin and observing its outcome (heads or tails). Given a fair coin, the probability of heads equals a proportion 0.5 of the total probability space, while the probability of tails equals the remaining 0.5 proportion.

The total probability space is 1, and 'heads' occupies a density of 0.5 of the total space. The remaining space, also 0.5, is allocated to 'tails'. When we add a second independent trial, we now need 2 dimensions of probability space (Figure 14.2). The probability of heads or tails for coin 1 (the horizontal axis of Figure 14.2) remains unchanged, but we add another axis (vertical this time) to think about the equivalent probability space of coin 2.

Now we can see that that the area in which both coin 1 and coin 2 land on heads has a proportion of 0.25 of the total area. This is a geometric representation of what we did when calculating $P(Coin_1 = heads \cap Coin_2 = heads) = 0.5 \times 0.5 = 0.25$. The multiplication works because multiplying probabilities carves out more specific regions of probability space. Note that the same pattern would apply if we flipped a third coin. In this case, the probability of all 3 coins landing on heads would be $0.5 \times 0.5 \times 0.5 = 0.125$, or $0.5^3 = 0.125$.

What about when we want to know the probability of one outcome **or** another outcome happening? Here is where we add. Note that the probability of a coin flip landing on heads or tails must be 1 (there are only 2 possibilities!). What about the probability of both coins landing on the same outcome; that is, either both coins landing on heads or both landing on tails? We know that the probability of both coins landing on heads is 0.25. The probability of both coins landing on tails is also 0.25, so the probability that

Figure 14.2.: Total probability space for flipping two coins and observing their different possible outcomes (heads-heads, heads-tails, tails-heads, and tails-tails). Given two fair coins, the probability of flipping each equals 0.25, which corresponds to the lower left square of the probability space.

both coins land on either heads **or** tails is $0.25 + 0.25 = 0.5$. The visual representation in Figure 14.2 works for this example too. Note that heads-heads and tails-tails outcomes are represented by the lower left and upper right areas of probability space, respectively. This is 0.5 (i.e., 50%) of the total probability space.

14.2. Biological applications

Coin flips are instructive, but the relevance for biological and environmental sciences might not be immediately clear. In fact, probability is extremely relevant in nearly all areas of the natural sciences. The following are just 2 hypothetical examples where the calculations in the previous section might be usefully applied:

1. From a recent report online, suppose you learn that 1 in 40 people in your local area are testing positive for Covid-19. You find yourself in a small shop with 6 other people. What is the probability that at least 1 of these 6 other people would test positive for Covid-19? To calculate this, note that the probability that any given person has Covid-19 is $1/40 = 0.025$, which means that the probability that a person does **not** must be $1 - 0.025 = 0.975$ (they either do or do not, and the probabilities must sum to 1). The probability that **all** 6 people *do not* have Covid-19 is therefore $(0.975)^6 = 0.859$. Consequently, the probability that at least 1 of the 6 people **does** have Covid-19 is $1 - 0.859 = 0.141$, or 14.1%.

2. Imagine you are studying a population of sexually reproducing, diploid (i.e., 2 sets of chromosomes), animals, and you find that a particular genetic locus has 3 alleles with frequencies $P(A_1) = 0.40$, $P(A_2) = 0.45$, and $P(A_3) = 0.15$. What is the probability that a randomly sampled animal will be heterozygous with 1 copy of the A_1 allele and 1 copy of the A_3 allele? Note that there are 2 ways for A_1 and A_3 to arise in an individual, just like there were 2 ways to get a heads and tails coin in the section 14.1 example (see Figure 14.2). The individual could either get an A_1 in the first position and A_3 in the second position, or an A_3 in the first position and A_1 in the second position. We can therefore calculate the probability as, $P(A_1) \times P(A_3) + P(A_3) \times P(A_1)$, which is $(0.40 \times 0.15) + (0.15 \times 0.4) = 0.12$, or 12% (in population genetics, we might use the notation $p = P(A_1)$ and $r = P(A_3)$, then note that $2pr = 0.12$).

In both of these examples, we made some assumptions, which might or might not be problematic. In the first example, we assumed that the 6 people in our shop were a random and independent sample from the local area (i.e., people with Covid-19 are not more or less likely to be in the shop, and the 6 people in the shop were not associated in a way that would affect their individual probabilities of having Covid-19). In the second example, we assumed that individuals mate randomly, and that there is no mutation, migration, or selection on genotypes (Hardy, 1908). It is important to recognise these

assumptions when we are making them because violations of assumptions could affect the probabilities of events!

14.3. Sampling with and without replacement

It is often important to make a distinction between sampling with or without replacement. Sampling with replacement just means that whatever has been sampled once gets put back into the population before sampling again. Sampling without replacement means that whatever has been sampled does not get put back into the population before sampling again. An example makes the distinction between sampling with and without replacement clearer.

Figure 14.3.: Playing cards can be useful for illustrating concepts in probability. Here we have 5 hearts (left) and 5 spades (right).

Figure 14.3 shows 10 playing cards, 5 hearts and 5 spades. If we shuffle these cards thoroughly and randomly select 1 card, what is the probability of selecting a heart? This is simply,

$$P(heart) = \frac{5\,hearts}{10\,total\,cards} = 0.5.$$

What is the probability of randomly selecting 2 hearts? This depends if we are sampling with or without replacement. If we sample 1 card, then put it back into the deck before

sampling the second card, then the probability of sampling a heart does not change (in both samples, we have 5 hearts and 10 cards). Hence, the probability of sampling two hearts with replacement is $P(heart) \times P(heart) = 0.5 \times 0.5 = 0.25$. If we do not put the first card back into the deck before sampling again, then we have changed the total number of cards. After sampling the first heart, we have one fewer hearts in the deck and one fewer cards, so the new probability for sampling a heart becomes,

$$P(heart) = \frac{4\,hearts}{9\,total\,cards} = 0.444.$$

Since the probability has changed after the first heart is sampled, we need to use this adjusted probability when sampling without replacement. In this case, the probability of sampling two hearts is $0.5 \times 0.444 = 0.222$. This is a bit lower than the probability of sampling with replacement because we have decreased the number of hearts that can be sampled. When sampling from a set, it is important to consider whether the sampling is done with or without replacement (in assessments, we will always make this clear).

14.4. Probability distributions

Up until this point, we have been considering the probabilities of specific outcomes. That is, we have considered the probability that a coin flip will be heads, that an animal will have a particular combination of alleles, or that we will randomly select a particular suit of card from a deck. Here we will move from specific outcomes and consider the *distribution* of outcomes. For example, instead of finding the probability that a flipped coin lands on heads, we might want to consider the distribution of the number of times that it does (in this case, 0 times or 1 time).

This is an extremely simple distribution. There are only two discrete possibilities for the number of times the coin will land on heads, 0 or 1. And the probability of both outcomes is 0.5, so the bars in Figure 14.4 are the same height. Next, we will consider some more interesting distributions.

14.4.1. Binomial distribution

The simple distribution with a single trial of a coin flip was actually an example of a binomial distribution. More generally, a binomial distribution describes the number of successes in some number of trials (Miller and Miller, 2004). The word 'success' should not be taken too literally here; it does not necessarily indicate a good outcome, or an accomplishment of some kind. A success in the context of a binomial distribution just means that an outcome *did* happen as opposed to it *not* happening. If we define a coin flip landing on heads as a success, we could consider the probability distribution of the number of successes over 10 trials (Figure 14.5)

Figure 14.4.: Probability distribution for the number of times that a flipped coin lands on heads in 1 trial.

Figure 14.5.: Probability distribution for the number of times that a flipped coin lands on heads in 10 trials.

Figure 14.5 shows that the most probable outcome is that 5 of the 10 coins flipped will land on heads. This makes some sense because the probability that any 1 flip lands on heads is 0.5, and 5 is 1/2 of 10. But 5 out of 10 heads happens only with a probability of about 0.25. There is also about a 0.2 probability that the outcome is 4 heads, and the same probability that the outcome is 6 heads. Hence, the probability that we get an outcome of between 4-6 heads is about $0.25 + 0.2 + 0.2 = 0.65$. In contrast, the probability of getting all heads is very low (about 0.00098).

More generally, we can define the number of successes using the random variable X. We can then use the notation $P(X = 5) = 0.25$ to indicate the probability of 5 successes, or $P(4 \leq X \leq 6) = 0.65$ as the probability that the number of successes is greater than or equal to 4 and less than or equal to 6.

Imagine that you were told a coin was fair, then flipped it 10 times. Imagine that 9 flips out of the 10 came up heads. Given the probability distribution shown in Figure 14.5, the probability of getting 9 or more heads in 10 flips given a fair coin is very low ($P(X \geq 9) \approx 0.011$). Would you still believe that the coin is fair after these 10 trials? How many, or how few, heads would it take to convince you that the coin was not fair? This question gets to the heart of a lot of hypothesis-testing in statistics, and we will discuss it more in Week 6.

Note that a binomial distribution does not need to involve a fair coin with equal probability of success and failure. We can consider again the first example in Section 14.2, in which 1 in 40 people in an area are testing positive for Covid-19, then ask what the probability is that 0-6 people in a small shop would test positive (Figure 14.6).

Note that the shape of this binomial distribution is different from the coin flipping trials in Figure 14.5. The distribution is skewed, with a high probability of 0 successes and a diminishing probability of 1 or more successes.

The shape of a statistical probability distribution can be defined mathematically. Depending on the details (more on this later), we call the equation defining the distribution either a probability mass function or a probability density function. This book is about statistical techniques, not statistical theory, so we will relegate these equations to footnotes.[2] What is important to know is that the shape of a distribution is modulated by

[2]For those interested, more technically, we can say that a random variable X has binomial distribution if and only if its probability mass function is defined by (Miller and Miller, 2004),

$$b\left(x; n, \theta\right) = \binom{n}{x} \theta^{x} \left(1 - \theta\right)^{n-x} .$$

In this binomial probability mass function, $x = 0, 1, 2, \ldots, n$ (i.e., x can take any integer value from 0 to n). Note that the n over the x in the first parentheses on the right hand side of the equation is a binomial coefficient, which can be read "n choose x". This can be written out as,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

Note that the exclamation mark indicates a factorial, such that $n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1$. That is, the factorial multiplies every decreasing integer down to 1. For example, $4! = 4 \times 3 \times 2 \times 1 =$

Figure 14.6.: Probability distribution for the number of people who have Covid-19 in a shop of 6 when the probability of testing positive is 0.025.

parameters. The shape of a binomial distribution is determined by 2 parameters, the number of trials (n) and the probability of success (θ). In Figure 14.5, there were 10 trials each with a success probability of 0.5 (i.e., $n = 10$ and $\theta = 0.5$). In Figure 14.6, there were 6 trials each with a success probability of 0.025 (i.e., $n = 6$ and $\theta = 0.025$). This difference in parameter values is why the two probability distributions have a different shape.

14.4.2. Poisson distribution

Imagine sitting outside on a park bench along a path that is a popular route for joggers. On this particular day, runners pass by the bench at a steady rate of about 4 per minute, on average. We might then want to know the *distribution* of the number of runners passing by per minute. That is, given that we see 4 runners per minute on average, what is the probability that we will see just 2 runners pass in any given minute. What is the probability that we will see 8 runners pass in a minute? This hypothetical example is modelled with a Poisson distribution. A Poisson distribution describes events happening over some interval (e.g., happening over time or space). There are a lot of situations where a Poisson distribution is relevant in biological and environmental sciences:

- Number of times a particular species will be encountered while walking a given distance.
- Number of animals a camera trap will record during a day.
- Number of floods or earthquakes that will occur in a given year.

The shape of a Poisson distribution is described by just 1 parameter, λ. This parameter is both the mean and the variance of the Poisson distribution. We can therefore get the probability that some number of events (x) will occur just by knowing λ (Figure 14.7).

Like the binomial distribution, the Poisson distribution can also be defined mathematically[3]. Also like the binomial distribution, probabilities in the Poisson distribution focus on **discrete** observations. This is, probabilities are assigned to a specific number of successes in a set of trials (binomial distribution) or the number of events over time (Poisson distribution). In both cases, the probability distribution focuses on countable numbers. In other words, it does not make any sense to talk about the probability of a coin landing on heads 3.75 times after 10 flips, nor the probability of 2.21 runners passing by a park bench in a given minute. The probability of either of these events

24. None of this is critical to know for applying statistical techniques to biological and environmental science data, but it demonstrates just a bit of the theory underlying statistical tools.

[3] A random variable X has a Poisson distribution if and only if its probability mass function is defined by (Miller and Miller, 2004),

$$p(x; \lambda) = \frac{\lambda^x e^x}{x!}.$$

Recall from Chapter 1 Euler's number, $e \approx 2.718282$, and from footnote 13 that the exclamation mark indicates a factorial. In the Poisson probability mass function, x can take any integer value greater than or equal to 0.

Figure 14.7.: Poisson probability distributions given different rate parameter values.

happening is zero, which is why the Figures 14.5-14.7 all have spaces between the vertical bars. These spaces indicate that values between the integers are impossible. When observations are discrete like this, they are defined by a *probability mass function*. In the next section, we consider distributions with a continuous range of possible sample values; these distributions are defined by a *probability density function*.

14.4.3. Uniform distribution

We now move on to continuous distributions, starting with the continuous uniform distribution. We introduce this distribution mainly to clarify the difference between a discrete and continuous distribution. While the uniform distribution is very important in a lot of statistical tools (notably, simulating pseudorandom numbers), it is not something that we come across much in biological or environmental science data. The continuous uniform distribution has two parameters, α and β (Miller and Miller, 2004).[4] Values of α and β can be any real number (not just integers). For example, suppose that $\alpha = 1$ and $\beta = 2.5$. In this case, Figure 14.8 shows the probability distribution for sampling some value x.

The height of the distribution in Figure 14.8 is $1/(\beta - \alpha) = 1/(2.5 - 1) \approx 0.667$. All values between 1 and 2.5 have equal probability of being sampled.

Here is a good place to point out the difference between the continuous distribution versus the discrete binomial and Poisson distributions. From the uniform distribution of Figure 14.8, we can, theoretically, sample *any* real value between 1 and 2.5 (e.g., 1.34532 or 2.21194; the sampled value can have as many decimals as our measuring device allows). There are uncountably infinite real numbers, so it no longer makes sense to ask what is the probability of sampling a specific number. For example, what is the probability of sampling a value of *exactly* 2, rather than, say, 1.999999 or 2.000001, or something else arbitrarily close to 2? The probability of sampling a specific number exactly is negligible. Instead, we need to think about the probability of sampling within intervals. For example, what is the probability of sampling a value between 1.9 and 2.1, or any value greater than 2.2? This is the nature of probability when we consider continuous distributions.

14.4.4. Normal distribution

The last distribution, the normal distribution (also known as the "Gaussian distribution" or the "bell curve") has a special place in statistics (Miller and Miller, 2004; Navarro

[4] A random variable X has a continuous uniform distribution if and only if its probability density function is defined by (Miller and Miller, 2004),

$$u(x; \alpha, \beta) = \frac{1}{\beta - \alpha},$$

where $\alpha < x < \beta$, and $u(x; \alpha, \beta) = 0$ everywhere else. The value x can take any real number.

Figure 14.8.: A continuous uniform distribution in which a random variable X takes a value between 1 and 2.5.

and Foxcroft, 2022). It appears in many places in the biological and environmental sciences and, partly due to the central limit theorem (see Chapter 15), is fundamental to many statistical tools. The normal distribution is continuous, just like the continuous uniform distribution from the previous section. Unlike the uniform distribution, with the normal distribution, it is possible (at least in theory) to sample *any* real value, $-\infty < x < \infty$. The distribution has a symmetrical, smooth bell shape (Figure 14.8), in which probability density peaks at the mean, which is also the median and mode of the distribution. The normal distribution has two parameters, the mean (μ) and the standard deviation (σ).[5] The mean determines where the peak of the distribution is, and the standard deviation determines the width or narrowness of the distribution. Note that we are using μ for the mean here instead of $\bar{x}$, and σ for the standard deviation instead of s, to differentiate between the *population* parameters from the *sample* estimates of Chapter 11 and Chapter 12.

Figure 14.9.: A standard normal probability distribution, which is defined by a mean value of 0 and a standard deviation of 1.

[5] A random variable X has a normal distribution if and only if its probability density function is defined by (Miller and Miller, 2004),

$$n(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

In the normal distribution, $-\infty < x < \infty$. Note the appearance of two irrational numbers introduced back in Chapter 1, π and e.

The normal distribution shown in Figure 14.9 is called the **standard normal distribution**, which means that it has a mean of 0 ($\mu = 0$) and a standard deviation of 1 ($\sigma = 1$). Note that because the standard deviation of a distribution is the square-root of the variance (see Chapter 12), and $\sqrt{1} = 1$, the variance of the standard normal distribution is also 1. We will look at the standard normal distribution more closely in Chapter 15.

14.5. Summary

This chapter has introduced probability models and different types of distributions. It has focused on the key points that are especially important for understanding and implementing statistical techniques. As such, a lot of details have been left out. For example, the probability distributions considered in Section 14.4 comprise only a small number of example distributions that are relevant for biological and environmental sciences. In Chapter 15, we will get an even closer look at the normal distribution and why it is especially important.

15. The Central Limit Theorem (CLT)

The previous chapter finished by introducing the normal distribution. This chapter focuses on the normal distribution in more detail and explains why it is so important in statistics.

15.1. The distribution of means is normal

The central limit theorem (CLT) is one of the most important theorems in statistics. It states that if we sample values from **any** distribution and calculate the mean, as we increase our sample size N, the distribution *of the mean* gets closer and closer to a normal distribution (Sokal and Rohlf, 1995; Miller and Miller, 2004; Spiegelhalter, 2019).[1] This statement is busy and potentially confusing at first, partly because it refers to two separate distributions, the sampling distribution and the distribution of the sample mean. We can take this step by step, starting with the sampling distribution.

The sampling distribution could be any of the 4 distributions introduced in Chapter 14 (binomial, Poisson, uniform, or normal). Suppose that we sample the binomial distribution from Figure 14.6, the one showing the number of people out of 6 who would test positive for Covid-19 if the probability of testing positive was 0.025. Assume that we sample a value from this distribution (i.e., a number from 0 to 6) 100 times (i.e., $N = 100$). If it helps, we can imagine going to 100 different shops, all of which are occupied by 6 people. From these 100 samples, we can calculate the sample mean $\bar{x}$. This would be the mean number of people in a shop who would test positive for Covid-19. If we were just collecting data to try to estimate the mean number of people with Covid-19 in shops of 6, this is where our calculations might stop. But here is where the second distribution becomes relevant.

Suppose that we could somehow go back out to collect *another* 100 samples from 100 completely different shops. We could then get the mean of this new sample of $N = 100$ shops. To differentiate, we can call the first sample mean $\bar{x}_1$ and this new sample mean $\bar{x}_2$. Will $\bar{x}_1$ and $\bar{x}_2$ be the exact same value? Probably not! Since our samples are independent and random from the binomial distribution (Figure 14.6), it is almost certain that the two sample means will be at least a bit different. We can therefore ask

[1] For those interested, a mathematical proof of the CLT can be found in Miller and Miller (2004). Here we will demonstrate the CLT by simulation. As an aside, the CLT also applies to the sum of sample values, which will also have a distribution that approaches normality as $N \to \infty$.

about the *distribution* of sample means. That is, what if we kept going back out to get more samples of 100, calculating additional sample means $\bar{x}_3$, $\bar{x}_4$, $\bar{x}_5$, and so forth? What would this distribution look like? It turns out, it would be a normal distribution!

Figure 15.1.: A simulated demonstration of the central limit theorem. (a) Recreation of Figure 14.6 showing the probability distribution for the number of people who have Covid-19 in a shop of 6 when the probability of testing positive is 0.025. (b) The distribution of 1000 means sampled from panel (a), where the sample size is 100.

To demonstrate the CLT in action, Figure 15.1 shows the two distributions side-by-side. The first (Figure 15.1a) shows the original distribution from Figure 14.6, from which samples are collected and sample means are calculated. The second (Figure 15.1b) shows the distribution of 1000 sample means (i.e., $\bar{x}_1, \bar{x}_2, ..., \bar{x}_{999}, \bar{x}_{1000}$). Each mean $\bar{x}_i$ is calculated from a sample of $N = 100$ from the distribution in Figure 15.1a. Sampling is simulated using a random number generator on the computer (the lab practical in Chapter 16 shows an example of how to do this in Jamovi).

The distribution of sample means shown in Figure 15.1b is not perfectly normal. We can try again with an even bigger sample size of $N = 1000$, this time with a Poisson distribution where $\lambda = 1$ in Figure 14.7. Figure 15.2 shows this result, with the original Poisson distribution shown in Figure 15.2a, and the corresponding distribution built

from 1000 sample means shown in Figure 15.2b.

Figure 15.2.: A simulated demonstration of the central limit theorem. (a) Recreation of Figure 14.7 showing the probability distribution for the number of events occurring in a Poisson distribution with a rate parameter of 1. (b) The distribution of 1000 means sampled from panel (a), where the sample size is 1000.

Finally, we can try the same approach with the continuous uniform distribution shown in Figure 14.8. This time, we will use an even larger sample size of $N = 10000$ to get our 1000 sample means. The simulated result is shown in Figure 15.3.

In all cases, regardless of the original sampling distribution (binomial, Poisson, or uniform), the distribution of sample *means* has the shape of a normal distribution. This normal distribution of sample means has important implications for statistical hypothesis testing. The CLT allows us to make inferences about the means of non-normally distributed distributions (Sokal and Rohlf, 1995), to create confidence intervals around sample means, and to apply statistical hypothesis tests that would otherwise not be possible. We will look at these statistical tools in future chapters.

Figure 15.3.: A simulated demonstration of the central limit theorem. (a) Recreation of Figure 14.8 showing a continuous uniform distribution with a minimum of 1 and a maximum of 2.5. (b) The distribution of 1000 means sampled from panel (a), where the sample size is 10000.

15.2. Probability and z-scores

We can calculate the probability of sampling some range of values from the normal distribution if we know the distribution's mean (μ) and standard deviation (σ). For example, because the normal distribution is symmetric around the mean, the probability of sampling a value greater than the mean will be 0.5 (i.e., $P(x > \mu) = 0.5$), and so will the probability of sampling a value less than the mean (i.e., $P(x < \mu) = 0.5$). Similarly, about 68.2% of the normal distribution's probability density lies within 1 standard deviation of the mean (shaded region of Figure 15.4), which means that the probability of randomly sampling a value x that is greater than $\mu - \sigma$ but less than $\mu + \sigma$ is $P(\mu - \sigma < x < \mu + \sigma) = 0.682$.

Figure 15.4.: A normal distribution in which the shaded region shows the area within one standard deviation of the mean (dotted line); that is, the shaded region starts on the left at the mean minus one standard deviation, then ends at the right at the mean plus one standard deviation. This shaded area encompases 68.2 per cent of the total area under the curve.

Remember that total probability always needs to equal 1. This remains true whether it is the binomial distribution that we saw with the coin flipping example in Chapter 14, or any other distribution. Consequently, the area under the curve of the normal distribution (i.e., under the curved line of Figure 15.4) must equal 1. When we say that the probability of sampling a value within 1 standard deviation of the mean is 0.682,

this also means that the *area* of this region under the curve equals 0.682 (i.e., the shaded area in Figure 15.4). And, again, because the whole area under the curve sums to 1, that must mean that the unshaded area of Figure 15.4 (where $x < \mu - \sigma$ or $x > \mu + \sigma$) has an area equal to $1 - 0.682 = 0.318$. That is, the probability of randomly sampling a value x in this region is $P(x < \mu - \sigma \cup x > \mu + \sigma) = 0.318$, or 31.8% (note that the $\cup$, is just a fancy way of saying 'or', in this case; technically, the *union* of two sets).

We can calculate other percentages using standard deviations too (Sokal and Rohlf, 1995). For example, about 95.4% of the probability density in a normal distribution lies between 2 standard deviations of the mean, i.e., $P(\mu - 2\sigma < x < \mu + 2\sigma) = 0.954$. And about 99.6% of the probability density in a normal distribution lies between 3 standard deviations of the mean, i.e., $P(\mu - 3\sigma < x < \mu + 3\sigma) = 0.996$. We could go on mapping percentages to standard deviations like this; for example, about 93.3% of the probability density in a normal distribution is less than $\mu + 1.5\sigma$ (i.e., less than 1.5 standard deviations greater than the mean; see Figure 15.5).

Figure 15.5.: A normal distribution in which the shaded region shows the area under 1.5 standard deviations of the mean (dotted line). This shaded area encompases about 93.3 per cent of the total area under the curve.

Notice that there are no numbers on the x-axes of Figure 15.4 or 15.5. This is deliberate; the relationship between standard deviations and probability density applies regardless of the scale. We could have a mean of $\mu = 100$ and standard deviation of $\sigma = 4$, or

$\mu = -12$ and $\sigma = 0.34$. It does not matter. Nevertheless, it would be very useful if we could work with some standard values of x when working out probabilities. This is where the standard normal distribution, first introduced in Chapter 14, becomes relevant. Recall that the standard normal distribution has a mean of $\mu = 0$ and a standard deviation (and variance) of $\sigma = 1$. With these standard values of μ and σ, we can start actually putting numbers on the x-axis and relating them to probabilities. We call these numbers **standard normal deviates**, or **z-scores** (Figure 15.6).

Figure 15.6.: A standard normal probability distribution with z-scores shown on the x-axis.

What z-scores allow us to do is map probabilities to deviations from the mean of a standard normal distribution (hence 'standard normal deviates'). We can say, e.g., that about 95% of the probability density lies between $z = -1.96$ and $z = 1.96$, or that about 99% lies between $z = -2.58$ and $z = 2.58$ (this will become relevant later). It is important to get a good sense of what this means, so we have written an interactive application (click here) that visually shows how probability density changes with changing z-score.

Click here for an interactive application to visualise z-scores

Of course, most variables that we measure in the biological and environmental sciences will not fit the standard normal distribution. Almost all variables will have a different mean and standard deviation, even if they are normally distributed. Nevertheless, we

can translate any normally distributed variable into a standard normal distribution by subtracting its mean and dividing by its standard deviation. We can see what this looks like visually in Figure 15.7.

Figure 15.7.: A visual representation of what happens when we subtract the sample mean from a dataset, then divide by its standard deviation. (a) A histogram (grey bars) show 10000 normally distributed values with a mean of 5 and a standard deviation of 2; the curved dotted line shows the standard normal distribution with a mean of 0 and standard deviation of 1. (b) Histogram after subtracting 5, then dividing by 2, from all values shown in panel (a).

In Figure 15.7a, we see the standard normal distribution curve represented by the dotted line, centered at $\mu = 0$ and with a standard deviation of $\sigma = 1$. To the right of this normal distribution we have 10000 values randomly sampled from a normal distribution with a mean of 5 and a standard deviation of 2 (note that the histogram peaks around 5 and is wider than the standard normal distribution because the standard deviation is higher). After subtracting 5 from all of the values in the histogram of Figure 15.7a, then dividing by 2, the data fit nicely within the standard normal curve, as shown in Figure 15.7b. By doing this transformation on the original dataset, z-scores can now be used with the data. Mathematically, here is how the calculation is made,

$$z = \frac{x - \mu}{\sigma}.$$

For example, if we had a value of $x = 9.1$ in our simulated dataset, in which $\mu = 5$ and

$\sigma = 2$, then we could calculate $z = (9.1 - 5)/2 = 2.05$. Since we almost never know what the true population mean (μ) and standard deviation (σ) are, we usually need to use the estimates made from our sample,

$$z = \frac{x - \bar{x}}{s}.$$

We could then use a statistical program such as Jamovi (The Jamovi Project, 2022), our interactive application, or an old-fashioned z-table[2] to find that only about 2% of values are expected to be higher than $x = 9.1$ in our original dataset. These z scores will become especially useful for calculating confidence intervals in Chapter 17. They can also be useful for comparing values from variables or statistics measured on different scales (Sokal and Rohlf, 1995; Cheadle et al., 2003; Adams and Collyer, 2016).

[2]Before the widespread availability of computers, which can easily be used to calculate probability densities on a normal distribution, the way to map z scores to probabilities was using a z table. The table would have rows and columns mapping to different z values, which could be used to find the appropriate probability densities. Such tables would be used for many different distributions, not just the normal distribution. The text Sokal and Rohlf (1995) comes with a nearly 200 page supplemental book that is just statistical tables. These tables are more or less obsolete nowadays, but some people still use them.

16. *Practical*. Probability and simulation

This practical focuses on applying the concepts from Chapter 14 and Chapter 15 in Jamovi (The Jamovi Project, 2022). There will be 3 exercises.

1. Calculating probabilities from a dataset.
2. Calculating probabilities from a normal distribution.
3. Demonstrating the central limit theorem (CLT).

To complete exercises 2 and 3, we will need to download and install two new Jamovi modules. Jamovi modules are add-ons that make it possible to run specialised statistical tools inside Jamovi. These tools are written by a community of statisticians, scientists, and educators and listed in the Jamovi library. Like Jamovi, these tools are open source and free to use.

The dataset for this practical is something a bit different. It comes from the Beacon Project, which is an interdisciplinary scientific research programme led by Dr Isabel Jones at the University of Stirling. This project focuses on large hydropower dams as a way to understand the trade-offs between different United Nations Sustainable Development Goals. It addresses challenging questions about environmental justice, biodiversity, and sustainable development.
The project works with people affected, and sometimes displaced, by dam construction in Brazil, Kazakhstan, India, USA, and the UK. Part of this project involves the use of mobile games to investigate how people make decisions about sustainable development.

Figure 16.1.: Welcome screen of the mobile game Power Up!

The game "Power Up!" is freely available as an Android and iPhone app (Figure 16.1). Data are collected from players' decisions and used to investigate social-ecological questions. We will use the power_up dataset in exercises 1 and 2 (right click on the links and choose 'Save Link As…', then save it with the extension '.csv'). To get started, first download the power_up dataset and open it in Jamovi. Note that these data are already in a tidy format, so we do not need to do any reorganising. The dataset includes columns for each player's ID, the OS that they use, the dam size that they decided to build in the game, their in-game investment in Biodiversity, Community, and Energy, and their final Score.

16.1. Probabilities from a dataset

Suppose that we want to estimate the probability that a new Power Up! game player will be an Android user. To estimate this probability, we can use the proportion of players in the dataset who are Android users. To get this proportion, we need to divide the number of Android users by the total number of players,

$$P(Android) = \frac{Number\ of\ Android\ users}{Number\ of\ players}.$$

In Jamovi, you could figure this out the long way by counting up the number of rows with 'Android' in the second column, then dividing by the total number of rows. But there is an easier way, which is faster and less prone to human error than manually tallying up items. To do this, go to the Analyses tab in Jamovi and navigate to Exploration, then Descriptives. Place the 'OS' variable in to the 'Variables' box. Next, find the check box called 'Frequency tables' just under the 'Split by' box and above the 'Statistics' drop down tab. Check this box to get a table of frequencies for Android versus iPhone users.

The table of frequencies shown in Figure 16.2 includes counts of Android versus iPhone users. We can see that 56 of the 74 total game players use Android, while 18 players use iPhone. To get the proportion of Android users, we could divide 56 by 74 to get 0.7567568. Similarly, for the proportion of iPhone users, we could calculate 18 / 74 = 0.2432432. But Jamovi already does this for us, with a bit of rounding. The second column of the Frequencies table gives us these proportions, but expressed as a percentage. The percentage of Android users is 75.7%, and the percentage of iPhone users is 24.3%. Percentages are out of a total of 100, so to get back to the proportions, we can just divide by 100%, 75.7 / 100 = 0.757 for Android and 24.3 / 100 = 0.243 for iPhone. To answer the original question, our best estimate of the probability that a new Power Up! game player will be an Android user is therefore 0.757.

Figure 16.2.: Jamovi Descriptives toolbar showing the OS column from the Power Up! dataset selected. The 'Frequency tables' checkbox builds a table of counts and percentages.

Next, use the same procedure to find the probability that a game player will make a small, medium, and large size dam. Now, fill in Table 16.1 with counts, percentage, and the estimated probability of a player selecting a small, medium, or large dam.

Table 16.1.: Statistics of Power Up! decisions for dam size.

Dam size	Counts	Percentage	Estimated Probability
Small			
Medium			
Large			

We can use these estimated probabilities of small, medium, and large dam size selection to predict what will happen in future games. Suppose that a new player decides to play the game. What is the probability that this player chooses a small **or** a large dam?

$P(small\ or\ large) =$ _____

Now suppose that 3 new players arrive and decide to play the game. What is the probability that all 3 of these new players choose a large dam?

$P(3\ large) =$ _____

What is the probability that the first player chooses a small dam, the second player chooses a medium dam, and the third player chooses a large dam?

$P(Player\ 1 = small, Player\ 2 = \ medium, Player\ 3 = large) =$ _____

Now consider a slightly different type of question. Instead of trying to predict the probability of new player decisions, we will focus on sampling from the existing power up

dataset. Imagine that you randomly choose one of the 74 players with equal probability (i.e., every player is equally likely to be chosen). What is the probability that you choose player 20?

$P(Player\ 20) =$ _____

What is the probability that you choose player 20, *then* choose a different player with a large dam? As a hint, remember that you are now sampling *without replacement*. The second choice cannot be player 20 again, so the probability of choosing a player with a large dam has changed from the estimated probability in Table 16.1.

$P(Player\ 20,\ Large) =$ _____

Now we can use the Descriptives tool in Jamovi to ask a slightly different question with the data. Suppose that we wanted to estimate the probability that an Android user will choose a large dam. We could multiply the proportion of Android users times the proportion of players who choose a large dam (i.e., find the probability of Android *and* large dam). But this assumes that the two characteristics are independent (i.e., that Android users are not more or less likely than iPhone users to build large dams). To estimate the probability that a player chooses a large dam *given* that they are using Android, we can keep Dam_size in the Variables box, but now put OS in the 'Split by' box. Figure 16.3 shows the output of Jamovi. A new frequency table breaks down dam choice for each OS.

Figure 16.3.: Jamovi Descriptives toolbar showing the dam size column from the Power Up! dataset selected as a variable split by OS. The 'Frequency tables' checkbox builds a table of counts for small, medium, and large dam size broken down by Android versus iPhone OS.

