Statistical Techniques for Biological and Environmental Sciences

Brad Duthie

2023-01-01

Contents

Preface

Welcome to the module. This workbook will be used throughout the semester and contain all of the information that you need for the statistical techniques (SCIU4T4) module.

What is statistics?

An explanation of the material, and what will be taught.

Why this module is important

Some discussion of module importance

Teaching overview

Here is how you will be taught, with online lectures, reading assignments, and face-to-face practicals.

Assessment overview

You will have one formative test and two summatitive tests. You will also have one mock exam and one exam exam.

Test 1F

Information about Test 1F

Test 1S

Information about Test 1S

Test 2S

Information about Test 1S

Mock Exam

Information about the mock exam

Exam

Information about the exam

Jamovi statistical software

Introduction to Jamovi, and why we are using it instead of other software.

Textbooks

Introduction to the primary textbook Learning statistics with jamovi, and a mention of other sources.

Canvas

How we will use Canvas, and how this book relates to it (Learning and Teaching content, where lectures, assessments, and discussions can be found).

Timetable

Table 1.: Table caption.

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)

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

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: Wickham (2014) (Download)
	Optional: Navarro and Foxcroft (2022) Section 2.1
Lectures	1.0: Introduction to Module (20 min.)
	1.1: Numbers and operations (18 min.)
	1.2: Orders of operations (7 min.)
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 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

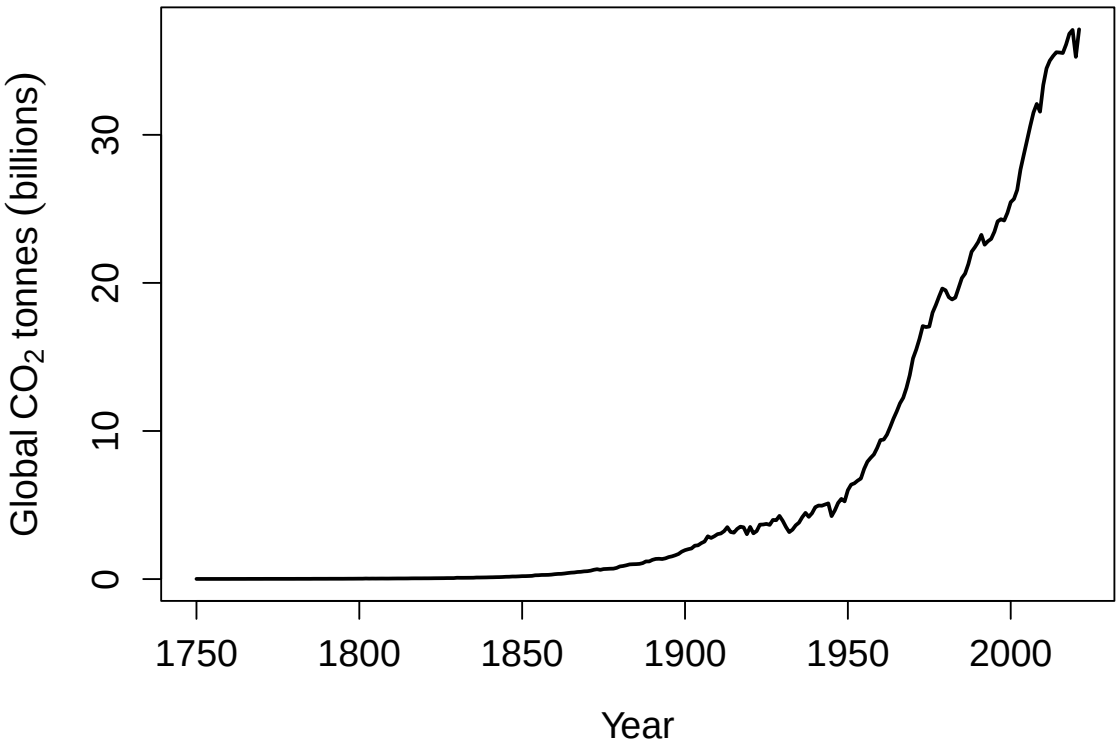


Figure 1.1.: Global carbon dioxide emissions from 1750-2021.

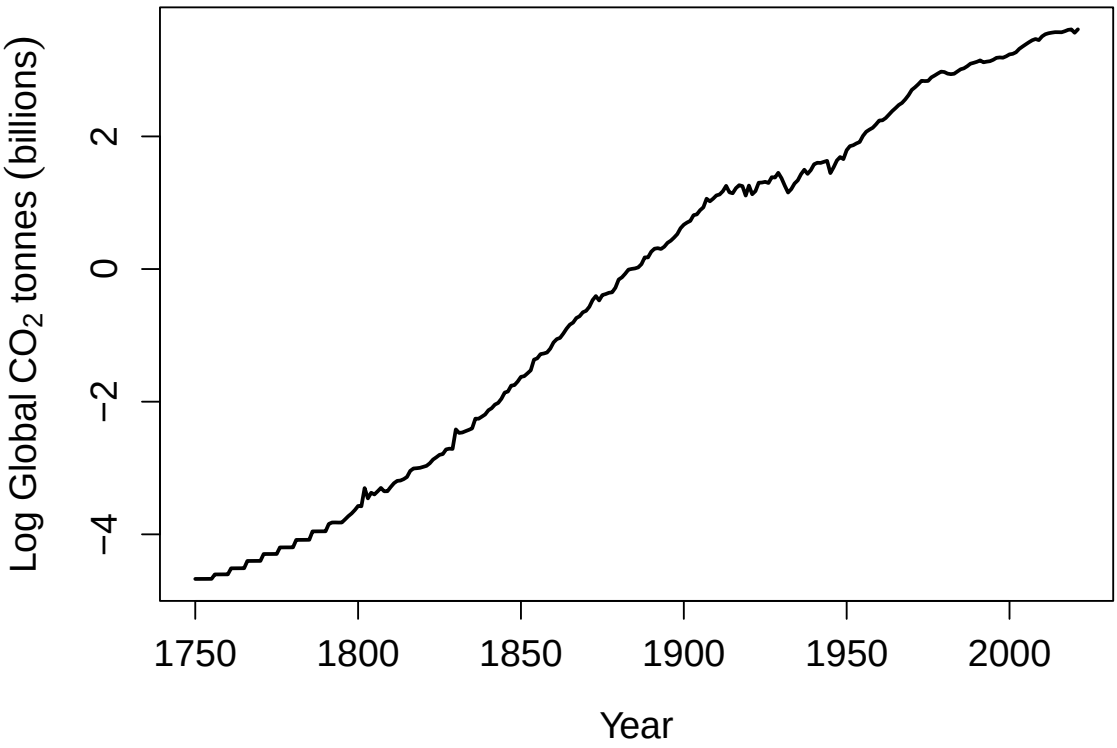


Figure 1.2.: Natural logarithm of global carbon dioxide emissions from 1750-2021.

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,

$$x = \sqrt{3 + 4^2} = \sqrt{(3 + 4^2)} \approx 3.59.$$

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.

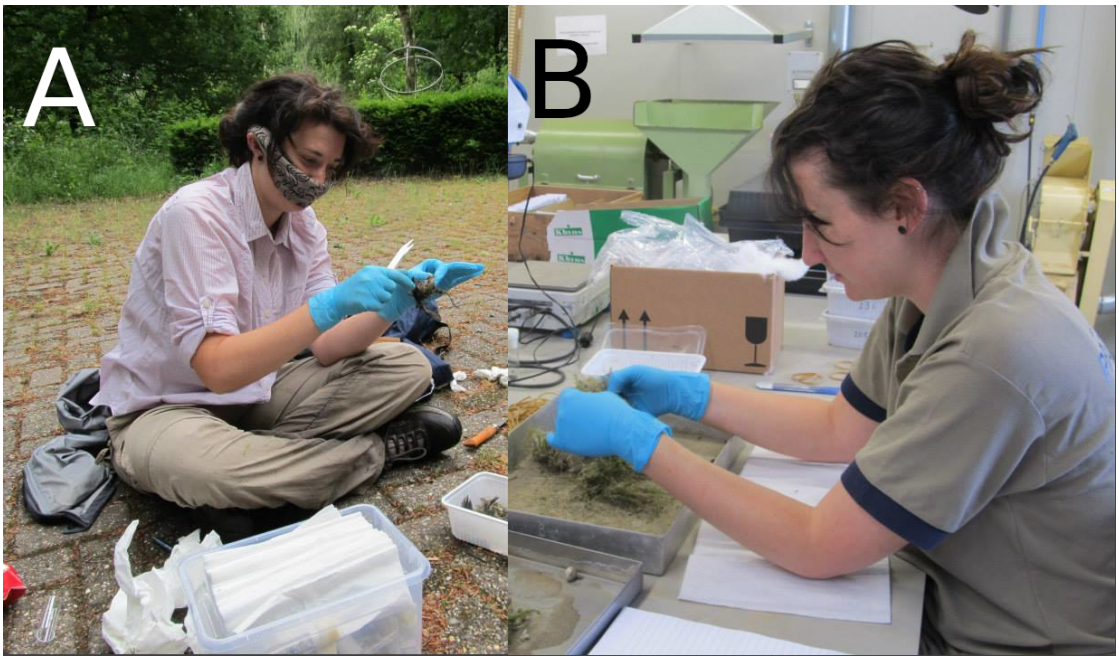


Figure 2.1.: Dr Becky Boulton collects data from nest boxes in the field (A), then processes nest material in the lab (B).

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. 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 file that can be read and used in statistical software such as Jamovi or R.

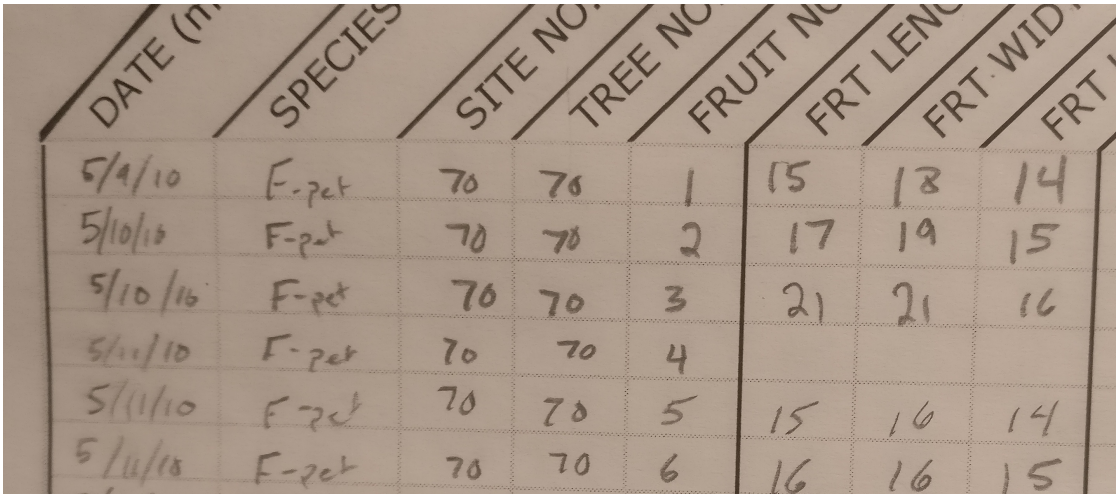
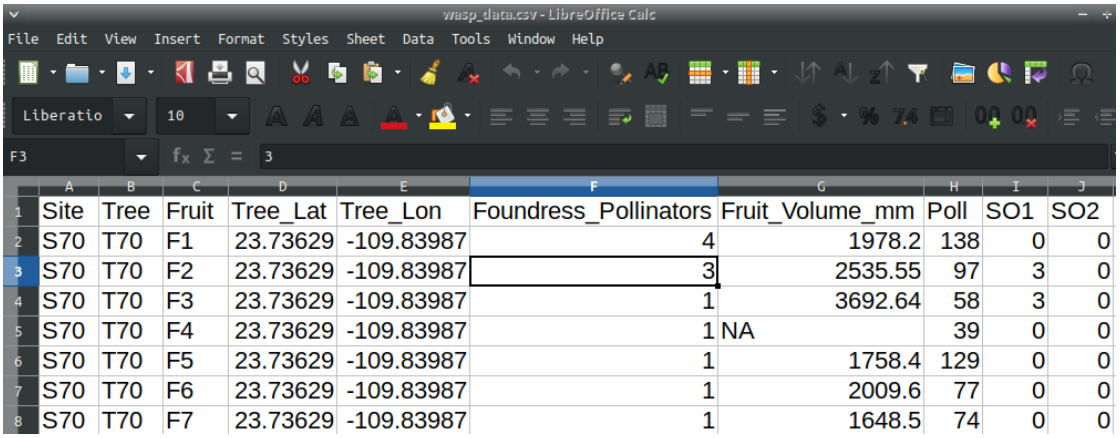


Figure 2.2.: A portion of a lab notebook used to record measurements of fig fruits from different trees in Baja, Mexico, in 2010.

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.



	A	B	C	D	E	F	G	H	I	J
1	Site	Tree	Fruit	Tree_Lat	Tree_Lon	Foundress_Pollinators	Fruit_Volume_mm	Poll	SO1	SO2
2	S70	T70	F1	23.73629	-109.83987	4	1978.2	138	0	0
3	S70	T70	F2	23.73629	-109.83987	3	2535.55	97	3	0
4	S70	T70	F3	23.73629	-109.83987	1	3692.64	58	3	0
5	S70	T70	F4	23.73629	-109.83987	1	NA	39	0	0
6	S70	T70	F5	23.73629	-109.83987	1	1758.4	129	0	0
7	S70	T70	F6	23.73629	-109.83987	1	2009.6	77	0	0
8	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.

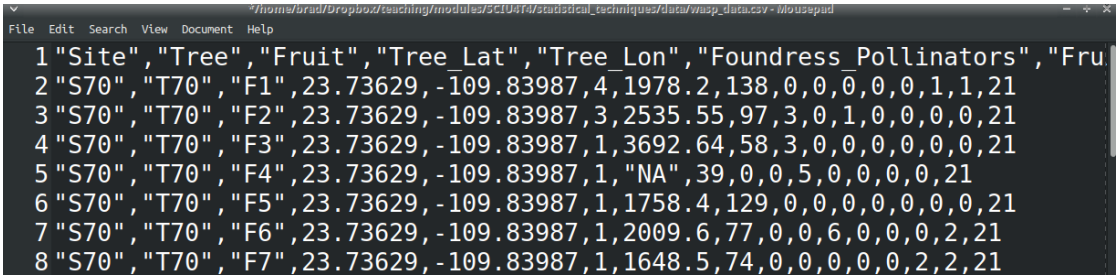


Figure 2.4.: A plain text comma-separated value (CSV) file showing data from fig fruits collected in 2010. Each line is a unique row and sample (fruit), and commas separate the data into columns in which the properties of fruit are recorded. The file has been opened in a program called 'Mousepad', but 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.

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 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.

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,

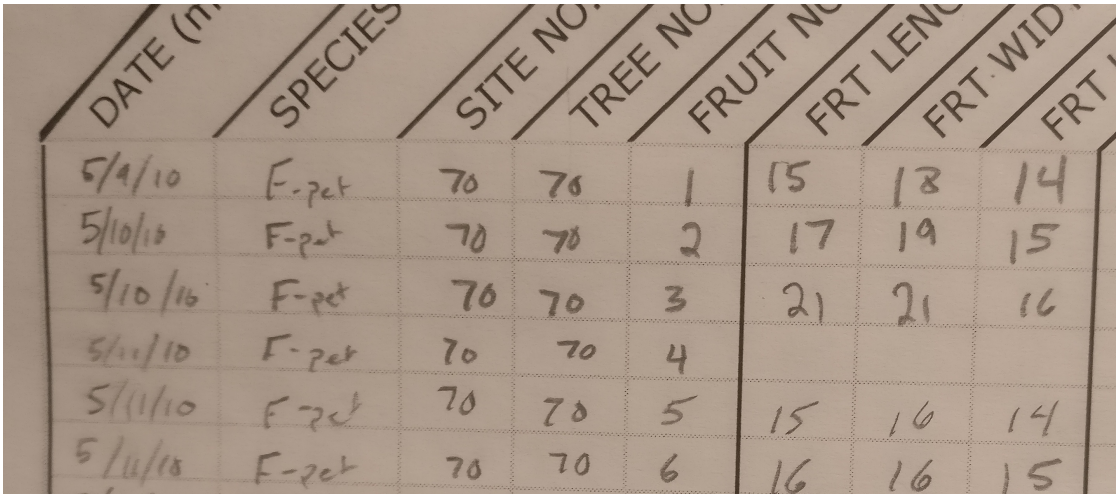


Figure 3.2.: A portion of a lab notebook used to record measurements of fig fruits from different trees in 2010.

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

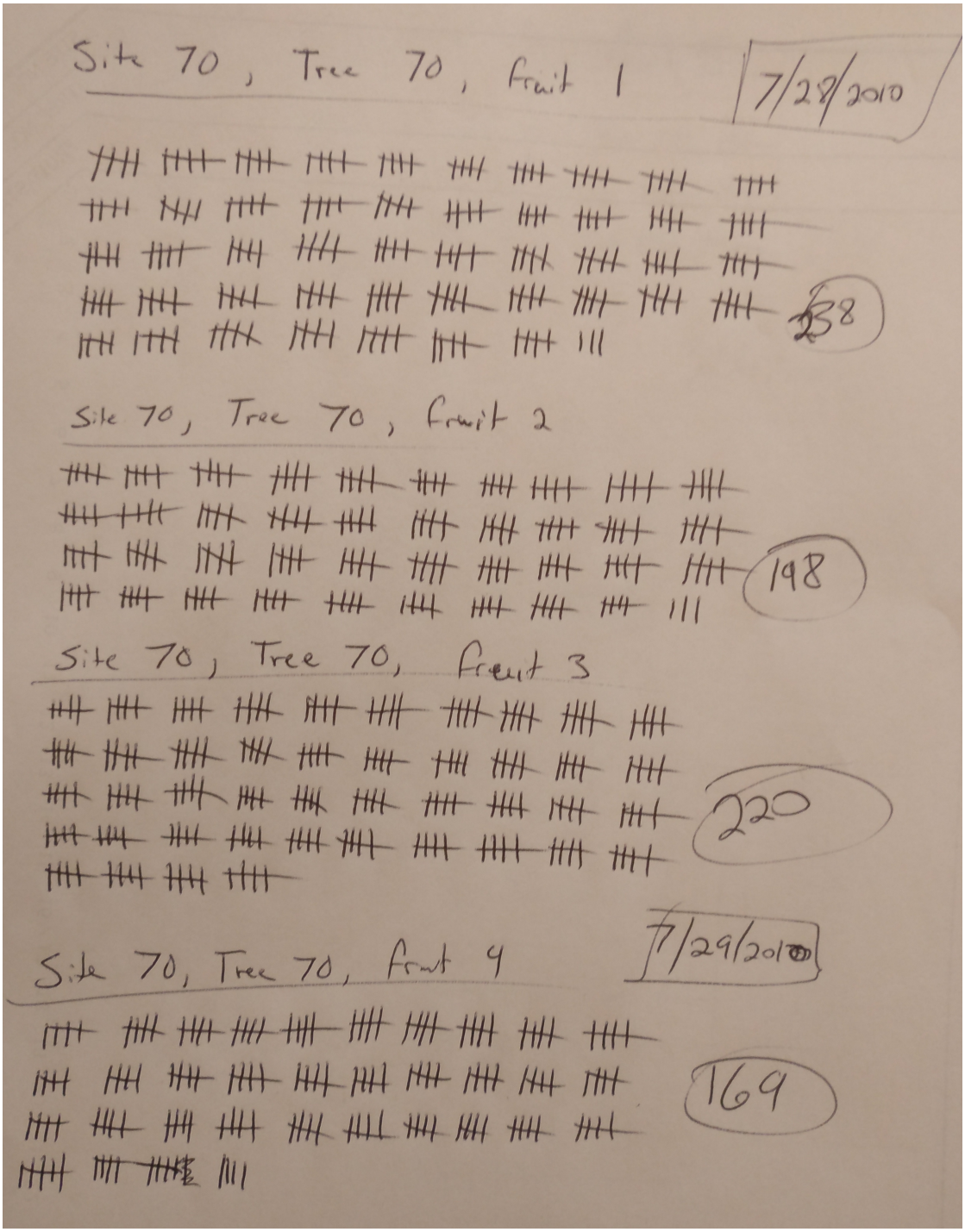


Figure 3.4.: Tallies of seed counts collected from 4 fig fruits in Baja, Mexico in 2010.