To get the proportion of Android users who choose to build a large dam, we just need to divide the number of Android users who chose the large dam size by the total number of Android users (i.e., sum of the first column in the Frequencies table; Figure 16.3). Note

that the vertical bar, |, in the equation below just means 'given' (or, rather, 'conditional up', so the number of players that chose a large dam *given* that they are Android users),

$$P(Large|Android) = \frac{Number\ of\ Android\ users\ choosing\ large\ dam}{Number\ of\ Android\ users}.$$

Now, recreate the table in Figure 16.3 and estimate the probability that an Android user will choose to build a large dam,

$P(Large|Android) =$ _____

Is $P(Large|Android)$ much different from the probability that *any* player chooses a large dam, as calculated in Table 16.1? Do you think that the difference is significant?

Next, we will move on to calculating probabilities from a normal distribution.

16.2. Probabilities from a normal distribution

In the example of the first exercise, we looked at OS and dam size choice. Players only use Android or iPhone, and they could only choose one of three sizes of dam. For these nominal variables, estimating the probability of a particular discrete outcome (e.g., Android versus iPhone) was just a matter of dividing counts. But we cannot use the same approach for calculating probabilities from continuous data. Consider, for example, the final score for each player in the column 'Score'. Because of how the game was designed, Score can potentially be any real number, although most scores are somewhere around 100. We can use a histogram to see the distribution of player scores (Figure 16.4).

In this case, it does not really make sense to ask what the probability is of a particular score. If the score can take *any* real value, out to as many decimals as we want, then what is the probability of a score being *exactly* 94.97 (i.e., 94.97 with infinite zeros after it, $94.97000000\bar{0}$)? The probability is infinitesimal, i.e., basically zero, because there are an infinite number of real numbers. Consequently, we are not really interested in the probabilities of specific values of continuous data. Instead, we want to focus on intervals. For example, what is the probability that a player scores higher than 120? What is the probability that a player scores lower than 100? What is the probability that a player scores between 100 and 120?

Figure 16.4.: Distribution of player scores in the game Power Up!

Take another look at Figure 16.4, then take a guess at each of these probabilities. As a hint, the y-axis of this histogram is showing density instead of frequency. What this means is that the total grey area (i.e., the histogram bars) sums to 1. Guessing the probability that a player scores higher than 120 is the same as guessing the proportion of grey space in the highest 4 bars of Figure 16.4 (i.e., grey space >120).

$P(Score > 120) =$ _____

$P(Score < 100) =$ _____

$P(100 < Score < 120) =$ _____

Trying to do this by looking at a histogram is not easy, and it is really not the best way to get the above probabilities. We can get much better estimates using Jamovi, but we need to make an assumption about the distribution of Player Score. Specifically, we need to assume that the distribution of Player Score has a specific shape. More technically, we must assume a specific probability density function that we can use to mathematically calculate probabilities of different ranges of player scores. Inspecting Figure 16.4, Player Score appears to be normally distributed. In other words, the shape of Player Score distribution appears to be normal, or 'Gaussian'. If we are willing to assume this, then we can calculate probabilities using its mean and standard deviation. Use Jamovi to find the mean and the standard deviation of player score (note, we can just say that score is unitless, so no need to include units).

Mean score: _____

Standard deviation score: _____

We will assume that the *sample* of scores shown in Figure 16.4 came from a *population* that is normally distributed with the mean and standard deviation that you wrote above (recall sample versus population from Chapter 4). We can overlay this distribution on the histogram above using a curved line (Figure 16.5).

Figure 16.5.: Distribution of player scores in the game Power Up! shown in histogram bars. The overlaid curve shows the probability density function for a normal distribution that has the same mean and standard deviation as the sample described by the histogram.

We can interpret the area under the curve in the same way that we interpret the area in the grey bars. As mentioned earlier, the total area of the histogram bars must sum to 1. The total area under the curve must also sum to 1. Both represent the probability of different ranges of player scores. Notice that the normal distribution is not a perfect match for the histogram bars. For example, the middle bar of values illustrating scores between 90 and 100 appears to be a bit low compared to a perfect normal distribution, and there are more scores between 40 and 50 than we might expect. Nevertheless, the two distributions broadly overlap, so we might be willing to assume that the player scores represented in the histogram bars are sampled from the population described by the curve.

Because the curve relating player score to probability density is described by an equation (see Chapter 14), we can use that equation to make inferences about the probabilities of different ranges of scores. The simplest example is the mean of the distribution. Because the normal distribution is symmetric, the area to the left of the mean must be the same

as the area to the right of the mean. And since the whole area under the curve must sum to 1, we can conclude that the probability of sampling a player score that is less than the mean is 1/2, and the probability of sampling a player score greater than the mean is also 1/2. Traditionally, we would need to do some maths to get other player score probabilities, but Jamovi can do this much more easily.

To get Jamovi to calculate probabilities from a normal distribution, we need to go to the Modules option and download a new module (Figure 16.6).

Figure 16.6.: Jamovi tool bar, which includes an option for downloading new Modules (right hand side)

Click on the 'Modules' button, and select the first option called 'jamovi library' from the pull-down menu. From the 'Available' tab, scroll down until you find the Module called 'distrACTION - Quantiles and Probabilities of Continuous and Discrete Distributions' (Rihs and Mayer, 2018). Click the 'Install' button to install it into Jamovi. A new button in the toolbar called 'distrACTION' should become visible (Figure 16.7).

Figure 16.7.: Jamovi tool bar, which includes an added module called distrACTION.

If the module is not there after installation, then it should be possible to find by again going to Modules and selecting distrACTION from the pulldown menu. Click on the module and choose 'Normal Distribution' from the pulldown menu. Next, we can see a box for the mean and standard deviation (SD) under the 'Parameters' subtitle in bold. Put the mean and the standard deviation calculated from above into these boxes. In the panel on the right, Jamovi will produce the same normal distribution that is in Figure 16.5 (note that the axes might be scaled a bit differently).

Given this normal distribution, we can compute the probability that a player scores less than x1 = 80 by checking the box 'Compute probability', which is located just under 'Function' (Figure 16.8). We can then select the first radio button to find the probability that a randomly sampled value X from this distribution is less than x1, $P(X \leq x1)$. Notice in the panel on the right that the probability is given as $P = 0.238$. This is also represented in the plot of the normal distribution, with the same proportion in the lower part of the distribution shaded ($P = 0.238$, i.e., about 23.8 per cent).

Figure 16.8.: Jamovi options for the distrACTION module for computing probability for a given normal distribution. The example shown here calculates the probability that a value sampled from the normal distribution of interest is less than 80.

To find the probability that a value is greater than 80, we could subtract our answer of 0.238 from 1, 1 - 0.238 = 0.762 (remember that the total area under the normal curve equals 1, so the shaded plus the unshaded region must also equal 1; hence, 1 minus the shaded region gives us the unshaded region). We could also just select the second radio button for $P(X \geq x1)$. Give this a try, and notice that the shaded and unshaded regions have flipped in the plot, and we get our answer in the table of 0.762.

Finally, to compute the probability of an interval, we can check the third radio button and set x2 in the bottom box (Figure 16.8). For example, to see the probability of a score between 80 and 120, we can choose select $P(x1 \leq X \leq x2)$, then set $x2 = 120$ in the bottom box. Notice where the shaded area is in the newly drawn plot. What is the probability of a player getting a score between 80 and 120?

$P(80 \leq X \leq 120) =$ _____

What is the probability of a player getting a score greater than 130?

$P(X \geq 130) =$ _____

Now try the following probabilities for different scores.

$P(X \geq 120) =$ _____

$P(X \leq 100) =$ _____

$P(100 \leq X \leq 120) =$ _____

Note, these last 3 were the same intervals that you guessed using the histogram. How close was your original guess to the calculations above?

One last one. What is the probability of a player getting a score lower than 70 or higher than 130?

$P(X \leq 70 \cup X \geq 130) =$ _____

There is more than one way to figure this last one out. How did you do it, and what was your reasoning?

We will now move on to the central limit theorem.

16.3. Central limit theorem

To demonstrate the central limit theorem, we need to download and install another module in Jamovi. This time, go to 'Modules', and from the 'Available' tab, scroll down until you find 'Rj' in the Jamovi library. Install 'Rj', then a new button 'R' should become available in the toolbar. This will allow us to run a bit of script using the coding language R. We will work with R a bit more in future practicals, but for now you will not need to do anymore than copying and pasting. For now, click on the new 'R' button in the toolbar and select 'Rj Editor' from the pulldown menu. You will see an open editor; this is where the code will go. If it has some code in it already (e.g., `# summary(data[1:3])`), just delete it so that we can start with a clean slate. Copy and paste the following lines into the Rjeditor.

```
v1  <- runif(n = 200, min = 0, max = 100);
v2  <- runif(n = 200, min = 0, max = 100);
v3  <- runif(n = 200, min = 0, max = 100);
v4  <- runif(n = 200, min = 0, max = 100);
v5  <- runif(n = 200, min = 0, max = 100);
v6  <- runif(n = 200, min = 0, max = 100);
v7  <- runif(n = 200, min = 0, max = 100);
v8  <- runif(n = 200, min = 0, max = 100);
```

```
v9  <- runif(n = 200, min = 0, max = 100);
v10 <- runif(n = 200, min = 0, max = 100);
v11 <- runif(n = 200, min = 0, max = 100);
v12 <- runif(n = 200, min = 0, max = 100);
v13 <- runif(n = 200, min = 0, max = 100);
v14 <- runif(n = 200, min = 0, max = 100);
v15 <- runif(n = 200, min = 0, max = 100);
v16 <- runif(n = 200, min = 0, max = 100);
v17 <- runif(n = 200, min = 0, max = 100);
v18 <- runif(n = 200, min = 0, max = 100);
v19 <- runif(n = 200, min = 0, max = 100);
v20 <- runif(n = 200, min = 0, max = 100);
v21 <- runif(n = 200, min = 0, max = 100);
v22 <- runif(n = 200, min = 0, max = 100);
v23 <- runif(n = 200, min = 0, max = 100);
v24 <- runif(n = 200, min = 0, max = 100);
v25 <- runif(n = 200, min = 0, max = 100);
v26 <- runif(n = 200, min = 0, max = 100);
v27 <- runif(n = 200, min = 0, max = 100);
v28 <- runif(n = 200, min = 0, max = 100);
v29 <- runif(n = 200, min = 0, max = 100);
v30 <- runif(n = 200, min = 0, max = 100);
v31 <- runif(n = 200, min = 0, max = 100);
v32 <- runif(n = 200, min = 0, max = 100);
v33 <- runif(n = 200, min = 0, max = 100);
v34 <- runif(n = 200, min = 0, max = 100);
v35 <- runif(n = 200, min = 0, max = 100);
v36 <- runif(n = 200, min = 0, max = 100);
v37 <- runif(n = 200, min = 0, max = 100);
v38 <- runif(n = 200, min = 0, max = 100);
v39 <- runif(n = 200, min = 0, max = 100);
v40 <- runif(n = 200, min = 0, max = 100);

hist(x = v1, main = "", xlab = "Random uniform variable");
```

What this code is doing is creating 40 different datasets of 200 random numbers from 0 to 100 (there is a way to do all of this in much fewer lines of code, but it requires a bit more advanced use of R). The `hist` function plots a histogram of the first variable. To run the code, find the green triangle in the upper right (Figure 16.9).

When you run the code, the 40 new variables will be created, each variable being made up of 200 random numbers. The histogram for `v1` is plotted to the right (to plot other variables, substitute `v1` in the `hist` function for some other variable). How would you describe the shape of the distribution of `v1`?

Figure 16.9.: Jamovi interface for the Rj Editor module. Code can be run by clicking on the green triangle in the upper right.

Next, we are going to get the mean value of each of the 40 variables. To do this, copy the code below and paste it at the bottom of the Rj Editor (somewhere below the `hist` function).

```
m1  <- mean(v1);
m2  <- mean(v2);
m3  <- mean(v3);
m4  <- mean(v4);
m5  <- mean(v5);
m6  <- mean(v6);
m7  <- mean(v7);
m8  <- mean(v8);
m9  <- mean(v9);
m10 <- mean(v10);
m11 <- mean(v11);
m12 <- mean(v12);
m13 <- mean(v13);
m14 <- mean(v14);
m15 <- mean(v15);
m16 <- mean(v16);
m17 <- mean(v17);
m18 <- mean(v18);
m19 <- mean(v19);
m20 <- mean(v20);
```

```
m21 <- mean(v21);
m22 <- mean(v22);
m23 <- mean(v23);
m24 <- mean(v24);
m25 <- mean(v25);
m26 <- mean(v26);
m27 <- mean(v27);
m28 <- mean(v28);
m29 <- mean(v29);
m30 <- mean(v30);
m31 <- mean(v31);
m32 <- mean(v32);
m33 <- mean(v33);
m34 <- mean(v34);
m35 <- mean(v35);
m36 <- mean(v36);
m37 <- mean(v37);
m38 <- mean(v38);
m39 <- mean(v39);
m40 <- mean(v40);

all_means <- c(m1,  m2,  m3,  m4,  m5,  m6,  m7,  m8,  m9,  m10,
               m11, m12, m13, m14, m15, m16, m17, m18, m19, m20,
               m21, m22, m23, m24, m25, m26, m27, m28, m29, m30,
               m31, m32, m33, m34, m35, m36, m37, m38, m39, m40);
```

Now we have calculated the mean for each variable. The last line of code defines `all_means`, which makes a new dataset that includes the mean value of each of our original variables. Think about what you think the distribution of these mean values will look like. Sketch what you predict the shape of its distribution will be below.

Now, add one more line of code to the very bottom of the Rj Editor.

```
hist(x = all_means, main = "", xlab = "All variable means");
```

This last line will make a histogram of the means of all 40 variables. Click the green button again to run the code. Compare the distribution of the original `v1` to the means of variables 1-40, and to your prediction above. Is this what you expected? As best you can, explain why the shapes of the two distributions differ.

We did all of this the long way to make it easier to see and think about the relationship between the original, uniformly distributed, variables and the distribution of their means. Now, we can repeat this more quickly using one more Jamovi module. Go to 'Modules', and from the 'Available' tab, download the 'clt - Demonstrations' module from the Jamovi library. Once it is downloaded, go to the 'Demonstrations' button in the Jamovi toolbar and select 'Central Limit Theorem' from the pulldown menu.

Figure 16.10.: Jamovi interface for the 'Demonstrations' module, which allows users to randomly generate data from a specific source distribution (normal, uniform, geometric, lognormal, and binary), sample size, and number of trials (i.e., variables)

To replicate what we did in the Rjeditor above, we just need to set the 'Source distribution' to 'uniform' using the pulldown menu, set the sample size to 200, and set the number of trials to 40 (Figure 16.10). Try doing this, then look at the histogram generated to the lower right. It should look similar, but not identical, to the histogram produced with the R code. Now try increasing the number of trials to 200. What happens to the histogram? What about when you increase the number of trials to 2000?

Try playing around with different source distributions, sample sizes, and numbers of trials. What general conclusion can you make about the distribution of sample means from the different distributions?

Part V.

Statistical inference

Week 5 Overview

Dates	20 February 2023 - 24 February 2023
Reading	**Required:** SCIU4T4 Workbook chapters 17 and 18 **Recommended:** None **Suggested:** Fedor-Freybergh and Mikulecký (2006) (Download) **Advanced:** None
Lectures	5.1: Some background for confidence intervals (6:36 min; Video) 5.2: Recap of z-scores (10:47 min; Video) 5.3: Confidence interval for the population mean (14:08 min Video) 5.4: The t-interval (10:24 min; Video) 5.5: Confidence interval for the population proportion (6:37 min; Video)
Practical	z- and t- intervals (Chapter 19) Room: Cottrell 2A17 Group A: 22 FEB 2023 (WED) 13:05-15:55 Group B: 23 FEB 2023 (THU) 09:05-11:55
Help hours	Ian Jones Room: Cottrell 1A13 24 FEB 2023 (FRI) 15:05-17:55
Assessments	Week 5 Practice quiz on Canvas

Week 5 focuses making statistical inferences using confidence intervals (CIs) and and introduces the t-interval.

Chapter 17 introduces what confidence intervals are and how to calculate them for normally and binomially distributed variables.

Chapter 18 introduces the t-interval and explains why this interval is usually necessary for calculating confidence intervals.

Chapter 19 guides you through the week 5 practical. The aim of this practical is to practice working with intervals and calculating confidence intervals.

17. Confidence intervals (CIs)

In Chapter 15, we saw how it is possible to calculate the probability of sampling values from a specific interval of the normal distribution (e.g., the probability of sampling a value within 1 standard deviation of the mean). In this chapter, we will see how to apply this knowledge to calculating intervals that express confidence in the mean value of a population.

Remember that we almost never really know the true mean value of a *population*, μ. Our best estimate of μ is the mean that we have calculated from a *sample*, $\bar{x}$ (see Chapter 4 for a review of the difference between populations and samples). But how good of an estimate is $\bar{x}$ of μ, really? Since we cannot know μ, one way of answering this question is to find an interval that expresses a degree of confidence about the value of μ. The idea is to calculate 2 numbers that we can say with some degree of confidence that μ is between (i.e., a lower confidence interval and an upper confidence interval). The wider this interval is, the more confident that we can be that the true mean μ is somewhere within it. The narrower the interval is, the less confident we can be that our confidence intervals (CIs) contain μ.

Confidence intervals are notoriously easy to misunderstand. We will explain this verbally first, focusing on the general ideas rather than the technical details. Then we will present the calculations before coming back to their interpretation again. The idea follows a similar logic to the standard error from Chapter 12.

Suppose that we want to know the mean body mass of all domestic cats (Figure 17.1). We cannot weigh every living cat in the world, but maybe we can find enough to get a sample of 20. From these 20 cats, we want to find some interval of masses (e.g., 3.9-4.3 kg) within which the *true* mean mass of the population is contained. The only way to be 100% certain that our proposed interval *definitely* contains the true mean would be to make the interval absurdly large. Instead, we might more sensibly ask what the interval would need to be to contain the mean with 95% confidence. What does "with 95% confidence" actually mean? It means when we do the calculation to get the interval, the true mean should be somewhere within the interval 95% of the time that a sample is collected.

In other words, if we were to go back out and collect another sample of 20 cats, and then another, and another (and so forth), calculating 95% CIs each time, then in 95% of our samples the true mean will be within our CIs (meaning that 5% of the time it will be outside the CIs). Note that this is slightly different than saying that there is

Figure 17.1.: Two domestic cats sitting side by side with much different body masses.

a 95% probability that the true mean is between our CIs.[1] Instead, the idea is that if we were to repeatedly resample from a population and calculate CIs each time, then 95% of the time the true mean would be within our CIs (Sokal and Rohlf, 1995). If this idea does not make sense at first, that is okay. The calculation is actually relatively straightforward, and we will come back to the statistical concept again afterwards to interpret it. First we will look at CIs assuming a normal distribution, then the special case of a binomial distribution.

17.1. Normal distribution CIs

Remember from the Central Limit Theorem in Chapter 15 that as our sample size N increases, the distribution of our sample mean $\bar{x}$ will start looking more and more like a normal distribution. Also from Chapter 15, we know that we can calculate the

[1] The reason that these two ideas are different has to do with the way that probability is defined in the frequentist approach to statistics (see Chapter 14). With this approach, there is no way to get the probability of the true mean being within an interval, strictly speaking. Other approaches to probability, such as Bayesian probability, do allow you to build intervals in which the true mean is contained with some probability. These are called "credible intervals" rather than "confidence intervals" (e.g., Ellison, 2004). The downside to credible intervals (or not, depending on your philosophy of statistics) is that Bayesian probability is at least partly subjective; i.e., based in some way on the subjective opinion of the individual researcher.

probability associated with any interval of values in a normal distribution. For example, we saw that about 68.2% of the probability density of a normal distribution is contained within a standard deviation of the mean. We can use this knowledge from Chapter 15 to set confidence intervals for any percentage of values around the sample mean ($\bar{x}$) using a standard error (SE) and z-score (z). Confidence intervals include 2 numbers. The **lower confidence interval** (LCI) is below the mean, and the **upper confidence interval** (UCI) is above the mean. Here is how they are calculated,

$$LCI = \bar{x} - (z \times SE),$$

$$UCI = \bar{x} + (z \times SE).$$

Note that the equations are the same, except that for the LCI, we are subtracting $z \times SE$, and for the UCI we are adding it. The specific value of z determines the confidence interval that we are calculating. For example, about 95% of the probability density of a standard normal distribution lies between $z = -1.96$ and $z = 1.96$ (Figure 17.2). Hence, if we use $z = 1.96$ to calculate LCI and UCI, we would be getting 95% confidence intervals around our mean.

Figure 17.2.: A standard normal probability distribution showing 95 per cent of probability density surrounding the mean.

An interactive application helps visualise the relationship between probability intervals and z-scores more generally (make sure to set 'Tailed' to 'Two-tailed' using the pulldown menu).

> Click here for an interactive application demonstrating the relationship between probability intervals and z-scores.

Now suppose that we want to calculate 95% CIs around the sample mean of our $N = 20$ domestic cats from earlier. We find that the mean body mass of cats in our sample is $\bar{x} = 4.1$ kg, and that the standard deviation is $s = 0.6$ kg (suppose that we are willing to assume, for now, that $s = \sigma$; that is, we know the true standard deviation of the population). Remember from Chapter 12 that the sample standard error can be calculated as $s/\sqrt{N}$. Our lower 95% confidence interval is therefore,

$$LCI_{95\%} = 4.1 - \left(1.96 \times \frac{0.6}{\sqrt{20}}\right) = 3.837$$

Our upper 95% confidence interval is,

$$UCI_{95\%} = 4.1 + \left(1.96 \times \frac{0.6}{\sqrt{20}}\right) = 4.363$$

Our 95% CIs are therefore $LCI = 3.837$ and $UCI = 4.363$. We can now come back to the statistical concept of what this actually means. If we were to go out and repeatedly collect new samples of 20 cats, and do the above calculations each time, then 95% of the time our true mean cat body mass would be somewhere between the LCI and UCI.

Ninety-five per cent confidence intervals are the most commonly used in biological and environmental sciences. In other words, we accept that about 5% of the time (1 in 20 times), our confidence intervals will not contain the true mean that we are trying to estimate. Suppose, however, that we wanted to be a bit more cautious. We could calculate 99% CIs; that is, CIs that contain the true mean in 99% of samples. To do this, we just need to find the z-score that corresponds with 99% of the probability density of the standard normal distribution. This value is about $z = 2.58$, which we could find with the interactive application, a z table, some maths, or a quick online search[2]. Consequently, the upper 99% confidence interval for our example of cat body masses would be,

$$LCI_{99\%} = 4.1 - \left(2.58 \times \frac{0.6}{\sqrt{20}}\right) = 3.754$$

Our upper 99% confidence interval is,

[2] While it is always important to be careful when searching, typing "z score 99 percent confidence interval" will almost always get the intended result.

$$UCI_{99\%} = 4.1 + \left(2.58 \times \frac{0.6}{\sqrt{20}}\right) = 4.446$$

Notice that the confidence intervals became wider around the sample mean. The 99% CI is now 3.754-4.446, while the 95% CI was 3.837-4.363 This is because if we want to be more confident about our interval containing the true mean, we need to make a bigger interval.

We could make CIs using any percentage that we want, but in practice it is very rare to see anything other than 90% ($z = 1.65$), 95% ($z = 1.96$), or 99% ($z = 2.58$). It is useful to see what these different intervals actually look like when calculated from actual data, so this interactive application illustrates CIs on a histogram with red dotted lines next to the LCI and UCI equations.

Click here for an interactive application demonstrating confidence intervals.

Unfortunately, the CI calculations from the this section are a bit of an idealised situation. We assumed that the sample means are normally distributed around the population mean. While we know that this *should* be the case as our sample size increases, it is not quite true when our sample is small. In practice, what this means is that our z-scores are usually not going to be the best values to use when calculating CIs, although they are often good enough when a sample size is large[3]. We will see what to do about this in Chapter 18, but first we turn to the special case of how to calculate CIs from binomial proportions.

17.2. Binomial distribution CIs

For a binomial distribution, our data are counts of successes and failures (see Chapter 14). For example, we might flip a coin 40 times and observe 22 heads and 18 tails. Suppose that we do not know in advance the coin is fair, so we cannot be sure that the probability of it landing on heads is $p = 0.5$. From our collected data, our estimated probability of landing on heads is, $\hat{p} = 22/40 = 0.55$.[4] But how would we calculate the CIs around this estimate? In this case, the formula is similar to ones for LCI and UCI from the normal distribution shown earlier. We just need to note that the variance of p for a binomial distribution is $\sigma^2 = p\,(1-p)$ (Box et al., 1978; Sokal and Rohlf, 1995).[5]

[3] What defines a 'small' or a 'large' sample is a bit arbitrary. A popular suggestion (e.g., Sokal and Rohlf, 1995, page 145) is that any $N < 30$ is too small to use z-scores, but any cut-off N is going to be somewhat arbitrary. Technically, the z score is not **completely** accurate until $N \to \infty$, but for all intents and purposes, it is usually only trivially inaccurate for sample sizes in the hundreds. Fortunately, you do not need to worry about any of this when calculating CIs from continuous data in Jamovi because Jamovi applies a correction for you, which we will look at in Chapter 18.

[4] The hat over the P, ($\hat{P}$) is just being used here to indicate the *estimate* of $P(heads)$, rather than the *true* $P(heads)$.

[5] Note, the variance of total *successes* is simply $np\,(1-p)$; i.e., just multiply the variance of p by n.

This means that the standard deviation of p is $\sigma = \sqrt{p\,(1-p)}$, and p has a standard error,

$$SE(p) = \sqrt{\frac{p\,(1-p)}{N}}.$$

We can use this standard error in the same equation from earlier for calculating confidence intervals. For example, if we wanted to calculate the lower 95% CI for $\hat{p} = 0.55$,

$$LCI_{95\%} = 0.55 - 1.96\sqrt{\frac{0.55\,(1-0.55)}{40}} = 0.396$$

Similarly, to calculate the upper 95% CI,

$$UCI_{95\%} = 0.55 + 1.96\sqrt{\frac{0.55\,(1-0.55)}{40}} = 0.704.$$

Our conclusion is that, based on our sample, 95% of the time we flip a coin 40 times, the true mean p will be somewhere between 0.396 and 0.704. These are quite wide CIs, which suggests that our flip of $\hat{p} = 0.55$ would not be particularly remarkable even if the coin was fair ($p = 0.5$).[6]

We can do another example, this time with our example of the probability of testing positive for Covid-19 at $\hat{p} = 0.025$. Suppose this value of $\hat{p}$ was calculated from a survey of 400 people ($N = 400$). We might want to be especially cautious about estimating CIs around such an important probability, so perhaps we prefer to use 99% CIs instead of 95% CIs. In this case, we use $z = 2.58$ as with the normal distribution example from earlier. But we apply this z score using the binomial standard error to get the LCI,

$$LCI_{99\%} = 0.025 - 2.58\sqrt{\frac{0.025\,(1-0.025)}{400}} = 0.00486$$

Similarly, we get the UCI,

$$UCI_{99\%} = 0.025 + 2.58\sqrt{\frac{0.025\,(1-0.025)}{400}} = 0.0451.$$

[6] You might ask, why are we doing all of this for the binomial distribution? The central limit theorem is supposed to work for the mean of any distribution, so should that not include the distribution of p too? Can we not just indicate success (heads) with a 1 and failures (tails) with a 0, then estimate the standard error of 22 values of 1 and 18 values of 0? Well, yes! That actually does work and gives an estimate of 0.079663, which is very close to the $\sqrt{\hat{p}(1-\hat{p})/N} = 0.078661$. The problem arises when the sample size is low, or when p is close to 0 or 1, and we are trying to map the z score to probability density. For this reason, it is best to stick with $\sqrt{\hat{p}(1-\hat{p})/N}$. There are other methods that attempt to give even better estimates (the one we are using is called the Wald method), but we will not consider these here.

Notice that the LCI and UCI differ here by about an order of magnitude (i.e., the UCI is about 10 times higher than the LCI).

In summary, this chapter has focused on what confidence intervals are and how to calculate them. Chapter 18 will turn to the t-interval, what it is and why it is used.

18. The t-interval

Chapter 14 introduced the binomial, Poisson, uniform, and normal distributions. In this chapter, we introduce another distribution, the t-distribution. Unlike the distributions of Chapter 14, the t-distribution arises from the need to make accurate statistical inferences, not from any particular kind of data (e.g., successes or failures in a binomial distribution, or events happening over time in a Poisson distribution). In Chapter 17, we calculated confidence intervals (CIs) using the normal distribution and z-scores. In doing so, we made the assumption that the sample standard deviation (s) was the same as the population standard deviation (σ), $s = \sigma$. In other words, we assumed that we knew what σ was, which is almost never true. For large enough sample sizes (i.e., high N), this is not generally a problem, but for lower sample sizes we need to be careful.

If there is a difference between s and σ, then our CIs will also be wrong. More specifically, the uncertainty between our sample estimate (s) and the true standard deviation (σ) is expected to increase the deviation of our sample mean ($\bar{x}$) from the true mean (μ). This means that if we are using the *sample* standard deviation instead of the *population* standard deviation (which is pretty much always), then the shape of the standard normal distribution from Chapter 17 (Figure 17.2) will be wrong. The correct shape will be "wider and flatter" (Sokal and Rohlf, 1995), with more probability density at the extremes and less in the middle of the distribution (Box et al., 1978). What this means is that if we use z-scores when calculating CIs using s, our CIs will not be wide enough, and we will think that we have more confidence in the mean than we really do. Instead of using the standard normal distribution, we need to use a t-distribution[1].

The difference between the standard normal distribution and t-distribution depends on our sample size, N. As N increases, we become more confident that the sample variance will be close to the true population variance (i.e., the deviation of s^2 from σ^2 decreases). At low N, our t-distribution is much wider and flatter than the standard normal distribution. As N becomes large[2], the t-distribution becomes basically indistinguishable

[1] This is also called the "Student's t-distribution". It was originally discovered by the head brewer of Guinness in Dublin in the early 20th century (Box et al., 1978). The brewer, W. S. Gosset, published under the pseudonym "A. Student" because Guinness had a policy of not allowing employees to publish (Miller and Miller, 2004).

[2] How large N needs to be for the t-distribution to considered close enough to the normal distribution is subjective. The two distributions get closer and closer as $N \to \infty$, but Sokal and Rohlf (1995) suggest that they are indistinguishable for all intents and purposes once $N > 30$. It is always safe to use the t-distribution when calculating confidence intervals, which is what all statistical programs such as Jamovi or R will do by default, so there is no need to worry about these kinds of arbitrary cutoffs in this case.

from the standard normal distribution. For calculating CIs from a sample, especially for small sample sizes, it is therefore best to use t-scores instead of z-scores. The idea is the same; we are just multiplying the standard errors by a different constant to get our CIs. For example, in Chapter 17, we multiplied the standard error of 20 cat masses by $z = 1.96$ because 95% of the probability density lies between $z = -1.96$ and $z = 1.96$ in the standard normal distribution. In truth, we should have multiplied by -2.093 because we only had a sample size of $N = 20$. Figure 18.1 shows the difference between the standard normal distribution and the more appropriate t-distribution[3].

Note that in Figure 18.1, a t-distribution with 19 degrees of freedom (df) is shown. The t-distribution is parameterised using df, and we lose a degree of freedom when calculating s^2 from a sample size of $N = 20$, so $df = 20 - 1 = 19$ is the correct value (see Chapter 12 for a brief explanation). For calculating CIs, df will always be $N - 1$, and this will be taken care of automatically in statistical programs such as Jamovi and R[4] (The Jamovi Project, 2022; R Core Team, 2022).

Recall from Chapter 17 that our body mass measurements of 20 cats had a sample mean of $\bar{x} = 4.1$ kg and sample standard deviation of $s = 0.6$ kg. We calculated the lower 95% CI to be $LCI_{95\%} = 4.041$ and the upper 95% CI to be $UCI_{95\%} = 4.159$. We can now repeat the calculation using the t-score 2.093 instead of the z-score 1.96. Our corrected lower 95% CI is,

$$LCI_{95\%} = 4.1 - \left(2.093 \times \frac{0.6}{\sqrt{20}}\right) = 3.819$$

Our upper 95% confidence interval is,

$$UCI_{95\%} = 4.1 + \left(2.093 \times \frac{0.6}{\sqrt{20}}\right) = 4.381$$

[3]We can define the t-distribution mathematically (Miller and Miller, 2004), but it is an absolute beast,

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v}\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}}.$$

In this equation, v is the degrees of freedom. The $\Gamma()$ is called a "gamma function", which is basically the equivalent of a factorial function, but for any number z (not just integers), such that $\Gamma(z+1) = \int_0^\infty x^z e^{-x} dx$ (where $z > -1$, or, even more technically, the real part of $z > -1$). If z is an integer n, then $\Gamma(n+1) = n!$ (Borowski and Borwein, 2005). What about the rest of the t probability density function? Why is it all so much? The reason is that it is the result of 2 different probability distributions affecting t independently, a standard normal distribution and a Chi-square distribution (Miller and Miller, 2004). We will look at the Chi-square in Week 9. Suffice to say that underlying mathematics of the t-distribution is not important for our purposes in applying statistical techniques.

[4]Another interesting caveat, which Jamovi and R will take care of automatically (so we do not actually have to worry about it), is that when we calculate s^2 to map t-scores to probability densities in the t-distribution, we multiply the sum of squares by $1/N$ instead of $1/(N-1)$ (Sokal and Rohlf, 1995). In other words, we no longer need to correct the sample variance s^2 to account for bias in estimating σ^2 because the t-distribution takes care of this for us.

Figure 18.1.: A standard normal probability distribution showing 95 per cent of probability density surrounding the mean (grey). On top of the standard normal distribution in grey, red dotted lines show a t-distribution with 19 degrees of freedom. Red shading shows 95 per cent of the probability density of the t-distribution.

The confidence intervals have not changed too much. By using the t-distribution, the LCI changed from 3.837 to 3.819, and the UCI changed from 4.363 to 4.381 In other words, we only needed our CIs to be a bit wider ($4.381 - 3.819 = 0.562$ for the using t-scores versus $4.363 - 3.837 = 0.526$ using z-scores). This is because a sample size of 20 is already large enough for the t-distribution and standard normal distribution to be very similar (Figure 18.1). But for lower sample sizes (N) and therefore fewer degrees of freedom ($df = N - 1$), the difference between the shapes of these distributions gets more obvious (Figure 18.2).

Figure 18.2.: A t-distribution with infinite degrees of freedom (df) is shown in black; this distribution is identical to the standard normal distribution. Other t-distributions with the same mean and standard deviation, but different degrees of freedom, are indicated by curves of different colours and line types.

The main point of Figure 18.2 is that as degrees of freedom decreases, the t-distribution becomes wider, with more probability density in the tails. Figure 18.2 is quite busy, so we have made an interactive application to make visualising the t-distribution easier.

Click here for an interactive application to visualise t-scores

Note that t-scores do not need to be used when making binomial confidence intervals. Using z-scores is fine.

The t-distribution is important throughout most of the rest of this module. It is not just used for calculating confidence intervals. The t-distribution also plays a critical role in hypothesis-testing, which is the subject of Week 6 and applied throughout the rest of the book. The t-distribution is therefore very important for understanding most of the statistical techniques presented in this book.

19. *Practical*. z- and t- intervals

This lab focuses on applying the concepts from Chapter 17 and Chapter 18 in Jamovi (The Jamovi Project, 2022). Specifically, we will practice calculating confidence intervals (CIs). There will be 4 exercises focusing on calculating confidence intervals in Jamovi. To complete the first 2 exercises, you will need the distrACTION module in Jamovi. We downloaded the distrACTION module in the Week 4 practical. If you need to download it again, the instructions to do this are in Chapter 16 Exercise 16.2 (briefly, go to the Modules option and select 'jamovi library', then scroll down until you find the 'distraACTION' module).

The data for this lab are inspired by ongoing work in the Woodland Creation and Ecological Networks (WrEN) project (Fuentes-Montemayor et al., 2022a,b). The Wren project is led by University of Stirling researchers Dr Elisa Fuentes-Montemayor, Dr Robbie Whytock, Dr Kevin Watts, and Prof Kirsty Park (https://www.wren-project.com/). It focuses on questions about what kinds of conservation actions should be prioritised to restore degraded ecological networks.

Figure 19.1.: Images from the WrEN project led by the University of Stirling

The WrEN project encompasses a huge amount of work and data collection from hundreds of surveyed secondary or ancient woodland sites. Here we will focus on observations of tree diameter at breast height (DBH) and grazing to calculate confidence intervals.

19.1. Confidence intervals with distrACTION

First, it is important to download the disrACTION module if it has not been downloaded already. If the distrACTION module has already been downloaded, it should appear in the toolbar of Jamovi (Figure 19.2) If it has not been downloaded, then see the instructions for downloading it with the 'Modules' option (see Figure 19.2) in Exercise 16.2.

Figure 19.2.: Jamovi tool bar, which includes an added module called distrACTION.

Once the distrACTION module has been made available, download the WrEN trees dataset wren_trees.xlsx dataset and open it in a spreadsheet. Notice that the dataset is not in a tidy format. There are 4 different sites represented by different columns in the dataset. The numbers under each column are measurements of tree diameter at breast height (DBH) in centimeteres. Before doing anything else, it is therefore necessary to put the WrEN dataset into a tidy format. The tidy dataset should include two columns, one for site and the other for DBH.

Once the WrEN trees dataset has been reorganised into a tidy format, save it as a CSV file and open it in Jamovi. In Jamovi, go to Exploration and Descriptives in the toolbar and build a histogram that shows the distribution of DBH. Do these data appear to be roughly normal? Why or why not?

Next, calculate the grand mean and standard deviation of tree DBH (i.e., the mean and standard deviation of trees across all sites).

Grand Mean: _____

Grand Standard Deviation: _____

We will use this mean and standard deviation to compute quantiles and obtain 95% z-scores. First, click on the distrACTION icon in the toolbar (see Figure 19.2). From the distrACTION pulldown menu, select 'Normal Distribution'. To the left, you should

see boxes to input parameter values for the mean and standard deviation (SD). Below the Parameters options, you should also see different functions for computing probability or quantiles. To the right, you should see a standard normal distribution (i.e., a normal distribution with a mean of 0 and a standard deviation of 1).

For this exercise, we will assume that the population of DBH from which our sample came is normally distributed. In other words, if we somehow had access to *all possible* DBH measurements in the woodland sites (not just the 120 trees sampled), we assume that DBH would be normally distributed. To find the probability of sampling a tree within a given interval of DBH (e.g., greater than 30), we therefore need to build this distribution with the correct mean and standard deviation. We do not know the *true* mean (μ) and standard deviation (σ) of the population, but our best estimate of these values are the mean ($\bar{x}$) and standard deviation (s) of the sample, as reported above (i.e., the grand mean and standard deviation). Using the Mean and SD parameter input boxes in distrACTION, we can build a normal distribution with the same mean and standard deviation as our sample. Do this now by inputting the calculated Grand Mean and Grand Standard Deviation from above in the appropriate boxes. Note that the normal distribution on the right has the same shape, but the table of parameters has been updated to reflect the mean mean and standard deviation.

In the previous practical from Chapter 16, we calculated the probability of sampling a value within a given interval of the normal distribution. If we wanted to do the same exercise here, we might find the probability of sampling a DBH < 30 using the Compute probability function (the answer is P = 0.265). Instead, we are now going to do the opposite using the Compute quantile(s) function. We might want to know, for example, what 75% of DBH values will be less than (i.e., what is the cutoff DBH, below which DBH values will be lower than this cutoff with a probability of 0.75). To find this, uncheck the 'Compute probability' box and check the 'Compute quantile(s)' box. Make sure that the 'cumulative quantile' radio button is selected, then set p = 0.75 (Figure 19.3).

From Figure 19.3, we can see that the cumulative 0.75 quantile is 44.3, so if DBH is normally distributed with the mean and standard deviation calculated above, 75% of DBH values in a population will be below 36.1 cm. Using the same principles, what is the cumulative 0.4 quantile for the DBH data?

Quantile: _____ cm

We can also use the Compute quantile(s) option in Jamovi to compute interval quantiles. For example, if we want to know the DBH values within which 95% of the probability density is contained, we can set p = 0.95, then select the radio button 'central interval quantiles'. Do this for the DBH data. From the Results table on the right, what interval of DBH values will contain 95% of the probability density around the mean?

Interval: _____ cm

Figure 19.3.: Jamovi interface for the 'distrACTION' module, in which quantiles have been computed to find the diameter at breast height (DBH) below which 75 per cent of DBHs will be given a normal distribution with a mean of 36.9 and standard deviation of 11. Data for these parameter values were collected from in Scotland as part of the Woodland Creation and Ecological Networks (WrEN) project.

Remember that we are looking at the full sample distribution of DBH. That is, getting intervals for the probability of sampling DBH values around the mean, *not* confidence intervals around the mean as introduced in Chapter 17. How would we get confidence intervals around the mean? That is, what if we want to say that we have 95% confidence that the *mean* lies between 2 values? We would need to use the standard deviation *of the sample mean* $\bar{x}$ around the true mean μ, rather than the sample standard deviation. Recall from Chapter 12.6 that the standard error is the standard deviation of $\bar{x}$ values around μ. We can therefore use the standard error to calculate confidence intervals around the mean value of DBH. From the Descriptives panel in Jamovi (recall that this is under the 'Exploration' button), find the standard error of DBH,

Std. error of Mean: _____

Now, go back to the distrACTION Normal Distribution and put the DBH mean into the parameters box as before. But this time, put the standard error calculated above into the box for SD. Next, choose the Compute quantile(s) option and set p = 0.95 to calculate a 95% confidence interval. Based on the Results table, what can you infer are the lower and upper 95% confidence intervals (CIs) around the mean?

Lower 95% CI: _____

Upper 95% CI: _____

Remember that this assumed that the sample means ($\bar{x}$) are normally distributed around the true mean (μ). But as we saw in Chapter 18, when we assume that our sample standard deviation (s) is the same as the population standard deviation (σ), then the shape of the normal distribution will be at least a bit off. Instead, we can get a more accurate estimate of CIs using a t-distribution. Jamovi usually does this automatically when calculating CIs outside of the distrACTION module. To get 95% CIs, go back to the Descriptives panel in Jamovi, then choose DBH (cm) as variable. Scroll down to the Statistics options and check 'Confidence interval for Mean' under the **Mean Dispersion** options, and make sure that the number in the box is 95 for 95% confidence. Confidence intervals will appear in the Descriptives table on the right. From this Descriptives table now, write the lower and upper 95% CIs below.

Lower 95% CI: _____

Upper 95% CI: _____

You might have been expecting a bit more of a difference, but remember, for sufficiently large sample sizes (around N = 30), the normal and t-distributions are very similar (see Chapter 18). We really do not expect much of a difference until sample sizes become small, which we will see in Exercise 19.3.

19.2. Confidence intervals from z- and t-scores

While Jamovi can be very useful for calculating confidence intervals from a dataset, you might also need to calculate CIs from just a set of summary statistics (e.g., the mean, standard error, and sample size). This activity will demonstrate how to calculate CIs from z- and t-scores. Recall the formula for lower and upper confidence intervals from Chapter 17.1,

$$LCI = \bar{x} - (z \times SE),$$

$$UCI = \bar{x} + (z \times SE).$$

We could therefore calculate 95% confidence intervals for DBH with just the sample mean ($\bar{x}$), z-score (z), and standard error (SE). We have already calculated $\bar{x}$ and SE for the DBH in Exercise 19.1 above, so we just need to figure out z. Recall that z-scores are *standard normal deviates*; that is, deviations from the mean given a standard normal distribution, in which the mean equals 0 and standard deviation equals 1. For example, $z = -1$ is 1 standard deviation below the mean of a standard normal distribution, and $z = 2$ is 2 standard deviations above the mean of a standard normal distribution. What values of z contain 95% of the probability density of a standard normal distribution? We can use the distrACTION module again to find this out. Select 'Normal Distribution' from the pulldown of the distrACTION module. Notice that by default, a standard normal distribution is already set (Mean = 0 and SD = 1). All that we need to do now is compute quantiles for p = 0.95. From these quantiles, what is the proper z-score to use in the equations for LCI and UCI above?

z-score: _____

Now, use the values of $\bar{x}$, z, and SE for DBH in the equations above to calculate lower and upper 95% confidence intervals again.

Lower 95% CI: _____

Upper 95% CI: _____

Are these confidence intervals the same as what you calculated in Exercise 19.1?

Lastly, instead of using the z-score, we can do the same with a t-score. We can find the appropriate t-score from the t-distribution in the distrACTION module. To get the t-score, click on the distrACTION module button and choose 'T-Distribution' from the pulldown. To get quantiles with the t-distribution, we need to know the degrees of

freedom (df) of the sample. Chapter 18 explains how to calculate df from the sample size N. What are the appropriate df for DBH?

df: _____

Put the df in the Parameters box. Ignore the box for lambda (λ); this is not needed. Under the **Function** options, choose 'Compute quantile(s)' as before to calculate Quantiles. From the Results table, what is the proper t-score to use in the equations for LCI and UCI?

t-score: _____

Again, use the values of $\bar{x}$, t, and SE for DBH in the equations above to calculate lower and upper 95% confidence intervals.

Lower 95% CI: _____

Upper 95% CI: _____

How similar are the estimates for lower and upper CIs when using z- versus t-scores. Reflect on any similarities or differences that you see in all of these different ways of calculating confidence intervals

19.3. Confidence intervals for different sample sizes (t- and z-)

In Exercises 19.1 and 19.2, the sample size of DBH was fairly large (N = 120). Now, we will calculate confidence intervals for the mean DBH of each of the 4 different sites using both z- and t-scores. These sites have much different sample sizes. From the Descriptives tool in Jamovi, write the sample sizes for DBH split by site below.

Site 1182: N = _____

Site 1223: N = _____

Site 3008: N = _____

Site 10922: N = _____

For which of these sites would you predict CIs calculated from z-scores versus t-scores to differ the most?

Site: _____

The next part of this exercise is self-guided. In Exercises 19.1 and 19.2, you used different approaches for calculating 95% CIs from the normal and t-distributions. Now, fill in the table below reporting 95% CIs calculated using each distribution from the 4 sites using any method you prefer.

Table 19.1.: 95 per cent Confidence intervals calculated for tree diameter at breast height (DBH) in cm. Data for these parameter values were collected in Scotland as part of the Woodland Creation and Ecological Networks (WrEN) project.

Site	N	95% CIs (Normal)	95% CIs (t-distribution)
1182			
1223			
3008			
10922			

Next, do the same, but now calculate 99% CIs instead of 95% CIs.

Table 19.2.: 99 per cent Confidence intervals calculated for tree diameter at breast height (DBH) in cm. Data for these parameter values were collected in Scotland as part of the Woodland Creation and Ecological Networks (WrEN) project.

Site	N	99% CIs (Normal)	99% CIs (t-distribution)
1182			
1223			
3008			
10922			

What do you notice about the difference between CIs calculated from the normal distribution versus the t-distribution across the different sites?

In your own words, based on this practical and what you have read from the lab workbook and any other material, what do these confidence intervals *actually mean*?

We will now move on to calculating confidence intervals for proportions.

19.4. Proportion confidence intervals

We will now try calculating confidence intervals for proportional data using the WrEN Sites dataset, which you can download here (right click on the links and choose 'Save Link As...', then save it with the extension '.csv'). Notice that there are more sites included than there were in the dataset used in Exercises 19.1-19.3, and that some of these sites are grazed while others are not (column 'Grazing'). From the Descriptives options, find the number of sites grazed versus not grazed (hint, remember from the lab practical in Chapter 16 to put 'Grazing' in the variable box and click the 'Frequency tables' checkbox).

Grazed: _____

Not Grazed: _____

From these counts above, what is the estimate (p, or more technically $\hat{p}$, with the hat indicating that it is an estimate) of the proportion of sites that are grazed?

p: _____

Chapter 17.2 explained how to calculate lower and upper CIs for binomial distributions (i.e., proportion data). To do this, we can use equations similar to the ones used for LCI and UCI from Exercise 19.2 above,

$$LCI = p - z \times SE(p),$$

$$UCI = p + z \times SE(p),$$

We have already calculated p, and we can find z-scores for confidence intervals in the same way that we did in Exercise 19.2 (i.e., the z-scores associated with 95% confidence intervals do not change just because we are working with proportions). All that leaves for calculating LCI and UCI are the standard errors of the proportions. Remember from Chapter 17.2 that these are calculated differently from a standard deviation of continuous values such as diameter breast height. The formula for standard error of a proportion is,

$$SE(p) = \sqrt{\frac{p\,(1-p)}{N}}.$$

We can estimate p using p, and N is the total sample size. Using the above equation, what is the standard error of p?

SE(p): _____

Using this standard error, what are the lower and upper 95% confidence intervals around p?

$LCI_{95\%}$ = _____

$UCI_{95\%}$ = _____

Next, find the lower and upper 99% CIs around p and report them below (hint: the only difference here from the calculation of the 95% CIs are the z-scores).

$LCI_{99\%}$ = _____

$UCI_{99\%}$ = _____

19.5. Another proportion confidence interval

If you have sufficient time during the lab practical, try one more proportional confidence interval. This time, find the 80%, 95%, and 99% CIs for the proportion of sites that are classified as Ancient woodland. First consider an 80% CI (hint, use the distrACTION module again to find the z-scores).

$LCI_{80\%}$ = _____

$UCI_{80\%}$ = _____

Next, calculate 95% CIs for the proportion of sites classified as Ancient woodland.

$LCI_{95\%}$ = _____

$UCI_{95\%}$ = _____

Finally, calculate 99% CIs for the proportion of sites classified as Ancient woodland.

$LCI_{99\%}$ = _____

$UCI_{99\%}$ = _____

Reflect again on what these values actually mean. For example, what does it mean to have 95% confidence that the proportion of sites classified as Ancient woodland are between two values? Are there any situations in which this might be useful, from a scientific or conservation standpoint? There is no right or wrong answer here, but confidence intervals are very challenging to understand conceptually, so having now done the calculations to get them, it is a good idea to think again about what they mean.

Part VI.

Hypothesis testing

Week 6 Overview

Dates	27 February 2023 - 03 MAR 2023
Reading	**Required:** SCIU4T4 Workbook chapters 20 and 21 **Recommended:** Navarro and Foxcroft (2022) Chapter 11 **Suggested:** None **Advanced:** Johnson (1995) (Download)
Lectures	6.1: What is hypothesis testing? (11:18 min; Video) 6.2: Making and using hypotheses (10:32 min; Video) 6.3: The example of the right-handed European toad (16:43 min; Video) 6.4: Errors, hypothesis testing and CI intervals (16:43 min; Video) 6.5: Student's t-distribution and 1 sample t-test (13:31 min; Video) 6.6: Independent and paired samples t-tests (18:21 min; Video) 6.7: Handling violations of assumptions (7:50 min; Video) 6.8: Non-parametric tests (13:30 min; Video)
Practical	Hypothesis testing and t-tests (Chapter 22) Room: Cottrell 2A17 Group A: 01 MAR 2023 (WED) 13:05-15:55 Group B: 02 MAR 2023 (THU) 09:05-11:55
Help hours	Martina Quaggiotto Room: Cottrell 1A13 03 MAR 2023 (FRI) 15:05-17:55
Assessments	Week 6 Practice quiz on Canvas

Week 6 introduces hypothesis testing, and how to use and interpret statistical tests that test whether or not the mean (or median) of a dataset is significantly different from some specific value, or whether two different groups in a dataset have the same mean (or median).

Chapter 20 introduces hypothesis testing, and what it means in statistics. This chapter discusses the general idea of hypothesis testing, null and alternative hypotheses, and how to correctly interpret p-values.

Chapter 21 introduces the t-test and its non-parametric alternatives. These tests include the one sample t-test, the independent samples t-test, the paired sample t-test, the Wilcoxon test, and the Mann-Whitney U test. The chapter also explains the assumptions underlying these different tests.

Chapter 22 guides you through the week 6 practical. The aim of this practical is to practice using and correctly interpreting the tests that are introduced in Chapter 21.

20. What is hypothesis testing?

Statistical hypotheses are different from scientific hypotheses. In science, a hypothesis should make some kind of testable statement about the relationship between two or more different concepts or observations (Bouma, 2000). For example, we might hypothesise that in a particular population of sparrows, juveniles that have higher body mass will also have higher survival rates. In contrast, statistical hypotheses compare a sample outcome to the outcome predicted given a relevant statistical distribution (Sokal and Rohlf, 1995). That is, we start with a hypothesis that our data are sampled from some distribution, then work out whether or not we should reject this hypothesis. This concept is counter-intuitive, but it is **absolutely fundamental for understanding the logic underlying most modern statistical techniques** (Sokal and Rohlf, 1995; Mayo, 1996; Greenland et al., 2016), including all subsequent chapters of this workbook, so we will focus on it here in-depth. The most instructive way to explain the general idea is with the example of coin flips (Mayo, 1996), as we looked at in Chapter 14.

20.1. How ridiculous is our hypothesis?

Imagine that a coin is flipped 100 times. We are told that the coin is fair, meaning that there is an equal probability of it landing on heads or tails (i.e., the probability is 0.5 for both heads and tails in any given flip). From Chapter 14.4.1, recall that the number of times out of 100 that the coin flip comes up heads will be described by a binomial distribution. The most probable outcome will be 50 heads and 50 tails, but frequencies that deviate from this perfect 50:50 ratio (e.g., 48 heads and 52 tails) are also expected to be fairly common (Figure 20.1).

The distribution in Figure 20.1 is what we expect to happen if the coin we are flipping 100 times is actually fair. In other words, it is the predicted distribution of outcomes *if our hypothesis that the coin is fair is true* (more on that later). Now, suppose that we actually run the experiment; we flip the coin in question 100 times. Perhaps we observe heads 30 times out of the 100 total flips. From the distribution in Figure 20.1, this result seems *very* unlikely if the coin is actually fair. If we do the maths, the probability of observing 30 heads or fewer (i.e., getting anywhere between 0 and 30 heads total) is only $P = 0.0000392507$. And the probability of getting this much of a deviation from 50 heads (i.e., either 20 less than or 20 more than 50) is $P = 0.0000785014$ (two times 0.0000392507, since the binomial distribution is symmetrical around 50). This seems a

Figure 20.1.: Probability distribution for the number of times that a flipped coin lands on heads in 100 trials. Note that some areas of parameter space on the x-axis are cut off because the pobabilities associated with this number of flips out of 100 being heads are so low.

bit ridiculous! Do we *really* believe that the coin is fair if the probability of getting a result this extreme is so low?

Getting 30 head flips is maybe a bit extreme. What if we flip the coin 100 times and get 45 heads? In this case, if the coin is fair, then we would predict this number heads or fewer with a probability of about $P = 0.0967$ (i.e., about 9.67% of the time, we would expect to get 45 or fewer heads). And we would predict a deviation as extreme as 5 from the 50:50 ratio of heads to tails with a probability of about $P = 0.193$ (i.e., about 19.3% of the time, we would get 45 heads or fewer, or 55 heads or more). This does not sound nearly so unrealistic. If a fair coin will give us this much of a deviation from the expected 50 heads and 50 tails about 20% of the time, then perhaps our hypothesis is not so ridiculous, and we can conclude the coin is indeed fair.

How improbable does our result need to be to cause us to reject our hypothesis that the coin is fair? There is no definitive answer to this question. In the biological and environmental sciences, we traditionally use a probability of 0.05, but this threshold is completely arbitrary[1]. All it means is that we are willing to reject our hypothesis (i.e., declare the coin to be unfair) when it is actually true (i.e., the coin really *is* fair) about 5% of the time. Note that we do need to decide on some finite threshold for rejecting our hypothesis because even extremely rare events, by definition, can sometimes happen. In the case of 100 coin flips, there is always a small probability of getting *any* number of heads from a fair coin (although getting zero heads would be extraordinarily rare, $P \approx 7.89 \times 10^{-31}$, i.e., a decimal followed by 30 zeros, then a 7). We can therefore never be *certain* about rejecting or not rejecting the hypothesis that we have a fair coin.

This was a very concrete example intended to provide an intuitive way of thinking about hypothesis testing in statistics. In the next section, we will look more generally at what hypothesis testing means in statistics and the terminology associated with it. But everything that follows basically relies on the same general logic as the coin-flipping example here; **if our hypothesis is true, then what is the probability of our result?**

20.2. Statistical hypothesis testing

A statistical test is used to decide if we should reject the hypothesis that some observed value or calculated statistic was sampled from a particular distribution (Sokal and Rohlf, 1995). In the case of the coin example in the previous section, the observed value was the number of heads, and the distribution was the binomial distribution. In other cases, we might, e.g., test the hypothesis that a value was sampled from a normal or t-distribution. In all of these cases, the hypothesis that we are testing is the **null hypothesis**, which we abbreviate as H_0 (e.g., the coin is fair). Typically, H_0 is associated

[1] I have heard many apocryphal stories about how a probability of 0.05 was decided upon, but I have no idea which, if any, of these stories are actually true.

with the lack of an interesting statistical pattern, such as when a coin is fair, when there is no difference between two groups of observations, or when two variables are not associated with each other. This null hypothesis contrasts an **alternative hypothesis**, which we abbreviate as H_A (e.g., the coin is not fair). Alternative hypotheses are always defined by some relationship to H_0 (Sokal and Rohlf, 1995). Typically, H_A is associated with something interesting happening, such as a biased coin, a difference between groups of observations, or an association between two variables. Table 20.1 below presents some null and alternative hypotheses that might be relevant in the biological or environmental sciences.

Table 20.1.: Hypothetical null and alternative hypotheses in the biological and environmental sciences.

Null hypothesis H_0	Alternative hypothesis H_A
There is no difference between juvenile and adult sparrow mortality	Mortality differs between juvenile and adult sparrows
Amphibian body size does not change with increasing latitude	Amphibian body size increases with latitude
Soil nitrogen concentration does not differ between agricultural and non-agricultural fields	Soil nitrogen concentration is lower in non-agricultural fields

Notice that alternative hypotheses can indicate direction (e.g., amphibian body size will increase, or nitrogen content will be lower in non-agricultural fields), or they can be non-directional (e.g., mortality will be different based on life-history stage). When our alternative hypothesis indicates direction, we say that the hypothesis is **one-sided**. This is because we are looking at one side of the null distribution. In the case of our coin example, a one-sided H_A might be that the probability of flipping heads is less than 0.5, meaning that we reject H_0 only given numbers on the left side of the distribution in Figure 20.1 (where the number of flips heads are fewer than 50). A different one-sided H_A would be that the probability of flipping heads is greater than 0.5, in which case we would reject H_0 only given numbers on the right side of the distribution. In contrast, when our alternative hypothesis does not indicate direction, we say that the hypothesis is **two-sided**. This is because we are looking at both sides of the null distribution. In the case of our coin example, we might not care in which direction the coin is biased (towards heads or tails), just that the probability of flipping heads does not equal 0.5. In this case, we reject H_0 at both extremes of the distribution of Figure 20.1.

20.3. P-values, false positives, and power

In our hypothetical coin flipping example, we used P to indicate the probability of getting a particular number of heads out of 100 total flips if our coin was fair. This P

(sometimes denoted with a lower-case p) is what we call a 'p-value'.

> **A p-value is the probability of getting a result as or more extreme than the one observed assuming H_0 is true.**[2]

This is separated and in bold because it is a very important concept in statistics, and it is one that is very, very easy to misinterpret[3]. A p-value is *not* the probability that the null hypothesis is true (we actually have no way of knowing this probability). It is also not the probability that an alternative hypothesis is false (we have no way of knowing this probability either). A p-value specifically *assumes that the null hypothesis is true*, then asks what the probability of an observed result would be *conditional upon this assumption*. In the case of our coin flipping example, we cannot really know the probability that the coin is fair or unfair (depending on your philosophy of statistics, this might not even make conceptual sense). But we can say that **if** the coin **is** fair, then an observation of ≤ 45 would occur with a probability of $P = 0.0967$.

Before actually calculating a p-value, we typically set a threshold level (α) below which we will conclude that our p-value is **statistically significant**[4]. As mentioned in section 20.1, we traditionally set $\alpha = 0.05$ in the biological and environmental sciences (although rarely $\alpha = 0.01$ is used). This means that if $P \leq 0.05$, then we reject H_0 and conclude that our observation is statistically significant. It also means that even when H_0 really is true (e.g., the coin is really fair), we will mistakenly reject H_0 with a probability of 0.05 (i.e., 5% of the time). This is called a **Type I error** (i.e., a false positive), and it typically means that we will infer a pattern of some kind (e.g., a difference between groups, or a relationship between variables) where none really exists. This is obviously an error that we want to avoid, which is why we set α to a low value.

In contrast, we can also fail to reject H_0 when H_A is actually true. That is, we might mistakenly conclude that there is no evidence to reject the null hypothesis when the null hypothesis really is false. This is called a **Type II error**. The probability that we commit a Type II error, i.e., that we fail to reject the null hypothesis when it is false, is given the symbol β. Since β is the probability that we fail to reject H_0 when it is false, $1 - \beta$ is the probability that that we *do* reject H_0 when it is false. This $1 - \beta$ is the **statistical power** of a statistical test. Note that α and β are not necessarily related to each other. Our α is whatever we set it to be (e.g., $\alpha = 0.05$). But statistical power will depend on the size of the effect that we are measuring (e.g., how much bias their is in a coin if we are testing whether or not it is fair), and on the size of our

[2] Technically, it also assumes that all of the assumptions of the model underlying the hypothesis test are true, but we will worry about this later.

[3] In fact, the p-value is so easy to mis-interpret and so widely mis-used, that some scientists have called for them to be abandoned entirely (Wasserstein and Lazar, 2016; but see Stanton-Geddes et al., 2014; Mayo, 2019).

[4] Like p-values, setting thresholds below which we consider P to be significant is at least somewhat controversial (McShane et al., 2019; Mayo, 2021). But the use of statistical significance thresholds is ubiquitous in the biological and environmental sciences, so we will use them throughout this lab workbook (it is important to understand them and interpret them).

sample. Increasing our sample size will always increase our statistical power, i.e., our ability to reject the null hypothesis when it is really false. Table 20.2 below illustrates the relationship between whether or not H_0 is true, and whether or not we reject it.

Table 20.2.: Summary of Type I and Type II errors in relation to a null hypothesis (H_0).

	Do not reject H_0	Reject H_0
H_0 is true	Correct decision	Type I error
H_0 is false	Type II error	Correct decision

Note that we never *accept* a null hypothesis; we just fail to reject it. Statistical tests are not really set up in a way that H_0 can be accepted[5]. The reason for this is subtle, but we can see the logic if we again consider the case of the fair coin. If H_0 is true, then the probability of flipping heads is $P(heads) = 0.5$ (i.e., $H_0 : P(heads) = 0.5$). But even if we fail to reject H_0, this does not mean that we can conclude with any real confidence that our null hypothesis $P(heads) = 0.5$ is true. What if we instead tested the null hypothesis that our coin was *very slightly* biased, such that $H_0 : P(heads) = 0.4999$? If we failed to reject the null hypothesis that $P(heads) = 0.5$, then we would probably also fail to reject a H_0 that $P(heads) = 0.4999$. There is no way to meaningfully distinguish between these two potential null hypotheses by just testing one of them. We therefore cannot conclude that a H_0 is correct; we can only find evidence to reject it. In contrast, we can reasonably accept an alternative hypothesis H_A when we reject H_0.

[5]Note that we might, for non-statistical reasons conclude the absence of a particular phenomenon or relationship between observations. For example, following a statistical test, we might become convinced that a coin really is fair, or that there is no relationship between sparrow body mass and survival. But these are conclusions about scientific hypotheses, not statistical hypotheses.

21. The t-test

A t-test is a simple and widely used statistical hypothesis test that relies on the t-distribution introduced in Chapter 18. In this chapter, we will look at 3 types of t-tests, (1) the one sample t-test, (2) the independent samples t-test, and (3) the paired samples t-test. We will also look at non-parametric alternatives to t-tests (Wilcoxon and Mann-Whitney tests), which become relevant when the assumptions of t-tests are violated. The use of all of these tests in Jamovi will be demonstrated in the lab practical in Chapter 22.

21.1. One sample t-test

Suppose that a biology teacher has created a new approach to teaching and wants to test whether or not their new approach results in student test scores that are higher than the reported national average of 60. This teacher should first define their null and alternative hypotheses.

- H_0: Student test scores equal 60
- H_A: Student test scores are greater than 60

Note that this is a one-sided hypothesis. The teacher is not interested in whether or not the mean test score of their students is below 60. They just want to find out if the mean test scores are greater than 60. Suppose the teacher has 10 students with the following test scores (out of 100).

```
49.3, 62.9, 73.7, 65.5, 69.6, 70.7, 61.5, 73.4, 61.1, 78.1
```

The teacher can use a one sample t-test to test H_0. The one sample t-test will test whether the sample mean of test scores ($\bar{y} = 66.58$) is significantly greater than the reported national average, $\mu_0 = 60$. How does this work? Recall from Chapter 15 that, due to the central limit theorem, the distribution of sample means ($\bar{y}$) will be normally distributed around the true mean μ as sample size N increases. At low N, when we need to estimate the true standard deviation (σ) from the sample standard deviation (s), we need to correct for a bias and use the t-distribution (see Chapter 18). The logic here is to use the t-distribution as the null distribution for $\bar{y}$. If we subtract μ_0 from $\bar{y}$, then we can centre the mean of the null distribution at 0. We can then divide by the standard

error of test scores so that we can compare the deviation of $\bar{y}$ from μ_0 in terms of the t-distribution. This is the same idea as calculating a z-score from Chapter 15.2. In fact, the equations look almost the same,

$$t_{\bar{y}} = \frac{\bar{y} - \mu_0}{SE(\bar{y})}$$

In the above equation, $SE(\bar{y})$ is the standard error of $\bar{y}$.

If the sample mean of test scores is really the same as the population mean $\mu_0 = 60$, then $\bar{y}$ should have a t-distribution. Consequently, very small or large values of $t_{\bar{y}}$ would suggest that the sample mean is improbable given the null distribution predicted if $H_0 : \mu_0 = \bar{y}$ is true. We can calculate $t_{\bar{y}}$ for our above sample (note, $SE(\bar{y}) = s/\sqrt{N} = 8.334373/\sqrt{10} = 2.63556$),

$$t_{\bar{y}} = \frac{66.58 - 60}{2.63556} = 2.496623.$$

Our t-statistic is therefore 2.496623 (note that a t-statistic can also be negative; this would just mean that our sample mean is less than μ_0, instead of greater than μ_0, but nothing about the t-test changes if this is the case). We can see where this value falls on the t-distribution with 9 degrees of freedom in Figure 21.1.

The t-distribution in Figure 21.1 is the probability distribution if H_0 is true (i.e., the student test scores were sampled from a distribution with a mean of $\mu_0 = 60$). The arrow pointing to the calculated $t_{\bar{y}} = 2.496623$ indicates that if H_0 is true, then the sample mean of student test scores $\bar{y} = 66.58$ would be very unlikely. This is because only a small proportion of the probability distribution in Figure 21.1 is greater than or equal to our t-statistic, $t_{\bar{y}} = 2.496623$. In fact, the proportion of t-statistics greater than 2.496623 is only about P = 0.017. Hence, if our null hypothesis is true, then the probability of getting a mean student test score of 66.58 or higher is P = 0.017 (this is our p-value).

It is important to understand the relationship between the t-statistic and the p-value. An interactive application helps visualise the relationship between the t-statistic and p-values.

> Click here for the shiny app of the t-distribution

Typically, we set a threshold level of $\alpha = 0.05$, below which we conclude that our p-value is statistically significant (see Chapter 20). Consequently, because our p-value is less than 0.05, we reject our null hypothesis and conclude that student test scores are higher than the reported national average.

Figure 21.1.: A t-distribution is shown with a calculated t-statistic of 2.49556 indicated with a downard arrow.