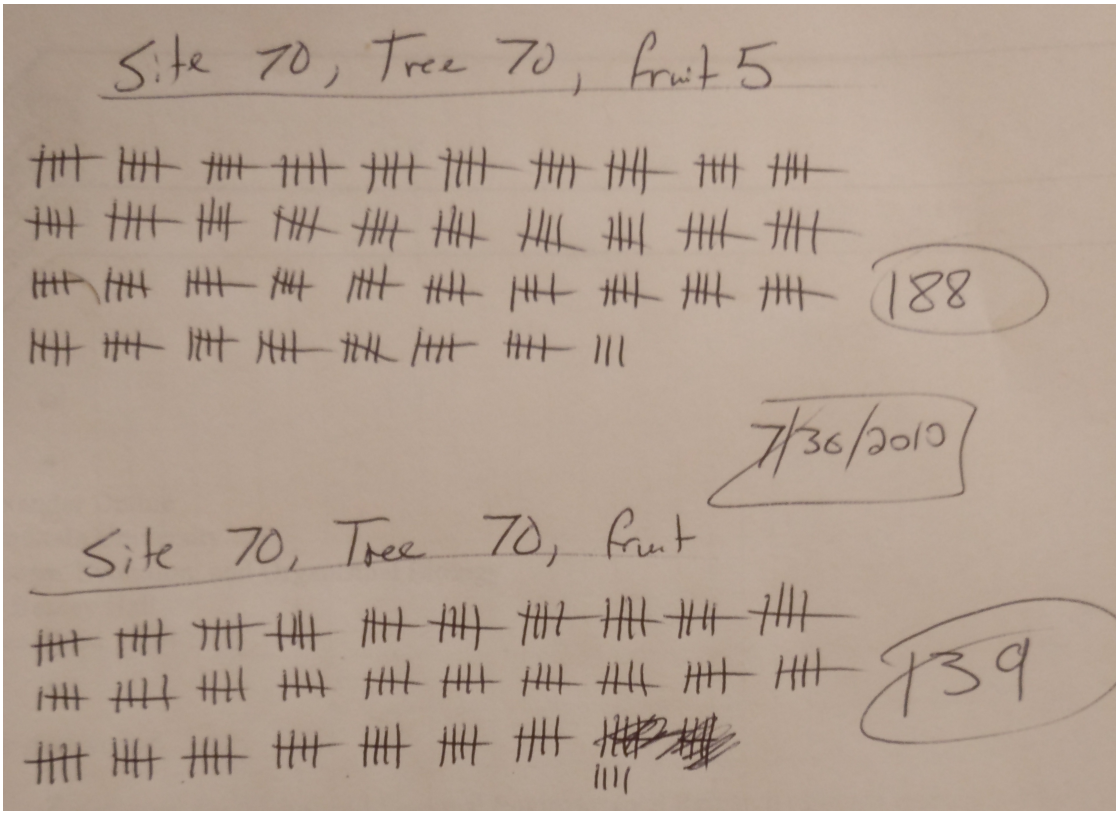


Figure 3.5.: Tallies of seed counts collected from 2 fig fruits in Baja, Mexico in 2010.

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? _____

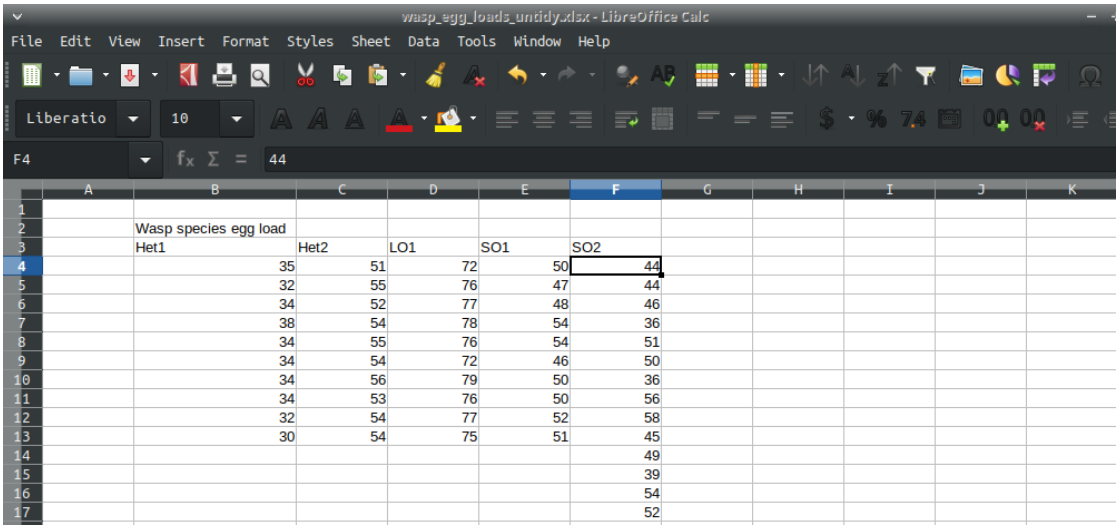


Figure 3.6.: An untidy dataset of egg loads from fig wasps of five different species, including two unnamed species of the genus *Heterandrium* (Het1 and Het2) and three unnamed species of the genus *Idarnes* (LO1, SO1, and SO2).

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.

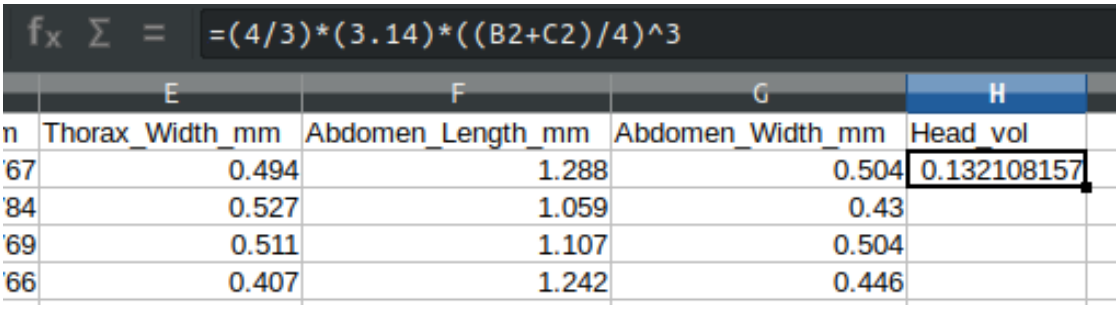
Symbol	Operation
`+`	Addition
`-`	Subtraction
`*`	Multiplication
`/`	Division
`^`	Exponent
`sqrt()`	Square-root

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,



	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
	Optional: Rowntree (2018) Chapter 2
	Advanced: Spiegelhalter (2019) Chapters 1-3
Lectures	2.0: Introduction to Week 2 (2 min.)
	2.1: Why study statistics? (18 min.)
	2.2: Populations and samples (7 min.)
	2.3: Types of variables (11 min.)
	2.4: Units, precision, and accuracy (9 min.)
	2.5. Uncertainty propagation (11 min.)
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.

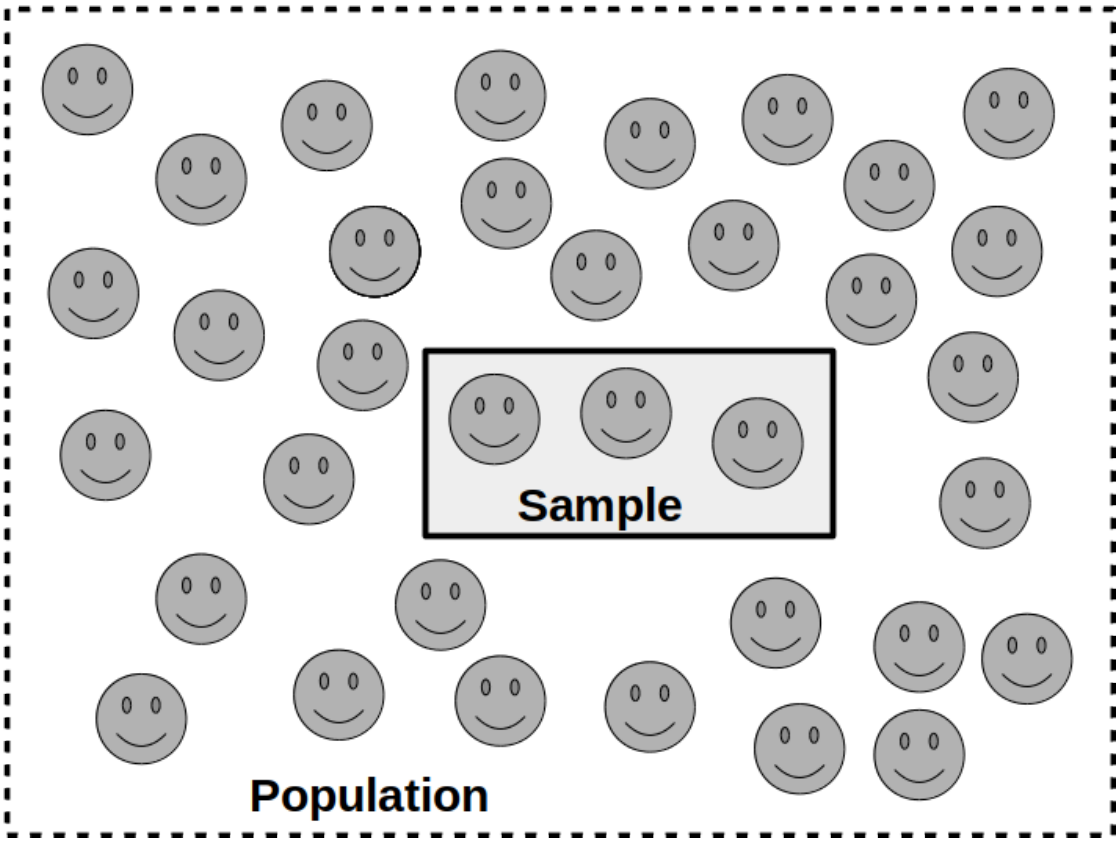


Figure 4.1.: A conceptual figure illustrating how a statistical population relates to a statistical sample. The population is represented by 35 smiling faces enclosed within a dashed box. The sample is represented by a solid box within the dashed box, within which there are 3 smiling faces. Hence, we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

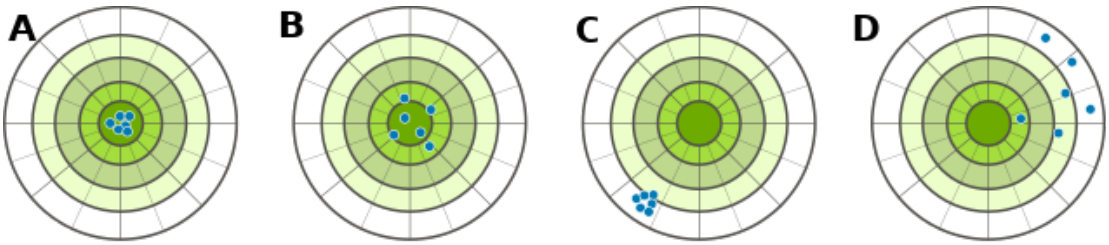


Figure 6.1.: A conceptual figure illustrating the difference between accuracy and precision. Points in (A) are both accurate and precise, points in (B) are accurate, but not precise, points in (C) are precise but not accurate, and points in (D) are neither accurate nor precise.

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.

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 $E_X^2 = 3$, 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 $E_W = 1$ 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. 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 module, 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**

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**

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, 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).

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.

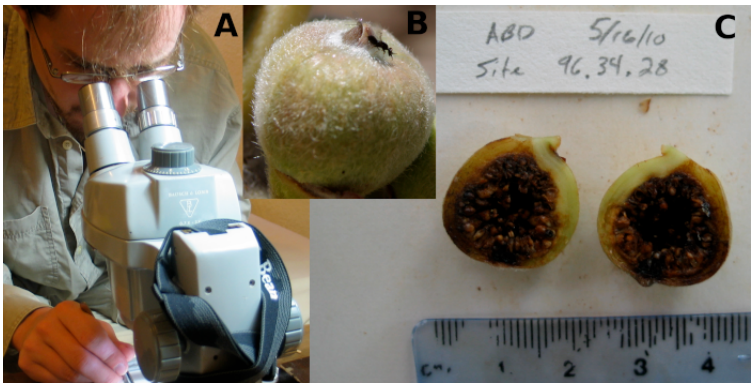


Figure 8.1.: Three images showing the process of collecting data for the dimensions of figs from trees of the Sonoran Desert Rock Fig in Baja, Mexico. (A) Processing fig fruits, which included measuring the diameter of figs along three diferent axes of length, width, and height, (B) a fig still attached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

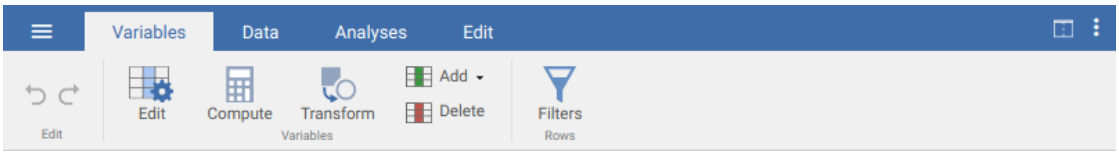


Figure 8.2.: The Jamovi toolbar including tabs for opening files, Variables, Data, Analyses, and Edit. To open a file, select the three horizontal lines in the upper right

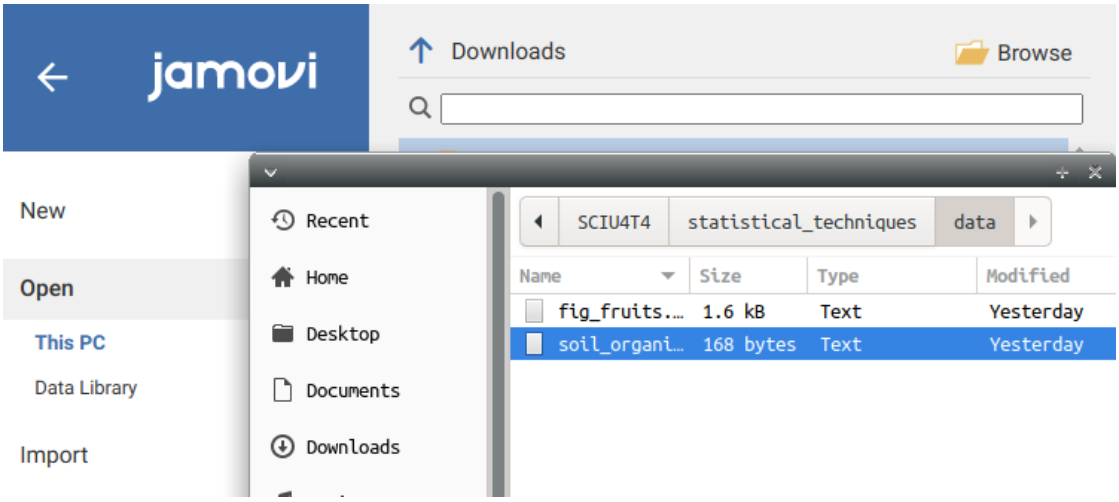


Figure 8.3.: The Jamovi interface for opening a file with the 'Import' tab selected. Options for browsing to a file on the computer are available in the upper right, which opens the window in the foreground.

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 Figure 8.2). 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).

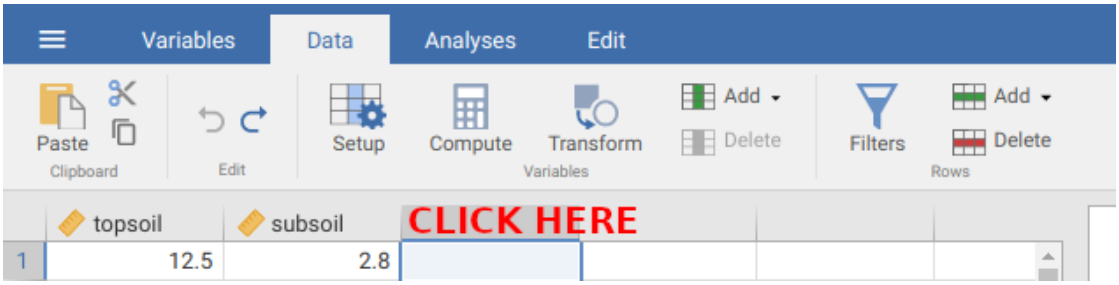


Figure 8.4.: The Jamovi toolbar is shown with the soil organic carbon dataset. In Jamovi, double-clicking above column three where it says 'CLICK HERE' will allow you to input a new variable.

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).