21.2. Independent samples t-test

Perhaps the biology teacher is not actually interested in comparing their students' test results with those of the reported national average. After all, there might be many reasons for why their students score differently from the national average that have nothing to do with their new approach to teaching. To see if their new approach is working, the teacher might instead decide that a better hypothesis to test is whether or not the mean test score from the current year is higher than the mean test score from the class that they taught in the previous year. We can use $\bar{y}_1$ to denote the mean of test scores from the current year, and $\bar{y}_2$ to denote the mean of test scores from the previous year. The test scores of the current year (y_1) therefore remain the same as in the example of the one sample t-test from the previous section.

```
49.3, 62.9, 73.7, 65.5, 69.6, 70.7, 61.5, 73.4, 61.1, 78.1
```

Suppose that in the previous year, there were 9 students in the class (i.e., one fewer than the current year). These 9 students received the following test scores (y_2).

```
57.4, 52.4, 70.5, 71.6, 46.1, 60.4, 70.0, 64.5, 58.8
```

The mean score from last year was $\bar{y}_2 = 61.30$, which does appear to be lower than the mean score of the current year, $\bar{y}_1 = 66.58$. But is the difference between these two means statistically significant? In other words, were the test scores from each year sampled from a population with the same mean, such that the population mean of the previous year (μ_2) and the current year (μ_1) are the same? This is the null hypothesis, $H_0 : \mu_2 = \mu_1$.

The general idea for testing this null hypothesis is the same as it was in the one sample t-test. In both cases, we want to calculate a t-statistic, then see where it falls along the t-distribution to decide whether or not to reject H_0. In this case, our t-statistic ($t_{\bar{y}_1-\bar{y}_2}$) is calculated slightly differently,

$$t_{\bar{y}_1-\bar{y}_2} = \frac{\bar{y}_1 - \bar{y}_2}{SE(\bar{y})}$$

The logic is the same as the one sample t-test. If $\mu_2 = \mu_1$, then we also would expected $\bar{y}_1 = \bar{y}_2$ (i.e., $\bar{y}_1 - \bar{y}_2 = 0$). Differences between $\bar{y}_1$ and $\bar{y}_2$ cause the t-statistic to be either above or below 0, and we can map this deviation of $t_{\bar{y}_1-\bar{y}_2}$ from 0 to the probability density of the t-distribution after standardising by the standard error ($SE(\bar{y})$).

What is $SE(\bar{y})$ in this case? After all, there are two different samples y_1 and y_2, so could the two samples not have *different* standard errors? This could indeed be the case, and how we actually conduct the independent samples t-test depends on whether or not

we are willing to assume that the two samples came from populations with the same variance (i.e., $\sigma_1 = \sigma_2$). If we are willing to make this assumption, then we can pool the variances (s_p^2) together to get a combined (more accurate) estimate of the standard error $SE(\bar{y})$ from both samples[1,2].

If we are unwilling to assume that y_1 and y_2 have the same variance, then we need to use an alternative version of the independent samples t-test. This alternative version is called the Welch's t-test (Welch, 1938), also known as the unequal variances t-test (Dytham, 2011; Ruxton, 2006). In contrast to the standard independent t-test (also called the 'Students independent t-test'), the Welch's t-test does not pool the variances of the samples together[3]. While there are some mathematical differences between the Student's and Welch's independent samples t-tests, the general concept is the same.

This raises the question, when is it acceptable to assume that y_1 and y_2 have the same variance? The sample variance of $s_1^2 = 69.46$ and $s_2^2 = 76.15$. Is this close enough to

[1] This is not a calculation that needs to be done by hand anymore. Statistical software such as Jamovi or R will calculate s_p automatically from a dataset (The Jamovi Project, 2022; R Core Team, 2022). For those interested, the formula that it uses to make this calculation is,

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_{y_1}^2 + (n_2 - 1)\, s_{y_2}^2}{n_1 + n_2 - 2}}.$$

This looks like a lot, but really all the equation is doing is add the two variances ($s_{y_1}^2$ and $s_{y_2}^2$) together, but weighing them by their degrees of freedom ($n_1 - 1$ and $n_2 - 1$), so that the one with a higher sample size has more influence on the pooled standard deviation. To get $SE(\bar{y})$, we then need to multiply by the square root of $(n_1 + n_2)/(n_1 n_2)$,

$$SE(\bar{y}) = s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}.$$

It might be useful to try this once by hand, but only to convince yourself that it matches the t-statistic produced by Jamovi or R.

[2] Note that the degrees of freedom for the t-statistic is not straightforward when we pool variances. If the sample sizes n_1 and n_2 are the same, then $df = 2(n - 1)$. If $n_1 \neq n_2$, then $df = n_1 + n_2 - 2$ (Sokal and Rohlf, 1995). The rationale here is that we lose an extra degree of freedom if we need to estimate the variance of each sample separately, which is necessary if we need to weigh each variance by its own degree of freedom (see footnote 34). Statistical software such as Jamovi and R handles all of this automatically (The Jamovi Project, 2022; R Core Team, 2022), so we will not dwell on it here.

[3] The equation for calculating the Welch's t-statistic is actually a bit simpler than the Student's t-test that uses the pooled estimate (Ruxton, 2006). Instead of using s_p (defined in footnote 34), we define our t-statistic as,

$$t_s = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

The degrees of freedom, however, is quite a bit messier than it is for the independent samples Student's t-test,

$$v = \frac{\left(\frac{1}{n_1} + \frac{(s_2^2/s_1^2)}{n_2}\right)^2}{\frac{1}{n_1^2(n_1-1)} + \frac{(s_2^2/s_1^2)^2}{n_2^2(n_2-1)}}.$$

Of course, Jamovi and R will make these calculations for us, so it is not necessary to do it by hand.

treat them as the same? Like a lot of choices in statistics, there is no clear right or wrong answer. In theory, if both samples do come from a population with the same variance ($\sigma_1^2 = \sigma_2^2$), then the pooled variance is better because it gives us a bit more statistical power; we can correctly reject the null hypothesis more often when it is actually false (i.e., it decreases the probability of a Type II error). Nevertheless, the increase in statistical power is quite low, and the risk of pooling the variances when they actually are not the same increases the risk that we reject the null hypothesis when it is actually true (i.e., it increases the probability of a Type I error, which we definitely do not want!). For this reason, some researchers advocate using the Welch's t-test by default, unless there is a very good reason to believe y_1 and y_2 are sampled from populations with the same variance (Ruxton, 2006; Delacre et al., 2017).

Here we will adopt the traditional approach of first testing the null hypothesis that $\sigma_1^2 = \sigma_2^2$ using a homogeneity of variances test. If we fail to reject this null hypothesis (i.e., $P > 0.05$), then we will use the Student's t-test, and if we reject it (i.e., $P \leq 0.05$), then we will use the Welch's t-test. This approach is mostly used for pedagogical reasons; in practice, defaulting to the Welch's t-test is fine (Ruxton, 2006; Delacre et al., 2017). Testing for homogeneity of variances is quite straightforward in most statistical programs, and we will save the conceptual and mathematical details of this for when we look at the F-distribution in Chapter 23. But the general idea is that if $\sigma_1^2 = \sigma_2^2$, then the ratio of variances (σ_1^2/σ_2^2) has its own null distribution (like the normal distribution, or the t-distribution), and we can see the probability of getting a deviation of σ_1^2/σ_2^2 from 1 if $\sigma_1^2 = \sigma_2^2$ is true.

In the case of the test scores from the two samples of students (y_1 and y_2), a homogeneity of variance test reveals that $s_1^2 = 69.46$ and $s_2^2 = 76.15$ are not significantly different (P = 0.834). We can therefore use the pooled variance and the Student's independent samples t-test. We can calculate $SE(\bar{y}) = 8.521033$ using the formula for s_p (again, this is not something that ever actually needs to be done by hand), then find $t_{\bar{y}_1-\bar{y}_2}$,

$$t_{\bar{y}_1-\bar{y}_2} = \frac{\bar{y}_1 - \bar{y}_2}{SE(\bar{y})} = \frac{66.58 - 61.3}{3.915144} = 1.348609.$$

As with the one-sample t-test, we can identify the position of $t_{\bar{y}_1-\bar{y}_2}$ on the t-distribution (Figure 21.2).

The proportion of t-scores that are higher than $t_{\bar{y}_1-\bar{y}_2} = 1.348609$ is about 0.098. In other words, given that the null hypothesis is true, the probability of getting a t-statistic this high is $P = 0.098$. Because this p-value exceeds our critical value of $\alpha = 0.05$, we do not reject the null hypothesis. We therefore should conclude that the mean of test scores from the current year ($\bar{y}_1$) is not significantly different from the mean of test scores in the previous year ($\bar{y}_2$). The biology teacher in our example might therefore conclude that mean test results have not improved from the previous year.

Figure 21.2.: A t-distribution is shown with a calculated t-statistic of 1.348609 indicated with a downard arrow.

Table 21.1.: Test scores from 10 students (S1-S10) for 2 different tests in a hypothetical biology education example.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Test 1	49.3	62.9	73.7	65.5	69.6	70.7	61.5	73.4	61.1	78.1
Test 2	46.6	62.7	73.8	58.3	66.8	69.7	64.5	71.3	64.5	78.8
Change	-2.7	-0.2	0.1	-7.2	-2.8	-1.0	3.0	-2.1	3.4	0.7

21.3. Paired sample t-test

There is one more type of t-test to consider. The paired sample t-test is applied when the data points in one sample can be naturally paired with those in another sample. In this case, data points between samples are not independent. For example, we can consider the student test scores (y_1) yet again.

```
49.3, 62.9, 73.7, 65.5, 69.6, 70.7, 61.5, 73.4, 61.1, 78.1
```

Suppose that the teacher gave these same students (S1-S10) a second test and wanted to see if the mean student score changed from one test to the next (i.e., a two-sided hypothesis).

In this case, what we are really interested in is the *change* in scores from Test 1 to Test 2. We want to test the null hypothesis that this change is zero. This is actually the same test as the one-sample t-test. We are just substituting the mean difference in values (i.e., 'Change' in the table above) for $\bar{y}$ and setting $\mu_0 = 0$. We can calculate $\bar{y} = -0.88$ and $SE(\bar{y}) = 0.9760237$, then set up the t-test as before,

$$t_{\bar{y}} = \frac{-0.88 - 0}{0.9760237} = -0.9016175.$$

Again, we can find the location of our t-statistic $t_{\bar{y}} = -0.9016175$ on the t-distribution (Figure 21.3).

Since this is a two-sided hypothesis, we want to know the probability of getting a t-statistic as extreme as -0.9016175 (i.e., either ± 0.9016175) given that the null distribution is true. In the above t-distribution, 95% of the probability density lies between $t = -2.26$ and $t = 2.26$. Consequently, our calculated $t_{\bar{y}} = -0.9016175$ is not sufficiently extreme to reject the null hypothesis. The p-value associated with $t_{\bar{y}} = -0.9016175$ is $P = 0.391$. We therefore fail to reject H_0 and conclude that there is no significant difference in student test scores from Test 1 to Test 2.

Figure 21.3.: A t-distribution is shown with a calculated t-statistic of -0.9016175 indicated with a downward arrow.

21.4. Assumptions of t-tests

We make some potentially important assumptions when using t-tests. A consequence of violating these assumptions is a potentially misleading Type I error rate. That is, if our data do not fit the assumptions of our statistical test, then we might not actually be rejecting our null hypothesis at the $\alpha = 0.05$ level. We might unknowingly be rejecting H_0 at a much higher α value, and therefore concluding that we have evidence supporting the alternative hypothesis H_A when really do not. It is therefore important to recognise the assumptions that we are making when using any statistical test (including t-tests). If our assumptions are violated, we might need to use a different test, or perhaps apply some kind of transformation on the data. Assumptions that we make when conducting a t-test are as follows:

- Data are continuous (i.e., not count or categorical data)
- Sample observations are a random sample from the population
- Sample means are normally distributed around the true mean

Note that if we are running a Student's independent samples t-test that pools sample variances (rather than a Welch's t-test), then we are also assuming that sample variances are the same (i.e., a homogeneity of variance). The last bullet point concerning normally distributed sample means is frequently misunderstood to mean that the *sample* data themselves need to be normally distributed. This is not the case (Johnson, 1995; Lumley et al., 2002). Instead, what we are really concerned with is the distribution of sample means ($\bar{y}$) around the true mean (μ). And given a sufficiently large sample size, a normal distribution is assured due to the central limit theorem (see Chapter 15).

Moreover, while a normally distributed variable is not *necessary* for satisfying the assumptions of a t-test (or many other tests we will introduce in this book), it is *sufficient*. In other words, if the variable being measured is normally distributed, then the sample means will also be normally distributed around the true mean (even at low sample size). So when is a sample size large enough, or close enough to being normally distributed, for the assumption of normality to be satisfied? There really is not a definitive answer to this question, and the truth is that most statisticians will prefer to use a histogram (or some other visualisation approach) and their best judgement to decide if the assumption of normality is likely to be violated.

In this book, we will take the traditional approach of running a statistical test called the Shapiro-Wilk test to test the null hypothesis that data are normally distributed. If we reject the null hypothesis (when $P < 0.05$), then we will conclude that the assumption of normality is violated and the t-test is not appropriate. The details of how the Shaprio-Wilk test work are not important for now, but the test can be easily run using Jamovi or R (The Jamovi Project, 2022; R Core Team, 2022). If we reject the null hypothesis that the data are normally distributed, then we can use one of two methods to run our statistical test.

1. Transform the data in some way (e.g., take the log of all values) to improve normality.
2. Use a non-parameteric alternative test.

The word 'non-parameteric' in this context just means that there are no assumptions (or very few) about the shape of the distribution (Dytham, 2011). We will consider the non-parametric equivalents of the one-sample t-test (the Wilcoxon test) and independent sample t-test (Mann-Whitney U test) in the next section. But first, we can show how transformations can be used to improve the fit of the data to model assumptions.

Often data will have a skewed distribution; this just means that the distribution will be asymmetrical. For example, in Figure 21.4A below, we have a dataset (sample size $N = 200$) with a large positive skew (i.e., it is right skewed). Most values are in the same general area, but with some values being especially high.

Figure 21.4.: A set of values with a high positive skew (A), which, when log-transformed (i.e., when we take the natural log of all values), have a normal distribution (A).

Using a t-test on the variable X shown in Figure 21.4A is probably not the good idea. But taking the natural log of all the values of X makes the dataset more normally distributed, thereby more convincingly satisfying the normality assumption required by the t-test. This might seem a bit suspicious at first. Is it really okay to just take the logarithm of

all the data instead of the actual values that were measured? Actually, there is no real scientific or statistical reason that we need to use the original scale (Sokal and Rohlf, 1995). Using the log or the square-root of a set of numbers is perfectly fine if it helps satisfy the assumptions of a statistical test.

21.5. Non-parametric alternatives

If we find that the assumption of normality is not satisfied, and a transformation of the data cannot help, then we can consider using non-parametric alternatives to a t-test. These alternatives include the Wilcoxon test and the Mann-Whitney U test.

21.5.1. Wilcoxon test

The Wilcoxon test (also called the Wilcoxon signed-rank test) is the non-parametric alternative to a one sample t-test (or a paired t-test). Instead of using the actual data, the Wilcoxon test ranks all of the values in the dataset, then sums up their signs (either positive or negative). The general idea is that we can compare the sum of the *ranks* of the actual data with what would be predicted by the null hypothesis. It tests the null hypothesis that the *median* (M) is significantly different from some given value[4]. An example will make it easier to see how it works. We can use the same hypothetical dataset on student test scores.

```
49.3, 62.9, 73.7, 65.5, 69.6, 70.7, 61.5, 73.4, 61.1, 78.1
```

The first step is to subtract the null hypothesis value ($M = 60$, if we again set H_0 to be that student test scores equal 60) from each value ($49.3 - 60 = -10.7$, $62.9 - 60 = 2.9$, etc.).

```
-10.7, 2.9, 13.7, 5.5, 9.6, 10.7, 1.5, 13.4, 1.1, 18.1
```

We need to note the sign of each value as negative ($-$) or positive ($+$).

```
-, +, +, +, +, +, +, +, +, +
```

[4] Actually, the Wilcoxon signed-rank test technically tests the null hypothesis that two distributions are identical, not just that the medians are identical (Johnson, 1995; Lumley et al., 2002). Note that if one distribution has a higher variance than another, or is differently skewed, it will also affect the ranks. In other words, if we want to compare medians, then we need to assume that the distributions have the same shape (Lumley et al., 2002). Because of this, and because the central limit theorem ensures that means will be normally distributed given a sufficiently high sample size (in which case, t-test assumptions are more likely to be satisfied), some researchers suggest that nonparametric statistics such as the Wilcoxon signed rank test should be used with caution (Johnson, 1995).

Next, we need to compute the absolute values of the numbers (i.e., $|-10.7| = 10.7$, $|2.9| = 2.9$, $|13.7| = 13.7$, etc.).

```
10.7, 2.9, 13.7, 5.5, 9.6, 10.7, 1.5, 13.4, 1.1, 18.1
```

We then rank these values from lowest to highest and record the sign of each value.

```
6.5, 3.0, 9.0, 4.0, 5.0, 6.5, 2.0, 8.0, 1.0, 10.0
```

Note that both the first and sixth position had the same value (10.7), so instead of ranking them as 6 and 7, we split the difference and rank both as 6.5. Now, we can calculate the sum of the negative ranks (W^-), and the positive ranks W^+. In this case, the negative ranks are easy; there is only one value (the first one), so the sum is just 6.5,

$$W^- = 6.5$$

The positive ranks are in positions 2-10, and the rank values in these positions are 3, 9, 4, 5, 6.5, 2, 8, 1, and 10. The sum of our positive ranks is therefore,

$$W^+ = 3.0 + 9.0 + 4.0 + 5.0 + 6.5 + 2.0 + 8.0 + 1.0 + 10.0 = 48.5$$

Note that the sum of W^- and W^+ (i.e., $6.5 + 48.5 = 55$ in the example here) must always be equal for a given sample size N (in this case, $N = 10$),

$$W = \frac{N\,(N+1)}{2}.$$

What the Wilcoxon test is doing is calculating the probability of getting a value of W^+ as or more extreme than would be the case if the null hypothesis is true. Note that if, e.g., if there are an equal number of values above and below the median, then both W^- and W^+ will be relatively low and about the same value. This is because the ranks of the values below 0 (which we times by -1) and above 0 (which we times by 1) will be about the same. But if there are a lot more values above the median than expected (as with the example above), then W^+ will be relatively high. And if there are a lot more values below the median than expected, then W^+ will be relatively low.

To find the probability of W^+ being as low or high as it is given that the null hypothesis is true (i.e., a p-value), we need to compare the test statistic W^+ to its distribution under the null hypothesis (note that we can also conduct a one-tailed hypothesis, in which case we are testing if W^+ is either higher or lower than expected given H_0). The old way of doing this is to compare the calculated W^+ threshold values to those

Table 21.2.: Test scores from 10 students (S1-S10) for 2 different tests in a hypothetical biology education example.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Test 1	49.3	62.9	73.7	65.5	69.6	70.7	61.5	73.4	61.1	78.1
Test 2	46.6	62.7	73.8	58.3	66.8	69.7	64.5	71.3	64.5	78.8
Change	-2.7	-0.2	0.1	-7.2	-2.8	-1.0	3.0	-2.1	3.4	0.7

from a Wilcoxon Signed-Ranks Table. The critical value table is no longer necessary, and statistical software such as Jamovi and R will calculate a p-value for us. For the example above, the p-value associated with $W^+ = 48.6$ and $N = 10$ is $P = 0.037$. Since this p-value is less than our threshold of $\alpha = 0.05$, we can reject the null hypothesis and conclude that the median of our dataset is significantly different from 60.

Note that we can also use a Wilcoxon signed rank test as a non-parametric equivalent to a paired t-test. In this case, instead of subtracting out the null hypothesis of our median value (e.g., $H_0 : M = 60$ in the example above), we just need to subtract the paired values. Consider again the example of the 2 different tests introduced for the paired sample t-test above.

If the 'Change' values were not normally distributed, then we could apply a Wilcoxon test to test the null hypothesis that the median value of $Test\,2 - Test\,1 = 0$ (note that a Shapiro-Wilk normality does not reject the null hypothesis that the difference between test scores is normally distributed, so the paired t-test would be preferred in this case). To do this, we would first note the sign of each value as negative or positive.

```
-, -, +, -, -, -, +, -, +, +
```

Next, we would rank the absolute values of the changes.

```
6, 2, 1, 10, 7, 4, 8, 5, 9, 3
```

If we then sum the ranks, we get $W^- = 6+2+10+7+4+5 = 34$ and $W^+ = 1+8+9+3 = 21$. The p-value associated with $W^- = 34$ and $N = 10$ in a two-tailed test is $P = 0.557$, so we do not reject the null hypothesis that Test 1 and Test 2 have the same median.

21.5.2. Mann-Whitney U test

A non-parametric alternative of the independent samples t-test is the Mann-Whitney U test. That is, a Mann-Whitney test can be used if we want to know whether the median of two independent groups is significantly different. Like the Wilcoxon test, the Mann-Whitney U test uses the ranks of values, rather than the values themselves. In the

Mann-Whitney U test, the general idea is rank all of the data across both groups, then see if the sum of the ranks is significantly different (Sokal and Rohlf, 1995). To demonstrate this, we can again consider the same hypothetical dataset used when demonstrating the independent samples t-test above. Test scores from the current year (y_1) are below.

```
49.3, 62.9, 73.7, 65.5, 69.6, 70.7, 61.5, 73.4, 61.1, 78.1
```

We want to know if the median of the above scores is significantly different from the median scores in the previous year (y_2) shown below.

```
57.4, 52.4, 70.5, 71.6, 46.1, 60.4, 70.0, 64.5, 58.8
```

There are 19 values in total, 10 values for y_1 and 9 values for y_2. We therefore rank *all* of the above values from 1 to 19. For y_1, the ranks are below.

```
2, 9, 18, 11, 12, 15, 8, 17, 7, 19
```

For y_2, the ranks are below.

```
4, 3, 14, 16, 1, 6, 13, 10, 5
```

This might be easier to see if we present it as a table showing the test Year (y_1 versus y_2), test Score, and test Rank.

What we need to do now is sum the ranks for y_1 and y_2. If we add up the y_1 ranks, then we get a value of $R_1 = 118$. If we add up the y_2 ranks, then we get a value of $R_2 = 72$. We can then calculate a value U_1 from R_1 and the sample size of y_1 (N_1),

$$U_1 = R_1 - \frac{N_1 (N_1 + 1)}{2}.$$

In the case of our example, $R_1 = 118$ and $N_1 = 10$, We therefore calculate U_1,

$$U_1 = 118 - \frac{10 (10 + 1)}{2} = 63.$$

We then calculate U_2 using the same general formula,

$$U_2 = R_2 - \frac{N_2 (N_2 + 1)}{2}.$$

From the above example of test scores,

Table 21.3.: Test scores from different students across 2 years, and the overall rank of each test score, in a hypothetical biology education example.

Year	Score	Rank
1	49.3	2
1	62.9	9
1	73.7	18
1	65.5	11
1	69.6	12
1	70.7	15
1	61.5	8
1	73.4	17
1	61.1	7
1	78.1	19
2	57.4	4
2	52.4	3
2	70.5	14
2	71.6	16
2	46.1	1
2	60.4	6
2	70.0	13
2	64.5	10
2	58.8	5

$$U_2 = 72 - \frac{9\,(9+1)}{2} = 27.$$

Whichever of U_1 or U_2 is lower[5] is set to U, so for our example, $U = 27$. Note that if y_1 and y_2 have similar medians, then U will be relatively low. But if y_1 and y_2 have much different medians, then U will be relatively high. One way to think about this is that if y_1 and y_2 are very different, then one of the two samples should have very low ranks, and the other should have very high ranks (Sokal and Rohlf, 1995). But if y_1 and y_2 are very similar, then their summed ranks should be nearly the same. Hence, if the deviation of U is greater than what is predicted given the null hypothesis in which the distributions (and therefore the medians) of y_1 and y_2 are identical, the we can reject H_0.

As with the Wilcoxon test, we could compare our U test statistic to a critical value from a table, but statistical software such as Jamovi or R will also just give us the test statistic U and associated p-value (The Jamovi Project, 2022; R Core Team, 2022). In our example above, $U = 27$, $N_1 = 10$, and $N_2 = 9$ has a p-value of $P = 0.156$, so we do not reject the null hypothesis. We therefore conclude that there is no evidence that there is a difference in the median test scores of the two years.

21.6. Summary

The main focus of this chapter was to provide a conceptual explanation of different statistical tests. In practice, running these tests is relatively straightforward in statistical software such as Jamovi or R (this is good, as long as the tests are understood and interpreted correctly). In general, the first step in approaching these tests is to determine if the data call for a one sample test, an independent samples test, or a paired samples test. In a one sample test, the objective is to test the null hypothesis that the mean (or median) equals a specific value. In an independent samples test, there are two different groups, and the objective is to test the null hypothesis that the groups have the same mean (or median). In a paired samples test, there are two groups, but individual values are naturally paired between each group, and the objective is to test the null hypothesis that the difference between paired values is zero[6]. If assumptions of the t-test are violated, then it might be necessary to use a data transformation or a non-parametric test such as the Wilcoxon test (in place of a one sample or paired sample t-test) or a Mann-Whitney U test (in place of an independent samples t-test).

[5] Note that you could also take the higher of the two values (Sokal and Rohlf, 1995), which is what R uses (Jamovi uses the lower of the two). The null distribution from which the p-value would be calculate would just be different in this case.

[6] Note, you could also test the null hypothesis that the difference between paired values is some value other than zero, but this is quite rare in practice.

22. *Practical*. Hypothesis testing and t-tests

This lab focuses on applying the concepts about hypothesis testing from Chapter 20 and the statistical tests from Chapter 21 in Jamovi (The Jamovi Project, 2022). There will be 5 exercises in total, which will focus on using t-tests and their non-parametric equivalents. The data for this lab focus on an example of biology and environmental sciences education (Figure 22.1[1]). Similar to the example in Chapter 21, it uses datasets of student test scores (note that these data are entirely fictional; we have not used scores from real students).

Figure 22.1.: The practical this week focuses on an example relevant to education. This image is in the public domain.

We will use test data from 2022 (student_scores_year_1.csv) and 2023 (student_scores_year_2.csv). To download these files, right click on the links and choose 'Save Link As…', then save it with the extension '.csv'. Variables in these datasets include student IDs and scores from 3 different tests.

[1]This figure has been released into the Public Domain.

Table 22.1.: Student test scores from 2022.

ID	Test_1	Test_2	Test_3
202201	46.9	58.1	60.1
202202	65.0	72.4	51.8
202203	41.1	53.0	57.2
202204	59.1	62.5	58.3
202205	47.6	63.0	58.3
202206	38.0	41.5	54.6
202207	52.9	66.4	61.2
202208	73.1	79.9	57.6
202209	67.9	78.1	63.8
202210	38.6	41.8	56.6
202211	58.9	67.3	62.0
202212	69.0	70.9	67.7
202213	73.5	84.6	56.8
202214	58.7	67.2	61.1
202215	41.1	41.2	54.1
202216	50.8	63.3	77.6
202217	62.3	69.0	60.6
202218	52.2	55.7	52.9
202219	54.3	63.9	54.9
202220	54.3	63.4	71.6
202221	32.1	50.5	52.0

22.1. Exercise on a simple one sample t-test

First we will consider test data from 2022 (student_scores_year_1.csv). These data are shown below, but you should be able to download the file in the link above.

Open this dataset in Jamovi. Before doing any analysis, compute a new column of data called 'Overall' that is each student's mean test score. To do this, double click on the fifth column of data, then when a new panel opens up, choose 'NEW COMPUTED VARIABLE' from the list of available options. Your input should like Figure 22.2.

After calculating the overall grade for each student, find the sample size (N), overall sample mean ($\bar{x}$), and sample standard deviation (s). Report these below.

N: _____

$\bar{x}$: _____

s: _____

Figure 22.2.: The Jamovi interface for adding a new computed variable called 'Overall' and calculating the mean of the three tests.

Suppose that you have been told that the national average of overall student test scores in classes like yours is $\mu_0 = 60.1$. A concerned colleague approaches you and expresses concern that your test scores are lower than the national average. You decide to test whether or not your students' scores really are lower than the national average, or if your colleague's concerns are unfounded. What kind(s) of statistical test would be most appropriate to use in this case, and what is the null hypothesis (H_0) of the test?

Test to use: _____

H_0: _____

What is the alternative hypothesis (H_A), and should you use a one or two-tailed test?

H_A: _____

One or two tailed? _____

To test if the mean of overall student scores in your class is significantly different from $\mu_0 = 60.1$ (technically, that your student's scores were sampled from a population with a mean lower than 60.1), we might use a one-tailed, one sample t-test. Recall from Chapter 21.4 the assumptions underlying a t-test:

- Data are continuous (i.e., not count or categorical data)
- Sample observations are a random sample from the population
- Sample means are normally distributed around the true mean

We will first focus on the first and third assumption. First, the sample data are indeed continuous. But are the sample means normally distributed around the true mean?

A sufficient condition to fulfil this assumption is that the sample data are normally distributed (Johnson, 1995; Lumley et al., 2002). We can test whether or not the overall mean student scores are normally distributed. In the Jamovi toolbar, find the button that says 'T-Tests' (second from the left), then choose 'One Sample T-test' from the pulldown menu. A new window will open up. In a one sample t-test, there is only 1 dependent variable, which in this case is 'Overall'. Move the variable 'Overall' to the Dependent Variables box with the right arrow. Below the boxes, find the checkbox options under '**Assumption Checks**'. Check both 'Normality test' (this is a Shapiro-Wilk test of normality) and 'Q-Q Plot' (Figure 22.3).

Figure 22.3.: The Jamovi interface for running a one sample t-test (right hand side). The variable 'Overall' has been brought in the Dependent Variables box, and two different options have been selected to check the assumption of normality (a normality hypothesis test and a Q-Q plot).

After checking the box for 'Normality test' and 'Q-Q Plot', new output will appear in the panel on the right hand side. The first output is a table showing the results a Shapiro-Wilk test. The Shapiro-Wilk test tests the null hypothesis that the dependent variable is normally distributed (note, this is *not* our t-test yet; we use the Shapiro-Wilk to see if the assumptions of the t-test are violated or not). If we reject the null hypothesis, then we will conclude that the data are **not** normally distributed. If we do not reject the null hypothesis, then we will will assume that the data **are** normally distributed. From the Normality Test table, what is the p-value of the Shapiro-Wilk test?

P = _____

Based on this p-value, should we reject the null hypothesis? _____

Now have a look at the Q-Q plot. We will learn about the details of how this plot works later, but for now, you just need to know that this plot can be used to assess whether or not the data are normally distributed. If the data are normally distributed, then the points in the Q-Q plot should fall roughly along the diagonal black line. These plots take some practice to read and interpret, but for now, just have a look at the Q-Q plot and think about how it relates to the results of the Shapiro-Wilk test of normality. It might be useful to plot a histogram of the overall scores too in order to judge whether or not they are normally distributed (most researchers will use both a visual inspection of the data and a normality test to evaluate whether or not the assumption of normality is satisfied)

Assuming that we have not rejected our null hypothesis, we can proceed with our one sample t-test. To do this, make sure the variable 'Overall' is still in the Dependent Variables box, then make sure the check box 'Student's' is checked below **Tests**. Underneath **Hypothesis**, change the test value to 60.1 (because $\mu_0 = 60.1$) and put the radio button to '< Test value' to test the null hypothesis that the mean overall score of students is the same as μ_0 against the alternative hypothesis that it is less than μ_0. On the right panel of Jamovi, you will see a table with the t-statistic, degrees of freedom, and p-value of the one sample t-test. Write these values down below.

t-statistic: _____

degrees of freedom: _____

p-value: _____

Remember how the t-statistic, degrees of freedom, and p-value are all related. For a reminder, use this interactive application. Slide the value for 't' until it matches the t-statistic that your reported above, then input the correct degrees of freedom in the box to the lower right. Look at the probability density in blue to confirm that it is approximately the same as your p-value (there might be a slight rounding error here). Based on the p-value, should you reject the null hypothesis that your students' mean overall grade is the same as the national average? Why or why not?

Based on this test, how would you respond to your colleague who is concerned that your students are performing below the national average?

Consider again the three assumptions of the t-test listed above. Is there an assumption that might be particularly suspect when comparing the scores of students in a single classroom with a national average? Why or why not?

We will now test whether students in this dataset improved their test scores from Test 1 to Test 2.

22.2. Exercise on a paired t-test

Suppose we want to test whether or not students in this dataset improved their grades from Test 1 to Test 2. In this case, we are interested in the variables 'Test_1' and 'Test_2', but we are not just interested in whether or not the means of the two test scores are the same. Instead, we want to know if the scores of individual students increased or not. The data are therefore naturally paired. Each test score belongs to a unique student, and each student has a score for Test 1 and Test 2. To test whether or not there has been an increase in test scores, go to the 'T-Tests' button in the Jamovi toolbar and select 'Paired Samples T-Test' from the pulldown menu. Place 'Test_1' in the Paired Variables box first, followed by 'Test_2'. Jamovi will interpret the first variable placed in the box as 'Measure 1', and the second variable placed in the box as 'Measure 2'. Before looking at any test results, first check to see if the difference between Test 1 and Test 2 scores is normally distributed using the same **Assumption Checks** as in the previous exercise. Is there any reason to believe that the data are not normally distributed?

We want to know if student grades have improved. What is the null hypothesis (H_0) and alternative hypothesis (H_A) in this case?

H_0: _____

H_A: _____

To test the null hypothesis against the appropriate alternative hypothesis, select the radio button 'Measure 1 < Measure 2'. As with the one sample t-test, on the right panel of Jamovi, you will see a table with the t-statistic, degrees of freedom, and p-value of the one sample t-test. Write these values down below.

t-statistic: _____

degrees of freedom: _____

p-value: _____

Based on this p-value, should you reject or fail to reject your null hypothesis? What can you then conclude about student test scores?

Note that this paired t-test is, mathematically, the exact same as a one sample t-test. If you want to convince yourself of this, you can create a new computed variable that is $Test\,1 - Test\,2$, then run a one sample t-test. You will see that the t-statistic, degrees of freedom, and p-value are all the exact same. Next, we will consider a case in which the assumption of normality is violated.

22.3. Wilcoxon test

Suppose that we want to test the null hypothesis that the scores from Test 3 of the dataset used in Exercises 1 and 2 (student_scores_year_1.csv) were sampled from a population with a mean of $\mu_0 = 62$. We are not interested in whether the scores are higher or lower than 62, just that they are different. Consequently, what should our alternative hypothesis (H_A) be?

H_A: _____

Use the same procedure that you did in Exercise 1 to test this new hypothesis concerning Test 3 scores. First, check the assumption that the Test 3 scores are normally distributed. What is the p-value of the Shaprio-Wilk test this time?