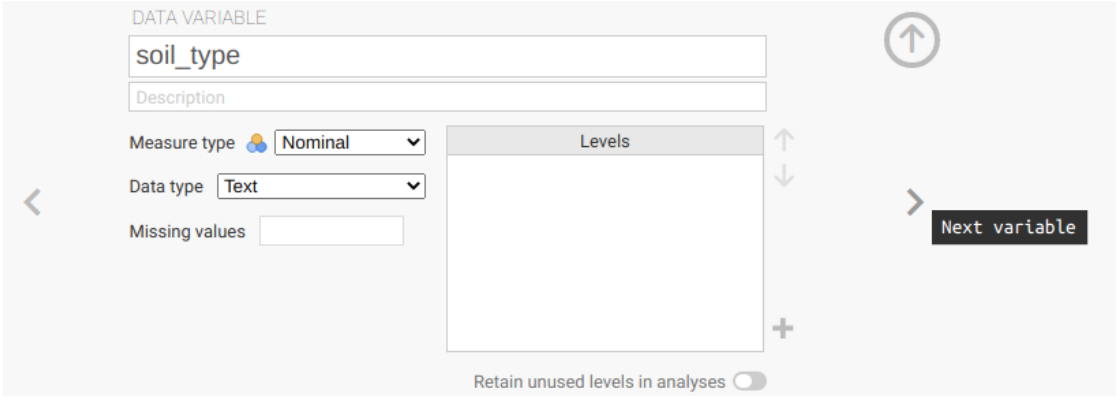


Figure 8.5.: The Jamovi toolbar is shown with the input for creating a new data variable. The new variable added is to indicate the soil type (topsoil or subsoil), so it needs to be a nominal variable with a data type of text.

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).

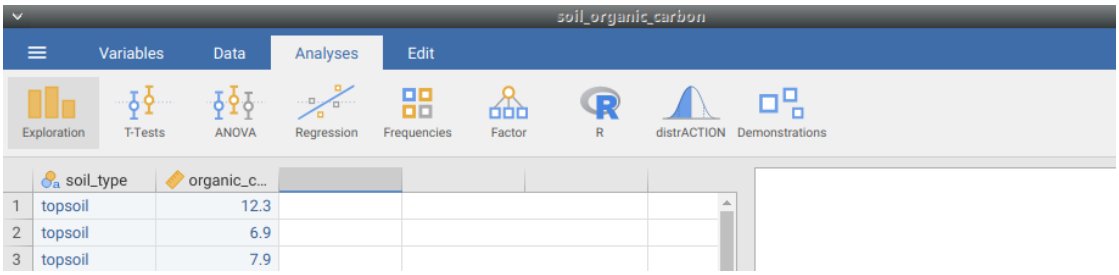


Figure 8.6.: The Jamovi toolbar where the tab 'Analyses' can be selected at the very top. Below this tab, the button 'Exploration' can be clicked to calculate descriptive statistics.

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.

Topsoil Mean: _____

Topsoil Minimum: _____

Topsoil 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. 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: _____

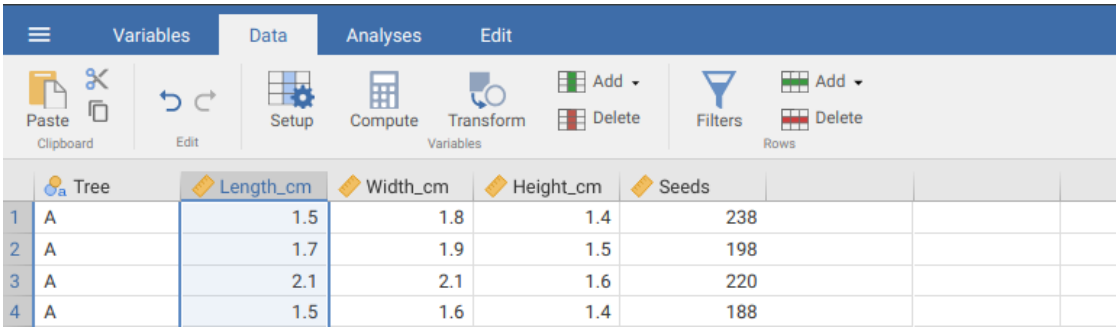
Grand Mean width: _____

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.



	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

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

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 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).

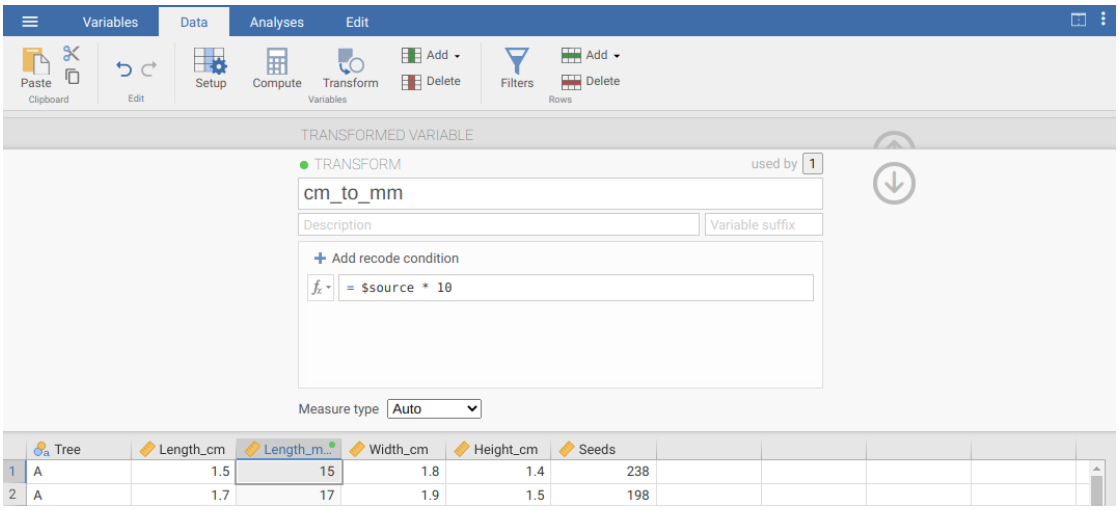


Figure 8.8.: The Jamovi toolbar where the tab 'Data' is selected. The box below shows the transform, which has been named 'cm to mm'. The transformation occurs by multiplying the source (Length mm) by 10. The dataset underneath shows the first few rows with the transformed column highlighted (note that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

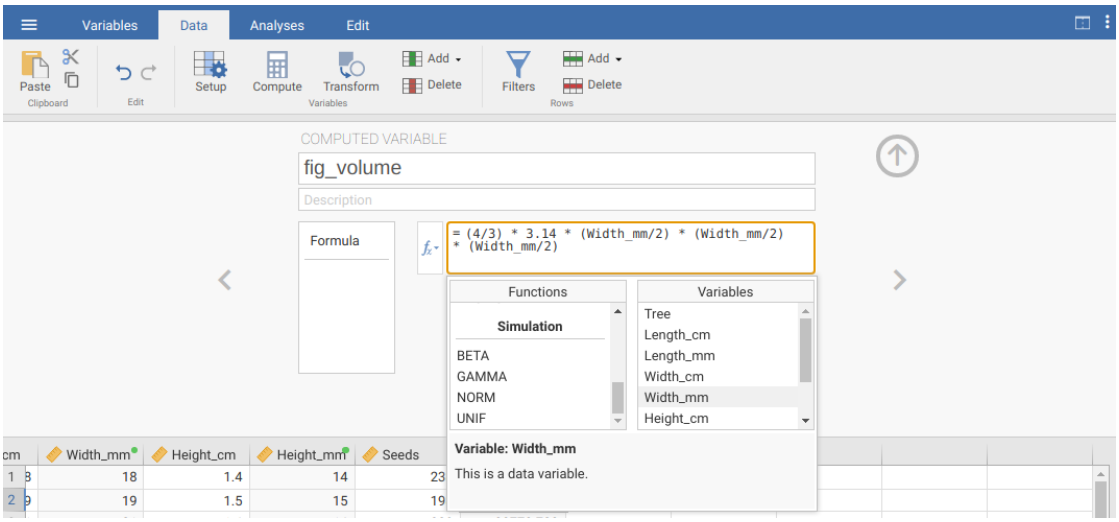


Figure 8.9.: The Jamovi toolbar where the tab 'Data' is selected. The box below shows the new computed variable 'fig volume', which has been created by calculating the product of 4/3, 3.14, and Width mm three times.

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).

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).

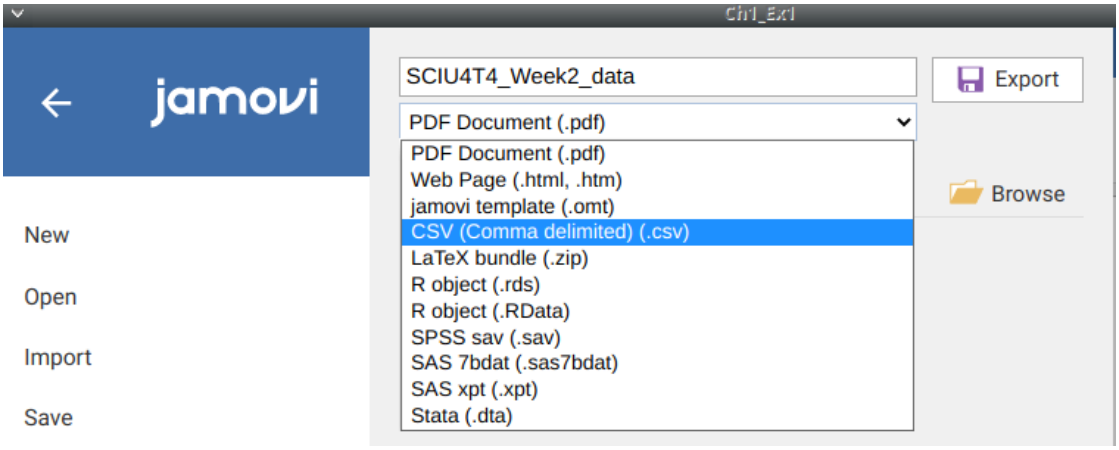


Figure 8.10.: The Jamovi Export menu in which data are be saved as a CSV using the pulldown menu below the filename

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 **Optional:** Rowntree (2018) Chapter 3
Lectures	3.0: Decimal places and significant figures part 1 (8 min.) 3.1: Decimal places and significant figures part 2 (7 min.) 3.2: Graphs (11 min.) 3.3: Box-whisker plots (8 min.) 3.4: The mean (17 min.) 3.5: The mode (7 min.) 3.6: The median and quantiles (8 min.) 3.7: Mean, mode, median, and resistance (9 min.)
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 2 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 *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 two 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 some 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	5
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 3.1415 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 3.146 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 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 four 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. Visualising a dataset 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 the person reading the graph has a clear understanding what is being presented. In practice, this means clearly labelling axes with meaningful descriptions and appropriate units, including a descripting 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 from the fig fruits data set from the practical in Chapter 8.

The histogram in Figure 10.1 shows how many fruits there are for different intervals of width (for a step-by-step demonstration of how a histogram is built, see this interactive application[1]). That is, 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 that has a width greater than 2.0 cm (the biggest is 2.1 cm), so we see that

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

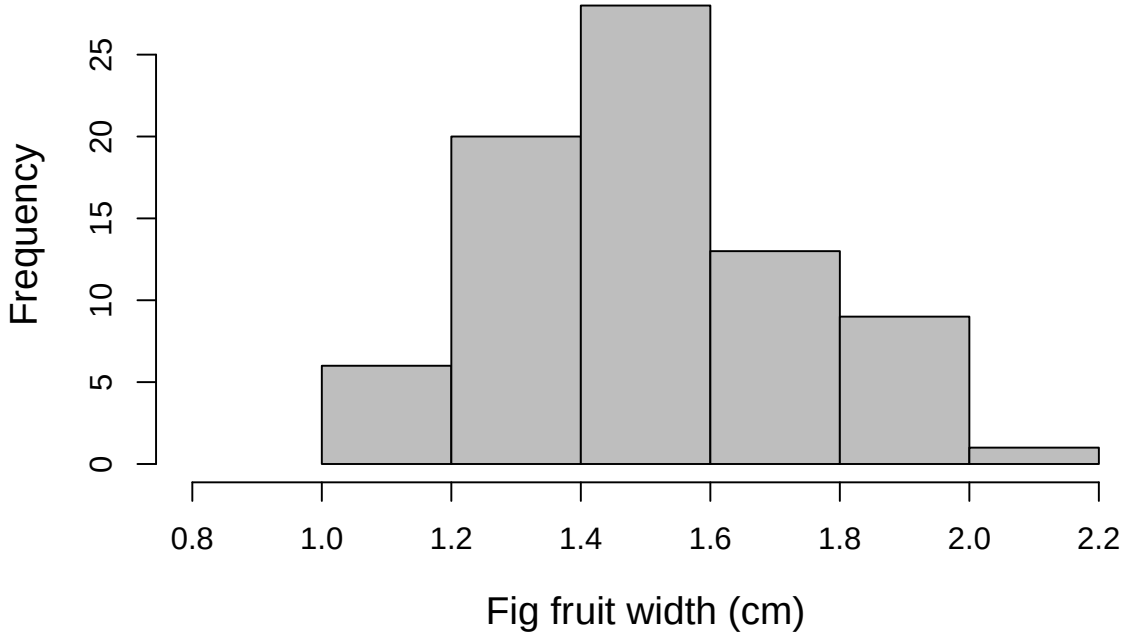


Figure 10.1.: Example histogram fig fruit width (cm) using data from 78 fig fruits collected in in 2010 from Baja, Mexico.

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 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 module, 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 Part IV (next week).

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 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).

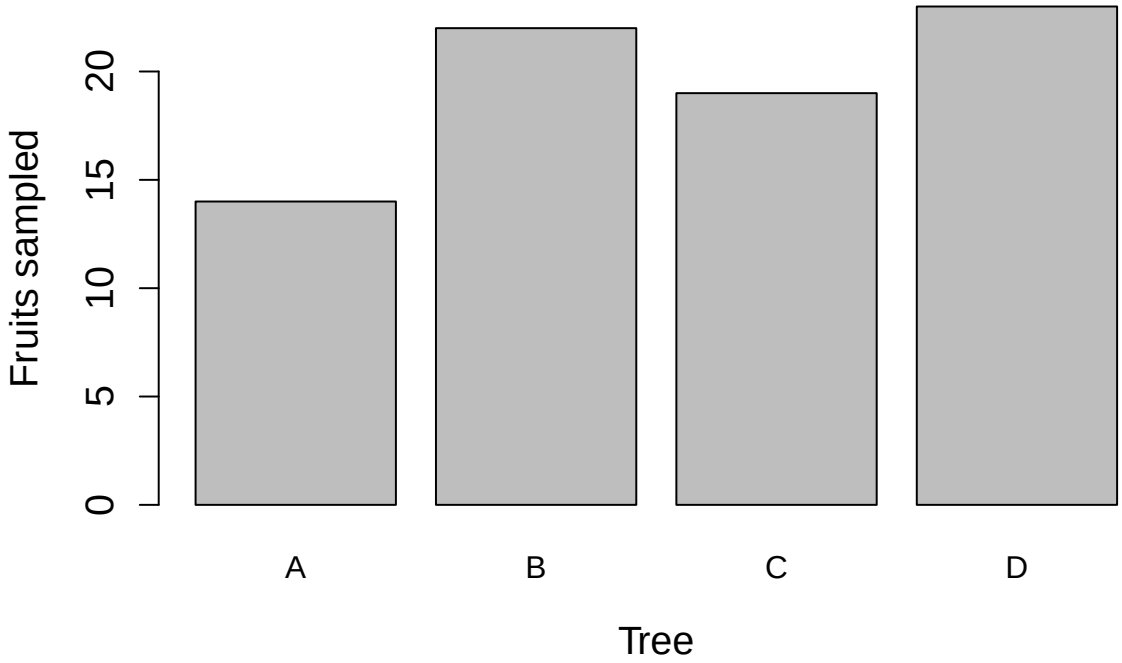


Figure 10.2.: Example bar plot showing how many fruits were collected from each of 4 trees (78 collected in total) in 2010 from Baja, Mexico.

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 that different categories of data are being shown (in this case, different trees). The height of 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 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 the total from the relative sizes of the bars. Pie charts, in contrast, only illustrate

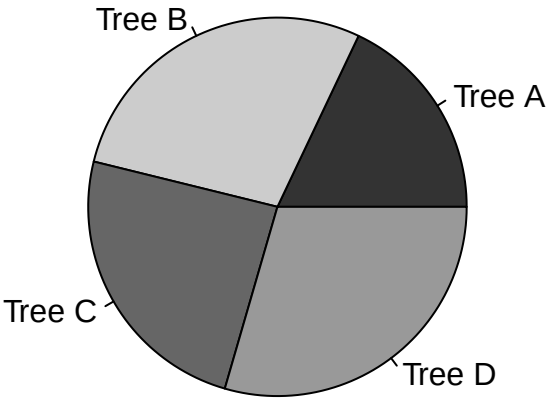


Figure 10.3.: Example pie plot showing the percentage of fruits that were collected from each of 4 trees (78 collected in total) in 2010 from Baja, Mexico.

relative percentages unless numbers are used to indicate absolute quantities. Unless only percentage 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 below). 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.

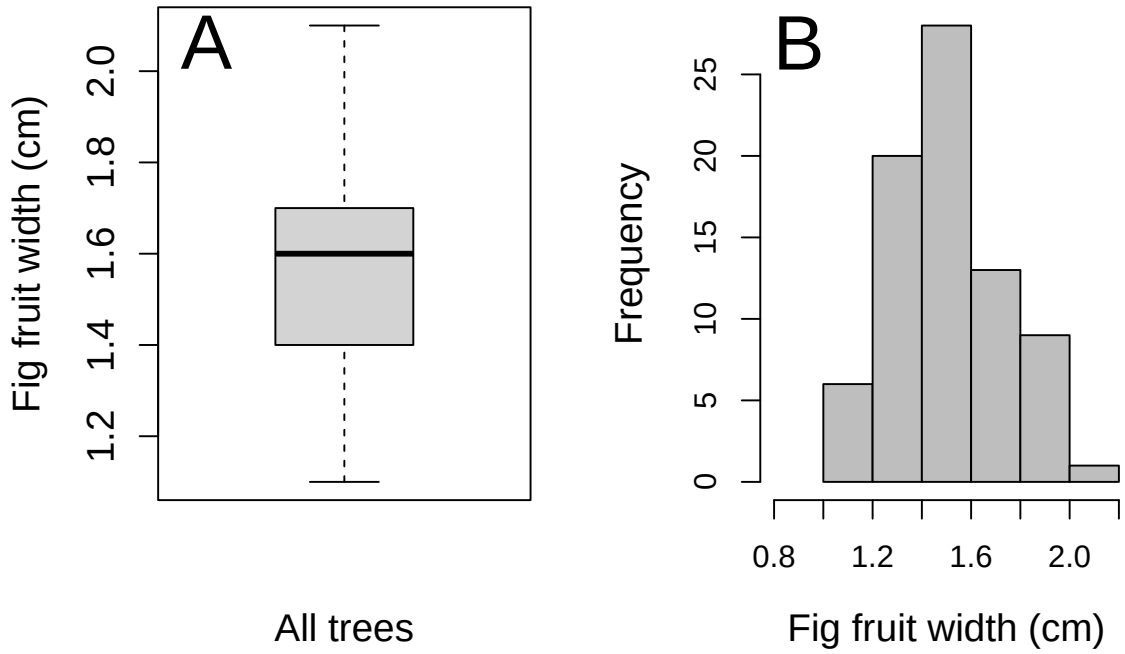


Figure 10.4.: Boxplot (A) of fig fruit widths (cm) for 78 fig fruits collected in 2010 in Baja, Mexico. Panel (B) presents the same data as a histogram.