P = _____

What inference can you make from the Q-Q plot? Do the points fall along the diagonal line?

Based on the Shaprio-Wilk test and Q-Q plot, is it safe to assume that the Test 3 scores are normally distributed?

Since the Test 3 scores are not normally distributed (and an assumption of the t-test is therefore potentially violated), we can use the non-parametric Wilcoxon signed-rank test instead, as was introduced in Chapter 21.5.1. To apply the Wilcoxon test, check the box 'Wilcoxon rank' under the **Tests** option. Next, make sure to set the 'Test value' to 62 under the **Hypothesis** option. What are the null and alternative hypotheses of this test?

H_0: _____

H_A: _____

The results of the Wilcoxon signed-rank test will appear in the 'One Sample T-Test' table in the right panel of Jamovi in a row called 'Wilcoxon W'. What is the test statistic (not the p-value) for the Wilcoxon test?

Test statistic: _____

Based on what you learned in Chapter 21.5.1, what does this test statistic actually mean?

Now look at the p-value for the Wilcoxon test. What is the p-value, and what should you conclude from it?

P: _____

Conclusion: _____

Next, we will introduce a new dataset to test hypotheses concerning the means of two different groups.

22.4. Independent samples t-test

In the first 3 exercises, we tested hypotheses using data from 2022 (student_scores_year_1.csv). In this exercise, we will test the differences in mean test scores between students from 2022 and 2023. To do this, first download the 2023 data (right click the link student_scores_year_2.csv and choose 'Save Link As...', then save as a file with the extension '.csv'). These data are shown below, but you should be able to download the file in the link above.

As might be expected, the 2022 and 2023 datasets are stored in separate files. What we need to do first is combine the data into a single tidy dataset. To do this, open the student_scores_year_2.csv dataset in a spreadsheet and copy all of the data (but

Table 22.2.: Student test scores from 2023.

ID	Test_1	Test_2	Test_3
202301	62.7	63.5	81.8
202302	57.9	77.4	57.3
202303	64.6	50.7	57.4
202304	56.1	70.5	57.0
202305	72.5	75.2	57.4
202306	51.9	45.7	59.5
202307	58.1	49.1	63.3
202308	59.5	52.4	57.2
202309	45.2	39.3	57.2
202310	61.8	45.6	57.3
202311	72.0	59.4	57.0
202312	58.4	75.8	57.9
202313	52.8	46.0	65.8
202314	71.6	68.9	62.7
202315	56.4	41.7	57.7
202316	66.1	50.7	61.3
202317	57.0	50.4	57.6
202318	57.1	67.7	59.5
202319	71.1	49.8	58.8
202320	66.3	53.6	57.3
202321	68.7	43.9	57.7
202322	61.0	77.3	57.6
202323	54.3	67.7	57.7
202324	56.8	47.6	78.9

not the column names). You can then paste these data directly into Jamovi in the next available row (row 22). Next, add in a new column for 'Year', so that you can differentiate 2022 students from 2023 students in the same dataset. To do this, you can right-click on the 'ID' column and choose 'Add Variable', then 'Insert'. A new column will appear to the left, which you can name 'Year'. Input '2022' for rows 1-21, then '2023' for the remaining rows that you just pasted into Jamovi (see Figure 22.4).

Figure 22.4.: The Jamovi interface for inserting a new variable called 'Year', and the value 2022 pasted into the first 21 rows. The value 2023 has been pasted into the remaining rows of the Year column.

Note that Year does not necessarily need to be in the first column. You could have added it in as the second column, or as the last (i.e., the location of the column will not affect any analyses).

With the new data now included in Jamovi, we can now compare student test scores between years. Suppose that we first want to test the null hypothesis that the overall student scores have the same mean, $\mu_{2022} = \mu_{2023}$, with the alternative hypothesis that $\mu_{2022} \neq \mu_{2023}$. Is this a one or a two tailed hypothesis?

One or two tailed? _____

It is generally a good idea to plot and summarise your data before running any statistical tests. If you have time, have a look at histograms of the Overall student scores from each year, and look at some summary statistics from the Descriptives output. This will give you a sense of what to expect when running your test diagnostics (e.g., tests of normality) and might alert you to any problems before actually running the analysis

(e.g., major outliers that do not make sense, such as a student having an overall score of over 1000 due to a data input error).

After you have had a look at the histograms and summary statistics, go to the Jamovi toolbar and navigate to 'T-Tests', then the 'Independent Samples T-Test' from the pulldown menu. Remember that our objective here is to test whether the means of two groups (2022 versus 2023) are the same. We can therefore place the 'Overall' variable in the Dependent Variables box, then 'Year' in the Grouping Variable box. Before running the independent samples t-test, we again need to check our assumptions. In addition to the assumption of normality that we checked for in the previous exercises, recall from Chapter 21.2 that the standard 'Student's' independent samples t-test also assumes that the variances of groups (i.e., the 2022 and 2023 scores) are the same, while the 'Welch's' t-test does not assume equal variances. In addition to the **Assumption Checks** options 'Normality test' and 'Q-Q plot', there is also now a test called 'Homogeneity test', which will test the null hypothesis that groups have the same variances (Figure 22.5). This is called a 'Levene's test', and we can interpret it in a similar way to the Shapiro-Wilk test of normality.

Figure 22.5.: The Jamovi interface for running the assumptions of an Independent Samples T-Test. Here, we are testing if the Overall grades of students differ by Year (2022 versus 2023). Assumption Checks include a test for homogeneity of variances (Homogeneity test) and normal distribution of Overall grades (Normality test and Q-Q plot).

If our p-value is sufficiently low ($P < 0.05$), then we reject the null hypothesis that the groups have the same variances. Based on the **Assumption Checks** in Jamovi (and Figure 22.5), what can you conclude about the t-test assumptions?

What is the p-value for the Levene's test?

P: _____

It appears from our test of assumptions that we do not reject the null hypothesis that the data are normally distributed, but we should reject the null hypothesis that the groups have equal variances. Based on what you learned in Chapter 21.2, what is the appropriate test to run in this case?

Test: _____

Note that the appropriate test should be available as a check box underneath the **Tests** options. Check the box for the correct test, then report the test statistic and p-value from the table that appears in the right panel.

Test statistic: _____

P = _____

What can you conclude from this t-test?

Next, we will compare mean Test 3 scores between years 2022 and 2023.

22.5. Mann-Whitney U Test

Suppose that we now want to use the data from the previous exercise to test whether or not Test 3 scores differ between years. Consequently, in this exercise, Test 3 will be our Dependent Variable and Year will again be the Grouping Variable (also called the 'Independent variable'). Below, summarise the hypotheses for this new test.

H_0 : _____

H_A : _____

Is this a one or two tailed test? _____

Next, begin a new Independent Samples T-Test in Jamovi and check the assumptions underlying the t-test. Do the variances appear to be the same for Test 3 scores in 2022 versus 2023? How can you make this conclusion?

Next check to see if the data are normally distributed. What is the p-value of the Shapiro-Wilk test?

P = _____

Now, have a look at the Q-Q plot. What can you infer from this plot about the normality of the data, and why?

Based on what you found from testing the model assumptions above, and the material in Chapter 21, what test is the most appropriate one to use?

Test: _____

Run the above test in Jamovi, then report the test statistic and p-value below.

Test statistic: _____

p-value: _____

Based on what you learned in Chapter 21.5.2, what does this test statistic actually mean?

Finally, what conclusions can you make about Test 3 scores in 2022 versus 2023?

Figure 22.6.: The Jamovi interface for creating a new computed variable that is the natural logarithm of Test 3 data scores in a fictional dataset of student grades.

For the sake of time, we have not introduced an example in which a transformation (such as a log transformation) is applied to normalise a dataset, as explained in Chapter 21.4. To do such a transformation, you could create a new computed variable in Jamovi and calculate it as the natural log (LN) of a variable (e.g., Test 3). Figure 22.6 illustrates what this should look like.

Lastly, if you have time, suppose that you wanted to test whether students from 2023 improved their scores from Test 1 to Test 2 more than students from 2022 did? Is there a way to do this with just the tools presented here and in Chapter 21? Think about how the paired t-test works, and how you could apply that logic to test for the difference in the change between two independent samples (2022 versus 2023). What could you do to test the null hypothesis that the change in scores from Test 1 to Test 2 is the same between years?

Part VII.

Review of parts I-V

Week 7 Overview (Reading week)

This is a special chapter for week 6, which is a reading week, and it will function as a very brief pause for review. It will also ensure that the numbers of parts in the book will correspond to weeks.

Part VIII.

Analysis of Variance (ANOVA)

Week 8 Overview

Dates	13 March 2023 - 17 MAR 2023
Reading	**Required:** SCIU4T4 Workbook chapters 23-25
	Recommended: Navarro and Foxcroft (2022) (Chapter 13)
	Suggested: Rowntree (2018) (Chapter 7)
	Advanced: Head et al. (2015) (Download)
Lectures	8.1: What is ANOVA? (15:35 min; Video)
	8.2: One Way ANOVA (16:51 min; Video)
	8.3: Two Way ANOVA (17:02 min; Video)
	8.4: Kruskal-Wallis H Test (6:37 min; Video)
Practical	ANOVA and associated tests (Chapter 27)
	Room: Cottrell 2A17
	Group A: 15 MAR 2023 (WED) 13:05-15:55
	Group B: 16 MAR 2023 (THU) 09:05-11:55
Help hours	Martina Quaggiotto
	Room: Cottrell 1A13
	17 MAR 2023 (FRI) 15:05-17:55
Assessments	Week 8 Practice quiz on Canvas

Week 8 introduces ANalysis Of VAriance (ANOVA) and related methods, all of which focus on testing whether or not multiple groups in a dataset have the same mean.

Chapter 23 introduces the general idea of the ANOVA, how it is calculated, and the F statistic that is used to test the null hypothesis that all groups have the same mean. This chapter also outlines the assumptions of ANOVA.

Chapter 24 introduces multiple comparisons tests. These tests can be used to find out which group means differ in a dataset when there are more than 2 groups.

Chapter 25 introduces the Kruskall-Wallis H test. This is the non-parametric equivalent of the one-way ANOVA, which can be used when assumptions of the ANOVA are violated. The Kruskall-Wallis is an extension of the Mann-Whitney U test from Chapter 21.5.2

Chapter 26 introduces the two-way ANOVA. With the two-way ANOVA, it also introduces the ideas of testing multiple null hypotheses with a single test and the concept of statistical interactions.

Chapter 27 guides you through the week 8 practical. The aim of this practical is to practice using and correctly interpreting the tests that are introduced in chapters 23-26.

23. Analysis of variance

An ANalysis Of VAriance (ANOVA) is, as the name implies, a method for analysing variances in a dataset. This is confusing, at first, because the most common application of an ANOVA is to test for differences among group *means*. That is, an ANOVA can be used to test the same null hypothesis as the independent samples student's t-test introduced Chapter 21.2; are 2 groups sampled from a population that has the same mean? The t-test works fine when we have only 2 groups, but it does not work when there are 3 or more groups and we want two know if the groups all have the same mean. An ANOVA can be used to test the null hypothesis that *all* groups in a dataset are sampled from a population with the same mean. For example, we might want to know if mean wing length is the same for 5 species of fig wasps sampled from the same area (Duthie et al., 2015). What follows is an explanation of why this can be done by looking at the variance within and between groups (note, 'groups' are also sometimes called 'factors' or 'levels'). Groups are categorical data (see Chapter 5). In the case of the fig wasp example, the groups are the different species (Table 23.1).

Table 23.1.: Wing lengths (mm) measured for 5 unnamed species of non-pollinating fig wasps collected from fig trees in 2010 near La Paz in Baja, Mexico. Note, for readability, this table is not presented in a tidy format.

Het1	Het2	LO1	SO1	SO2
2.122	1.810	1.869	1.557	1.635
1.938	1.821	1.957	1.493	1.700
1.765	1.653	1.589	1.470	1.407
1.700	1.547	1.430	1.541	1.378

Why is any of this necessary? If we want to know if the 5 species of fig wasps in Table 23.1 have the same mean wing length, can we not just use t-tests to compare the means between each species? There are a couple problems with this approach. First, there are a lot of group combinations to compare (Het1 vs Het2, Het1 vs LO1, Het1 vs SO1, etc.). For the 5 fig wasp species in Table 21.2, there are 10 pair-wise combinations that would need to be tested. And the number of combinations grows exponentially[1] with each new group added to the dataset (Table 23.2)

[1]Technically polynomially, but the distinction really is not important for understanding the concept.

Table 23.2.: The number of individual t-tests that would need to be run to compare the means given different numbers of groups (e.g., if a dataset had measurements from 2-10 species)

Groups	2	3	4	5	6	7	8	9	10
Tests	1	3	6	10	15	21	28	36	45

Aside from the tedium of testing every possible combination of group means, there is a more serious problem having to do with the Type I error. Recall from Chapter 20.3 that a Type I error occurs when we reject the null hypothesis (H_0) and erroneously conclude that H_0 is false when it is actually true (i.e., a false positive). If we reject H_0 at a threshold level of $\alpha = 0.05$ (i.e., reject H_0 when $P < 0.05$, as usual), then we will erroneously reject the null hypothesis about 5% of the time that we run a statistical test and H_0 is true. But if we run 10 separate t-tests to see if the fig wasp species in Table 23.1 have different mean wing lengths, then the probability of making an error increases considerably. The probability of erroneously rejecting **at least 1** of the 10 null hypotheses increases from 0.05 to about 0.40. In other words, about 40% of the time, we would conclude that at least 2 species differ in their mean wing lengths[2], even when all species *really do* have the same wing length. This is not a mistake that we want to make, which is why we should first test if all of the means are equal:

- H_0 : All mean species wing lengths are the same
- H_A : Mean species wing lengths are not all the same

We can use an ANOVA to test the null hypothesis above against the alternative hypothesis. If we reject H_0, then we can start comparing pairs of group means (more on this in Chapter 25).

How do we test the above H_0 by looking at *variances* instead of *means*? Before getting into the details of how an ANOVA works, we will first look at the F-distribution. This is relevant because the test statistic calculated in an ANOVA is called an F-statistic, which is compared to an F-distribution in the same way that a t-statistic is compared to a t-distribution for a t-test (see Chapter 21).

In general, the number of possible comparisons between groups is described by a binomial coefficient,

$$\binom{g}{2} = \frac{g!}{2\,(g-2)!}.$$

The number of combinations therefore increases with increasing group number (g).

[2]To get the 0.4, we can first calculate the probability that we (correctly) do not reject H_0 for all 10 pair-wise species combinations $(1-0.05)^{10} \approx 0.60$, then subtract from 1, $P(Do\,not\,reject\,H_0) = 1-(1-0.05)^{10} \approx 0.4$. That is, we find the probability of there not being a Type I error in the first test (1 - 0.05), **and** the second test (1 - 0.05), and so forth, thereby multiplying (1 - 0.05) by itself 10 times. This gives the probability of not committing any Type I error across all 10 tests, so the probability that we commit at least 1 Type I error is 1 minus this probability.

23.1. The F-distribution

If we want to test whether or not 2 variances are the same, then we need to know what the null distribution should be if 2 different samples came from a population with the same variance. The general idea is the same as it was for the distributions introduced in Chapter 14.4. For example, if we wanted to test whether or not a coin is fair, then we could flip it 10 times and compare the number of times it comes up heads to probabilities predicted by the binomial distribution when $P(Heads) = 0.5$ and $N = 10$ (see Chapter 14.4.1 Figure 14.5). To test variances, we will calculate the ratio of variances (F), then compare it to the F probability density function[3]. For example, the ratio of the variances for samples 1 and 2 is (Sokal and Rohlf, 1995),

$$F = \frac{Variance\,1}{Variance\,2}.$$

Note that if the variances of samples 1 and 2 are the exact same, then F = 1. If the variances are very different, then F is either very low (if Variance 1 < Variance 2) or very high (if Variance 1 > Variance 2). To test the null hypothesis that samples 1 and 2 have the same variance, we therefore need to map the calculated F to the probability density of the F distribution. Again, the general idea is the same as comparing a t-score to the t-distribution in Chapter 21.1. Recall that the shape of the t-distribution is slightly different for different degrees of freedom (df). As df increases, the t-distribution starts to resemble the normal distribution. For the F-distribution, there are actually 2 degrees of freedom to consider. One degree of freedom is needed for Variance 1, and a second degree of freedom is needed for Variance 2. Together, these 2 degrees of freedom will determine the shape of the F-distribution (Figure 23.1).

Figure 23.1 shows an F distribution for 3 different combinations of degrees of freedom. The F distribution changes its shape considerably given different df values. Visualising this is much, much easier using an interactive application.

> Click here for an interactive application demonstrating how the F distribution changes with different degrees of freedom.

It is not necessary to memorise how the F distribution changes with different degrees of freedom. The point is that the probability distribution changes given different degrees of freedom, and that the relationship between probability and the value on the x-axis (F) works like other distributions such as the normal or t-distribution. The entire area under the curve must sum to 1, and we can calculate the area above and below any F value (rather, we can get statistical programs such as Jamovi and R to do this for us). Consequently, we can use the F-distribution as the null distribution for the ratio of two

[3] The F-distribution was originally discovered in the context of the ratio of random variables with chi-square distributions, with each variable being divided by its own degree of freedom (Miller and Miller, 2004). We will look at the Chi-square distribution in Week 9.

Figure 23.1.: Probability density functions for 3 different F distributions, each of which have different degrees of freedom for the variances in the numerator (df1) and denominator (df2).

variances. If the null hypothesis that the 2 variances are the same is true (i.e., F = 1), then the F-distribution gives us the probability of the ratio of 2 variances being as or more extreme (i.e., further from 1) than a specific value.

23.2. One-way ANOVA

We can use the F-distribution to test the null hypothesis mentioned at the beginning of the chapter (that fig wasp species have the same mean wing length). The general idea is to compare the mean variance among groups to the mean variance within groups, so our F value (i.e., "F statistic") is calculated,

$$F = \frac{Mean\ variance\ among\ \ groups}{Mean\ variance\ within\ \ groups}.$$

The rest of this section works through the details of how to calculate this F statistic. It is easy to get lost in these details, but the calculations that follow do not need to be done by hand. As usual, Jamovi or R will do all of this work for us (The Jamovi Project, 2022; R Core Team, 2022). The reason for going through the ANOVA step-by-step is to show how the total variation in the dataset is being partitioned into the variance among versus within groups, and to provide some conceptual understanding of what the numbers in ANOVA output actually mean.

23.2.1. ANOVA mean variance among groups

To get the mean variance among groups (i.e., mean squares; MS_{among}), we need to use the sum of squares (SS). The SS was introduced to show how the variance is calculated in Chapter 12.3,

$$SS = \sum_{i=1}^{N} (x_i - \bar{x})^2 .$$

Instead of dividing SS by N - 1 (i.e., the total df), as we would do to get a sample variance, we will need to divide it by the df *among groups* (df_{among}) and df within groups (df_{within}). We can then use these SS_{among}/df_{among} and SS_{within}/df_{within} values to calculate our F[4].

This all sounds a bit abstract at first, so an example will be helpful. We can again consider the wing length measurements from the 5 species of fig wasps shown in Table

[4]Note that the SS divided by the degrees of freedom (N - 1) is a variance. For technical reasons (Sokal and Rohlf, 1995), we cannot simply calculate the mean variance of groups (i.e., the mean of s^2_{Het1}, s^2_{Het2}, etc.). We need to sum up all the squared deviations from group means *before* dividing by the relevant degrees of freedom (i.e., dfs for the among and within group variation).

23.1. First, note that the **grand mean** (i.e., the mean across all species) is $\bar{\bar{x}} = 1.6691$. We can also get the sample mean values of each group, individually. For example, for Het1,

$$\bar{x}_{Het1} = \frac{2.122 + 1.938 + 1.765 + 1.7}{4} = 1.88125$$

We can calculate the means for all 5 fig wasps (Table 23.3).

Table 23.3.: Mean wing lengths (mm) from 5 unnamed species of non-pollinating fig wasps collected from fig trees in 2010 near La Paz in Baja, Mexico. Each species mean was calculated from 4 wasps (N = 4).

Het1	Het2	LO1	SO1	SO2
1.88125	1.70775	1.71125	1.51525	1.53

To get the mean variance among groups, we need to calculate the sum of the squared deviations of each species wing length ($\bar{x}_{Het1} = 1.88125$, $\bar{x}_{Het2} = 1.70775$, etc.) from the grand mean ($\bar{\bar{x}} = 1.6691$). We also need to weigh the squared deviation of each species by the number of samples for each species[5]. For example, for Het1, the squared deviation would be $4(1.88125 - 1.6691)^2$ because there are 4 fig wasps, so we multiply the squared deviation from the mean by 4. We can then calculate the sum of squared deviations of the species means from the grand mean,

$$SS_{among} = 4(1.88125 - 1.6691)^2 + 4(1.70775 - 1.6691)^2 \ + \ ... \ + \ 4(1.53 - 1.6691)^2.$$

Calculating the above across the 5 species of wasps gives a value of $SS_{among} = 0.3651868$. To get our mean variance among groups, we now just need to divide by the appropriate degrees of freedom (df_{among}). Because there are 5 total species ($N_{species} = 5$), $df_{among} = 5 - 1 = 4$. The mean variance among groups is therefore $MS_{among} = 0.3651868/4 = 0.0912967$.

23.2.2. ANOVA mean variance within groups

To get the mean variance within groups (MS_{within}), we need to calculate the sum of squared deviations of wing lengths from *species means*. That is, we need to take the wing length of each wasp, subtract the mean species wing length, then square it. For example, for Het1, we calculate,

[5] In this case, weighing by sample size is not so important because each species has the same number of samples. But when different groups have different numbers of samples, we need to multiply by sample number so that each group contributes proportionally to the SS.

$$SS_{Het1} = (2.122 - 1.88125)^2 + (1.938 - 1.88125)^2 + (1.765 - 1.88125)^2 + (1.7 - 1.88125)^2.$$

If we subtract the mean and square each term of the above,

$$SS_{Het1} = 0.0579606 + 0.0032206 + 0.0135141 + 0.0328516 = 0.1075467.$$

Table 23.4 shows what happens after taking the wing lengths from Table 22.1, subtracting the means, then squaring.

Table 23.4.: The squared deviations from species means for each wing length presented in Table 23.1

Het1	Het2	LO1	SO1	SO2
0.0579606	0.0104551	0.0248851	0.0017431	0.011025
0.0032206	0.0128256	0.0603931	0.0004951	0.028900
0.0135141	0.0029976	0.0149451	0.0020476	0.015129
0.0328516	0.0258406	0.0791016	0.0006631	0.023104

If we sum each column (i.e., do what we did for SS_{Het1} for each species), then we get the SS for each species (Table 23.5).

Table 23.5.: The sum of squared deviations from species means for each wing length presented in Table 23.1

Het1	Het2	LO1	SO1	SO2
0.1075467	0.0521188	0.1793248	0.0049487	0.078158

If we sum the squared deviations in Table 23.5, we get a $SS_{within} = 0.422097$. Note that each species included 4 wing lengths. We lose a degree of freedom for each of the 5 species (because we had to calculate the species mean), so our total df is 3 for each species, and $5 \times 3 = 15$ degrees of freedom within groups (df_{within}). To get the mean variance within groups (denominator of F), we calculate $MS_{within} = SS_{within}/df_{within} = 0.0281398$.

23.2.3. ANOVA F statistic calculation

From Chapter 23.2.1, we have the mean variance among groups,

$$MS_{among} = 0.0912967.$$

From Chapter 23.2.2, we have the mean variance within groups,

$$MS_{within} = 0.0281398$$

To calculate F, we just need to divide MS_{among} by MS_{within},

$$F = \frac{0.0912967}{0.0281398} = 3.2443976.$$

Remember that if the mean variance among groups is the same as the mean variance within groups (i.e., $MS_{among} = MS_{within}$), then F = 1. We can test the null hypothesis that $MS_{among} = MS_{within}$ against the alternative hypothesis that there is more variation among groups than within groups ($H_A : MS_{among} > MS_{within}$) using the F distribution (note that this is a one-tailed test). In the example of 5 fig wasp species, $df_{among} = 4$ and $df_{within} = 15$, so we need an F distribution with 4 degrees of freedom in the numerator and 15 degrees of freedom in the denominator[6]. We can use the interactive app to get the F-distribution and p-value (Figure 23.2).

The area shaded in grey in Figure 23.2, where F > 3.2443976, is approximately $P = 0.041762$. This is our p-value. Since $P < 0.05$, we can reject the null hypothesis that all mean species wing lengths are the same because the variance among species wing lengths is significantly higher than the variance within species wing lengths. Note that the critical value of F (i.e., for which $P = 0.05$) is 3.0555683, so for any F value above this (for df1 = 5 and df2 = 19), we would reject H_0.

When running an ANOVA in a statistical program, output includes (at least) the calculated F statistic, degrees of freedom, and the p-value. Figure 23.3 shows the one-way ANOVA output of the test of fig wasp wing lengths.