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.

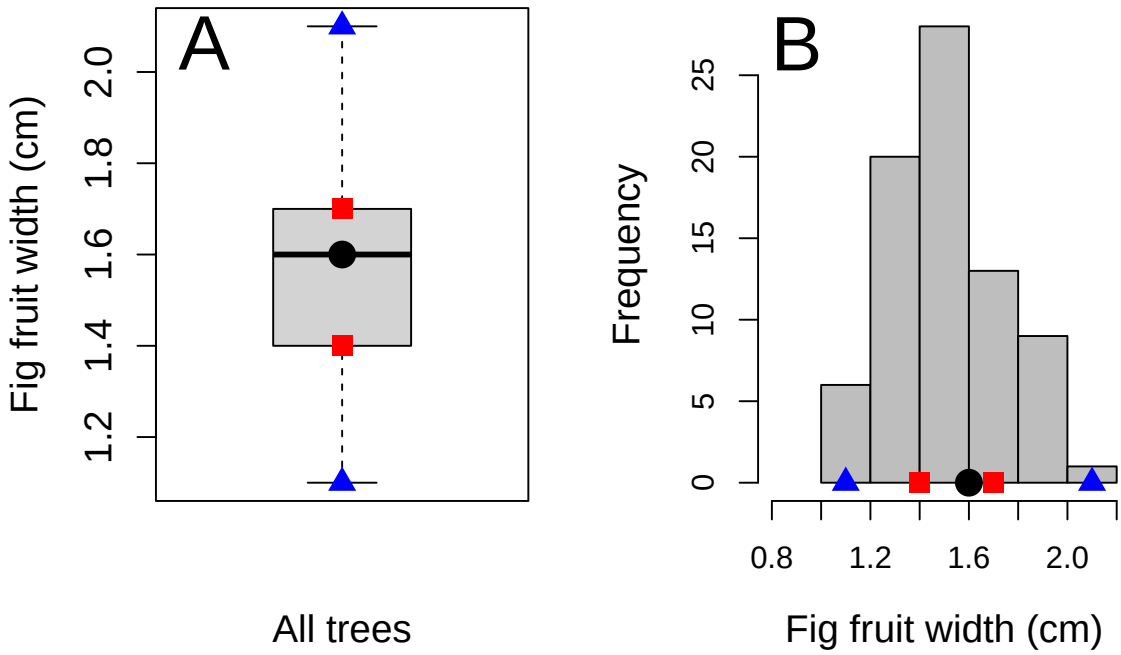


Figure 10.5.: Boxplot (A) of fig fruit widths (cm) for 78 fig fruits collected in 2010 in Baja, Mexico. Panel (B) presents the same data as a histogram. Points in the boxplot indicate the median (black circle), first and third quartiles (red squares), and the limits of the distribution (blue triangles).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

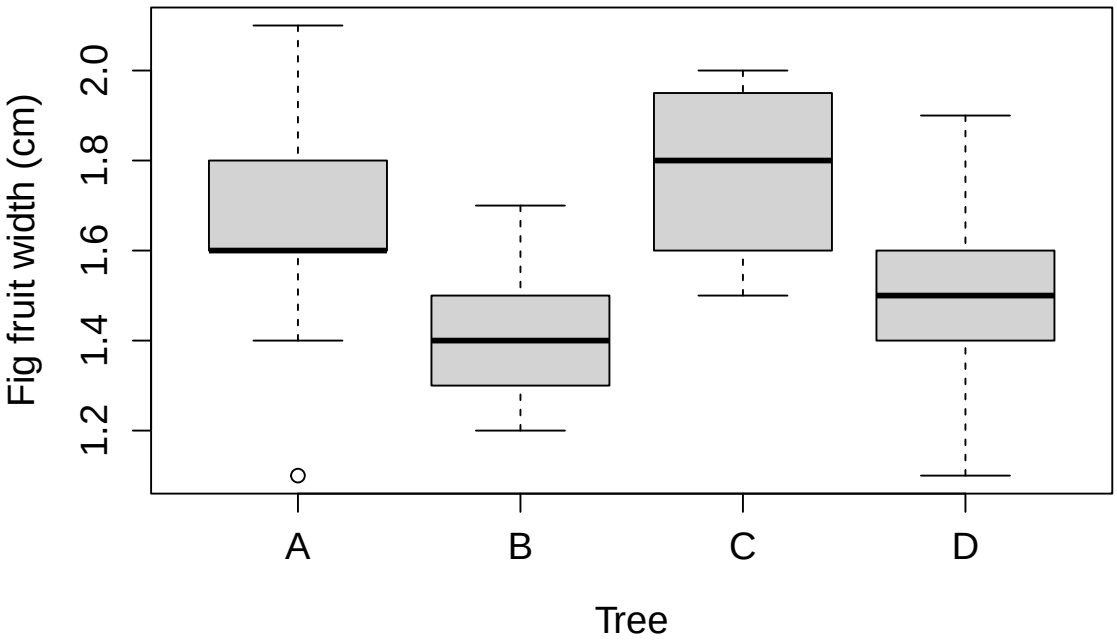


Figure 10.6.: Boxplot of fig fruit widths (cm) collected from 4 separate trees sampled in 2010 from Baja, Mexico.

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 centre of a distribution. Figure 10.7 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). Note, Miller and Miller (2004) will not be useful for this module.

[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.

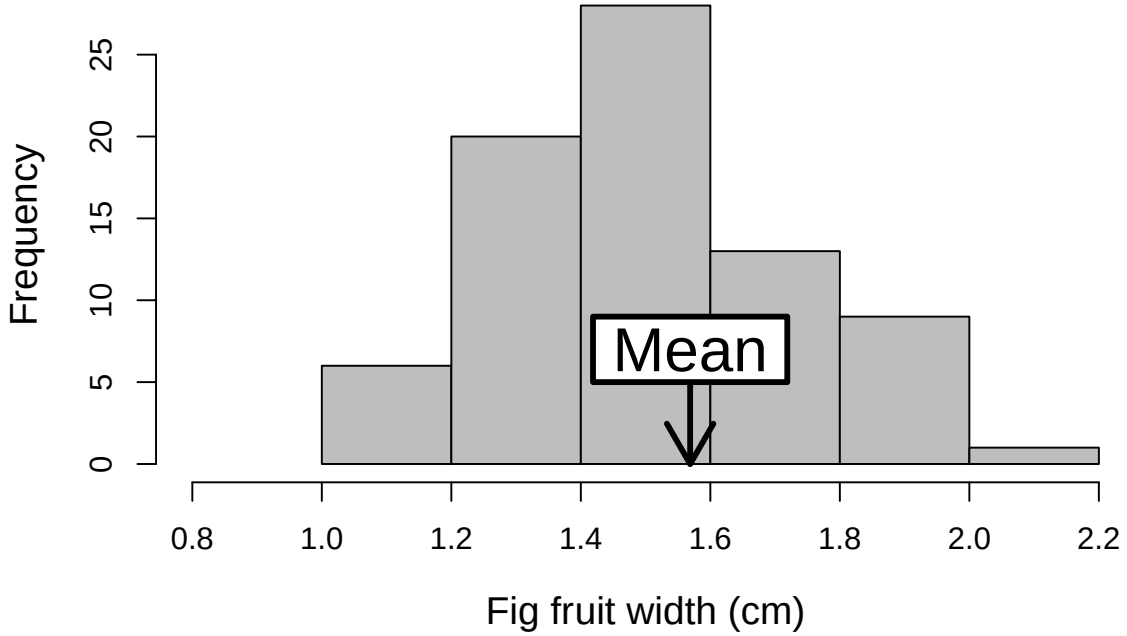


Figure 11.1.: Example histogram fig fruit width (cm) using data from 78 fig fruits collected in in 2010 from Baja, Mexico.

```
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 I 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 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 is the starting point and the top of it is 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 therefore 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).

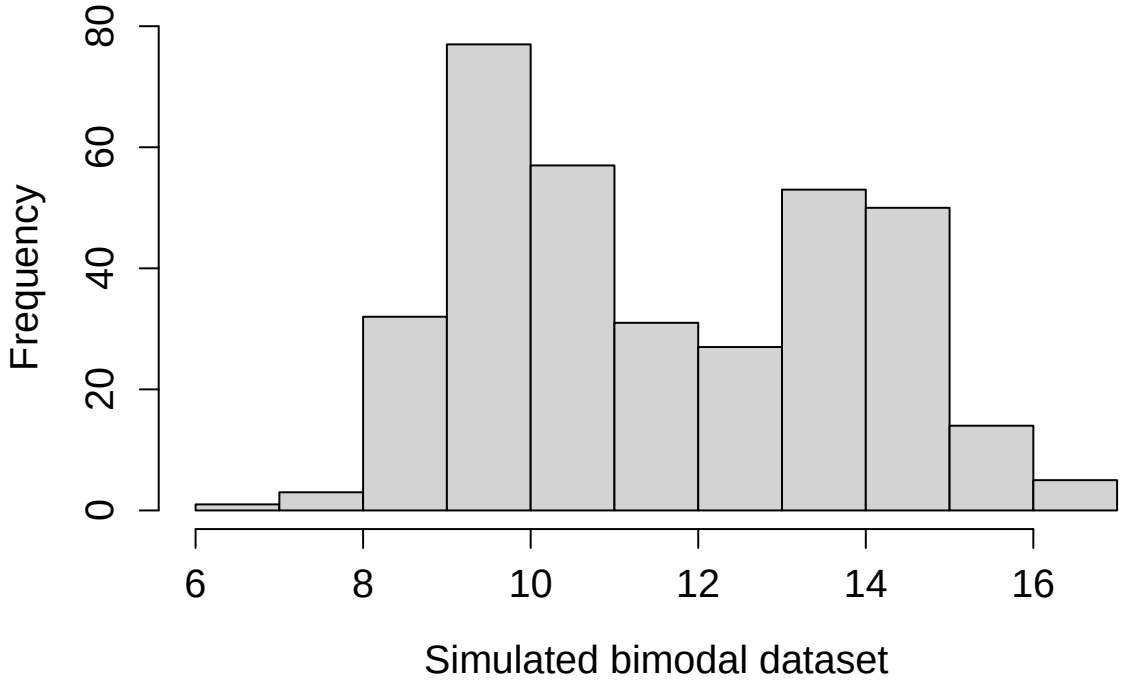


Figure 11.2.: Example histogram of a hypothetical dataset that has a bimodal distribution.

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 of 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 quartiles 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.

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 (three 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 to 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 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 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 $19.7 - 4.7 = 14$. 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 into 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. The variance of a dataset (is a measure of the expected 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 of 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 4.7, subtract the sample mean of 5.0 ($4.7 - 5.0 = -0.3$), then square the result ($(-0.3)^2 = 0.09$). We do the same for 4.9, $(4.9-5.0)^2 = 0.1$, and add it to the 0.09, then continue down the list of numbers finishing with 5.3. 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 (the 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 Chapter 41 and makes the squaring of differences make a bit more sense.

We sum up all of the squared deviations to get 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. This is really all that it is necessary

[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.

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 Evolution and Genetics when measuring heritability in quantitative genetics. 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 heritability ($H^2 = V_G/V_P$) can be used to predict how a population will respond to

[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 $Var(A + B)$.

natural selection. This is just one place where variance reveals itself to be a highly useful statistic in practice. 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]. Here is another interactive tool showing how the shape of a histogram changes when the standard deviation of a distribution is changed[6]. The practical in Chapter 13 explains how to calculate the standard deviation in Jamovi.

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

12.5. The coefficient of variation

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

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

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

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 8 measurements (assume that the nitrogen level of the lake has not changed in the meantime). We would expect to get similar values as our first 8 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 the variance and standard deviation is *of the sample means*. In other words, suppose that we kept going back out to the loch, collecting 8 new samples, and recording the sample mean each time? The standard deviation of those sample means would be the standard error. It 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 Chapter 19, 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	0.87	0.60	0.69	0.60	0.70	0.72	0.78	0.56	0.77	0.61
Sample_2	0.78	0.75	0.72	0.82	0.81	0.69	0.70	0.84	0.72	0.61
Sample_3	0.84	0.72	0.65	0.72	0.81	0.71	0.80	0.66	0.68	0.75
Sample_4	0.91	0.61	0.64	0.59	0.73	0.69	0.61	0.69	0.54	0.93
Sample_5	0.82	0.74	0.77	0.72	0.59	0.80	0.72	0.79	0.66	0.78
Sample_6	0.75	0.62	0.60	0.71	0.78	0.63	0.55	0.74	0.56	0.73
Sample_7	0.49	0.79	0.73	0.63	0.65	0.66	0.72	0.66	0.63	0.76
Sample_8	0.83	0.88	0.82	0.71	0.62	0.85	0.71	0.65	0.82	0.82
Sample_9	0.79	0.62	0.70	0.90	0.86	0.50	0.83	0.78	0.68	0.69
Sample_10	0.50	0.67	0.86	0.54	0.63	0.72	0.77	0.90	0.87	0.65
Sample_11	0.58	0.63	0.79	0.69	0.57	0.58	0.64	0.52	0.63	0.69
Sample_12	0.62	0.68	0.73	0.67	0.81	0.67	0.70	0.60	0.79	0.79
Sample_13	0.60	0.59	0.66	0.74	0.64	0.71	0.78	0.63	0.76	0.66
Sample_14	0.76	0.77	0.67	0.74	0.88	0.66	0.57	0.73	0.53	0.87
Sample_15	0.80	0.67	0.71	0.73	0.76	0.65	0.84	0.74	0.68	0.68
Sample_16	0.71	0.67	0.66	0.54	0.83	0.60	0.63	0.64	0.75	0.75
Sample_17	0.62	0.65	0.77	0.85	0.72	0.50	0.85	0.62	0.80	0.73
Sample_18	0.65	0.65	0.82	0.62	0.73	0.67	0.82	0.67	0.72	0.66
Sample_19	0.70	0.76	0.70	0.77	0.73	0.71	0.77	0.74	0.63	0.67
Sample_20	0.78	0.70	0.70	0.66	0.58	0.47	0.63	0.75	0.66	0.70

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

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
## [1,] 0.690 0.744 0.734 0.694 0.739 0.667 0.672 0.771 0.735 0.711
## [2,] 0.632 0.706 0.677 0.718 0.726 0.678 0.711 0.701 0.718 0.663
```

The standard deviation of the sample means reported above is 0.0333392. Now suppose that we only had Sample 1 (i.e., the top row of data). The standard deviation of Sample 1 is $s = 0.098545$. We can calculate the standard error from these sample values below,

$$s = \frac{0.098545}{\sqrt{10}} = 0.0311627.$$

The estimate of the standard error from calculating the standard deviation of the sample means is therefore 0.0333392, and the estimate from just using the standard error formula and data from only Sample 1 is 0.0311627. 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 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].

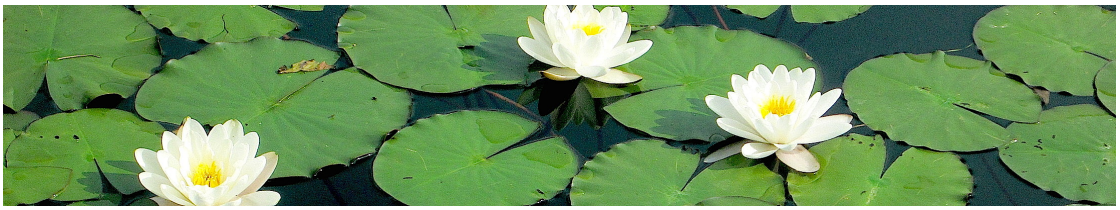


Figure 13.1.: Photo of white water lillies on the water.

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. 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 an independent sample. 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, 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
```

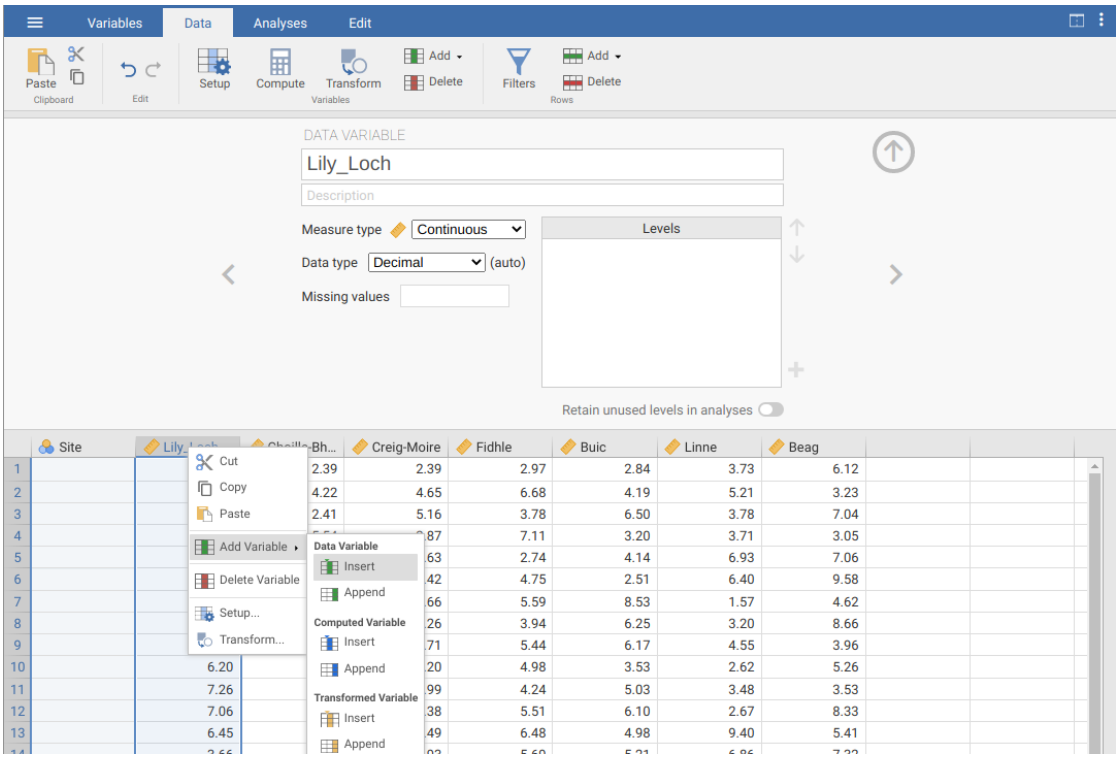


Figure 13.2.: Tidying the raw data of petiole diameters from lily pad measurements across 7 sites in Scotland. A new column of data is created by right clicking on an existing column and choosing 'Add Variable'.

```
## 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 'Analyses' tab of the Jamovi toolbar, then select the 'Exploration' button.

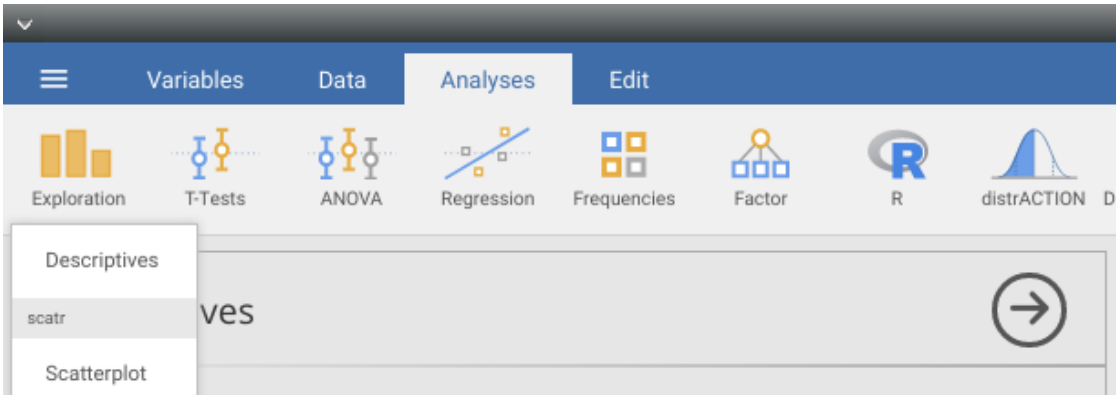


Figure 13.3.: Jamovi toolbar after having selected on the Analyses tab followed by the Exploration button.

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: _____

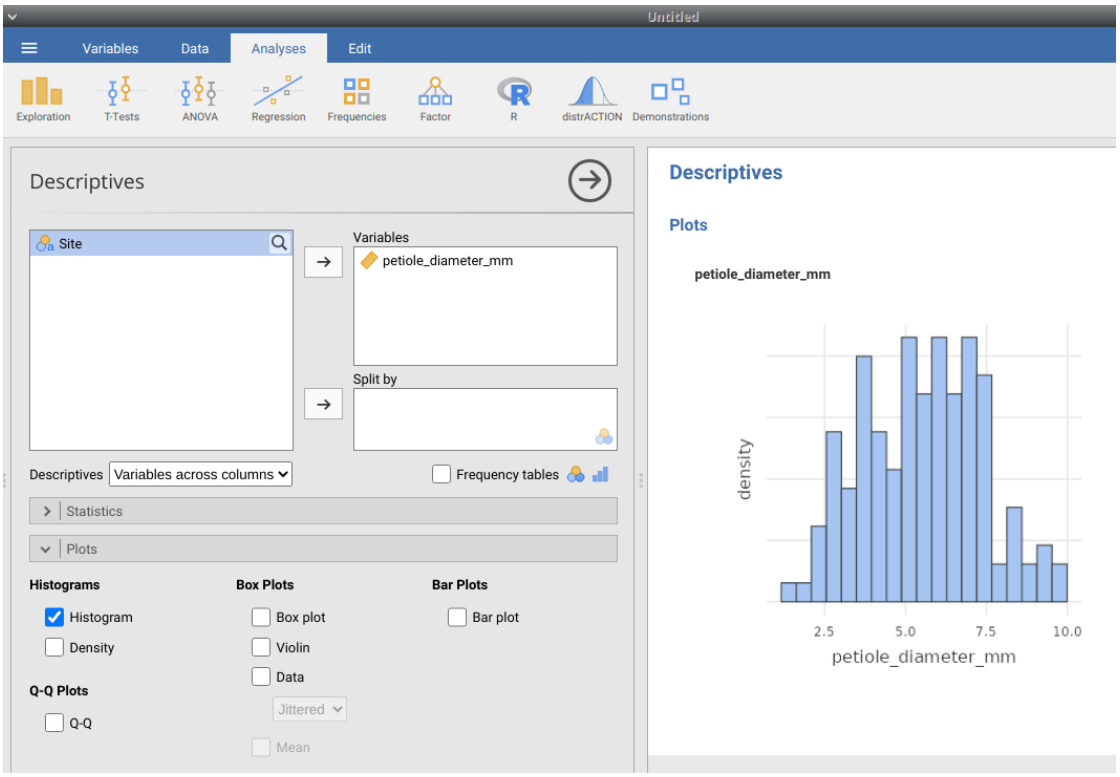


Figure 13.4.: Jamovi Descriptives toolbar with petiole diameter selected and a histogram produced in the plotting window.

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: _____
- Range: _____
- 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.

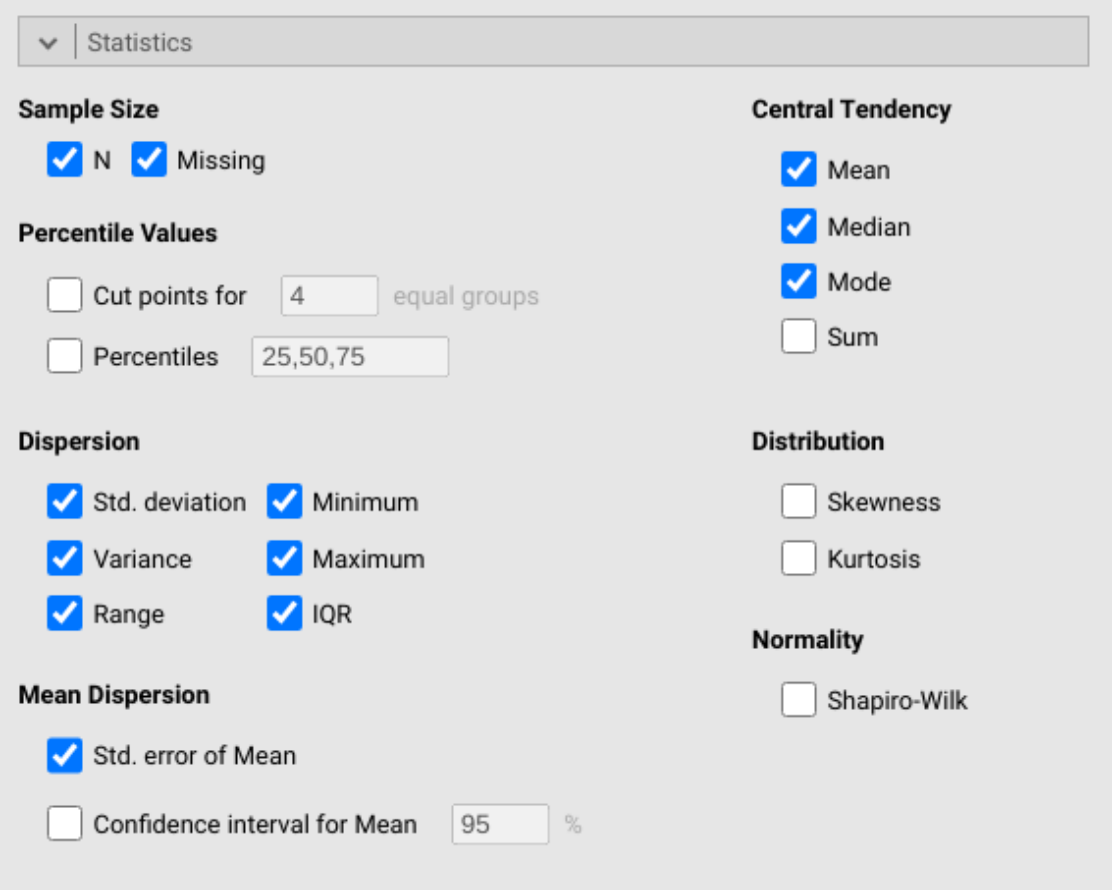


Figure 13.5.: Jamovi Descriptives toolbar showing the summary statistics available to report.

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 variation in petiole diameter, and what is its variation?

Site: _____

Variation: _____

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: _____
- Range: _____
- 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 petiol 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 14-15 **Recommended:** Navarro and Foxcroft (2022) Chapter 7 **Optional:** Rowntree (2018) Chapter 4
Lectures	4.1: Probability models (17 min.) 4.2: Probabilities to make predictions (12 min.) 4.3: Probability distributions (11 min.) 4.4: Predictions using sample statistics (16 min.) 4.5: Jamovi procedures
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 1A13 17 FEB 2023 (FRI) 15:05-17:55
Assessments	Week 4 Practice quiz on Canvas

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 knowledge, 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 best guess 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 practicals example 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 an 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 $P(X)$ to define the probability (P) of some event (X) happening. Note that the number of times an outcome occurs cannot be less than 0, so $P(X) \geq 0$ 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, $P(X) \leq 1$ 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 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 equals 0.25.

But why are we multiplying to get the joint probability of both coins landing on heads? Why not add, for example? Well, we could take it as a given that multiplication is the correct operation to use when calculating the probability that multiple events will occur. We could also 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 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, the 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).

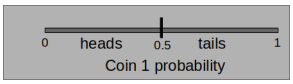


Figure 14.1.: Total probability space for flipping a single coin and observing its outcome (heads or tails). Given a fair coin, the probability of heads equals a proportion 0.5 of the total probability space, while the probability of tails equals the remaining 0.5 proportion.

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'. If 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

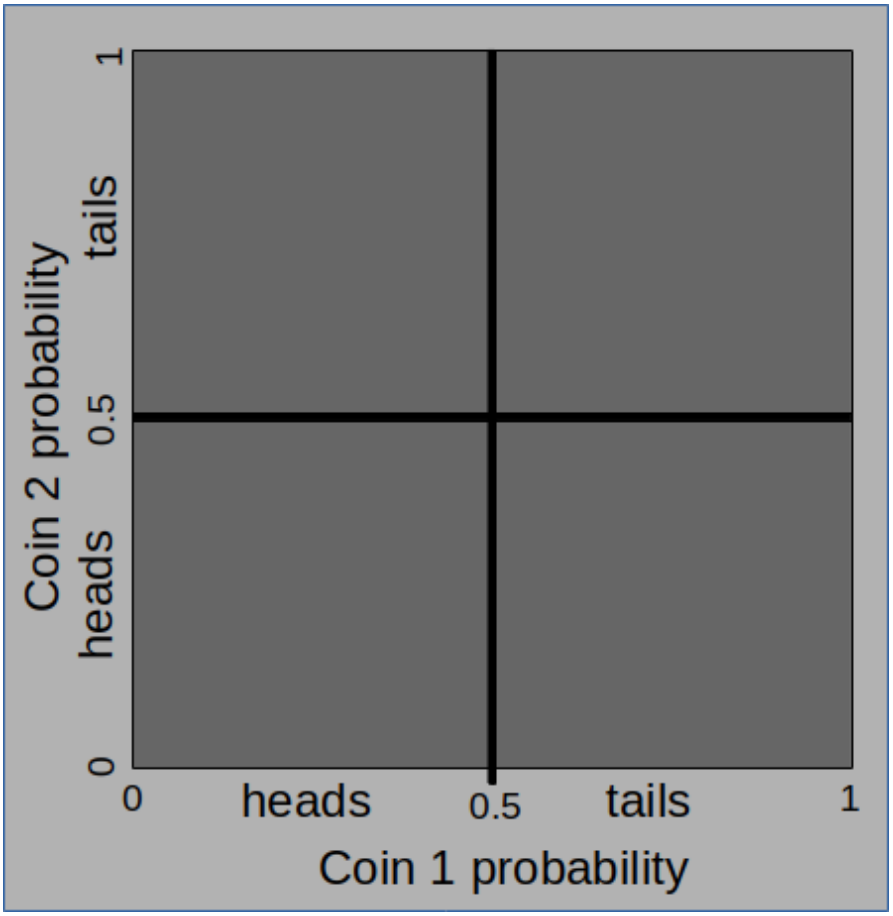


Figure 14.2.: Total probability space for flipping two coins and observing their different possible outcomes (heads-heads, heads-tails, tails-heads, and tails-tails). Given two fair coins, the probability of flipping each equals 0.25, which corresponds to the lower left square of the probability space.

both coins 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 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, 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 they do **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, 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 two ways for A_1 and A_3 to be arise in an individual, just like there were two 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, as 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 the whatever has been sampled does not get put back into the population. 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 clubs (right).

Figure 14.3 shows 10 playing cards, 5 hearts and 5 clubs. 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 randomly selecting two 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 $P(heart) \times P(heart) = 0.5 \times 0.5 = 0.25$. 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

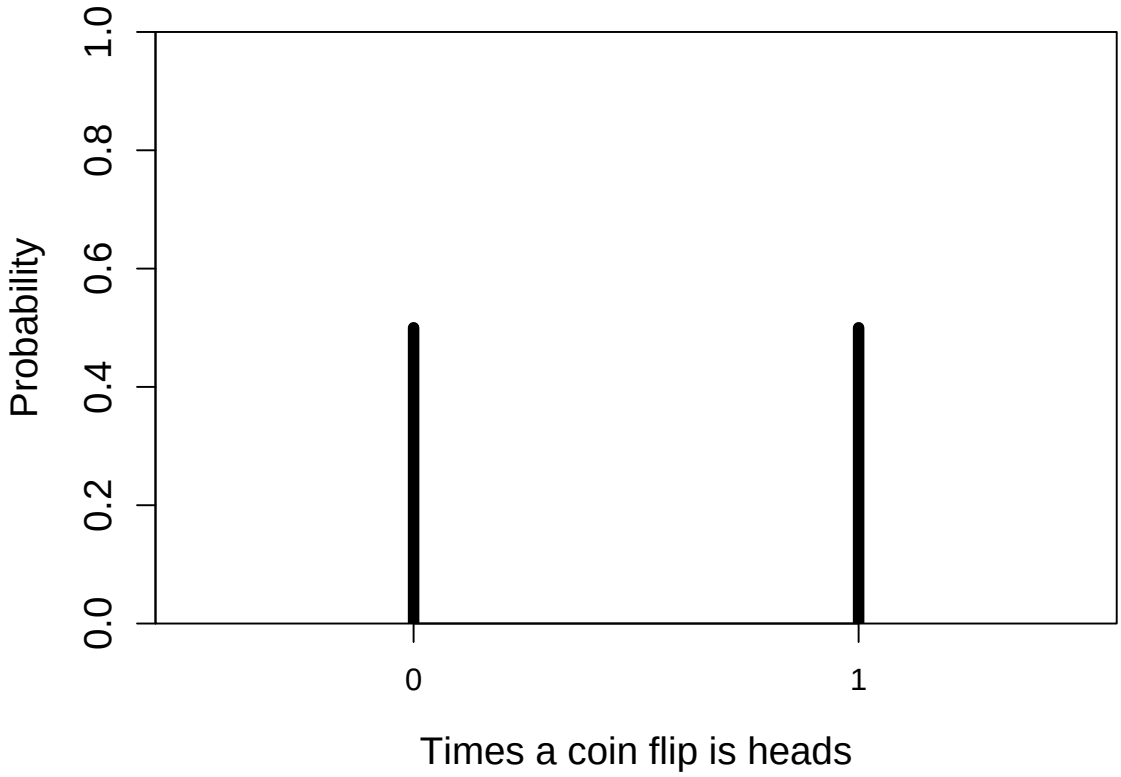


Figure 14.4.: Probability distribution for the number of times that a flipped coin lands on heads in 1 trial.

flip landing on heads as a success, we could consider the probability distribution of the number of success over 10 trials (Figure 14.5)

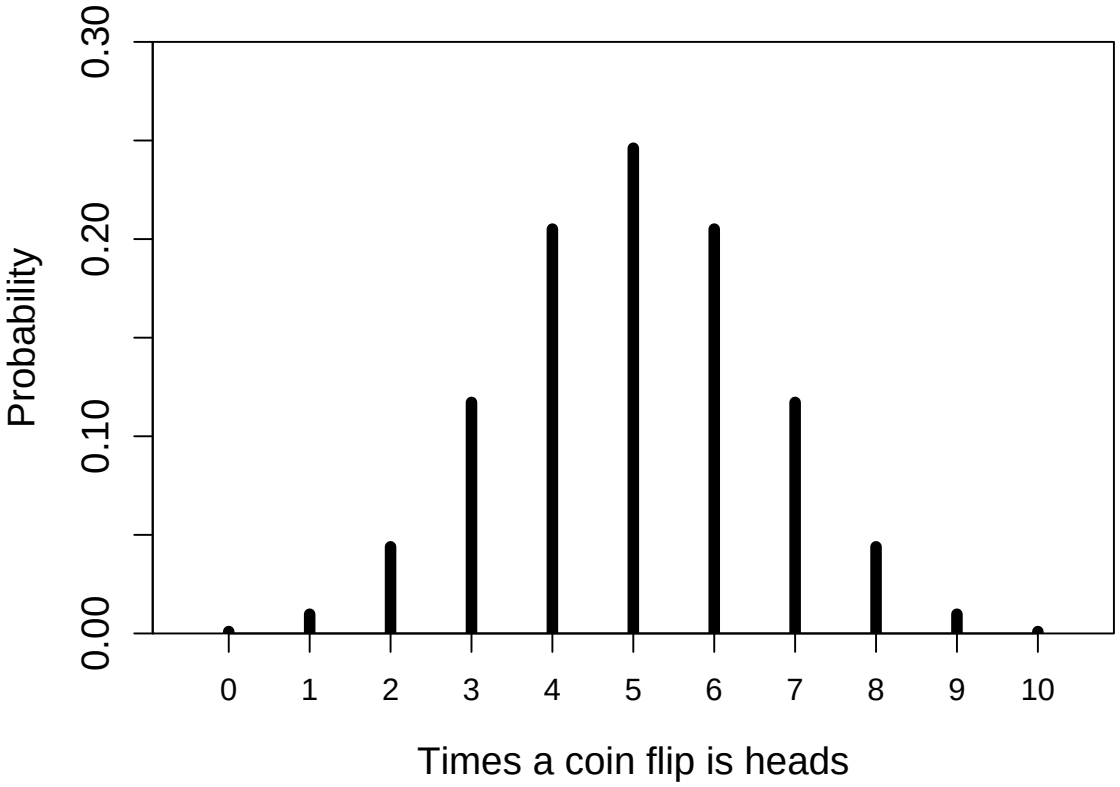


Figure 14.5.: Probability distribution for the number of times that a flipped coin lands on heads in 10 trials.