Jamovi is quite minimalist for a one-way ANOVA (The Jamovi Project, 2022), but these 4 statistics (F, df1, df2, and p) are all that is really needed. Most statistical programs will show ANOVA output that includes the SS and mean squares among (MS_{among}) and within (MS_{within}) groups.

```
## Analysis of Variance Table
##
## Response: wing_length
##           Df  Sum Sq  Mean Sq F value  Pr(>F)
## Species    4 0.36519 0.091297  3.2444 0.04176 *
## Residuals 15 0.42210 0.028140
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[6]Note that $df_{among} = 4$ and $df_{within} = 15$ sum to 19, which is the total df of the entire dataset ($N - 1 = 20 - 1 = 19$). This is always the case for the ANOVA; the overall df constrains the degrees of freedom among and within groups.

Figure 23.2.: F distribution with df = 4 for the numerator and df = 15 for the denominator. The arrow indicates an F value calculated from fig wasp species wing length measurements for 5 different species and 4 measurements per species. Fig wasp wing lengths were collected from a site near La Paz in Baja, Mexico 2010.

One-Way ANOVA (Fisher's)

	F	df1	df2	p
wing_length	3.24440	4	15	0.042

Figure 23.3.: Jamovi output for a one-way ANOVA of wing length measurements in 5 species of fig wasps collected in 2010 near La Paz in Baja, Mexico.

The above output, taken from R, includes the same information as Jamovi (F, df1, df2, and p), but also includes SS and mean variances. Note that we can also get this information from Jamovi if we want it (see Chapter 26).

23.3. Assumptions of ANOVA

As with the t-test (see Chapter 21.4), there are some important assumptions that we make when using an ANOVA. Violating these assumptions will mean that our Type I error rate (α) is, again, potentially misleading. Assumptions of ANOVA include the following (Box et al., 1978; Sokal and Rohlf, 1995):

1. Observations are sampled randomly
2. Observations are independent of one another
3. Groups have the same variance
4. Errors are normally distributed

Assumption 1 just means that observations are not biased in any particular way. For example, if the fig wasps introduced at the start of this chapter were used because they were the largest wasps that were collected for each species, then this would violate the assumption that the wing lengths were sampled randomly from the population.

Assumption 2 means that observations are not related to one another in some confounding way. For example, if all of the Het1 wasps came from one fig tree, and all of the Het2 wasps came from a different fig tree, then wing length measurements are not really independent within species. In this case, we could not attribute differences in mean wing length to species. The differences could instead be attributable to wasps being sampled from different trees (more on this in Chapter 26).

Assumption 3 is fairly self-explanatory. The ANOVA assumes that all of the groups in the dataset (e.g., species in the case of the fig wasp wing measurements) have the same variance. That is, we assume homogeneity of variances (as opposed to heterogeneity of variances). In general, ANOVA is reasonably robust to deviations from homogeneity, especially if groups have similar sample sizes (Blanca et al., 2018). This means that the Type I error rate is about what we want it to be (usually $\alpha = 0.05$), even when the assumption of homogeneity of variances is violated. In other words, we are not rejecting the null hypothesis when it is true more frequently than we intend! We can test the assumption that group variances are the same using a Levene's test in the same way that we did for the independent samples t-test in Chapter 22. If we reject the null hypothesis that groups have the same variance, then we should potentially consider a non-parametric alternative test such as the Kruskall-Wallis H test (see Chapter 25).

Assumption 4 is the equivalent of the t-test assumption from Chapter 21.4 that sample means are normally distributed around the true mean. What the assumption means is that if we were to repeatedly resample data from a population, the sample means that

we calculate would be normally distributed. For the fig wasp wing measurements, this means that if we were to go back out and repeatedly collect 4 fig wasps from each of the 5 species, then sample means of species wing length and overall wing length would be normally distributed around the true means. Due to the central limit theorem (see Chapter 15), this becomes less problematic with increasing sample size. We can test if the sample data are normally distributed using a Q-Q plot or Shapiro-Wilk test (the same procedure used for the t-test). Fortunately, the ANOVA is quite robust to deviations from normality (Schmider et al., 2010), but if data are not normally distributed, we should again consider a non-parametric alternative test such as the Kruskall-Wallis H test (see Chapter 25).

24. Multiple comparisons

In the Chapter 23.2 ANOVA example, we rejected the null hypothesis that all fig wasp species have the same mean wing length. We can therefore conclude that at least one species has a different mean wing length than the rest; can we determine which one(s)? We can try to find this out using a post hoc comparison (*post hoc* is Latin for 'after the event'). That is, after we have reject the null hypothesis in the one-way ANOVA, we can start comparing individual groups (Het1 vs Het2, Het1 vs LO1, etc.). Nevertheless, we need some way to correct for the Type I error problem explained at the beginning of Chapter 23. That is, if we run a large enough number of t-tests, then we are almost guaranteed that we will find a significant difference between means (P $<$ 0.05) where none really exists. A way to avoid this inflated Type I error rate is to set our significance threshold to be lower than the usual $\alpha = 0.05$. We can, for example, divide our α value by the total number of pair-wise t-tests tests that we run. This is called a Bonferonni correction (Dytham, 2011), and it is an especially cautious approach to post hoc comparisons between groups (Narum, 2006). For the fig wasp wing lengths, recall that there are 10 possible pairwise comparisons between the 5 species. This means that if we were to apply a Bonferonni correction and run 10 separate t-tests, then we would only conclude that species mean wing lengths were different when $P < 0.005$ instead of $P < 0.05$.

Another approach to correcting for multiple comparisons is a Tukey's honestly significant difference test (Tukey's HSD, or just a 'Tukey's test'). The general idea of a Tukey's test is the same as the Bonferonni. Multiple t-tests are run in a way that controls the Type I error rate so that the probability of making a Type I error across the whole set of comparisons is fixed (e.g., at $\alpha = 0.05$). The Tukey's test does this by using a modified t-test, with a t-distribution called the "studentised range distribution" that applies the range of mean group values (i.e., $max(\bar{x}) - min(\bar{x})$) and uses the sample variance across the groups with the highest and lowest sample means (Tukey, 1949; Box et al., 1978).

Multiple comparisons tests can be run automatically in statistical programs such as Jamovi and R. Figure 24.1 shows a post-hoc comparisons table for all pair-wise combinations of fig wasp species wing length means.

The column 'p' in Figure 24.1 is the uncorrected p-value, i.e., the p-value that a t-test would produce without any correction for multiple comparisons. The columns p_{tukey} and $p_{bonferonni}$ show corrected p-values for the Tukey's test and Bonferroni corrected t-test, respectively. We can interpret these p-values as usual, concluding that two species have different means if $P < 0.05$ (i.e., Jamovi does the correction for us; we do not need

Post Hoc Comparisons - Species

Comparison									
Species		Species	Mean Difference	SE	df	t	p	p_{tukey}	$p_{bonferroni}$
Het1	-	Het2	0.17350	0.11862	15.00000	1.46270	0.164	0.600	1.000
	-	LO1	0.17000	0.11862	15.00000	1.43319	0.172	0.617	1.000
	-	SO1	0.36600	0.11862	15.00000	3.08557	0.008	0.050	0.075
	-	SO2	0.35125	0.11862	15.00000	2.96122	0.010	0.063	0.097
Het2	-	LO1	−0.00350	0.11862	15.00000	−0.02951	0.977	1.000	1.000
	-	SO1	0.19250	0.11862	15.00000	1.62288	0.125	0.506	1.000
	-	SO2	0.17775	0.11862	15.00000	1.49853	0.155	0.579	1.000
LO1	-	SO1	0.19600	0.11862	15.00000	1.65238	0.119	0.489	1.000
	-	SO2	0.18125	0.11862	15.00000	1.52803	0.147	0.561	1.000
SO1	-	SO2	−0.01475	0.11862	15.00000	−0.12435	0.903	1.000	1.000

Note. Comparisons are based on estimated marginal means

Figure 24.1.: Jamovi output showing a table of 10 post hoc comparisons between species mean wing lengths for 5 different species of fig wasps. The last 3 columns show the uncorrected p-value (p), a p-value obtained from a Tukey's test, and a p-value obtained from a Bonferroni correction. Species wing length measurements were collected in 2010 near La Paz in Baja, Mexico.

to divide $\alpha = 0.05$ to figure out what the significance threshold should be given the Bonferonni correction).

Note that from Figure 24.1, it appears that both the Tukey's test and the Bonferonni correction fail to find that any pair of species have significantly different means. This does not mean that we have done the test incorrectly. The multiple comparisons tests are asking a slightly different question than the one-way ANOVA. The multiple comparisons tests are testing the null hypothesis that two *particular* species have the same mean wing lengths. The one-way ANOVA tested the null hypothesis that all species have the same mean, and our result for the ANOVA was barely below the $\alpha = 0.05$ threshold (P = 0.042). The ANOVA also has more statistical power because it makes use of all 20 measurements in the dataset, not just a subset of measurements between 2 of the 5 species. It is therefore not particularly surprising or concerning that we rejected H_0 for the ANOVA, but the multiple comparisons tests failed to find any significant difference between group means.

25. Kruskall-Wallis H test

If the assumptions of the one-way ANOVA are violated, then we can consider using a Kruskall-Wallis test. The Kruskall-Wallis test is essentially an extension of the Mann-Whitney U test from Chapter 21.5.2 for samples with more than 2 groups (Kruskal and Wallis, 1952). Like the Mann-Whitney U test, it uses the *ranks* of values instead of the actual values in the dataset. The idea is to rank all values in the dataset, ignoring group, then calculate a test statistic (H) that increases as the difference among group ranks increases, relative to the difference within group ranks. We can again use the example of the fig wasp wing lengths introduced in Chapter 23. For convenience, Table 23.1 is reproduced here as Table 25.1.

Table 25.1.: Reproduction of Table 23.1. Wing lengths (mm) measured for 5 unnamed species of non-pollinating fig wasps collected from fig trees in 2010 near La Paz in Baja, Mexico. Note, for readability, this table is not presented in a tidy format.

Het1	Het2	LO1	SO1	SO2
2.122	1.810	1.869	1.557	1.635
1.938	1.821	1.957	1.493	1.700
1.765	1.653	1.589	1.470	1.407
1.700	1.547	1.430	1.541	1.378

Recall that in the one-way ANOVA from Chapter 23.2, we reject the null hypothesis that all species had the same mean wing length (P = 0.042). But we had not actually tested the assumptions of the one-way ANOVA before running the test! If we had tested the ANOVA assumptions, we would not reject the null hypothesis that wing length is normally distributed (Shapiro-Wilk test P = 0.698). But a Levene's test of homogeneity of variances convincingly rejects the null hypothesis that the group variances are equal (P = 0.008). Consequently, we should probably have used the non-parametric Kruskall-Wallis H test instead. To do this, we first need to rank all of the values in Table 25.1. There are 20 total values, so we rank them from 1 to 20 (Table 25.2).

Table 25.2.: Ranks of wing lengths (mm) measured for 5 unnamed species of non-pollinating fig wasps collected from fig trees in 2010 near La Paz in Baja, Mexico.

Het1	Het2	LO1	SO1	SO2
20.0	15	17	8	10.0
18.0	16	19	5	12.5
14.0	11	9	4	2.0
12.5	7	3	6	1.0

From these ranks, we calculate a test statistic H from the overall sample size (N), the sample size of each group (N_i), the mean of group ranks ($\bar{R}_i$), and the overall mean rank ($\bar{\bar{R}}$). Of course, we do not need to do this by hand. But the formula shows how a statistical program will do the calculation (Kruskal, 1952),

$$H = (N-1)\frac{\sum_{i=1}^{g} N_i \left(\bar{R}_i - \bar{\bar{R}}\right)^2}{\sum_{i=1}^{g}\sum_{j=1}^{N_i}\left(R_{ij} - \bar{\bar{R}}\right)^2}.$$

For our purposes, the mathematical details are not important. The equation is included here only to show the similarity between the calculation of H versus F from Chapter 23.2. In the numerator of the equation for H, we are calculating the squared deviation of mean group ranks from the overall mean rank $(\bar{R}_i - \bar{\bar{R}})^2$, as weighted by the group sample size N_i. We are then dividing by the sum of squared deviations of all ranks (R_{ij}) from the overall mean rank $\bar{\bar{R}}$. All of this gets multiplied by $N-1$ to give the test statistic H. We can then compare H to a suitable null distribution, which might be calculated precisely using a specific algorithm (e.g., Choi et al., 2003). But most statistical programs such as Jamovi and R compare H to a Chi-square distribution (see Week 9), which is an effective approximation (Miller and Miller, 2004).

The output of a Kruskall-Wallis H test is quite minimal and easy to read. Jamovi and R report a Chi-square (χ^2) test statistic, degrees of freedom, and p-value (The Jamovi Project, 2022; R Core Team, 2022). Figure 25.1 shows the output of a Kruskall-Wallis H test on the fig wasp wing lengths data.

Given the Kruskall-Wallis H test output (Figure 25.1), we should not reject the null hypothesis that species have different wing lengths because $P > 0.05$. This is in contrast to our one-way ANOVA result, for which we did reject the same null hypothesis. The Kruskall-Wallis H test does not assume that group variances are the same, unlike the one-way ANOVA. Since we know that the homogeneity of variances assumption is violated for the fig wasp data, it is probably best to be cautious and conclude that there is no evidence that mean wing lengths differ among species.

Kruskal-Wallis

	x^2	df	p
wing_length	8.75301	4	0.068

Figure 25.1.: Jamovi output table for a non-parametric Kruskall-Wallis H test, which tests the null hypothesis that species mean wing lengths are the same for 5 different species of fig wasps. Species wing length measurements were collected in 2010 near La Paz in Baja, Mexico.

26. Two-way ANOVA

The one-way ANOVA tested the null hypothesis that all group means are the same. But we might also have a dataset in which there is more than 1 type of group. For example, suppose we know that the fig wasps used in Chapters 23-25 actually came from 2 different trees (Tree A and Tree B). This would mean that there could be a confounding variable affecting wing length, violating assumption 2 from Chapter 23.3. If, for whatever reason, fig wasps on different trees have different wing lengths, then we should include tree as an explanatory variable in the ANOVA. Because we would then have 2 group types (species and tree), we would need a two-way ANOVA[1]. Here, to illustrate the key concepts as clearly as possible, we will run an example of a two-way ANOVA with just 2 of the 5 species (Het1 and SO2). Table 26.1 shows the fig wasp wing length dataset for the 2 species in a tidy format, which includes columns for 2 group types, species and tree.

Table 26.1.: Wing lengths (mm) measured for 2 unnamed species of non-pollinating fig wasps collected from 2 fig trees in 2010 near La Paz in Baja, Mexico.

Species	Tree	wing_length
Het1	A	2.122
Het1	A	1.938
Het1	B	1.765
Het1	B	1.700
SO2	A	1.635
SO2	A	1.700
SO2	B	1.407
SO2	B	1.378

We could run a one-way ANOVA (or t-test) to see if wing lengths differ between species, then run another one-way ANOVA to see if wing lengths differ between trees. But by including both group types in the same model (species and tree), we can test how one group affects wing length in the context of the other group (e.g., how tree affects wing length, while also accounting for any effects of species on wing length). We can also see if there is any synergistic effect between groups, which is called an **interaction effect**. For example, if Het1 fig wasps had longer wing lengths than SO1 on Tree A, but shorter

[1] Yes, it is possible to have a three-way or a four-way ANOVA, but having so many group types can get very messy very quickly, especially if we want to look at interactions between group types.

wing lengths than SO1 on Tree B, then we would say that there is an interaction between species and tree. Given this kind of interaction effect, it would not make sense to say that Het1 fig wasps have longer or shorter wings than SO1 because this would depend on the tree!

This chapter will not delve into the mathematics of the two-way ANOVA[2]. Working out a two-way ANOVA by hand requires similar, albeit more laborious, calculations of sum of squares and degrees of freedom than is needed for the one-way ANOVA in Chapter 23.2. But the general concept is the same. The idea is to calculate the amount of the total variation attributable to different groups, or the interaction among group types. Note that the assumptions for the two-way ANOVA are the same as for the one-way ANOVA.

Unlike previous statistical tests in this book, we can actually test 3 separate null hypotheses with a two-way ANOVA. The first test is the same as the one-way ANOVA in Chapter 23, which focuses on mean species wing lengths:

- H_0 : Both mean species wing lengths are the same
- H_A : Mean species wing lengths are not the same

The second test focuses on the other group type, Tree:

- H_0 : Wing lengths are the same in both trees
- H_A : Wing lengths are not the same in both trees

These two hypotheses address the **main effects** of the independent variables (species and tree) on the dependent variable (wing length). In other words, the mean effect of one group type (either species or tree) by itself, holding the other constant. The third hypothesis addresses the interaction effect:

- H_0 : There is no interaction between species and tree
- H_A : There is an interaction between species and tree

Interaction effects are difficult to understand at first, so we will look at 2 concrete examples of two-way ANOVAs. The first will use the Table 26.1 data in which no interaction effect exists. The second will use a different species pairing in which an interaction does exist.

In a two-way ANOVA, the output (Figure 26.1) appears similar to that of a one-way ANOVA (see Chapter 23.2.3), but there are 4 rows of output. Rows 1 and 2 test the null hypotheses associated with the main effects of species and tree, respectively. Row 3 tests the interaction effect. And row 4 shows us the variation in the data that cannot be attributed to either the main effects or the interaction (i.e., the residual variation). This is equivalent to the variation within groups in the one-way ANOVA.

[2]For those interested, chapter 11 of Sokal and Rohlf (1995) works all of this out step by step.

ANOVA - wing_length

	Sum of Squares	df	Mean Square	F	p
Species	0.24675	1	0.24675	45.75115	0.002
Tree	0.16388	1	0.16388	30.38508	0.005
Species ✻ Tree	0.00025	1	0.00025	0.04693	0.839
Residuals	0.02157	4	0.00539		

Figure 26.1.: Jamovi output table for a two-way ANOVA, which tests for the effects of species (Het1 and SO2), tree, and their interaction on fig wasp wing lengths. Species wing length measurements were collected in 2010 near La Paz in Baja, Mexico.

This is a lot more information than the one-way ANOVA, but it helps to think of each row separately as a test of a different null hypothesis. From the first row, given the high F = 45.75 and the low P = 0.002, we can reject the null hypothesis that species mean wing lengths are the same. Similarly, from the second row, we can reject the null hypothesis that wing lengths are the same in both trees (F = 30.39; P = 0.005). In contrast, from the third row, we should not reject the null hypothesis that there is no interaction between species and trees (F = 0.05, P = 0.839). Figure 26.2 shows the these results visually (The Jamovi Project, 2022).

In Figure 26.2, there is no interaction between species and tree. How can we infer this from Figure 26.2? Wing length is always significantly higher in Tree A than it is in Tree B, regardless of the species of wasp. Similarly, wing length is always higher for Het1 than it is for SO2, regardless of tree. Consequently, while wasp species is important for predicting wing length, as is the tree from which the wasp was collected, we do not need to consider one variable to know the effect of the other. This is reflected in the orange and blue lines having a similar slope, or, more technically, slopes that are not significantly different from one another. If the slopes of the two lines were significantly different, then this would be evidence for an interaction between species and tree.

What would Figure 26.2 look like if there was an interaction effect between species and tree? We can run another two-way ANOVA, this time with a different pair of species, SO1 and SO2. Table 26.2 shows the data.

Table 26.2.: Wing lengths (mm) measured for 2 unnamed species of non-pollinating fig wasps collected from 2 fig trees in 2010 near La Paz in Baja, Mexico.

Species	Tree	wing_length
SO1	A	1.557
SO1	A	1.493

Species	Tree	wing_length
SO1	B	1.470
SO1	B	1.541
SO2	A	1.635
SO2	A	1.700
SO2	B	1.407
SO2	B	1.378

If we run the two-way ANOVA in Jamovi, then we can see the ANOVA output table (Figure 26.3).

Figure 26.3 shows that the effect of species is not significant ($P > 0.05$), but the effect of tree is significant ($P < 0.05$). Consequently, in terms of the main effects, it appears that species does not affect wing length, but tree does affect wing length. Unlike the example test between Het1 and SO2, the interaction between species and tree is significant ($P < 0.05$) in Figure 26.3. This means that we cannot interpret the main effects in isolation because the effect of each on wing length will change in the presence of the other. In other words, the effect of tree on wing length will depend on the species. Figure 26.4 shows this interaction visually.

Note that for SO1, wing length is basically the same for Tree A and B (the 2 left points in Figure 26.4). In contrast, for SO2, tree has a very noticeable effect on wing length! Fig wasps in species SO2 clearly have higher wing lengths on Tree A than they do on Tree B. *Overall*, however, the mean wing length for SO2 appears to be about the same as it is for SO1 (i.e., the mid-way point between the blue and orange points for SO2 is at about the same wing length as it is for SO1, around 1.5 mm). Species therefore does not really have an effect on wing length, at least not by itself. But species *does* have a clear effect if you also take tree into account. In other words, to predict how tree will affect wing length (not at all for SO1, or a lot for SO2), it is necessary to know what species of fig wasp is being considered. This is how interactions work in ANOVA, and more generally in statistics.

There is one last point that is relevant to make for the two-way ANOVA. Recall that we tested 3 null hypotheses simultaneously. Should we not then apply some kind of correction for multiple comparisons, as we did in Chapter 24? This is actually not necessary. The reason is subtle. With the multiple t-tests, we wanted to know if *any* of the pair-wise differences between groups were significant. Each t-test was not really a separate question in this case. We just tried all possible combinations in search of a pair of species with significantly different wing lengths. With the two-way ANOVA, we are asking 3 separate questions, and accepting a Type I error rate of 0.05 for each of them, individually.

Figure 26.2.: An interaction plot for a two-way ANOVA of fig wasp wing measurements as affected by species and tree. Points show mean values for the 4 species and tree combinations (e.g., Het1 from Tree A in the upper left). Error bars show standard errors around the means. Species wing length measurements were collected in 2010 near La Paz in Baja, Mexico.

ANOVA - Wing length (mm)

	Sum of Squares	df	Mean Square	F	p
Species	0.00044	1	0.00044	0.24509	0.647
Tree	0.04337	1	0.04337	24.42590	0.008
Species ✻ Tree	0.03264	1	0.03264	18.38492	0.013
Residuals	0.00710	4	0.00178		

Figure 26.3.: Jamovi output table for a two-way ANOVA, which tests for the effects of species (SO1 and SO2), tree, and their interaction on fig wasp wing lengths. Species wing length measurements were collected in 2010 near La Paz in Baja, Mexico.

Figure 26.4.: An interaction plot for a two-way ANOVA of fig wasp wing measurements as affected by species and tree. Points show mean values for the 4 species and tree combinations (e.g., SO2 from Tree A in the upper right). Error bars show standard errors around the means. Species wing length measurements were collected in 2010 near La Paz in Baja, Mexico.

27. *Practical*. ANOVA and associated tests

This lab focuses on applying the concepts from Chapter 23, Chapter 24, Chapter 25, and Chapter 26 in Jamovi (The Jamovi Project, 2022). The focus will be on ANOVA and associated tests, with 5 exercises in total. The data for this lab are inspired by the doctoral work of Dr Lidia De Sousa Teixeira at the University of Stirling (Figure 27.1). This work included a nutrient analysis of agricultural soil in different regions of Angola. Measuring soil nutrient concentrations is essential for assessing soil quality, and these data include measurements of Nitrogen (N), Phosphorus (P), and Potassium (K) concentrations.

Figure 27.1.: Images from a project led by Dr Lidia De Sousa Teixeira at the University of Stirling

Here we will focus on testing whether or not the concentrations of N, P, and K differ among 2 different sites and 3 soil profiles (upper, middle, and lower). To complete this lab, download the Angola_soils.csv dataset (right click and "Save Link As…", then save it with the extension '.csv'). All concentrations of Nitrogen, Phosphorus, and Potassium are given in parts per million (ppm).

27.1. One-way ANOVA (site)

Suppose that we first want to test whether or not mean Nitrogen concentration is the same in different sites. Notice that there are only 2 sites in the dataset, Funda and Bailundo. We could therefore also use an Independent samples t-test. We will do this

first, then compare the t-test output to the ANOVA output. What are the null (H_0) and alternative (H_A) hypotheses for the t-test.

H_0 : _____

H_A : _____

Before running a t-test, remember that we need to check the assumptions of a t-test to see if any of them are violated (see Chapter 21.4). Use the **Assumption Checks** in Jamovi as we did in the previous practical (Chapter 22.4) to test for normality and homogeneity of variances in Nitrogen concentration. What can you conclude from these 2 tests?

Normality conclusion: _____

Homogeneity of variances conclusion: _____

Given the conclusions from the checks of normality and homogeneity of variances above, what kind of test should you use to see if the mean Nitrogen concentration is significantly different in Funda versus Bailundo?

Test: _____

Run the test above in Jamovi. What is the p-value of the test, and what conclusion do you make about Nitrogen concentration at the two sites?

P: _____

Conclusion: _____

Now we will use an ANOVA to test if the mean Nitrogen concentration differs between sites. Remember from Chapter 23 that the ANOVA compares the variance among groups to the variance within groups, calculating an F statistic and finding where F is in the null F distribution. To run an ANOVA, navigate to the 'Analyses' tab in Jamovi, then select the 'ANOVA' button in the toolbar. From the ANOVA pulldown, select 'One-Way ANOVA' (Figure 27.2).

Figure 27.2.: Jamovi toolbar in the Analyses tab, which includes an option for running an ANOVA.

After selecting the one-way ANOVA, a familiar interface will open up. Place 'Nitrogen' in the Dependent Variables box and 'site' in the Grouping Variable box. Although we have already checked the assumptions of normality and homogeneity of variances when we ran the t-test, check these boxes under **Assumption Checks** too (Figure 27.3).

Figure 27.3.: Jamovi interface for running a one-way ANOVA to test if Nitrogen concentration (ppm) differs among sites in soils of Angola. Data for this test were inspired by the doctoral thesis of Dr Lidia De Sousa Teixeira.

Confirm that the Shapiro-Wilk test of normality and the Levene's test of homogeneity of variances are consistent with what you concluded when testing the assumptions of the t-test above. Since there is no reason to reject the null hypothesis that group variances are equal, we can use the Fisher's One-Way ANOVA by checking 'Assume equal (Fisher's)' under **Variances** (see Figure 27.3). A table called 'One-Way ANOVA' will appear in the panel on the right. Write down the test statistic (F), degrees of freedom, and p-values from this table below.

F: _____

df1: _____

df2: _____

P: _____

Remember from Chapter 23 that the ANOVA calculates an F statistic (mean variance among groups divided by mean variance withing groups). This F statistic is then compared to the null F distribution with the correct degrees of freedom to calculate the p-value. Use the interactive app to visualise this from the above Jamovi output. To do this, move the 'Variance 1' slider in the app until it is approximately equal to F, then change df1 and df2 to the values above. From the interactive app, what is the approximate area under the curve (i.e., the orange area) where the F value on the x-axis is greater than your calculated F?

P: _____

Slide the 'Variance 1' to the left now until you find the F value where the probability density in the tail (orange) is about $P = 0.05$. Approximately, what is this threshold F value above which we will reject the null hypothesis?

Approximate threshold F: _____

What should you conclude regarding the null hypothesis that sites have the same mean?

Conclusion: _____

Look again at the p-value from the one-way ANOVA output and the Student's t-test output. Are the two values the same, or different? Why might this be?

Next, we will run a one-way ANOVA to test the null hypothesis that all profiles have the same mean Nitrogen concentration.

27.2. One-way ANOVA (profile)

We will now run an ANOVA to see if Nitrogen concentration differs among profiles. In this dataset, there are lower, middle, and upper profiles, which refer to the location on along a slope from which soil samples were obtained. Using the same approach as the previous Exercise 27.1, we will run a one-way ANOVA with profile as the independent variable instead of site. Again, navigate to the 'Analyses' tab in Jamovi, then select the 'ANOVA' button in the toolbar. From the ANOVA pulldown, select 'One-Way ANOVA' (Figure 27.2). First check the assumptions of normality and homogeneity of variances. What can you conclude?

Normality conclusion: _____

Homogeneity of variances conclusion: _____

It appears that the assumptions of normality and homogeneity of variances are met. We can therefore proceed with the one-way ANOVA. Run the one-way ANOVA with the assumption of equal variances (i.e., Fisher's test). What are the output statistics in the One-Way ANOVA table?

F: _____

df1: _____

df2: _____

P: _____

From these statistics, what do you conclude about the difference in Nitrogen concentration among profiles?

Conclusion: _____

In the previous Exercise 27.1, we used an interactive app to visualise the relationship between the F statistic and the p-value. We can do the same thing with the distrACTION module in Jamovi. To do this, go to the distrACTION option in the Jamovi toolbar and select 'F-Distribution' from the pulldown menu. Place the df1 and df2 from the One-Way ANOVA table into the df1 and df2 boxes under **Parameters** (ignore λ). Under **Function**, select 'Compute probability', then place the F value from the One-Way ANOVA table in the box for x1. We want the upper tail of the F probability distribution, so choose $P(X \geq x1)$ from the radio buttons below. Write down the 'Probability' value from the Results table in the panel to the right.

Probability: _____

Note that this is the same value (perhaps with a rounding error) as the p-value from the One-Way ANOVA table above. We can also find the threshold value of F, above which we will reject the null hypothesis. To do this, check the 'Compute quantile' box and set p = 0.95 in the box below. From the Results table, what is the critical F value

('Quantile'), above which we would reject the null hypothesis that all groups have the same mean?

Critical F value: _____

Note that the objective of working this out in the distrACTION module (and with the interactive app) is to help explain what these different values in the One-Way ANOVA table actually mean. To actually test the null hypothesis, the One-Way ANOVA output table is all that we really need.

Finally, note that in the ANOVA pulldown from the Jamovi toolbar, the option 'ANOVA' is just below the 'One-way ANOVA' that we used in this exercise and Exercise 27.1. This is just a more general tool for running an ANOVA, which includes the two-way ANOVA that we will use in Exercise 27.5 below. For now, give this a try by selecting 'ANOVA' from the pull down menu. In the ANOVA interface, place 'Nitrogen' into the 'Dependent Variable' box and 'Profile' in the 'Fixed Factors' box (Figure 27.4).

Figure 27.4.: Jamovi interface for running an ANOVA to test if Nitrogen concentration (ppm) differs among soil profiles in Angola. Data for this test were inspired by the doctoral thesis of Dr Lidia De Sousa Teixeira.

The output in the right panel shows an ANOVA table. It includes the sum of squares of the among-group (Profile) and within-group (Residuals) sum of squares and mean square. This is often how ANOVA results are presented in the literature. Fill in the table below (Table 27.1) with the information for degrees of freedom, F, and P.

Table 27.1.: ANOVA output testing the null hypothesis that mean Nitrogen concentration is the same across 3 different soil profiles in Angola. Data for this test were inspired by the doctoral thesis of Dr Lidia De Sousa Teixeira.

	Sum of Squares	df	Mean Square	F	p
Profile	16888.18606		8444.09303		
Residuals	118092.02927		2460.25061		

Now that we have established from the one-way ANOVA that mean Nitrogen concentration is not the same across all soil profiles, we can use a test of multiple comparisons to test which profile(s) are significantly different from one another.

27.3. Multiple comparisons

In this exercise, we will pick up where we left of in the ANOVA of Exercise 27.2. We have established that not all soil profiles have the same mean. Next, we will run a post hoc multiple comparisons test to evaluate which, if any, soil profiles have different means. In the ANOVA input panel, scroll down to the pulldown option called 'Post Hoc Tests' (Figure 27.5).

Move 'Profile' to the box to the right, then select the 'Tukey' checkbox under **Correction**, as shown in Figure 27.5. Doing this will run the Tukey's honestly significant difference (HSD) test introduced in Chapter 24. The output will appear in the panel on the right in a table called 'Post Hoc Tests'. Note that these post-hoc tests use the t-distribution to test for significance. Find the p-values associated with the Tukey's HSD (P_{tukey}) for each profile pairing. Report these below.

Tukey's HSD Lower - Middle: P = _____

Tukey's HSD Lower - Upper: P = _____

Tukey's HSD Middle - Upper: P = _____

From this output, what can we conclude about the difference among soil profiles?

Next, instead of running a Tukey's HSD test, we will use a series of t-tests with a Bonferonni correction. Check the box for 'Bonferonni' in the ANOVA Post Hoc Tests input panel, then find the p-values for the Bonferonni correction ($p_{bonferonni}$). Note that we do not need to change the α threshold ourselves (i.e., we do not need to see if P is less than $\alpha = 0.05/3 = 0.016667$ instead of $\alpha = 0.05$). Jamovi modifies the p-values

Figure 27.5.: Jamovi interface for running post hoc tests following an ANOVA.

appropriately for the Bonferonni correction (we can see the difference by clicking the checkbox for 'No correction' in the Post Hocs Tests input panel). Report the p-values for the Bonferonni correction below.

Bonferonni Lower - Middle: P = _____

Bonferonni Lower - Upper: P = _____

Bonferonni Middle - Upper: P = _____

In general, how are the p-values different between the Tukey's HSD and the Boneferroni correction? Are they about the same, higher, or lower?

What does this difference mean in terms of making a Type I error? In other words, based on this output, are we more likely to make a Type I error with the Tukey's HSD test or the Bonferroni test?

Note that we ran the Tukey's HSD test and the Bonferroni test separately. This is because, when doing a post-hoc test, we should choose which test to use in advance. This will avoid biasing our results to get the conclusion that we *want* rather than the conclusion that is *accurate*. If, for example, we first decided to use a Bonferroni correction, but then found that none of our p-values were below 0.05, it would not be okay to try a Tukey's HSD test instead in hopes of changing this result. This kind of practice is colloquially called 'p-hacking' (or 'data dredging'), and it causes an elevated risk of Type I error and a potential for bias in scientific results. Put more simply, trying to game the system to get results in which $P < 0.05$ can lead to mistakes in science (Head et al., 2015). Specifically, p-hacking can lead us to believe that there are patterns in nature where none really exist, which is definitely something that we want to avoid!

27.4. Kruskall-Wallis H test

In this exercise, we will apply the non-parametric equivalent of the one-way ANOVA, the Kruskall-Wallis H test. Suppose that we now want to know if Potassium concentration differs among soil profiles. We therefore want to test the null hypothesis that the mean Potassium concentration is the same for all soil profiles. Before opening the ANOVA input panel, have a look at a histogram of Potassium concentration. How would you describe the distribution? Do the data appear to be normally distributed?

We can test the assumption of normality using a Shaprio-Wilk test. This can be done in the Descriptives panel of Jamovi, or we can do it in the One-Way ANOVA panel. To do it in the one-way ANOVA panel, first select 'ANOVA' from the pull down menu as we did at the end of Exercise 27.2. In the ANOVA interface, place 'Potassium' into the 'Dependent Variable' box and 'Profile' in the 'Fixed Factors' box. Next, scroll down to the 'Assumption Checks' pulldown menu and select all 3 options. From the Levene's test, the Shapiro-Wilk test, and the Q-Q plot, what assumptions of ANOVA might be violated?

Given the violation of ANOVA assumptions, we should consider a non-parametric option. As introduced in Chapter 25, the Kruskall-Wallis H test is a non-parametric alternative to a one-way ANOVA. Like other non-parametric tests introduced in this book, the Kruskall-Wallis H test uses the ranks of a dataset instead of the actual values. To run a Kruskall-Wallis H test, select the Analyses tab, then the 'ANOVA' button from the Jamovi toolbar. In the pulldown ANOVA menu, choose 'One-Way ANOVA: Kruskall-Wallis' (second to last one down the list; Figure 27.6).

Figure 27.6.: Jamovi toolbar for selecting a Kruskall-Wallis test.

The Kruskall-Wallis input is basically the same as the one-way ANOVA input. We just need to put 'Potassium' in the dependent variable list and 'Profile' in the Grouping Variable box. The output table includes the test statistic (Jaomvi uses a χ^2 value as a test statistic, which we will introduce in Week 9, degrees of freedom, and p-value). Report these values below.

χ^2 : _____

df: _____

P: _____

From the above output, should we reject or not reject our null hypothesis?

H_0 : _____

Note that the Kruskall-Wallis test in Jamovi also includes a type of multiple comparisons tests (DSCF pairwise comparisons checkbox). We will not use the Dwass-Steel-Critchlow-Fligner pairwise comparisons, but the general idea is the same as the Tukey's HSD test for post hoc multiple comparisons in the ANOVA.

27.5. Two-way ANOVA

Since we have two types of categorical variables (site and profile), we might want to know if either has a significant effect on the concentration of an element, and if there is any interaction between site and profile. The two-way ANOVA was introduced in Chapter 26 with an example of fig wasp wing lengths. Here we will test the effects of site, profile, and their interaction on Nitrogen concentration. Recall from Chapter 26 that a two-way ANOVA actually tests 3 separate null hypotheses. Write these null hypotheses down below (the order does not matter).