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 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 success being 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).

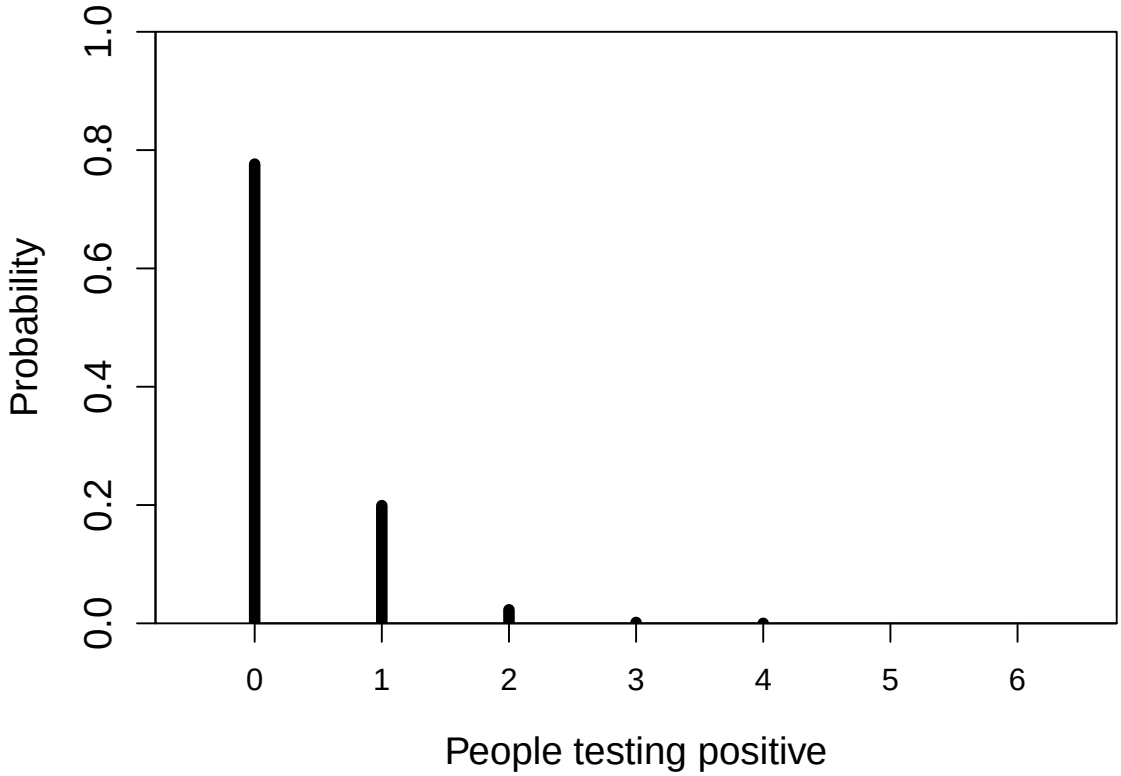


Figure 14.6.: Probability distribution for the number of people who have Covid-19 in a shop of 6 when the probability of testing positive is 0.025.

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 success and a diminishing probability of 1 or more success.

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

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 14.6).

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

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, $4! = 4 \times 3 \times 2 \times 1 = 24$. 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

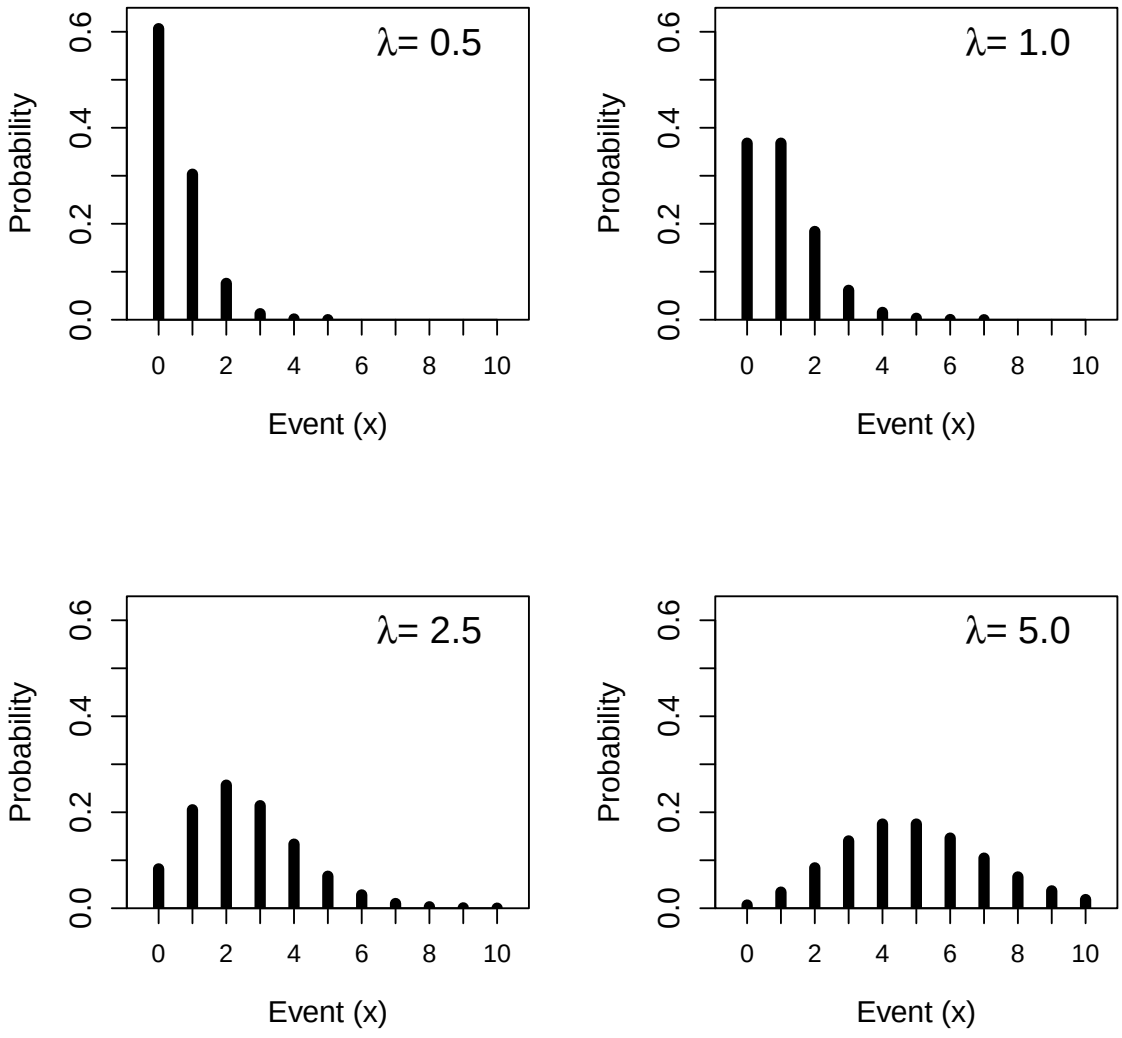


Figure 14.7.: Poisson probability distributions given different rate parameter values.

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 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 continues 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 14.7 shows the probability distribution for sampling some value x.

The height of the distribution in Figure 14.7 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 14.7, 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 the probability is 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

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.

[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.

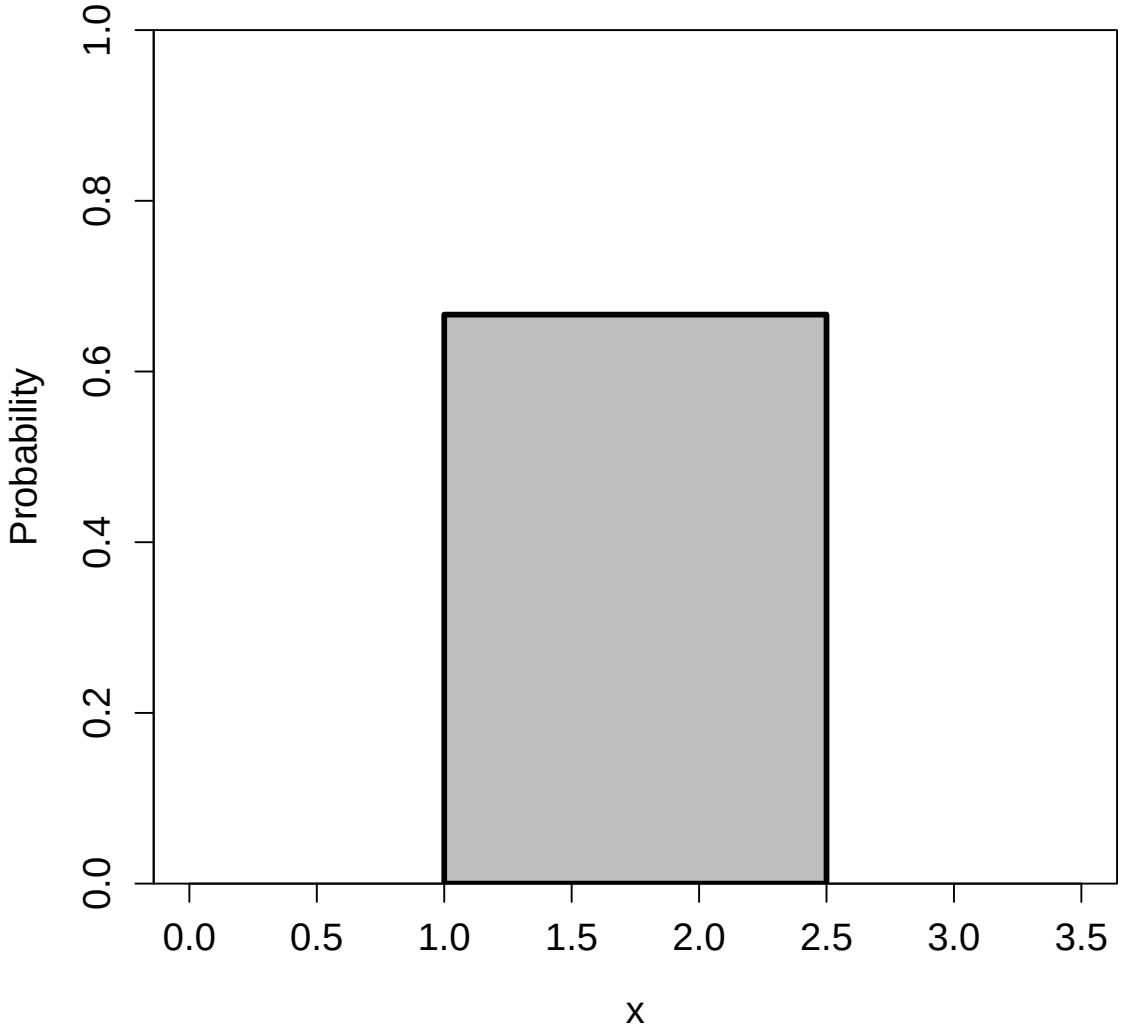


Figure 14.8.: A continuous uniform distribution in which a random variable X takes a value between 1 and 2.5.

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 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 sample estimates of Chapter 11 and Chapter 12.

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 that are especially important for understanding and implementing statistical techniques. As such, a lot of details have been left out. For example, the

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

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

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

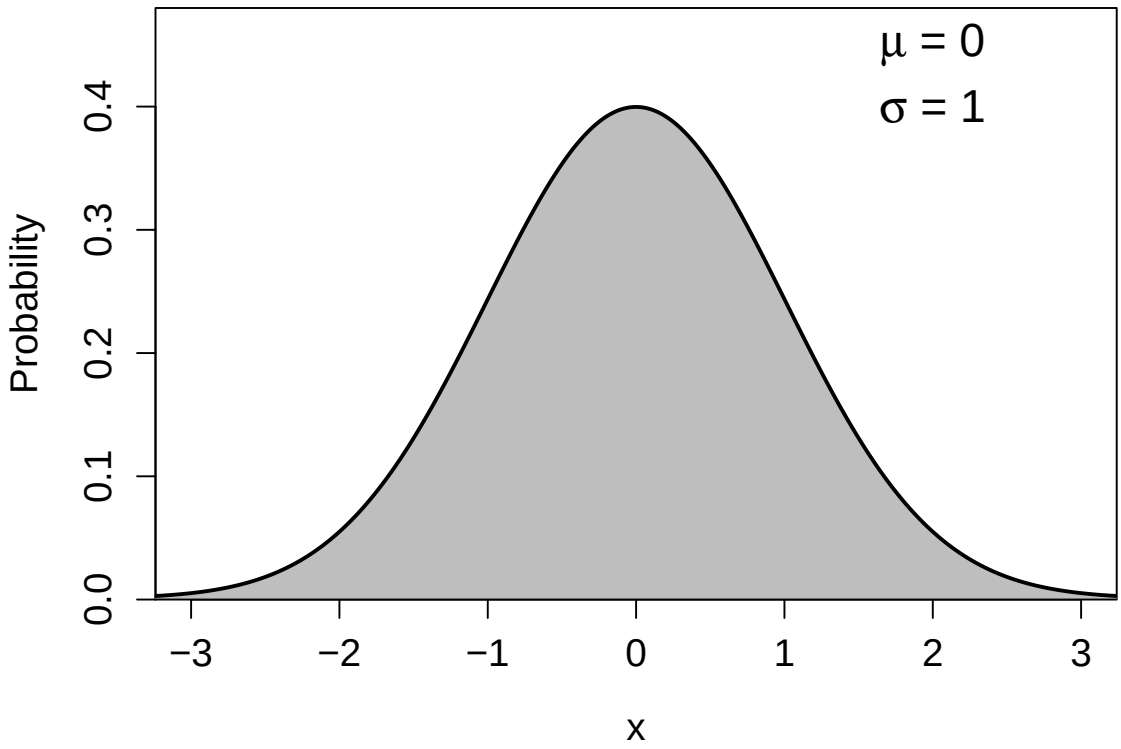


Figure 14.9.: A standard normal probability distribution, which is defined by a mean value of 0 and a standard deviation of 1.

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 useful.

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 four distributions introduced in Chapter 15 (binomial, poisson, uniform, or normal). Suppose that we sample the binomial distribution from Figure 14.6, the one describing the probability distribution of 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 get 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 a completely different set of 100 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 bit different. We can therefore

[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 the the sum of sample values, which will also have a distribution that approaches normality as $N \to \infty$.

ask about the *distribution* of these 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!

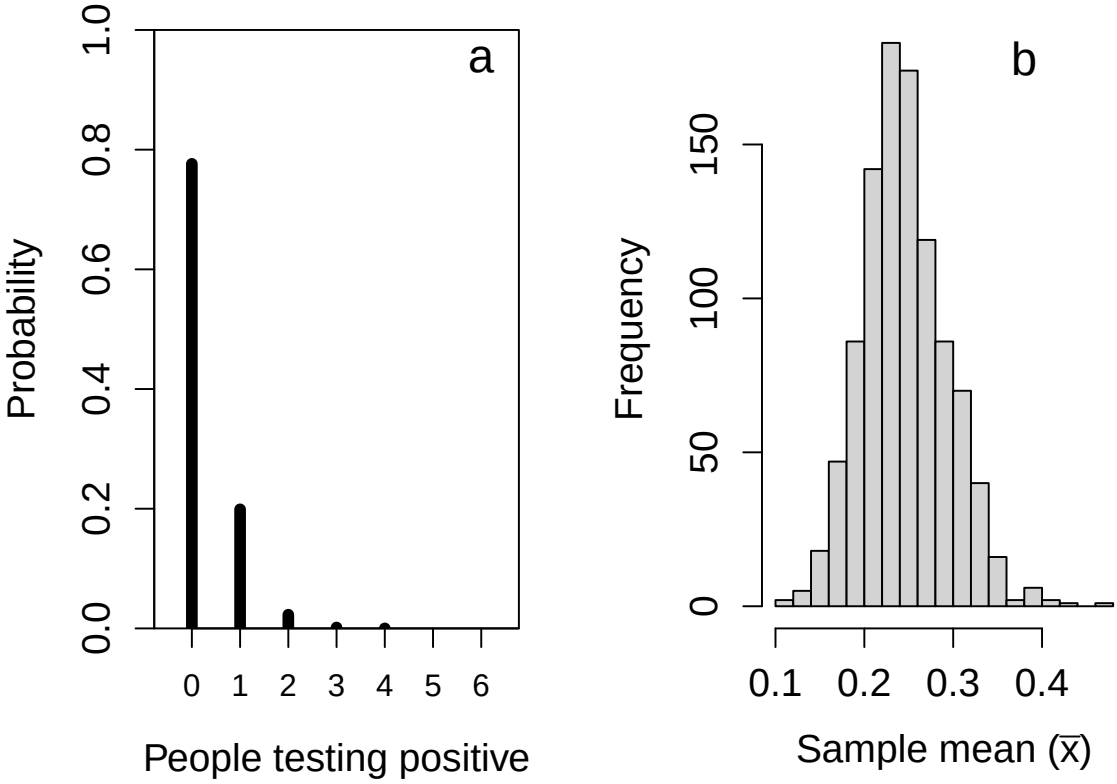


Figure 15.1.: A simulated demonstration of the central limit theorem. (a) Recreation of Figure 14.6 showing the probability distribution for the number of people who have Covid-19 in a shop of 6 when the probability of testing positive is 0.025. (b) The distribution of 1000 means sampled from panel (a), 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 151b) 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.

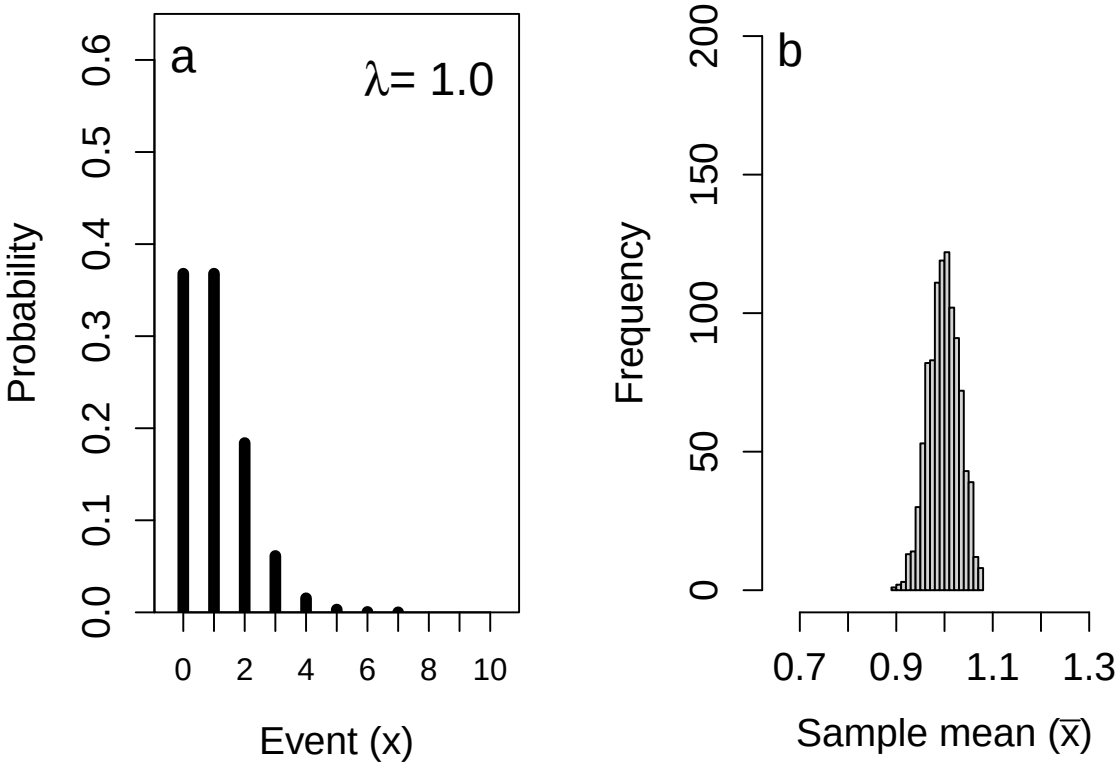


Figure 15.2.: A simulated demonstration of the central limit theorem. (a) Recreation of Figure 14.7 showing the probability distribution for the number of events occurring in a poisson distribution with a rate parameter of 1. (b) The distribution of 1000 means sampled from panel (a), where the sample 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 14.9.

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 estimates sample means, and to apply statistical hypothesis tests that would otherwise not be possible. We will look at these statistical tools in future chapters.

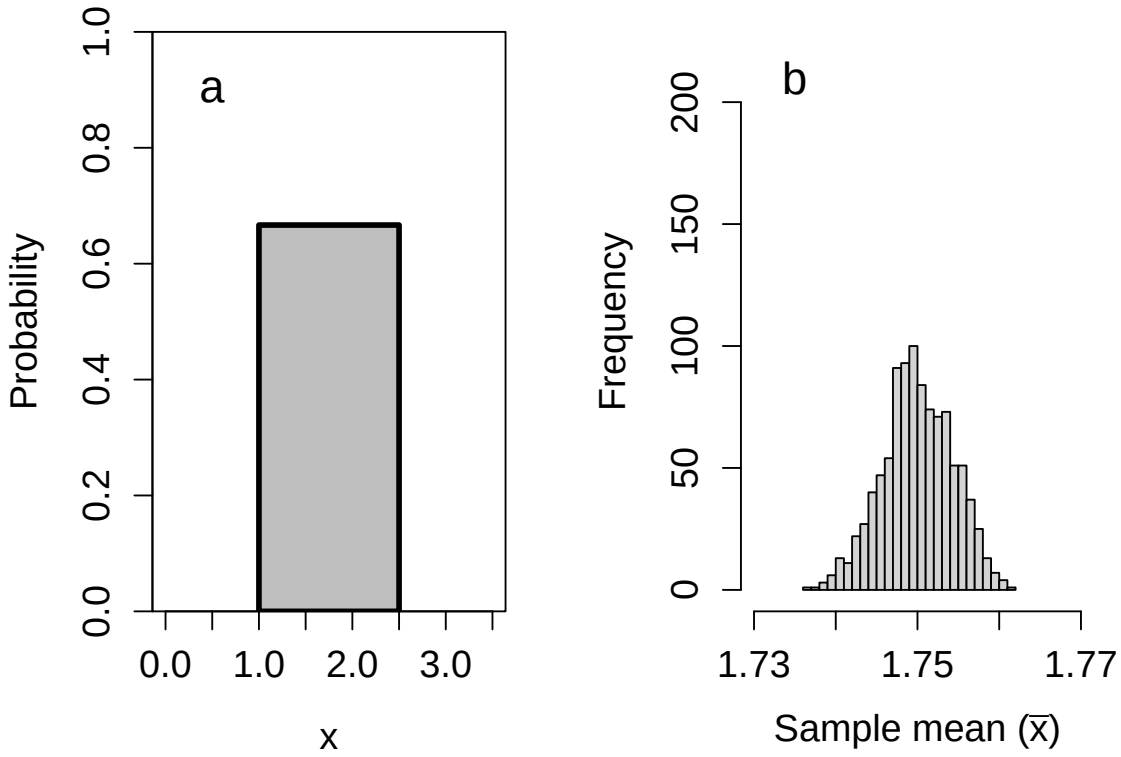


Figure 15.3.: A simulated demonstration of the central limit theorem. (a) Recreation of Figure 14.8 showing a continuous uniform distribution with a minimum of 1 and a maximum of 2.5. (b) The distribution of 1000 means sampled from panel (a), where the sample size is 10000.

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 (Figure 15.4), 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 $P(\mu - \sigma < x < \mu + \sigma) = 0.682$.

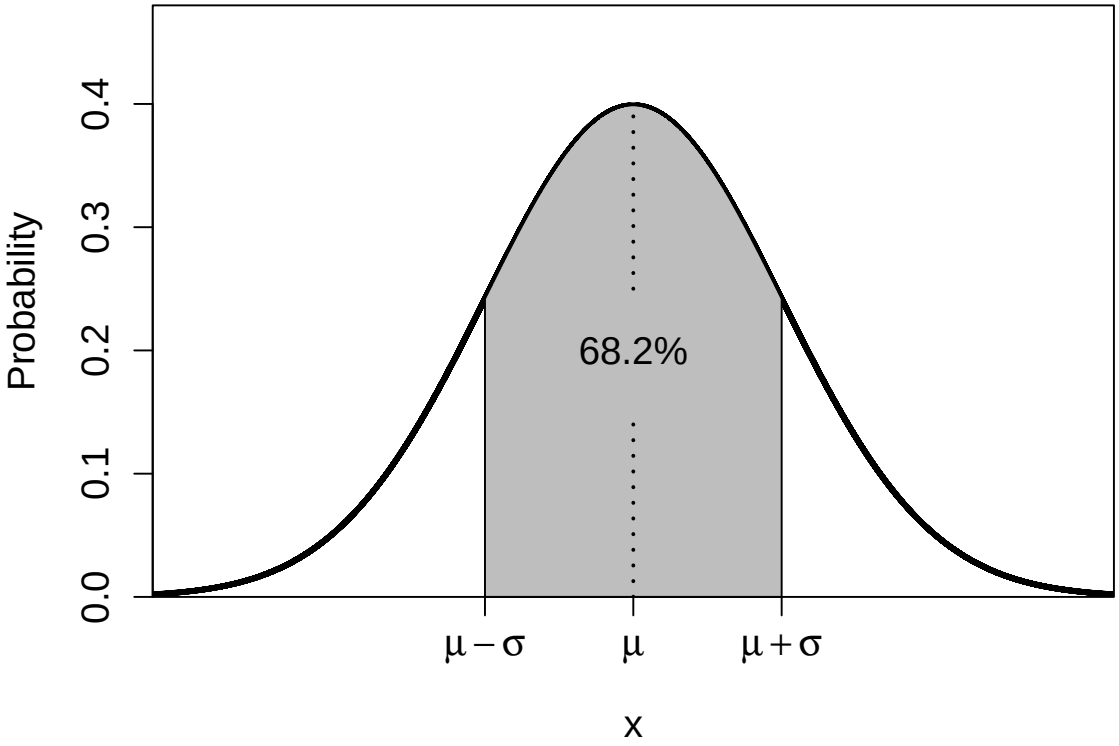


Figure 15.4.: A normal distribution in which the shaded region shows the area within one standard deviation of the mean (dotted line); that is, the shaded region starts on the left at the mean minus one standard deviation, then ends at the right at the mean plus one standard deviation. This shaded area 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 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 $P(x < \mu - \sigma \mid x > \mu + \sigma) = 0.318$, or 31.8% (note that the vertical bar, |, is just a fancy way of saying 'or', in this case).

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).

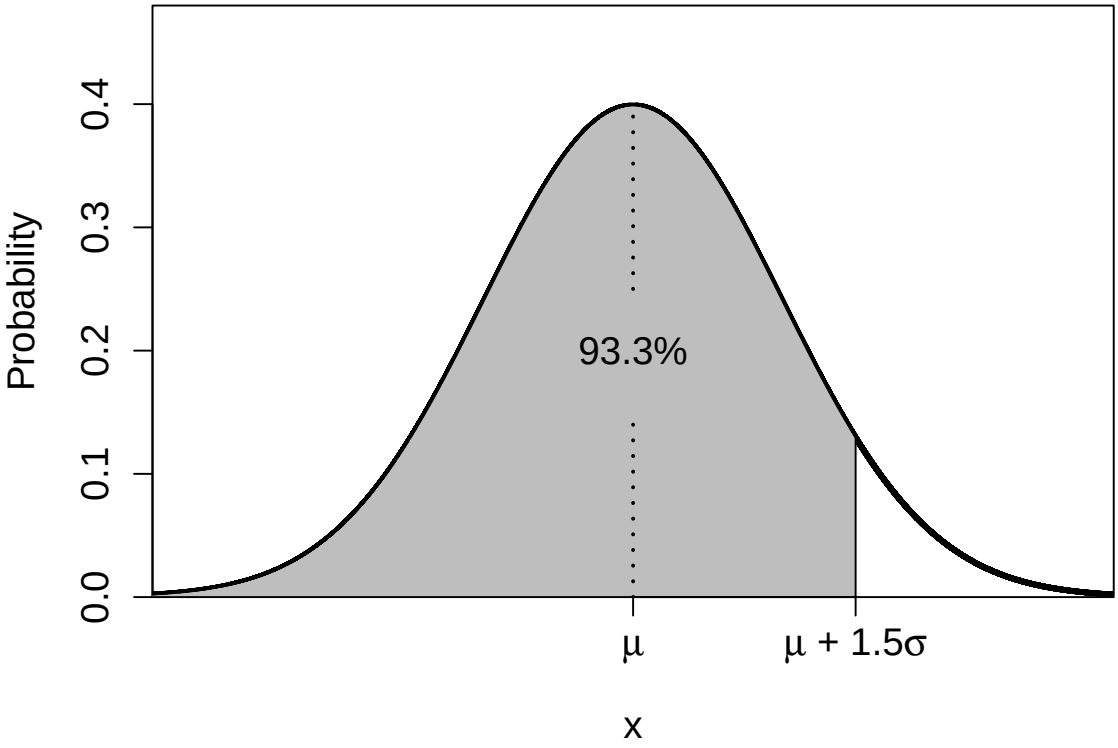


Figure 15.5.: A normal distribution in which the shaded region shows the area under 1.5 standard deviations of the mean (dotted line). This shaded area encompases about 93.3 per cent of the total area under the curve.

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 percentage of 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).

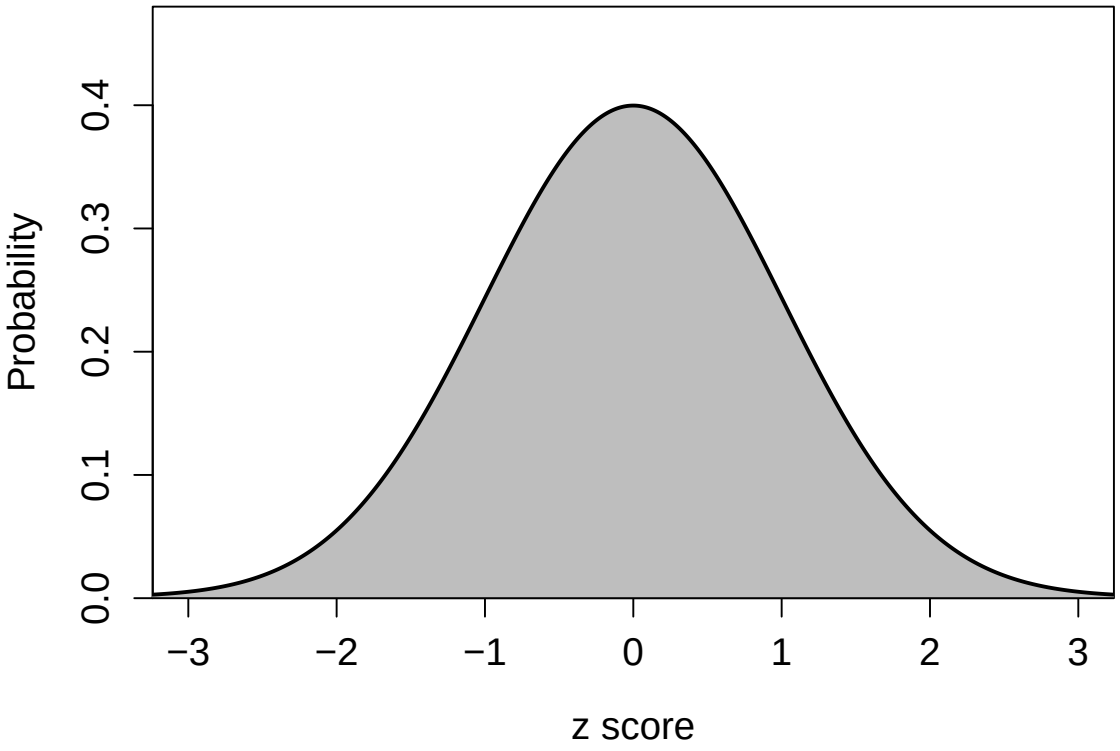


Figure 15.6.: A standard normal probability distribution with z-scores shown on the x-axis.

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.

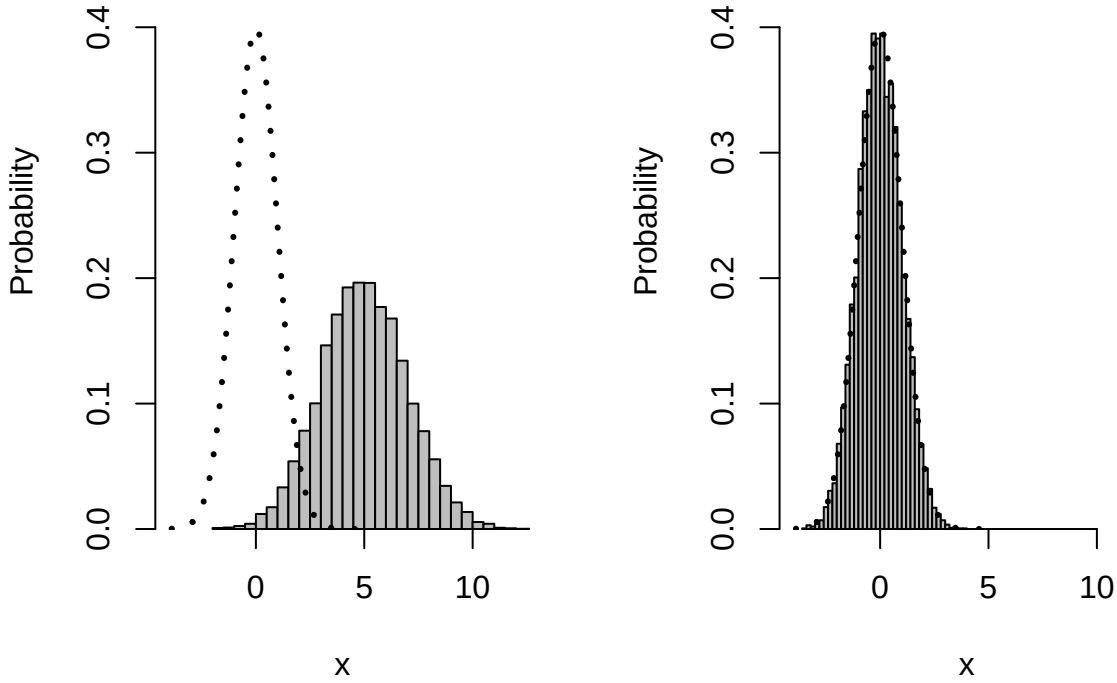


Figure 15.7.: A visual representation of what happens when we subtract the sample mean from a dataset, then divide by its standard deviation. (a) A histogram (grey bars) show 10000 normally distributed values with a mean of 5 and a standard deviation of 2; the curved dotted line shows the standard normal 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 $z = (9.1 - 5)/2 = 2.05$. We could then use a statistical program such as Jamovi, 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 normally distributed data. 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 15 in Jamovi. 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.

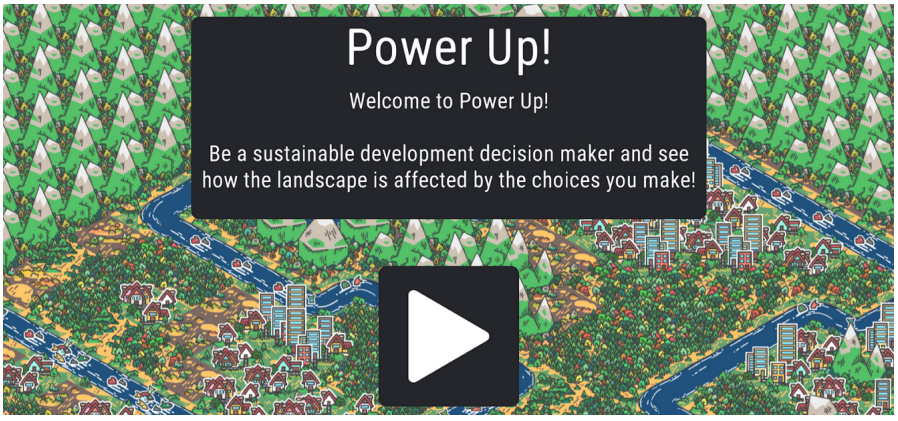


Figure 16.1.: Welcome screen of the mobile game Power Up!

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. To get started, first download the power_up dataset and open them 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, the proportion of iPhone users, we could calculate 18 / 74 = 0.2432432. But Jamovi already does that 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.

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.