First H_0: _____

Second H_0: _____

Third H_0: _____

To test these null hypotheses again select 'ANOVA' from the pull down menu as we did at the end of Exercise 27.2. In the ANOVA interface, place 'Nitrogen' into the 'Dependent Variable' box and both 'Site' and 'Profile' in the 'Fixed Factors' box. Next, scroll down to the 'Assumption Checks' pulldown menu and select all 3 options. From the assumption checks output tables, is there any reason to be concerned about using the two-way ANOVA?

In the two-way ANOVA output, we see the same ANOVA table as in Exercise 27.2 (Table 27.1). This time, however, there are 4 rows in total. The first 2 rows correspond with tests of the main effects of Site and Profile, and the third row tests the interaction between these two variables. Fill in Table 27.2 with the relevant information from the two-way ANOVA output.

Table 27.2.: Two-way ANOVA output testing the effects of 2 sites and 3 different soil profiles on soil Nitrogen concentration in Angola. Data for this test were inspired by the doctoral thesis of Dr Lidia De Sousa Teixeira.

	Sum of Squares	df	Mean Square	F	p
Site	21522.18384		21522.18384		
Profile	22811.1368		11405.5684		
Site * Profile	16209.13035		8104.56517		
Residuals	80497.68348		1788.83741		

From this output table, should you reject or not reject your null hypotheses?

Reject First H_0?: _____

Reject Second H_0?: _____

Reject Third H_0?: _____

In non-technical language, what should you conclude from this two-way ANOVA?

Lastly, we can look at the interaction effect between Site and Profile visually. To do this, scroll down to the 'Estimated Marginal Means' pulldown option. Move 'Site' and 'Profile' from the box on the left to the 'Marginal Means' box on the right (Figure 27.7).

In the panel on the right hand side, a plot will appear under the heading 'Estimated Marginal Means'. Based on what you learned in Chapter 26 about interaction effects, what can you say about the interaction between Site and Profile? Does one Profile, in particular, appear to be causing the interaction to be significant? How can you infer this from the Estimated Marginal Means plot?

Figure 27.7.: Jamovi two-way ANOVA test with the pulldown menu for Estimated Marginal Means, which will produce a plot showing the interaction effect of the two-way ANOVA.

If you have time, try running a two-way ANOVA to test the effects of Site and Profile on Phosphorus concentration. Based on the ANOVA output, what can you conclude?

Part IX.

Counts and Correlation

Week 9 Overview

Dates	20 March 2023 - 24 MAR 2023
Reading	**Required:** SCIU4T4 Workbook chapters 28 and 29
	Recommended: Navarro and Foxcroft (2022) (Chapter 12.1 and 12.2)
	Suggested: Rowntree (2018) (Chapter 8)
	Advanced: Rodgers and Nicewander (1988) (Download)
Lectures	9.1: Frequency and count data (13:19 min; Video)
	9.2: Chi-squared goodness of fit test (16:34 min; Video)
	9.3: Chi-squared test of association (0:00 min; Video)
	9.4: Correlation key concepts (7:02 min; Video)
	9.5: Correlation underlying mathematics (12:06 min; Video)
	9.6: Correlation hypothesis testing (27:28 min; Video)
Practical	Analysis of counts and correlations (Chapter 30)
	Room: Cottrell 2A17
	Group A: 22 MAR 2023 (WED) 13:05-15:55
	Group B: 23 MAR 2023 (THU) 09:05-11:55
Help hours	Martina Quaggiotto and Brad Duthie
	Room: Cottrell 1A13
	24 MAR 2023 (FRI) 15:05-17:55
Assessments	Week 9 Practice quiz on Canvas

28. Frequency and count data

In this book, we have introduced hypothesis testing as a tool to determine if variables were sampled from a population with a specific mean (one sample t-test in Chapter 21.1), or if different groups of variables were sampled from a population with the same mean (the independent samples t-test in Chapter 21.2 and ANOVA in Chapter 23). In these tests, the variables for which we calculated the means were always continuous (e.g., fig wasp wing lengths, nitrogen concentration in parts per million). That is, the variables of the t-test and ANOVA could always, at least in theory, take any real value (i.e., any decimal). And the comparison was always between the means of categorical groups (e.g., fig wasp species or study sites). But not every variable that we measure will be continuous. For example, in Chapter 5, we also introduced discrete variables, which can only take discrete counts (1, 2, 3, 4, and so forth). Examples of such **count data** might include the number of species of birds in a forest or the number of days in the year for which an extreme temperature is recorded. Chapter 14 included some examples of count data when introducing probability distributions (e.g., counts of heads or tails in coin flips, or the number of people testing positive for Covid-19). Count data are discrete because they can only take integer values. For example, there cannot be 14.24 bird species in a forest; it needs to be a whole number.

In the biological and environmental sciences, we often want to test whether or not observed counts are significantly different from some expectation. For example, we might hypothesise that the probability of flowers being red versus blue in a particular field is the same. In other words, $Pr(flower = red) = 0.5$ and $Pr(flower = Blue) = 0.5$. By this logic, if we were to collect 100 flowers at random from the field, we would expect 50 to be red and 50 to be blue. If we actually went out and collected 100 flowers at random, but found 46 to be red and 54 to be blue, would this be sufficiently different from our expectation to reject the null hypothesis that the probability of sampling a red versus blue flower is the same? We could test this null hypothesis using a Chi-square goodness of fit test (Chapter 28.1). Similarly, we might want to test if 2 different count variables (e.g., flower colour and flower species) are associated with one another (e.g., if blue flowers are more common in one species than another species). We could test this kind of hypothesis using a Chi-squared test of association (Chapter 30).

Before introducing the Chi-square goodness of fit test or the Chi-square test of association, it makes sense to first introduce the Chi-square (χ^2) distribution. The general motivation for introducing the Chi-square distribution is the same as it was for the t-distribution (Chapter 18) or F-distribution (Chapter 23.1). We need some probability density distribution that is our null distribution, which is what we predict if our null

hypothesis is true. We then compare this null distribution of our test statistic to find the probability of sampling a test statistic as or more extreme if the null hypothesis is really true (i.e., a p-value).

28.1. The Chi-square distribution

The Chi-square (χ^2) distribution is a continuous distribution in which values of χ^2 can be any real number greater than or equal to 0. We can generate a χ^2 distribution by adding up squared values that are sampled from a standard normal distribution (Sokal and Rohlf, 1995), hence the 'square' in 'Chi-square'. There is a lot to unpack in the previous sentence, so we can go through it step by step. First, we can take another look at the standard normal distribution from Chapter 14.4.4 (Figure 28.1).

Figure 28.1.: Recreation of Figure 14.9, a standard normal probability distribution

Suppose that we randomly sampled 4 values from the standard normal distribution shown in Figure 28.1.

- $x_1 = -1.244$
- $x_2 = 0.162$
- $x_3 = -2.214$

- $x_4 = 2.071$

We can square all of these values, then add up the squares,

$$\chi^2 = (-1.244)^2 + (0.162)^2 + (-2.214)^2 + (2.071)^2.$$

Note that χ^2 cannot be negative because when we square a number that is either positive or negative, we always end up with a positive value (e.g., $-2^2 = 4$, see Chapter 1.1). The final value is $\chi^2 = 10.76462$. Of course, this χ^2 value would have been different if our x_i values (x_1, x_2, x_3, and x_4) had been different. And if we are sampling randomly from the normal distribution, we should not expect to get the same χ^2 value from 4 random standard normal deviates. We can therefore ask, if we were to keep sampling 4 standard normal deviates and calculating new χ^2 values, what would be the distribution of these χ^2 values? The answer is shown in Figure 28.2.

Figure 28.2.: A Chi-square distribution, which is the expected sum of 4 squared standard normal deviates, i.e., the sum of 4 values sampled from a standard normal distribution and squared.

Looking at the shape of Figure 28.2, we can see that most of the time, the sum of deviations from the mean of $\mu = 0$ will be about 2. But sometimes we will get a much lower or higher value of χ^2 by chance, if we sample particularly low or high values of x_i.

If we summed a different number of squared x_i values, then we would expect the distribution of χ^2 to change. Had we sampled fewer than 4 x_i values, the expected χ^2 would be lower just because we are adding up fewer numbers. Similarly, had we sampled more than 4 x_i values, the expected χ^2 would be higher just because we are adding up more numbers. The shape of the χ^2 distribution[1] is therefore determined by the number of values sampled (N), or more specifically the degrees of freedom (df, or sometimes v), which in a sample is $df = N-1$. This is the same idea as the t-distribution from Chapter 18, which also changed shape depending on the degrees of freedom. Figure 28.3 shows the different χ^2 probability density distributions for different degrees of freedom.

Figure 28.3.: Probability density functions for 3 different Chi-square distributions, each of which have different degrees of freedom (df).

As with the F distribution from Chapter 23.1, visualising the χ^2 distribution is much, much easier using an interactive application.

Click here for an interactive application demonstrating how the Chi-square

[1] A random variable X has a χ^2 distribution if and only if its probability density function is defined by (Miller and Miller, 2004),

$$f(x) = \begin{cases} \frac{1}{2^{\frac{2}{v}}\Gamma(\frac{v}{2})} x^{\frac{v-2}{2}} e^{-\frac{x}{2}} & for\ x > 0 \\ 0 & elsewhere \end{cases}$$

In this equation, v is the degrees of freedom of the distribution.

distribution changes with different degrees of freedom.

And as with the F distribution, it is not necessary to memorise how the χ^2 distribution changes with different degrees of freedom. The important point is that the distribution changes with different degrees of freedom, and we can map probabilities to the χ^2 value on the x-axis in the same way as any other distribution.

What does any of this have to do with count data? It actually is a bit messy. The χ^2 distribution is not a perfect tool for comparing observed and expected counts (Sokal and Rohlf, 1995). After all, counts are integer values, and the χ^2 distribution is clearly continuous (unlike, e.g., the binomial or Poisson distributions from Chapter 14.4). The χ^2 distribution is in fact a useful approximation for testing counts, and one that becomes less accurate when sample size (Slakter, 1968) or expected count size (Tate and Hyer, 1973) is small. Nevertheless, we can use the χ^2 distribution as a tool for testing whether observed counts are significantly different from expected counts. The first test that we will look at is the goodness of fit test.

28.2. Chi-squared goodness of fit

The first kind of test that we will consider for count data is the goodness of fit test. In this test, we have some number of counts that we expect to observe (e.g., expected counts of red versus blue flowers), then compare this expectation to the counts that we actually observe. If the expected and observed counts differ by a lot, then we will get a large test statistic and reject the null hypothesis. A simple concrete example will make this a bit more clear.

Recall the practical in Chapter 16, in which players of the mobile app game Power Up! chose a small, medium, or large dam at the start of the game. Suppose that we are interested in the size of dam that policy-makers choose to build when playing the game, so we find 60 such people in Scotland and ask them to play the game. Perhaps we do not think that the policy-makers will have any preference for a particular dam size (and therefore just pick 1 of the 3 dam sizes at random). We would therefore expect an equal number of small, medium, and large dams to be selected among the 60 players. That is, for our expected counts of each dam size (E_{size}), we expect 20 small ($E_{small} = 20$), 20 medium ($E_{medium} = 20$), and 20 large ($E_{large} = 20$) dams in total (because $60/3 = 20$).

Of course, even if our players have no preference for a particular dam size, the number of small, medium, and large dams will not always be *exactly* the same. The expected counts might still be a bit different from the observed counts of each dam size (O_{size}). Suppose, for example, we find that out of our total 60 policy-makers, we observe 18 small ($O_{small} = 18$), 24 medium ($O_{medium} = 24$), and 18 large ($O_{large} = 18$), dams were actually chosen by game players. What we want to test is the null hypothesis that there is no significant difference between expected and observed counts.

- H_0: There is no significant difference between expected and observed counts.
- H_A: There is a significant difference between expected and observed counts.

To get our test statistic[2], we now just need to take each observed count, subtract the expected count, square this difference, divide by the expected count, then add everything up,

$$\chi^2 = \frac{(18-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(18-20)^2}{20}.$$

We can calculate the values in the numerator. Note that all of these numbers must be positive (e.g., $18 - 20 = -2$, but $-2^2 = 4$),

$$\chi^2 = \frac{4}{20} + \frac{16}{20} + \frac{4}{20}.$$

When we sum the 3 terms, we get a value of $\chi^2 = 1.2$. Note that if all of our observed values had been the same as the expected values (i.e., 20 small, medium, and large dams actually chosen), then we would get a χ^2 value of 0. The more the observed values differ from the expectation of 20, the higher the χ^2 will be. We can now check to see if the test statistic $\chi^2 = 1.2$ is sufficiently large to reject the null hypothesis that our policy-makers have no preference for small, medium, or large dams. There are $N = 3$ categories of counts (small, medium, and large), meaning that there are $df = 3 - 1 = 2$ degrees of freedom. The interactive application, can be used to compare our test statistic with the null distribution by setting df = 2 and the Chi-square value to 1.2. As it turns out, if the null hypothesis is true, then the probability of observing a value of $\chi^2 = 1.2$ or higher (i.e., the p-value) is $P = 0.5488$. Figure 28.4 shows the appropriate χ^2 distribution plotted, with the area above the test statistic $\chi^2 = 1.2$ shaded in grey.

Because $P > 0.05$, we do not reject the null hypothesis that there is no significant difference between expected and observed counts of chosen dam sizes.

Note that this was a simple example. For a goodness of fit test, we can have any number of different count categories (at least, any number greater than 2). The expectations also do not need to be integers. For example, if we only managed to find 59 policy makers instead of 60, then our expected counts would have been $59/3 = 19.33$ instead of $60/3 = 20$. The expectations also do not *need* to be the same. For example, we could have tested the null hypothesis that twice as many policy-makers would choose large dams (i.e., $E_{large} = 40$, $E_{medium} = 10$, and $E_{small} = 10$). For n categories, the more general equation for the χ^2 statistic is,

[2] A lot of statisticians will use X^2 to represent the test statistic here instead of χ^2 (Sokal and Rohlf, 1995). The difference is the upper case 'X' versus the Greek letter Chi, 'χ'. The X is used since the test statistic we calculate here is not *technically* from the χ^2 distribution, just an approximation. We will not worry about the distinction here, and to avoid confusion, we will just go with χ^2.

Figure 28.4.: A Chi-square distribution, which is the expected sum of 4 squared standard normal deviates, i.e., the sum of 4 values sampled from a standard normal distribution and squared.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}.$$

We can therefore use this general equation to calculate a χ^2 for any number of categories (n). Next, we will look at testing associations between counts in different types of categories.

28.3. Chi-squared test of association

The second kind of test that we will consider for count data is the Chi-square test of association. While the goodness of fit test focused on a single categorical variable (dam sizes in the example above), the Chi-square test of association focuses on 2 different categorical variables. What we want to know is whether or not the 2 categorical variables are independent of one another (Box et al., 1978). In other words, does knowing something about one variable tell us anything about the other variable? A concrete example will make it easier to explain. We can again make use of the Chapter 16 game Power Up!. As mentioned in the previous section, game players choose a small, medium, or large dam at the start of the game. Players can play the game on either an Android or MacOS mobile device. We therefore have 2 categorical variables, dam size and OS type. We might want to know, do Android users choose the same dam sizes as MacOS users? In other words, are dam size and OS type associated? We can state this as a null and alternative hypothesis.

- H_0: There is no association between OS and dam size choice.
- H_A: There is an association between OS and dam size choice.

Consider the data in Table 28.1, which show counts of Android versus MacOS users and their dam choices.

Table 28.1.: Counts (N = 60) from a mobile game called 'Power Up!', in which players are confronted with trade-offs between energy output, energy justice, and biodiversity. Players can use 1 of 2 types of Operating System (Android or MacOS) and build one of 3 types of dam in the game (Small, Medium, or Large).

	Small	Medium	Large
Android	8	16	6
MacOS	10	8	12

Just looking at the counts in Table 28.1, it appears that there might be an association between the 2 variables. For example, Android users appear to be more likely to choose a

medium dam than MacOS users. Medium dams are the most popular choice for Android users, but they are the least popular choice for MacOS users. Nevertheless, could this just be due to chance? If it were due to chance, then how unlikely are the counts in Table 28.1? In other words, if Android and MacOS users in the whole population really do choose dam sizes at the same frequencies, then what is the probability of getting a sample of 60 players in which the choices are as or more unbalanced as this? This is what we want to answer with our Chi-squared test of association.

The general idea is the same as with the Chi-squared goodness of fit test. We have our observed values (Table 28.1). We now need to find the expected values to calculate a χ^2 value. But the expected values are now a bit more complicated. With the goodness of fit test in Chapter 28.2, we just assumed that all categories were equally likely (i.e., the probability of choosing each size dam was the same). There were 60 players and 3 dam sizes, so the expected frequency of each dam choice was $60/3 = 20$. Now it is different. We are not testing if dam sizes or OS choices are the same. We want to know of they are *associated* with one another. That is, regardless of the popularity of Android versus MacOS, or the frequency with which small, medium and large dams are selected, do Android users choose different dam sizes than MacOS users? If dam size is not associated with OS, then we would predict that the relative frequency of small, medium, and large dams would be the same for both Android and MacOS.

To find the expected counts of each variable combination (e.g., Android and Small, or MacOS and Large), we need to get the probability that each category is selected independently. For example, what is the probability of a player selecting a large dam, regardless of the OS that they are using? Table 28.2 shows these probabilities as additional rows and columns added onto Table 28.1

Table 28.2.: Counts (N = 60) from a mobile game called 'Power Up!', in which players are confronted with trade-offs between energy output, energy justice, and biodiversity. Players can use 1 of 2 types of Operating System (Android or MacOS) and build one of 3 types of dam in the game (Small, Medium, or Large). Outer rows and columns show the probabilities of categories being selected

	Small	Medium	Large	**Probability**
Android	8	16	6	0.5
MacOS	10	8	12	0.5
Probability	0.3	0.4	0.3	–

Since there are 30 total Android users ($8 + 16 + 6 = 30$) and 30 total MacOS users ($10+8+12 = 30$), the probability of a player having an Android OS is $30/60 = 0.5$, and the probability of a player having a MacOS is also $30/60 = 0.5$. Similarly, there are 18 small, 24 medium, and 18 large dam choices in total. Hence, the probability of a player choosing a small dam is $18/60 = 0.3$, medium is $24/60 = 0.4$, and large is $18/60 = 0.3$.

If these probabilities combine independently[3], then we can multiply them to find the probability of a particular combination of categories. For example, the probability of a player using Android is 0.5 and choosing a small dam is 0.3, so the probability of a player having both Android **and** a small dam is $0.5 \times 0.3 = 0.15$ (see Chapter 15 for an introduction to probability models). The probability of a player using Android **and** choosing a medium dam is $0.5 \times 0.4 = 0.2$. We can fill in all of these joint probabilities in a new Table 28.3.

Table 28.3.: Probabilities for each combination of categorical variables from a dataset in which players on either an Android or MacOS choose a dam size in the mobile app game 'Power Up!', assuming that variables are independent of one another.

	Small	Medium	Large	**Probability**
Android	0.15	0.2	0.15	0.5
MacOS	0.15	0.2	0.15	0.5
Probability	0.3	0.4	0.3	–

From Table 28.3, we now have the probability of each combination of variables. Note that all of these probabilities sum to 1.

$$0.15 + 0.2 + 0.15 + 0.15 + 0.2 + 0.15 = 1.$$

To get the expected count of each combination, we just need to multiply the probability by the sample size, i.e., the total number of players (N = 60). For example, the expected count of players who use Android and choose a small dam will be $0.15 \times 60 = 9$. Table 28.4 fills in all of the expected counts. Note that the sum of all the counts equals our sample size of 60.

Table 28.4.: Expected counts for each combination of categorical variables from a dataset in which players on either an Android or MacOS choose a dam size in the mobile app game 'Power Up!', assuming that variables are independent of one another.

	Small	Medium	Large	**Sum**
Android	9	12	9	30
MacOS	9	12	9	30
Sum	18	24	18	–

[3]We can call these the 'marginal probabilities'.

We now have both the observed (Table 28.2) and expected (Table 28.4) counts (remember that the expected counts do not *need* to be integers). To get our χ^2 test statistic, we use the same formula as in Chapter 28.2,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}.$$

There are 9 total combinations of OS and dam size, so there are $n = 9$ values to sum up,

$$\chi^2 = \frac{(8-9)^2}{9} + \frac{(16-12)^2}{12} + ... + \frac{(16-12)^2}{12} + \frac{(8-9)^2}{9}.$$

If we sum all of the 9 terms, we get a value of $\chi^2 = 4.889$. We can compare this to the null χ^2 distribution as we did in the Chapter 28.2 goodness of fit test, but we need to know the correct degrees of freedom. The correct degrees of freedom[4] is the number of categories in variable 1 (n_1) minus 1, times the number of categories in variable 2 (n_2) minus 1,

$$df = (n_1 - 1) \times (n_2 - 1).$$

In our example, the degrees of freedom equals the number of dam types minus 1 ($n_{dam} = 3-1$) times the number of operating systems minus 1 ($n_{OS} = 2-1$). The correct degrees of freedom is therefore $df = 2 \times 1 = 2$. We now just need to find the p-value for a Chi-square distribution with 2 degrees of freedom and a test statistic of $\chi^2 = 4.889$. From the interactive app (set df to 2 and slide the Chi-square value to 4.9), we get a value of about $P = 0.0868$. In other words, if H_0 is true, then the probability of getting a χ^2 of 4.889 or higher is $P = 0.0868$. Consequently, because $P > 0.05$, we would not reject the null hypothesis. We should therefore conclude that there is no evidence for an association between OS and dam size choice.

Statistical programs such as R and Jamovi will calculate the χ^2 value and get the p-value for the appropriate degrees of freedom (The Jamovi Project, 2022; R Core Team, 2022). To do this in Jamovi, it is necessary to input the categorical data (e.g., Android, MacOS) in a tidy format, which will be a focus of the practical Chapter 30.

[4] This formula works due to a bit of a mathematical trick (Sokal and Rohlf, 1995). The actual logic of the degrees of freedom is a bit more involved. From our total of $k = 6$ different combinations, we actually need to subtract 1 degree of freedom for the total sample size ($N = 60$), then a degree of freedom for each variable probability estimated (i.e., subtract $n_1 - 1$ and $n_2 - 1$ because we need this many degrees of freedom to get the n_1 and n_2 probabilities, respectively; if we have all but 1 probability, then we know the last probability because the probabilities must sum to 1). Since we lose $n_1 - 1$ and $n_2 - 1$ degrees of freedom, and 1 for the sample size, this results in $df = k - (n_1 - 1) - (n_2 - 1) - 1$. In the case of the "Power Up!" example, we get $df = 6 - (3-1) - (2-1) - 1 = 2$. The $df = (n_1 - 1) \times (n_2 - 1)$ formulation is possible because $k = n_1 \times n_2$ (Sokal and Rohlf, 1995).

There is one final point regarding expected and observed values of the Chi-square test of association. There is another way of getting these expected values that is a bit faster (and more widely taught), but does not demonstrate the logic of expected counts as clearly. If we wanted to, we could sum the rows and columns of our original observations. Table 28.5 shows the original observations with the sum of each row and column.

Table 28.5.: Observed counts for each combination of categorical variables from a dataset in which players on either an Android or MacOS choose a dam size in the mobile app game 'Power Up!'. The last row and column shows the sum of observed dam sizes and OS users, respectively.

	Small	Medium	Large	**Sum**
Android	8	16	6	30
MacOS	10	8	12	30
Sum	18	24	18	–

We can get the expected counts from Table 28.5 if we multiply each row sum by each column sum, then divide by the total sample size ($N = 60$). For example, to get the expected counts of Android users who choose a small dam, we can multiply $(18 \times 30)/60 = 9$. To get the expected counts of MacOS users who choose a medium dam, we can multiply $(30 \times 24)/60 = 12$. This works for all of combinations of rows and columns, so we could do it to find all of the expected counts from Table 28.4.

29. Correlation

This chapter focuses on the association between types of variables that are quantitative (i.e., represented by numbers). It is similar to the Chi-squared test of association from Chapter 28 in the sense that it is about how variables are associated. The focus of the Chi-squared test of association was on the association when data were categorical (e.g., 'Android' or 'MacOS' operating system). Here we focus instead on the association when data are numeric. But the concept is generally the same; are variables independent, or does knowing something about one variable tell us something about the other variable? For example, does knowing something about the latitude of a location tell us something about its average yearly temperature?

29.1. Scatterplots

The easiest way to visualise the concept of a correlation is by using a scatterplot. Scatterplots are useful for visualising the association between 2 quantitative variables. In a scatterplot the values of one variable are plotted on the x-axis, and the values of a second variable are plotted on the y-axis. Consider two fig wasp species of the genus *Heterandrium* (Figure 29.1).

Both fig wasp species in Figure 29.1 are unnamed. We can call the species in Figure 29.1A 'Het1' and the species in Figure 29.1B 'Het2'. We might want to collect morphological measurements of fig wasp head, thorax, and abdomen lengths in these 2 species (Duthie et al., 2015). Table 29.1 shows these measurements for 11 wasps.

Table 29.1.: Body segment length measurements (mm) from fig wasps of 2 species. Data were collected from Baja, Mexico.

Species	Head	Thorax	Abdomen
Het1	0.566	0.767	1.288
Het1	0.505	0.784	1.059
Het1	0.511	0.769	1.107
Het1	0.479	0.766	1.242
Het1	0.545	0.828	1.367
Het1	0.525	0.852	1.408
Het2	0.497	0.781	1.248
Het2	0.450	0.696	1.092

Species	Head	Thorax	Abdomen
Het2	0.557	0.792	1.240
Het2	0.519	0.814	1.221
Het2	0.430	0.621	1.034

Intuitively, we might expect most of these measurements to be associated with one another. For example, if a fig wasp has a relatively long thorax, then it probably also has a relatively long abdomen (i.e., it could just be a big wasp). We can check this visually by plotting one variable on the x-axis and the other on the y-axis. Figure 29.2 does this for wasp thorax length (x-axis) and abdomen length (y-axis).

In Figure 29.2, each point is a different wasp from Table 29.1. For example, in the last row of Table 29.1, there is a wasp with a particularly low thorax length (0.621 mm) and abdomen length (1.034 mm). In the scatterplot, we can see this wasp represented by the point that is lowest and furthest to the left (Figure 29.2).

There is a clear association between thorax length and abdomen length in Figure 29.2. Fig wasps that have low thorax lengths also tend to have low abdomen lengths, and wasps that high thorax lengths also tend to have high abdomen lengths. In this sense, the 2 variables are associated. More specifically, they are positively correlated. As thorax length increases, so does abdomen length.

29.2. The correlation coefficient

The **correlation coefficient** formalises the association described in the previous section. It gives us a single number that defines how 2 variables are correlated. We represent this number with the letter 'r', which can range from values of -1 to 1[1]. Positive values indicate that 2 variables are positively correlated, such that a higher value of one variable is associated with higher values of the other variable (as was the case with fig wasp thorax and abdomen measurements). Negative values indicate that 2 variables are negatively correlated, such that a higher values of one variable are associated with lower values of the other variable. Values of zero (or not significantly different from zero, more on this later) indicate that 2 variables are uncorrelated (i.e., independent of one another). Figure 29.3 shows scatterplots for 4 different correlation coefficients between values of x and y.

We will look at 2 types of correlation coefficient, the Pearson product moment correlation coefficient and the Spearman rank correlation coefficient. The 2 are basically the same; the Spearman rank correlation coefficient is just a correlation of the ranks of values instead of the actual values.

[1]Note that r is the sample correlation coefficient, which is an estimate of the population correlation coefficient. The population correlation coefficient is represented by the Greek letter 'ρ' ('rho'), and sometimes the sample correlation coefficient is represented as $\hat{\rho}$.

Figure 29.1.: Fig wasps from 2 different species are shown. Wasps were collected from Baja, Mexico.

Figure 29.2.: An example of a scatterplot in which fig wasp thorax length (x-axis) is plotted against fig wasp abdomen length (y-axis). Each point is a different fig wasp. Wasps were collected in 2010 in Baja, Mexico.

Figure 29.3.: Examples of scatterplots with different correlation coefficients (r) between 2 variables (x and y).

29.2.1. Pearson product moment correlation coefficient

To understand the correlation coefficient, we need to first understand covariance. Chapter 12.3 introduced the variance (s^2) as a measure of spread in some variable x,

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 .$$

The variance is actually just a special case of a covariance. The variance describes how a variable x covaries with itself. The covariance ($cov_{x,y}$) describes how a variable x covaries with another variable y,

$$cov_{x,y} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) .$$

The $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively. Note that if $x_i = y_i$, then the equation for $cov_{x,y}$ is identical to the equation for s^2 because $(x_i - \bar{x})(x_i - \bar{x}) = (x_i - \bar{x})^2$.

What the equation for $cov_{x,y}$ is describing is how the variation in x relates to variation in y. If a value of x_i is much higher than the mean $\bar{x}$, and a value of y_i is much higher than the mean $\bar{y}$, then the product of $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will be especially high because we will be multiplying 2 large positive numbers together. If a value of x_i is much higher than the mean $\bar{x}$, but the corresponding y_i is much lower than the mean $\bar{y}$, then the product of $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will be especially low because we will be multiplying a large positive number and a large negative number. Consequently, when x_i and y_i tend to deviate from their means $\bar{x}$ and $\bar{y}$ in a consistent way, we get either high or low values of $cov_{x,y}$. If there is no such relationship between x and y, then we will get $cov_{x,y}$ values closer to zero.

The covariance can, at least in theory, be any real number. How low or high it is will depend on the magnitudes of x and y, just like the variance. To get the Pearson product moment correlation coefficient[2]: (r), we need to standardise the covariance so that the minimum possible value of r is -1 and the maximum possible value of r is 1. We can do this by dividing $cov_{x,y}$ by the product of the standard deviation of x (s_x) and the standard deviation of y (s_y),

$$r = \frac{cov_{x,y}}{s_x \times s_y}.$$

This works because $s_x \times s_y$ describes the total variation between the 2 variables, so the absolute value of $cov_{x,y}$ cannot be larger $s_x \times s_y$. We can again think about the special case in which $x = y$. Since the covariance between x and itself is just the variance of

[2] We can usually just call this the 'correlation coefficient'.

x (s_x^2), and $s_x \times s_x = s_x^2$, we end up with the same value on the top and bottom and $r = 1$.

We can expand $cov_{x,y}$, s_x, and s_y to see the details of the equation for r,

$$r = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}}.$$

This looks like a lot, but we can clean the equation up a bit because the $1/(N-1)$ expressions cancel on the top and bottom of the equation,

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}.$$

As with other statistics defined in this book, it is almost never necessary to calculate r by hand. Statistical programs such as Jamovi and R will make these calculations for us (The Jamovi Project, 2022; R Core Team, 2022). The reason for working through all of these equations is to help make the conceptual link between r and the variance of the variables of interest (Rodgers and Nicewander, 1988). To make this link a bit more clear, we can calculate the correlation coefficient of thorax and abdomen length from Table 29.1. We can set thorax to be the x variable and abdomen to be the y variable. Mean thorax length is $\bar{x} = 0.770$, and mean abdomen length is $\bar{y} = 1.210$. The standard deviation of thorax length is $s_x = 0.064$, and the standard deviation of abdomen length is $s_y = 0.123$. This gives us the numbers that we need to calculate the bottom of the fraction for r, which is $s_x \times s_y = 0.007872$. We now need to calculate the covariance on the top. To get the covariance, we first need to calculate $(x_i - \bar{x})(y_i - \bar{y})$ for each row (i) in Table 29.1. For example, for row 1, $(0.767 - 0.770)(1.288 - 1.210) = -0.000234$. For row 2, $(0.784 - 0.770)(1.059 - 1.210) = -0.002114$. We continue this for all rows. Table 29.2 shows the thorax length (x_i), abdomen length (y_i), and $(x_i - \bar{x})(y_i - \bar{y})$ for rows $i = 1$ to $i = 11$.

Table 29.2.: Measurements of 11 fig wasp thorax and abdomen lengths (mm). The fourth column shows the product of the deviations of each measurement from the mean, where mean thorax length is 0.770 and mean abdomen length is 1.210.

Row (i)	Thorax	Abdomen	Squared Deviation
1	0.767	1.288	-0.000234
2	0.784	1.059	-0.002114
3	0.769	1.107	0.000103
4	0.766	1.242	-0.000128
5	0.828	1.367	0.009106
6	0.852	1.408	0.016236

Row (i)	Thorax	Abdomen	Squared Deviation
7	0.781	1.248	0.000418
8	0.696	1.092	0.008732
9	0.792	1.240	0.000660
10	0.814	1.221	0.000484
11	0.621	1.034	0.026224

If we sum up all of the values in the column "Squared Deviation" from Table 29.2, we get a value of 0.059487. We can multiply this value by $1/(N-1)$ to get the top of the equation for r, $(1/(11-1)) \times 0.059487 = 0.0059487$. We now have all of the values we need to calculate r between fig wasp thorax and abdomen length,

$$r_{x,y} = \frac{0.0059487}{0.064 \times 0.123}.$$

Our final value is $r_{x,y} = 0.7556784$. As suggested by the scatterplot in Figure 29.2, thorax and abdomen lengths are highly correlated. We will test whether or not this value of r is statistically significant in Chapter 23.3 below, but first we will introduce the Spearman's rank correlation coefficient.

29.2.2. Spearman rank correlation coefficient

Throughout this book, we have seen how the *ranks* of data can be substituted in place of the actual values. This has been useful when data violate the assumptions of a statistical test, and we need a nonparametric test instead (e.g., the Wilcoxon signed rank test, the Mann-Whitney U test, or the Kruskall-Wallis H test). We can use the same trick for the correlation coefficient. The Spearman rank correlation coefficient is calculated the exact same way as the Pearson product moment correlation coefficient, except on the ranks of values. To calculate the Spearman rank correlation coefficient for the fig wasp example in the previous section, we just need to rank the thorax and abdomen lengths from 1 to 11, then calculate r using the rank values instead of the actual measurements of length. Figure 29.3 shows the same 11 fig wasp measurements as Figures 29.1 and 29.2, but with columns added to show the ranks of thorax and abdomen lengths.

Table 29.3.: Measurements of 11 fig wasp thorax and abdomen lengths (mm) and their ranks.

Wasp (i)	Thorax	Thorax rank	Abdomen	Abdomen rank
1	0.767	4	1.288	9
2	0.784	7	1.059	2
3	0.769	5	1.107	4

Wasp (i)	Thorax	Thorax rank	Abdomen	Abdomen rank
4	0.766	3	1.242	7
5	0.828	10	1.367	10
6	0.852	11	1.408	11
7	0.781	6	1.248	8
8	0.696	2	1.092	3
9	0.792	8	1.240	6
10	0.814	9	1.221	5
11	0.621	1	1.034	1

Note from Table 29.3 that the lowest value of Thorax is 0.621, so it gets a rank of 1. The highest value of Thorax is 0.852, so it gets a rank of 11. We do the same for abdomen ranks. To get the Spearman rank correlation coefficient, we just calculate r using the ranks. The ranks number from 1 to 11 for both variables, so the mean rank is 6 and the standard deviation is 3.317 for both thorax and abdomen ranks. We can then go through each row and calculate $(x_i - \bar{x}) \times (y_i - \bar{y})$ using the ranks. For the first row, this gives us $(4-6)\,(9-6) = -6$. If we do this same calculation for each row and sum them up, then multiply by 1/(N-1), we get a value of 6.4. To calculate r,

$$r_{rank(x),rank(y)} = \frac{6.4}{3.317 \times 3.317}$$

Our Spearman rank correlation coefficient is therefore $r = 0.582$, which is a bit lower than our Pearson product moment correlation was. The key point here is that the definition of the correlation coefficient has not changed; we are just using the ranks of our measurements instead of the measurements themselves. The reason why we might want to use the Spearman rank correlation coefficient instead of the Pearson product moment correlation coefficient is explained in the next section.

29.3. Correlation hypothesis testing

We often want to test if 2 variables are correlated. In other words, is r significantly different from zero? We therefore want to test the null hypothesis that r is not significantly different from zero.

- H_0 : The population correlation coefficient is zero.
- H_A : The correlation coefficient is significantly different from zero.

Note that H_A above is for a two-tailed test, in which we do not care about direction. We could also use a one-tailed test if our H_A is that the correlation coefficient is greater than (or less than) zero.