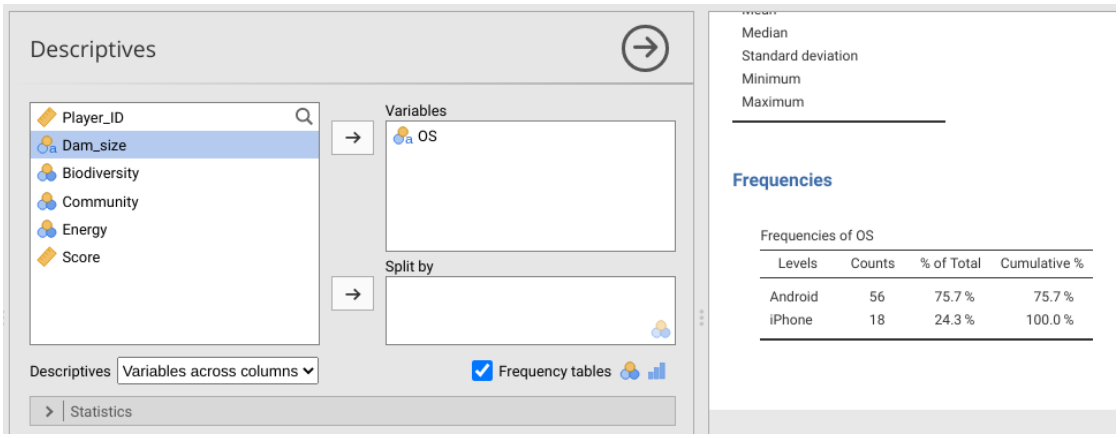


Figure 16.2.: Jamovi Descriptives toolbar showing the OS column from the Power Up! dataset selected. The 'Frequency tables' checkbox builds a table of counts and percentages.

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?

$Pr_{(3\ large)} =:$ _____

What is the probability that all 3 of these new players choose *different* dam sizes?

$P(small,\ medium,\ large) =:$ _____

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.

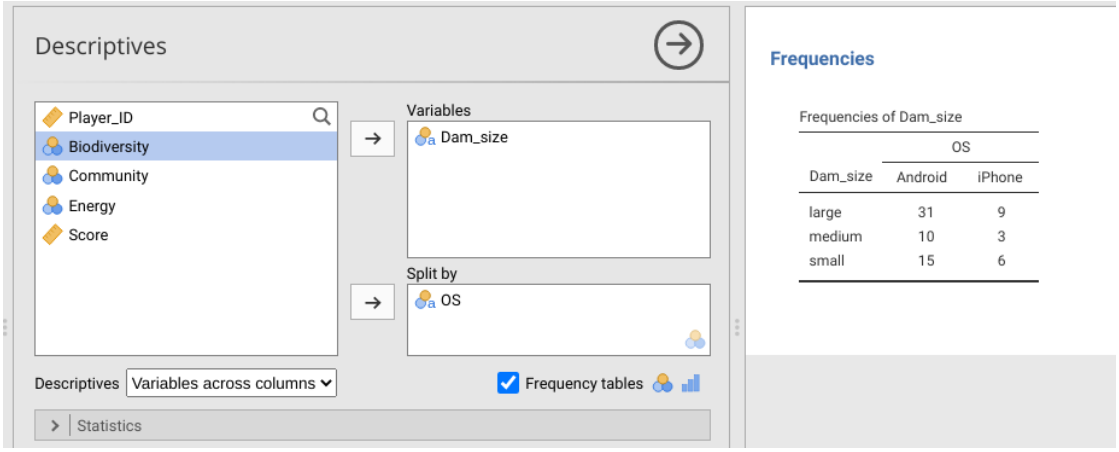


Figure 16.3.: Jamovi Descriptives toolbar showing the dam size column from the Power Up! dataset selected as a variable split by OS. The 'Frequency tables' checkbox builds a table of counts for small, medium, and large dam size broken down by Android versus iPhone OS.

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).

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

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, $94.9700000\bar{0}$)? 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 120 and 100?

Take another look at Figure 16.4 above, 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).

$P(Score > 120) =$: _____

$P(Score < 100) =$: _____

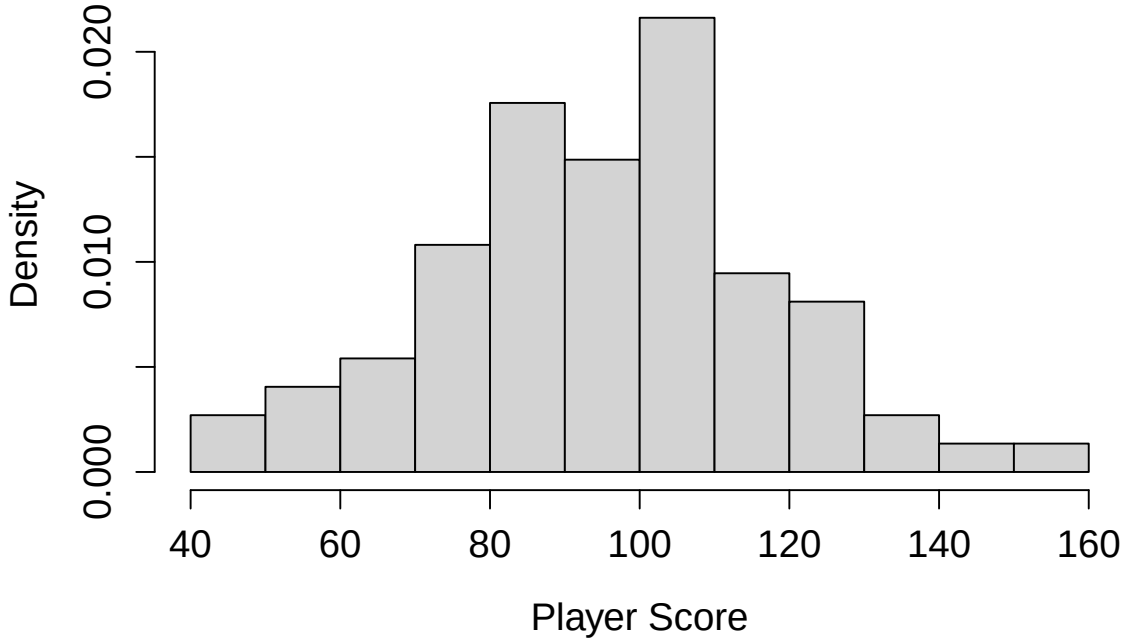


Figure 16.4.: Distribution of player scores in the game Power Up!

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

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 above that you wrote above (recall sample versus population from Chapter 4). We can overlay his distribution on the histogram above using a curved line (Figure 16.5).

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

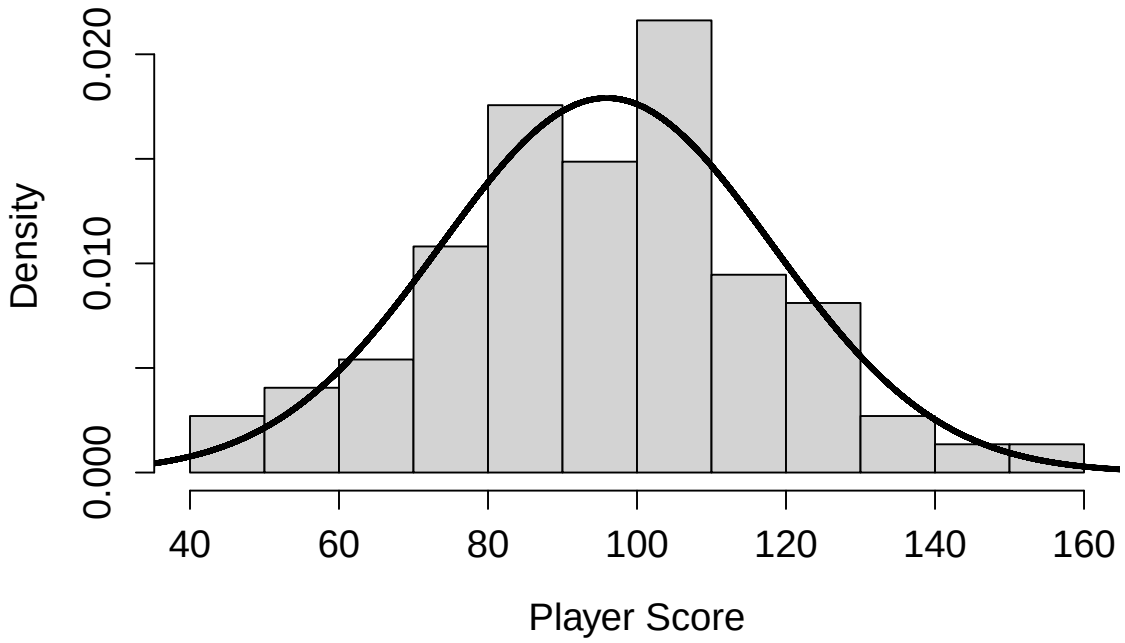


Figure 16.5.: Distribution of player scores in the game Power Up! shown in histogram bars. The overlaid curve shows the probability density function for a normal distribution that has the same mean and standard deviation as the sample described by the histogram.

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 16.5).

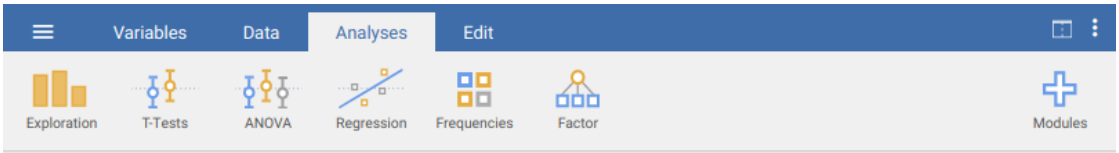


Figure 16.6.: Jamovi tool bar, which includes an option for downloading new Modules (right hand side)

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 16.6).

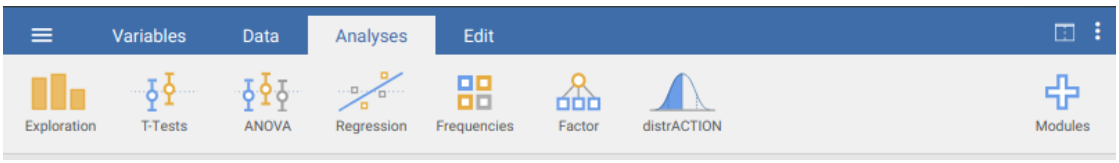


Figure 16.7.: Jamovi tool bar, which includes an added module called distrACTION.

If the module is not there, 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 $P = 0.237$. This is also represented in the plot of the normal distribution, with the same proportion in the lower part of the distribution shaded ($P = 0.237$, i.e., ca 23.7 per cent).

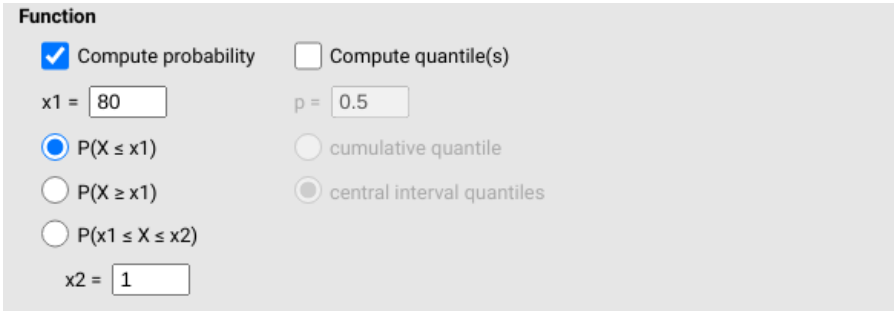


Figure 16.8.: Jamovi options for the distrACTION module for computing probability for a given normal distribution. The example shown here calculates the probability that a value sampled from the normal distribution of interest is less than 80.

To find the probability that a value is greater than 80, we could subtract our answer of 0.237 from 1, $1 - 0.237 = 0.763$ (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 $P = 0.763$.

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 setting 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 three 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?

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

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 code. 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).

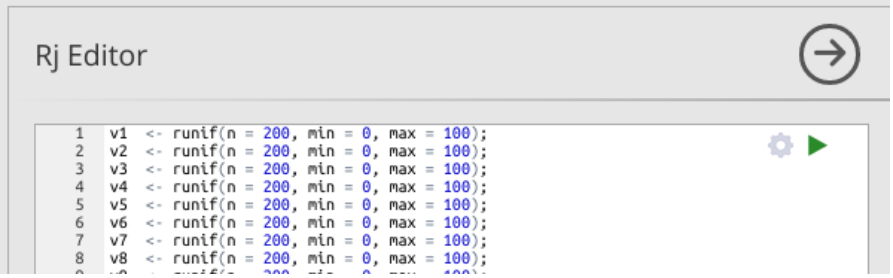


Figure 16.9.: Jamovi interface for the Rj Editor module. Code can be run by clicking on the green triangle in the upper left.

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`?

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.

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?

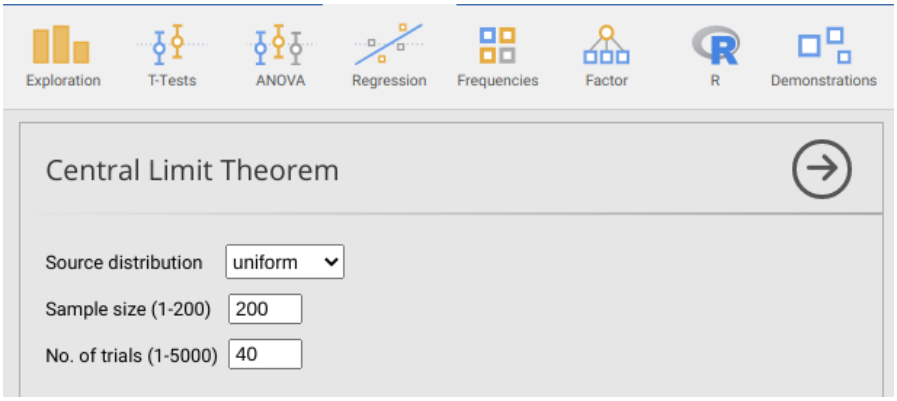


Figure 16.10.: Jamovi interface for the 'Demonstrations' module, which allows users to randomly generate data from a specific source distribution (normal, uniform, geometric, lognormal, and binary), sample size, and number of trials (i.e., variables)

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

General overview of what will be the focus of this week.

Week: 5 Dates: Suggested Readings: Textbook intro to probability Assessments: Practice quiz Practical:

17. Sample statistics and population parameters

An explanation of this

18. Standard Normal Distribution

What this means, and why it is important.

19. Confidence intervals

How these are calculated, and how to interpret them

20. The t-interval

What this is and how it relates to the normal distribution, and why it is important.

21. *Practical*. z- and t- intervals

21.1. Example constructing confidence intervals

21.2. Confidence interval for different levels (t- and z-)

21.3. Proportion confidence intervals

21.4. Another confidence interval example?

Part VI.

Hypothesis testing

Week 6 Overview

General overview of what will be the focus of this week.

Week: 6 Dates: Suggested Readings: Textbook Assessments: Practice quiz Practical:

22. What is hypothesis testing?

An explanation of this, and that we are starting to get into some of the more interesting bits of inferential statistics.

23. Making and using hypotheses and types of tests

What this means, and why it is important.

24. An example of hypothesis testing

Errors

25. Hypothesis testing and confidence intervals

Relationship between these two.

26. Student t-distribution and one sample t-test

What this is and how to do it in Jamovi.

27. Another example of a one sample t-test

From the lectures

28. Independent t-test

What this is and how to use it in Jamovi.

29. Paired sample t-test

Another explanation, example, and how to do it in Jamovi.

30. Violations of assumptions

What to do in this case

31. Non-parametric tests, and what these are.

Explanation of how to do them in Jamovi.

32. *Practical*. Hypothesis testing and t-tests

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 chapters will correspond to weeks.

Part VIII.

Analysis of Variance (ANOVA)

Week 8 Overview

General overview of what will be the focus of this week.

Week: 8 Dates: Suggested Readings: Textbook Assessments: Practice quiz Practical:

33. What is ANOVA?

General explanation

34. One-way ANOVA

Explain what this is.

35. Two-way ANOVA

More explanation

36. Kruskall-Wallis H test

Non-parametric explanation

37. *Practical*. ANOVA and associated tests

37.1. ANOVA Exercise 1

37.2. ANOVA Exercise 2

37.3. ANOVA Exercise 3

37.4. ANOVA Exercise 4

Part IX.

Counts and Correlation

Week 9 Overview

General overview of what will be the focus of this week.

Week: 9 Dates: Suggested Readings: Textbook Assessments: Practice quiz Practical:

38. Frequency and count data

General explanation

39. Chi-squared goodness of fit

Explain what this is.

40. Chi-squared test of association

More explanation

41. Correlation key concepts

42. Correlation mathematics

43. Correlation hypothesis testing

44. *Practical*. Analysis of count data, correlation, and regression

44.1. Chi-Square Exercise 1

44.2. Chi-Square association Exercise 2

44.3. Correlation Exercise 3

44.4. Correlation Exercise 4

Part X.

Linear Regression

Week 10 Overview

Week: 10 Dates: Suggested Readings: Textbook Assessments: Practice quiz Practical:

45. Regression key concepts

46. Regression validity

47. Introduction to multiple regression

General explanation

48. Model selection (maybe remove this?)

Seriously consider moving the regression into this week. and ease the amount of material in previous weeks.

49. *Practical*. Using regression

49.1. Regression Exercise 1

49.2. Regression Exercise 2

49.3. Regression Exercise 3

49.4. 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?

50. Introduction to randomisation

General explanation

51. Assumptions of randomisation

How these differ

52. Bootstrapping

What this is and why we use it.

53. Monte Carlo

54. *Practical*. Using R

54.1. R Exercise 1

54.2. R Exercise 2

54.3. 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?

55. Reporting statistics

General explanation

56. More introduction to R

How these differ

57. More getting started with R

Just more to do.

58. *Practical*. Using R

58.1. R Exercise 1

58.2. R Exercise 2

58.3. R Exercise 3

Part XIII.

Review of parts (VII-XII)

Module summary

This chapter will be specifically to prepare for exam.

A. Statistical units

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

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.

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.

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.

Courant, R., Robbins, H., and Stewart, I. (1996). *What is Mathematics?* Oxford University Press, Oxford, 2 edition.

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.

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.

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.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28(706):49–50.

Hyndman, R. J. and Fan, Y. (1996). Sample Quantiles in Statistical Packages. *American Statistician*, 50(4):361–365.

Kelleher, C. and Wagener, T. (2011). Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling and Software*, 26(6):822–827.

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.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. University of Chicago Press, Chicago.

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.

Miller, I. and Miller, M. (2004). *John E. Freund's mathematical statistics*. Pearson Prentice Hall, Upper Saddle River, New Jersey, 7 edition.

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.

Rabinovich, S. G. (2013). *Evaluating Measurement Accuracy*.

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.

Rowntree, D. (2018). *Statistics Without Tears*. Penguin.

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.

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.

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.

Weiblen, G. D. (2002). How to be a fig wasp. *Annual Review of Entomology*, 47:299–330.

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.