How do we actually test the null hypothesis? As it turns out, the sample correlation coefficient (r) will be approximately t-distributed around a true mean (ρ) with a t-score defined by $r - \rho$ divided by its standard error ($SE(r)$)[3],

$$t = \frac{r - \rho}{SE(r)}.$$

Since our null hypothesis is that variables are uncorrelated, $\rho = 0$. Statistical programs such as Jamovi or R will use this equation to test whether or not the correlation coefficient is significantly different from zero (The Jamovi Project, 2022; R Core Team, 2022). The reason for presenting it here is to show the conceptual link to other hypothesis tests in earlier chapters. In Chapter 21.1, we saw that the one sample t-test defined t as the deviation of the sample mean from the true mean, divided by the standard error. Here we are doing the same for the correlation coefficient. One consequence of this is that, like the one sample t-test, the test of the correlation coefficient assumes that r will be normally distributed around the true mean ρ. If this is not the case, and the assumption of normality is violated, then the test might have a misleading Type I error rate.

To be cautious, we should check whether or not the variables that we are correlating are normally distributed (especially if the sample size is small). If they are normally distributed, then we can use the Pearson's product moment correlation to test the null hypothesis. If the assumption of normality is violated, then we might consider using the nonparametric Spearman rank correlation coefficient instead. The fig wasp thorax and abdomen lengths from Table 29.1 are normally distributed, so we can use the Pearson product moment correlation coefficient to test whether or not the correlation between these two variables is significant. In Jamovi, the t-score is not even reported as output when using a correlation test. We only see r and the p-value (Figure 29.4).

From Figure 29.4, we can see that the sample $r = 0.75858$, and the p-value is $P = 0.00680$. Since our p-value is less than 0.05, we can reject the null hypothesis that fig wasp thorax and abdomen lengths are not correlated. To get a more intuitive sense of how the correlation coefficient works, we can use an interactive application.

> Click here for an interactive application showing the relationship between a scatterplot, a correlation coefficient, and statistical significance.

Points can be added to the interactive application by clicking anywhere within the plot. A table of x and y points is shown to the right, and the correlation coefficient and p-value can be shown or hidden using the buttons on the top.

[3] We can calculate the standard error of r as (Rahman, 1968),

$$SE(r) = \sqrt{\frac{1 - r^2}{N - 2}}.$$

But this is not necessary to ever do by hand. Note that we lose 2 degrees of freedom (N - 2), one for calculating each variable mean.

Correlation Matrix

Correlation Matrix

		Thorax_Length_mm	Abdomen_Length_mm
Thorax_Length_mm	Pearson's r	—	
	p-value	—	
Abdomen_Length_mm	Pearson's r	0.75858	—
	p-value	0.00680	—

Figure 29.4.: Jamovi output for a test of the null hypothesis that thorax length and abdomen length are not significantly correlated in a sample of fig wasps collected in 2010 from Baja, Mexico.

30. *Practical*. Analysis of counts and correlations

This lab focuses on applying the concepts from Chapter 28 and Chapter 29 in Jamovi (The Jamovi Project, 2022). Exercises in this practical will use the Chi-squared goodness of fit test, the Chi-squared test of association, and the correlation coefficient. For the chi-square tests, this practical will use a dataset inspired by the doctoral work of Dr Jessica Burrows (Burrows et al., 2022). This experimental work tested the effects of radiation on bumblebee nectar consumption, carbon dioxide output, and body mass in different bee colonies (Figure 30.1)[1].

Figure 30.1.: This lab practical focuses on the effects of radiation on bumblebees.

The practical will use the bumblebee.csv dataset (right click and "Save Link As…", then save it with the extension '.csv'). This dataset includes variables for the radiation level experienced by the bee (radiation), the colony from which the bee came (colony), whether or not the bee survived to the end of the 30 day experiment (survived), the mass of the bee in grams at the beginning of the experiment (mass), the output of carbon dioxide put out by the bee (CO_2) in micromoles per minute, and the daily volume of nector consumed by the be in ml (nectar).

[1] This figure was released into the public domain by Marco Almbauer on 4 May 2018. Please note that the data in this practical are for educational purposes only. They are not the data that were actually collected by Burrows et al. (2022).

30.1. Survival goodness of fit

Suppose that we want to run a simple goodness of fit test to determine whether or not bees are equally likely to survive versus die in the experiment. If this is the case, then we would expect to see the same number of living and dead bees in the dataset. We can use a Chi-square goodness of fit test to answer this question. What are the null and alternative hypotheses for this χ^2 goodness of fit test?

- H_0: _____
- H_A: _____

What is the sample size (N) of the dataset?

N: _____

Based on this sample size, what are the expected counts for bees that survived and died?

Survived (E_{surv}): _____

Died (E_{died}): _____

Next, we can find the observed counts of bees that survived and died. To do this, we need to use the Frequency tables option in Jamovi. We did this once in Chapter 16 for calculating probabilities. As a reminder, to find the counts of bumblebees that survived (Yes) or did not survive (No), we need to go to the Exploration toolbar in Jamovi, then choose 'Descriptives'. Place 'Survival' in the Variables box, then check the box for 'Frequency tables' below. A Frequencies table will appear in the panel on the right. Write down the observed counts of bees that survived and died.

Survived (O_{surv}): _____

Died (O_{died}): _____

Try to use the formula in Chapter 28.2 to calculate the χ^2 test statistic. Here is what it should look like for the two counts in this dataset,

$$\chi^2 = \frac{(O_{surv} - E_{surv})^2}{E_{surv}} + \frac{(O_{died} - E_{died})^2}{E_{died}}.$$

What is the χ^2 value?

χ^2 : _____

There are 2 categories for the survival category (Yes and No). How many degrees of freedom are there?

df: _____

Using the χ^2 and df values, find the approximate p-value using this interactive application.

P: _____

Now we can try to use Jamovi to replicate the analysis above. In Jamovi Analyses tab, select 'Frequencies' from the toolbar, then select 'N outcomes'. A new window will open up (Figure 30.2).

Figure 30.2.: Jamovi toolbar showing the location of the Frequencies options, with the Chi-square goodness of fit test available under a choice called 'N Outcomes' in the pulldown menu.

After selecting the option 'N Outcomes: χ^2 Goodness of fit', a new window will appear called 'Proportion Test (N Outcomes)'. To run a χ^2 goodness of fit test on bee survival, move the 'survived' variable in to the 'Variable' box. Leave the Counts box empty (Figure 30.3).

Figure 30.3.: Jamovi interface for running a Chi-square goodness of fit test on bumblebee survival in a dataset.

The χ^2 Goodness of Fit table will appear in the panel to the right. From this table, we can see the χ^2 test statistic, degrees of freedom (df), and the p-value (p). Do these values

match the values that you calculated by hand and with the interactive application?

Next, we will try another goodness of fit test, but this time to test whether or not bees were taken from all colonies with the same probability.

30.2. Colony goodness of fit

Next, suppose that we want to know if bees were sampled from the colonies with the same expected frequency. What are the null and alternative hypotheses in this scenario?

- H_0: _____
- H_A: _____

How many colonies are there in this dataset?

Colonies: _____

Run the χ^2 goodness of fit test using the same procedure in Jamovi that you used in the previous exercise. What is the output from the Goodness of Fit table?

χ^2: _____

df: _____

p: _____

From this output, what can you conclude about how bees were taken from the colonies?

Note that the distrACTION module in Jamovi includes a χ^2 distribution (called 'x2-Distribution'), which you can use to compute probabilities and quantiles in the same way as we did for previous distributions in the module. Next, we will move on to a χ^2 test of association between

30.3. Chi-Square test of association

Suppose we want to know if there is an association between bee colony and bee survival. We can use a χ^2 test of association to investigate this question. What are the null and alternative hypotheses for this test of association?

- H_0: _____
- H_A: _____

To run the χ^2 test of association, choose 'Frequencies' from the Jamovi toolbar (Figure 30.2), but this time select 'Independent Samples: χ^2 test of association' from the pull-down menu. To test for an association between bee colony and survival, place 'colony' in the 'Rows' box and 'survived' in the 'Columns' box. Leave the rest of the boxes blank (Figure 30.4).

Figure 30.4.: Jamovi interface for running a Chi-square test of association on bumblebee survival versus colony in a dataset.

There is a pulldown called 'Statistics' below the Contingency Tables input. Make sure that that the χ^2 checkbox is selected. Output from the χ^2 test of association will appear in the panel to the right. Report the key statistics in the output table below.

χ^2: _____

df: _____

p: _____

From these statistics, should you reject or not reject the null hypothesis?

H_0: _____

Note that scrolling down further in the left panel (Contingency Tables) reveals an option for plotting. Have a look at this and create a barplot by checking 'Bar Plot' under **Plots**. Note that there are various options for changing bar types (side by side or stacked), y-axis limits (counts versus percentages), and bar groupings (by rows or columns).

Now try running a χ^2 test of association to see if there is an association between radiation and bee survival (hint, you just need to swap 'colony' for 'radiation' in the Rows box). What can you conclude from this test? Explain your conclusion as if were reporting the results of the test to someone who was unfamiliar with statistical hypothesis testing.

Lastly, did the order which you placed the two variables matter? What if you switched Rows and Columns? In other words, put 'survived' in the Rows box and 'radiation' in the columns box. Does this give you the same answer?

Next, we will look at correlations between variables.

30.4. Pearson product moment correlation test

Suppose that we want to test if bumblebee mass at the start of the experiment (mass) is associated with the amount of nectar consumed (nectar). Specifically, we want to know if more massive bees also output less carbon dioxide (CO_2). Before running any test, it is a good idea to plot the two variables using a scatterplot. To do this, select the 'Exploration' button from the toolbar in Jamovi, but instead of choosing 'Descriptives' as usual from the pulldown menu, select 'Scatterplot'. A new window will open up that allows you to build a scatterplot by selecting the variables that you place on the x-axis and the y-axis. Put mass on the x-axis and nectar on the y-axis, as shown in Figure 30.5.

Notice that the scatterplot appears in the panel on the right. Each point in the scatterplot is a different bee (i.e., row). Just looking at the scatterplot, does it appear as though bee mass and CO_2 output are correlated? Why or why not?

Figure 30.5.: Jamovi interface for building a scatterplot with bumblebee mass on the x-axis and nectar consumption on the y-axis

Note that it is possible to separate points in the scatterplot by group., Try placing 'survived' in the box 'Group'.

Now we can test whether or not bee mass and CO_2 output are negatively correlated. What are the null and alternative hypotheses of this test?

- H_0: _____
- H_A: _____

Before we test whether or not the correlation coefficient (r) is significant, we need to know which correlation coefficient to use. Remember from Chapter 29.3 that a test of the Pearson product moment correlation assumes that the sample r is normally distributed around the true correlation coefficient. If both of our variables (mass and CO_2) are normally distributed, then we can be confident that this assumption will not be violated. But if one or both variables are not normally distributed, then we should consider using the Spearman rank correlation coefficient instead. To test if mass and CO_2 are normally distributed, navigate to the Descriptives panel in Jamovi (where we usually find the summary statistics of variables). Place mass and CO_2 in the 'Variables' box, then scroll down and notice that there is a checkbox under **Normality** for 'Shapiro-Wilk'. Check this box, then find the p-values for the Shapiro-Wilk test of normality in the panel to the right. Write these p-values down below.

Mass P: _____

CO_2 P: _____

Based on these p-values, which type of correlation coefficient should we use to test H_0, and why?

To run the correlation coefficient test, choose the button in the Jamovi toolbar called 'Regression', then select the first option 'Correlation Matrix' from the pulldown menu (Figure 30.6).

Figure 30.6.: Jamovi toolbar showing the location of the Regression options, with the Correlation Matrix option available in the pulldown menu.

The Correlation Matrix option will pull up a new window in Jamovi (Figure 30.7).

Figure 30.7.: Jamovi interface for testing correlation coefficients.

Notice that the Pearson product moment correlation is selected in the checkbox of Figure 30.7 ('Pearson'). Immediatley below this checkbox is a box called 'Spearman', which

would report the Spearman rank correlation coefficient test. Below the **Correlation Coefficients** options, there are options for **Hypothesis**. Remember that we are interested in the alternative hypothesis that mass and CO_2 are negatively correlated, so we should select the radio button 'Correlated negatively'.

The output of the correlation test appears in the panel on the right in the form of a table called 'Correlation Matrix'. This table reports both the correlation coefficient (here called "Pearson's r") and the p-value. Write these values below.

r: _____

P: _____

Based on this output what should we conclude about the association between bumblebee mass and carbon dioxide output?

Next, we will test the whether or not bee mass is associated with nectar consumption.

30.5. Spearman rank correlation test

Next, we will test whether or not bee mass and nectar consumption are correlated. What are the null and alternative hypotheses of this test?

- H_0: _____
- H_A: _____

Run a Shapiro-Wilk test of normality on each of the two variables, as was done in the previous exercise. Based on the output of these tests, what kind of correlation coefficient should we use for testing the null hypothesis?

Correlation coefficient: _____

Test whether or not bee mass and nectar consumption are correlated. What is the correlation coefficient and p-value from this test?

r: _____

P: _____

Based on these results, do should we reject or not reject the null hypothesis?

H_0: _____

Suppose that we had used the Pearson product moment correlation coefficient instead of the Spearman rank correlation coefficient. Would we have made the same conclusion

about the correlation (or lack thereof) between bee mass and CO_2 output? Why or why not?

If you have sufficient time, move on to the final exercise, which will demonstrate an additional way to run Chi-square tests.

30.6. Untidy goodness of fit

In Exercise 30.1, we ran a χ^2 test using data in a tidy format, in which each row corresponded to a single observation and categorical data were listed over $N = 256$ rows. For the 'survived' variable, this meant 256 rows of 'Yes' or 'No'. But there is a shortcut in Jamovi if we do not have a full tidy dataset. If you know that the dataset included 139 'Yes' counts and 117 'No' counts, you could set up the data as a table of counts (Table 30.1).

Table 30.1.: Counts of bees that did not survive (No) or did survive (Yes) in an experiment involving radiation.

Survived	Count
No	117
Yes	139

Open a new data frame in Jamovi, then recreate the small dataset in Table 30.1. Column names should be 'Survived' and 'Count', as shown in Figure 30.8.

Figure 30.8.: Jamovi data frame with a simple organisation of count data.

Next, navigate to the 'Analyses' tab and choose 'N Outcomes' to do a goodness of fit test. Place 'Survived' in the Variable box, then place 'Count' in the Counts (optional) box. Notice that you will get the same χ^2, df, and p values in the output table as you did in Exercise 30.1

We could do the same for a χ^2 test of association, although it would be a bit more complicated. To test for an association between radiation and survival, as we did at the end of Exercise 30.3, we would need 3 columns and 8 rows of data (Table 30.2).

Table 30.2.: Counts of bees that did not survive (No) or did survive (Yes) for different levels of radiation.

Survived	Radiation	Count
No	None	12
Yes	Low	52
No	Medium	29
Yes	High	35
No	None	39
Yes	Low	25
No	Medium	37
Yes	High	27

If we put Table 30.2 into Jamovi, we can run a χ^2 test of association by navigating to the 'Frequencies' button in the Jamovi toolbar and selecting 'Independent Samples: χ^2 test of assocation' from the pulldown. In the Contingency Tables input panel, we can put 'Survived' in the Rows box, 'Radiation' in the Columns box, then place 'Count' in the Counts (optional) box. The panel on the right will give us the output of the χ^2 test of association.

Part X.

Linear Regression

Week 10 Overview

Week: 10 Dates: Suggested Readings: Textbook Assessments: Practice quiz Practical:

31. Regression key concepts

32. Regression validity

33. Introduction to multiple regression

General explanation

34. Model selection (maybe remove this?)

Seriously consider moving the regression into this week. and ease the amount of material in previous weeks.

35. *Practical*. Using regression

35.1. Regression Exercise 1

35.2. Regression Exercise 2

35.3. Regression Exercise 3

35.4. Regression Exercise 4

Part XI.

Randomisation approaches

Week 11 Overview

The aim of this lecture is to introduce the randomisation approach to statistical hypothesis testing. We will first introduce the general idea of what randomisation is and how it relates to the hypothesis testing that we have been doing since week five. We will then consider an instructive example in which a randomisation approach is used in place of a traditional t-test to test whether or not the mean values of two different groups are identical. We will then compare the assumptions underlying randomisation and how they differ slightly from the assumptions of traditional hypothesis testing. We will then look at how randomisation can be used to build confidence intervals and test hypotheses that would difficult to test with other approaches. In learning about randomisation approaches, we will also review some key concepts from earlier in the module. The aim is not to understand all of the nuances of randomisation, but to understand, conceptually, what is going on in the methods described below.

Week: 11 Dates: Suggested Readings: Textbook Assessments: Practice quiz Practical: R starts creeping in now?

36. Introduction to randomisation

General explanation

37. Assumptions of randomisation

How these differ

38. Bootstrapping

What this is and why we use it.

39. Monte Carlo

40. *Practical*. Using R

40.1. R Exercise 1

40.2. R Exercise 2

40.3. R Exercise 3

Part XII.

Statistical Reporting

Week 12 Overview

Week: 12 Dates: Suggested Readings: Textbook Assessments: Practice quiz Practical: R starts creeping in now?

41. Reporting statistics

General explanation

42. More introduction to R

How these differ

43. More getting started with R

Just more to do.

44. *Practical*. Using R

44.1. R Exercise 1

44.2. R Exercise 2

44.3. R Exercise 3

Part XIII.

Review of parts (VII-XII)

Module summary

This chapter will be specifically to prepare for exam.

A. Common Marking Scheme

This appendix is a recreation of Appendix 1: Undergraduate Common Marking Scheme on the University of Stirling website.

Table A.1.: **Undergraduate Common Marking Scheme Information**: Column one lists the mark out of 100, column two lists the equivalent grade, column three describes the result as pass or fail, and column four describes the attainment of the learning outcome.

Mark	Equivalent Grade	Result	Descriptor of Attainment of Learning Outcomes
90+	1st	Pass	Meets all the requirements to attain 80-89 but in addition demonstrates an exceptional degree of originality and exceptional analytical, problem-solving and/or creative skills.
80-89			Meets all the requirements to attain 70-79 but in addition demonstrates outstanding quality evidenced by an ability to engage critically and analytically with source material, exhibits independent lines of argument, is highly original and uses an extremely wide range of relevant sources where appropriate.
70-79			Excellent range and depth of attainment of intended learning outcomes, secured by discriminating command of a comprehensive range of relevant materials and analyses, and by deployment of considered judgement relating to key issues, concepts or procedures.
60-69	2:1	Pass	Attainment of virtually all intended learning outcomes, clearly grounded on a close familiarity with a wide range of supporting evidence, constructively utilised to reveal an appreciable depth of understanding.

Mark	Equivalent Grade	Result	Descriptor of Attainment of Learning Outcomes
50-59	2:2	Pass	Attainment of most of the intended learning outcomes, some more securely grasped than others, resting on a circumscribed range of evidence and displaying a variable depth of understanding.
40-49	3rd	Pass	Acceptable attainment of most intended learning outcomes, displaying a qualified familiarity with a minimally sufficient range of relevant materials, and a grasp of the analytical issues and concepts which are generally reasonable, albeit insecure.
30-39	Fail - Marginal	Fail	Appreciable deficiencies in the attainment of intended learning outcomes, perhaps lacking a secure basis in relevant factual or analytical dimensions.
0-29	Fail - Clear	Fail	No convincing evidence of attainment of intended learning outcomes, such treatment of the subject as is in evidence being directionless and fragmentary.
X	Fail	Fail	Failure to comply with Compulsory Module Requirements or engage with the module, leading to no automatic right to reassessment.

B. Uncertainty derivation

It is not necessary to be able to derive the equations for propagating error from week 2, but working through the below might be interesting, and provide a better appreciation for why these formulas make sense. Another derivation is available in Box et al. (1978) (page 563), but this derivation is expressed in terms of variances and covariances, which is likely to be less helpful for this module.

Propagation of error for addition and subtraction.

For adding and subtracting error, we know that we get our variable Z by adding X and Y. This is just how Z is defined. We also know that Z is going to have some error E_Z, and we know that Z plus or minus its error will equal X plus or minus its error plus Y plus or minus its error,

$$(Z \pm E_Z) = (X \pm E_X) + (Y \pm E_Y).$$

Again, this is just our starting definition, but double-check to make sure it makes sense. We can now note that we know,

$$Z = X + Y.$$

If it is not intuitive as to why, just imagine that there is no error associated with the measurement of X an Y (i.e., $E_X = 0$ and $E_Y = 0$). In this case, there cannot be any error in Z. So, if we substitute $X + Y$ for Z, we have the below,

$$((X + Y) \pm E_Z) = (X \pm E_X) + (Y \pm E_Y).$$

By the associative property, we can get rid of the parenthesis for addition and subtraction, giving us the below,

$$X + Y \pm E_Z = X \pm E_X + Y \pm E_Y.$$

Now we can subtract X and Y from both sides and see that we just have the errors of X, Y, and Z,

$$\pm E_Z = \pm E_X \pm E_Y.$$

The plus/minus is a bother. Note, however, that for any real number m, $m^2 = (-m)^2$. For example, if $m = 4$, then $(4)2 = 16$ and $(-4)2 = 16$, so we can square both sides to get positive numbers and make things easier,

$$E_Z^2 = (\pm E_X \pm E_Y)^2.$$

We can expand the above,

$$E_Z^2 = E_X^2 + E_Y^2 \pm 2E_X E_Y.$$

Now here is an assumption that we have not told you about elsewhere in the module. With the formulas that we have given you, we are assuming that the errors of X and Y are independent. To put it in more statistical terms, the covariance between the errors of X and Y is assumed to be zero. Without going into the details (covariance will be introduced later in the module), if we assume that the covariance between these errors is zero, then we can also assume the last term of the above is zero, so we can get rid of it (i.e., $2E_X E_Y = 0$),

$$E_Z^2 = E_X^2 + E_Y^2.$$

If we take the square root of both sides, then we have the equation from Chapter 7,

$$E_Z = \sqrt{E_X^2 + E_Y^2}.$$

Propagation of error for multiplication and division.

Now that we have seen the logic for propagating errors in addition and subtraction, we can do the same for multiplication and division. We can start with the same point that we are getting our new variable Z by multiplying X and Y together, $Z = XY$. So, if both X and Y have errors, the errors will be multiplicative as below,

$$Z \pm E_Z = (X \pm E_X)(Y \pm E_Y).$$

Again, all we are doing here is substituting Z, X, and Y, for an expression in parentheses that includes the variable plus or minus its associated error. Now we can expand the right hand side of the equation,

$$Z \pm E_Z = XY + YE_X + XE_Y + E_X E_Y.$$

As with our propagation of error in addition, here we are also going to assume that the sources of error for X and Y are independent (i.e., their covariance is zero). This allows us to set $E_X E_Y = 0$, which leaves us with the below,

$$Z \pm E_Z = XY + YE_X + XE_Y.$$

Now, because $Z = XY$, we can substitute on the left hand side of the equation,

$$XY \pm E_Z = XY + YE_X + XE_Y.$$

Now we can subtract the XY from both sides of the equation,

$$\pm E_Z = YE_X + XE_Y.$$

Next, let us divide both sides by XY,

$$\frac{\pm E_Z}{XY} = \frac{YE_X + XE_Y}{XY}.$$

We can expand the right hand side,

$$\frac{\pm E_Z}{XY} = \frac{YE_X}{XY} + \frac{XE_Y}{XY}.$$

This allows us to cancel out the Y variables in the first term of the right hand side, and the X variables in second term of the right hand side,

$$\frac{\pm E_Z}{XY} = \frac{E_X}{X} + \frac{E_Y}{Y}.$$

Again, we have the plus/minus on the left, so let us square both sides,

$$\left(\frac{\pm E_Z}{XY}\right)^2 = \left(\frac{E_X}{X} + \frac{E_Y}{Y}\right)^2.$$

We can expand the right hand side,

$$\left(\frac{\pm E_Z}{XY}\right)^2 = \left(\frac{E_X}{X}\right)^2 + \left(\frac{E_Y}{Y}\right)^2 + 2\left(\frac{E_X}{X}\right)\left(\frac{E_Y}{Y}\right).$$

Again, because we are assuming that the errors of X and Y are independent, we can set the third term on the right hand side of the equation to zero. This leaves,

$$\left(\frac{\pm E_Z}{XY}\right)^2 = \left(\frac{E_X}{X}\right)^2 + \left(\frac{E_Y}{Y}\right)^2.$$

Note that $XY = Z$, so we can substitute in the left hand side,

$$\left(\frac{\pm E_Z}{Z}\right)^2 = \left(\frac{E_X}{X}\right)^2 + \left(\frac{E_Y}{Y}\right)^2.$$

Now we can apply the square on the left hand side to the top and bottom, which gets rid of the plus/minus,

$$\frac{E_Z^2}{Z^2} = \left(\frac{E_X}{X}\right)^2 + \left(\frac{E_Y}{Y}\right)^2.$$

We can now multiply both sides of the equation by Z^2,

$$E_Z^2 = Z^2\left(\left(\frac{E_X}{X}\right)^2 + \left(\frac{E_Y}{Y}\right)^2\right).$$

We can now take the square root of both sides,

$$E_Z = \sqrt{Z^2\left(\left(\frac{E_X}{X}\right)^2 + \left(\frac{E_Y}{Y}\right)^2\right)}.$$

We can pull the Z^2 out of the square root,

$$E_Z = Z\sqrt{\left(\frac{E_X}{X}\right)^2 + \left(\frac{E_Y}{Y}\right)^2}.$$

That leaves us with the equation that was given in Chapter 7.

C. Statistical tables

C.1. Wilcoxon signed rank critical values

N	alpha == 0.1	alpha == 0.05	alpha == 0.01
5	0	–	–
6	2	0	–
7	3	2	–
8	5	3	0
9	8	5	1
10	10	8	3
11	13	10	5
12	17	13	7
13	21	17	9
14	25	21	12
15	30	25	15
16	35	29	19
17	41	34	23
18	47	40	27
19	53	46	32
20	60	52	37
21	67	58	42
22	75	65	48
23	83	73	54
24	91	81	61
25	100	89	68
26	110	98	75
27	119	107	83
28	130	116	91
29	140	126	100
30	151	137	109
31	163	147	118
32	175	159	128
33	187	170	138
34	200	182	148
35	213	195	159
36	227	208	171
37	241	221	182
38	256	235	194
39	271	249	207
40	286	264	220
41	302	279	233
42	319	294	247
43	336	310	261
44	353	327	276
45	371	343	291
46	389	361	307
47	407	378	322
48	426	396	339
49	446	415	355
50	466	434	373

C.2. Mann-Whitney U critical values

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	–	–	–	–	–	–	0	0	0	0	1	1	1	1	1	2	2	2	2
3	–	–	–	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	–	–	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
5	–	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	–	1	2	3	5	6	7	10	11	13	14	16	17	19	21	22	24	25	27
7	–	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	0	2	4	6	7	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
21	3	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119	126	134
22	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125	133	141
23	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132	140	149
24	3	10	17	25	33	42	50	59	67	76	85	94	102	111	120	129	138	147	156
25	3	10	18	27	35	44	53	62	71	80	89	98	107	117	126	135	145	154	163
26	4	11	19	28	37	46	55	64	74	83	93	102	112	122	132	141	151	161	171
27	4	11	20	29	38	48	57	67	77	87	97	107	117	127	137	147	158	168	178
28	4	12	21	30	40	50	60	70	80	90	101	111	122	132	143	154	164	175	186
29	4	13	22	32	42	52	62	73	83	94	105	116	127	138	149	160	171	182	193
30	5	13	23	33	43	54	65	76	87	98	109	120	131	143	154	166	177	189	200
31	5	14	24	34	45	56	67	78	90	101	113	125	136	148	160	172	184	196	208
32	5	14	24	35	46	58	69	81	93	105	117	129	141	153	166	178	190	203	215
33	5	15	25	37	48	60	72	84	96	108	121	133	146	159	171	184	197	210	222
34	5	15	26	38	50	62	74	87	99	112	125	138	151	164	177	190	203	217	230
35	6	16	27	39	51	64	77	89	103	116	129	142	156	169	183	196	210	224	237
36	6	16	28	40	53	66	79	92	106	119	133	147	161	174	188	202	216	231	245
37	6	17	29	41	55	68	81	95	109	123	137	151	165	180	194	209	223	238	252
38	6	17	30	43	56	70	84	98	112	127	141	156	170	185	200	215	230	245	259
39	7	18	31	44	58	72	86	101	115	130	145	160	175	190	206	221	236	252	267
40	7	18	31	45	59	74	89	103	119	134	149	165	180	196	211	227	243	258	274

Bibliography

Adams, D. C. and Collyer, M. L. (2016). On the comparison of the strength of morphological integration across morphometric datasets. *Evolution*, 70(11):2623–2631.

Askey, R. (1999). Why does a negative x a negative = a positive? *American Educator*, pages 4–5.

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., and Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods*, 50(3):937–962.

Borowski, E. J. and Borwein, J. M. (2005). *Collins Dictionary of Mathematics*. HarperCollins Publishers, London, 2 edition.

Bouma, G. D. (2000). *The Research Process*. Oxford University Press, Oxford, UK, 4 edition.

Box, G. E. P., Hunter, W. G., and S, H. J. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. John Wiley & Sons, New York.

Burrows, J. E., Copplestone, D., Raines, K. E., Beresford, N. A., and Tinsley, M. C. (2022). Ecologically relevant radiation exposure triggers elevated metabolic rate and nectar consumption in bumblebees. *Functional Ecology*, 36(8):1822–1833.

Cheadle, C., Vawter, M. P., Freed, W. J., and Becker, K. G. (2003). Analysis of microarray data using Z score transformation. *Journal of Molecular Diagnostics*, 5(2):73–81.

Chernoff, E. J. and Zazkis, R. (2022). The simple reason a viral math equation stumped the internet.

Choi, W., Lee, J. W., Huh, M. H., and Kang, S. H. (2003). An Algorithm for Computing the Exact Distribution of the Kruskal-Wallis Test. *Communications in Statistics Part B: Simulation and Computation*, 32(4):1029–1040.

Courant, R., Robbins, H., and Stewart, I. (1996). *What is Mathematics?* Oxford University Press, Oxford, 2 edition.

Delacre, M., Lakens, D., and Leys, C. (2017). Why psychologists should by default use welch's t-Test instead of student's t-Test. *International Review of Social Psychology*, 30(1):92–101.

Duthie, A. B., Abbott, K. C., and Nason, J. D. (2015). Trade-offs and coexistence in fluctuating environments: evidence for a key dispersal-fecundity trade-off in five nonpollinating fig wasps. *American Naturalist*, 186(1):151–158.

Duthie, A. B. and Nason, J. D. (2016). Plant connectivity underlies plant-pollinator-exploiter distributions in Ficus petiolaris and associated pollinating and non-pollinating fig wasps. *Oikos*.

Dytham, C. (2011). *Choosing and Using Statistics: A Biologist's Guide*. John Wiley & Sons.

Edwards, A. W. F. (1972). *Likelihood: An account of the statistical concept of likelihood and its application to scientific inference*. Cambridge University Press, Cambridge.

Elavsky, F., Bennett, C., and Moritz, D. (2022). How accessible is my visualization? Evaluating visualization accessibility with Chartability. *Computer Graphics Forum*, 41(3):57–70.

Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology Letters*, 7(6):509–520.

Fedor-Freybergh, P. G. and Mikulecký, M. (2006). Erratum: From the descriptive towards inferential statistics: Hundred years since conception of the Student's t-distribution. *Neuroendocrinology Letters*, 27(3):292.

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Gregor, L., Hauck, J., Le Quéré, C., Luijkx, I. T., Olsen, A., Peters, G. P., Peters, W., Pongratz, J., Schwingshackl, C., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Alkama, R., Arneth, A., Arora, V. K., Bates, N. R., Becker, M., Bellouin, N., Bittig, H. C., Bopp, L., Chevallier, F., Chini, L. P., Cronin, M., Evans, W., Falk, S., Feely, R. A., Gasser, T., Gehlen, M., Gkritzalis, T., Gloege, L., Grassi, G., Gruber, N., Gürses, O., Harris, I., Hefner, M., Houghton, R. A., Hurtt, G. C., Iida, Y., Ilyina, T., Jain, A. K., Jersild, A., Kadono, K., Kato, E., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J. I., Landschützer, P., Lefèvre, N., Lindsay, K., Liu, J., Liu, Z., Marland, G., Mayot, N., McGrath, M. J., Metzl, N., Monacci, N. M., Munro, D. R., Nakaoka, S.-I., Niwa, Y., O'Brien, K., Ono, T., Palmer, P. I., Pan, N., Pierrot, D., Pocock, K., Poulter, B., Resplandy, L., Robertson, E., Rödenbeck, C., Rodriguez, C., Rosan, T. M., Schwinger, J., Séférian, R., Shutler, J. D., Skjelvan, I., Steinhoff, T., Sun, Q., Sutton, A. J., Sweeney, C., Takao, S., Tanhua, T., Tans, P. P., Tian, X., Tian, H., Tilbrook, B., Tsujino, H., Tubiello, F., van der Werf, G. R., Walker, A. P., Wanninkhof, R., Whitehead, C., Willstrand Wranne, A., Wright, R., Yuan, W., Yue, C., Yue, X., Zaehle, S., Zeng, J., and Zheng, B. (2022). Global carbon budget 2022. *Earth System Science Data*, 14(11):4811–4900.

Fuentes-Montemayor, E., Park, K. J., Cordts, K., and Watts, K. (2022a). The long-term development of temperate woodland creation sites: from tree saplings to mature woodlands. *Forestry*, 95(June 2021):28–37.

Fuentes-Montemayor, E., Watts, K., Sansum, P., Scott, W., and Park, K. J. (2022b). Moth community responses to woodland creation: The influence of woodland age, patch characteristics and landscape attributes. *Diversity and Distributions*, 28(9):1993–2007.

Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38.

Gotelli, N. J. (2001). *Gotelli*. Sinauer Associates, Inc., Sunderland, Massachusetts, 3 edition.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28(706):49–50.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3):e1002106.

Hyndman, R. J. and Fan, Y. (1996). Sample Quantiles in Statistical Packages. *American Statistician*, 50(4):361–365.

Johnson, D. H. (1995). Statistical sirens: The allure of nonparametrics. *Ecology*, 76(6):1998–2000.

Kelleher, C. and Wagener, T. (2011). Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling and Software*, 26(6):822–827.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, 23(4):525–540.

Kruskal, W. H. and Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Lande, R. (1977). On comparing coefficients of variation. *Systematic Zoology*, 26(2):214–217.

Law, A., Bunnefeld, N., and Willby, N. J. (2014). Beavers and lilies: Selective herbivory and adaptive foraging behaviour. *Freshwater Biology*, 59(2):224–232.

Lumley, T., Diehr, P., Emerson, S., and Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23:151–169.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. University of Chicago Press, Chicago.

Mayo, D. G. (2019). P-value thresholds: Forfeit at your peril. *European Journal of Clinical Investigation*, 49(10):1–4.

Mayo, D. G. (2021). Significance Tests: Vitiated or Vindicated by the Replication Crisis in Psychology? *Review of Philosophy and Psychology*, 12(1):101–120.

Mclean, R. A., Sanders, W. L., and Stroup, W. W. (1991). A unified approach to mixed linear models. *American Statistician*, 45(1):54–64.

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019). Abandon Statistical Significance. *American Statistician*, 73:235–245.

Miller, I. and Miller, M. (2004). *John E. Freund's mathematical statistics*. Pearson Prentice Hall, Upper Saddle River, New Jersey, 7 edition.

Narum, S. R. (2006). Beyond Bonferroni: Less conservative analyses for conservation genetics. *Conservation Genetics*, 7(5):783–787.

Navarro, D. J. and Foxcroft, D. R. (2022). *Learning Statistics with Jamovi*. (Version 0.75).

Pastor, J. (2008). *Mathematical Ecoogy of Populations and Ecosystems*. John Wiley & Sons, Inc, West Sussex, England.

Pélabon, C., Hilde, C. H., Einum, S., and Gamelon, M. (2020). On the use of the coefficient of variation to quantify and compare trait variation. *Evolution Letters*, 4(3):180–188.

Preston, C. M. and Schmidt, M. W. (2006). Black (pyrogenic) carbon: A synthesis of current knowledge and uncertainties with special consideration of boreal regions. *Biogeosciences*, 3(4):397–420.

Quinn, T. J. (1995). Base units of the Système International d'Unités, their accuracy, dissemination and international traceability. *Metrologia*, 31(6):515–527.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabinovich, S. G. (2013). *Evaluating Measurement Accuracy*.

Rahman, N. A. (1968). *A Course in Theoretical Statistics*. Charles Griffin and Company.

Rencher, A. C. (2000). *Linear Models in Statistics*. John Wiley & Sons, Inc, Provo, Utah.

Rihs, M. and Mayer, B. (2018). distraction-calculating and plotting distributions.

Rodgers, J. L. and Nicewander, W. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

Rowntree, D. (2018). *Statistics Without Tears*. Penguin.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, 17(4):688–690.

Santín, C., Doerr, S. H., Kane, E. S., Masiello, C. A., Ohlson, M., de la Rosa, J. M., Preston, C. M., and Dittmar, T. (2016). Towards a global assessment of pyrogenic carbon from vegetation fires. *Global Change Biology*, 22(1):76–91.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., and Bühner, M. (2010). Is It Really Robust?: Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology*, 6(4):147–151.

Slakter, M. J. (1968). Accuracy of an Approximation to the Power of the Chi-Square Goodness of Fit Test with Small but Equal Expected Frequencies. *Journal of the American Statistical Association*, 63(323):912–918.

Sokal, R. R. and Rohlf, F. J. (1995). *Biometry*. W. H. Freeman and Company, New York, 3rd edition.

Spiegelhalter, D. (2019). *The Art of Statistics Learning from Data*. Penguin UK.

Stanton-Geddes, J., De Freitas, C. G., and De Sales Dambros, C. (2014). In defense of P values: Comment on the statistical methods actually used by ecologists. *Ecology*, 95(3):637–642.

Stewart, I. (2008). *Taming the Infinite*. Quercus, London.

Stock, M., Davis, R., De Mirandés, E., and Milton, M. J. (2019). Corrigendum: The revision of the SI - The result of three decades of progress in metrology (Metrologia (2019) 56 (022001) DOI: 10.1088/1681-7575/ab0013). *Metrologia*, 56(4).

Suárez, M. (2020). *Philosophy of Probability and Statistical Modelling*. Cambridge University Press, Cambridge.

Tate, M. W. and Hyer, L. A. (1973). Inaccuracy of the X2 test of goodness of fit when expected frequencies are small. *Journal of the American Statistical Association*, 68(344):836–841.

The Jamovi Project (2022). jamovi.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *American Statistician*, 70(2):129–133.

Weiblen, G. D. (2002). How to be a fig wasp. *Annual Review of Entomology*, 47:299–330.

Welch, B. L. (1938). The Significance of the Difference Between Two Means when the Population Variances are Unequal. *Biometrika*, 29(3/4):350.

Wickham, H. (2014). Tiday Data. *Journal of Statistical Software*, 59(10):1–23.

Yee, A. J. (2019). Google Cloud Topples the Pi Record. Technical report.