PHCM9795: Foundations of Biostatistics

Timothy Dobbins

24 July, 2025

Table of contents

Course introduction

Welcome to PHCM9795 Foundations of Biostatistics.

This introductory course in biostatistics aims to provide students with core biostatistical skills to analyse and present quantitative data from different study types. These are essential skills required in your degree and throughout your career.

We hope you enjoy the course and will value your feedback and comment throughout the course.

Course information

Biostatistics is a foundational discipline needed for the analysis and interpretation of quantitative information and its application to population health policy and practice.

This course is central to becoming a population health practitioner as the concepts and techniques developed in the course are fundamental to your studies and practice in population health. In this course you will develop an understanding of, and skills in, the core concepts of biostatistics that are necessary for analysis and interpretation of population health data and health literature.

In designing this course, we provide a learning sequence that will allow you to obtain the required graduate capabilities identified for your program. This course is taught with an emphasis on formulating a hypothesis and quantifying the evidence in relation to a specific research question. You will have the opportunity to analyse data from different study types commonly seen in population health research.

The course will allow those of you who have covered some of this material in your undergraduate and other professional education to consolidate your knowledge and skills. Students exposed to biostatistics for the first time may find the course challenging at times. Based on student feedback, the key to success in this course is to devote time to it every week. We recommend that you spend an average of 10-15 hours per week on the course, including the time spent reading the course notes and readings, listening to lectures, and working through learning activities and completing your assessments. Please use the resources provided to assist you, including online support.

Units of credit

This course is a core course of the Master of Public Health, Master of Global Health and Master of Infectious Diseases Intelligence programs and associated dual degrees, comprising 6 units of credit towards the total required for completion of the study program. A value of 6 UOC requires a minimum of 150 hours work for the average student across the term.

Course aim

This course aims to provide students with the core biostatistical skills to apply appropriate statistical techniques to analyse and present population health data.

Learning outcomes

On successful completion of this course, you will be able to:

1. Summarise and visualise data using statistical software.
2. Demonstrate an understanding of statistical inference by interpreting p-values and confidence intervals.
3. Apply appropriate statistical tests for different types of variables given a research question, and interpret computer output of these tests appropriately.
4. Determine the appropriate sample size when planning a research study.
5. Present and interpret statistical findings appropriate for a population health audience.

Change log

24 July

- Added suggested reference for the course notes
- Added Module 10 notes

11 June 2025

- Section 2.15 jamovi: Added instructions for creating a stacked relative frequency bar chart using Survey Plots command
- Section 2.19 R: Added instructions for creating a stacked relative frequency bar chart using surveyPlot function

2 June 2025

- Section 1.4: corrected the figure labels for Figure 1.2 and Figure 1.3
- Section 1.8: included hyperlink to download jamovi
- Section 1.13: Added a note about using `na.rm=TRUE` when using the `density()` function in the presence of missing data
- Activity 2.6: moved to Activity 3.7
- Section 8.13: corrected typo - `install.packages(scatr)`

Module 1

Introduction to biostatistics and data fundamentals

Learning objectives

By the end of this module, you will be able to:

- Understand the difference between descriptive and inferential statistics
- Distinguish between different types of variables
- Present and report continuous data numerically and graphically
- Compute summary statistics to describe the centre and spread of data

Optional readings

Kirkwood and Sterne (2001); Chapters 2, 3 and 4. [UNSW Library Link]

Bland (2015); Chapter 4. [UNSW Library Link]

1.1 An introduction to statistics

The dictionary of statistics (Upton and Cook, 2008) defines statistics simply as: "The science of collecting, displaying, and analysing data."

Statistics is a branch of mathematics, and there are two main divisions within the field of statistics: mathematical statistics and applied statistics. Mathematical statistics deals with development of new methods of statistical inference and requires detailed knowledge of abstract mathematics for its implementation. Applied statistics applies the methods of mathematical statistics to specific subject areas, such as business, psychology, medicine and sociology.

Biostatistics can be considered as the "application of statistical techniques to the medical and health fields". However, biostatistics sometimes overlaps with mathematical statistics. For instance, given a certain biostatistical problem, if the standard methods do not apply then existing methods must be modified to develop a new method.

Scope of Biostatistics

Research is essential in the practice of health care. Biostatistical knowledge helps health professionals in deciding whether to prescribe a new drug for the treatment of a disease or to advise a patient to give up drinking alcohol. To practice evidence-based healthcare, health professionals must keep abreast of the latest research, which requires understanding how the studies were designed, how data were collected and analysed, and how the results were interpreted. In clinical medicine, biostatistical methods are used to determine the accuracy of a measurement, the efficacy of a drug in treating a disease, in comparing different measurement techniques, assessing diagnostic tests, determining normal values, estimating prognosis and

monitoring patients. Public health professionals are concerned about the administration of medical services or ensuring that an intervention program reduces exposure to certain risk factors for disease such as life-style factors (e.g. smoking, obesity) or environmental contaminants. Knowledge of biostatistics helps determine them make decisions by understanding, from research findings, whether the prevalence of a disease is increasing or whether there is a causal association between an environmental factor and a disease.

The value of biostatistics is to transform (sometimes vast amounts of) data into meaningful information, that can be used to solve problems, and then be translated into practice (i.e. to inform public health policy and decision making). When undertaking research having a biostatistician as part of a multidisciplinary team from the outset, together with scientists, clinicians, epidemiologists, healthcare specialists is vital, to ensure the validity of the research being undertaken and that information is interpreted appropriately.

1.2 What are data?

According to the Australian Bureau of Statistics, "data are measurements or observations that are collected as a source of information".[1] Note that technically, the word *data* is a plural noun. This may sound a little odd, but it means that we say "data are …" when discussing a set of measurements.

Other definitions that we use in this course are:

- **observation**, (or **record**, or **unit record**): one individual in the population being studied
- **variable**: a characteristic of an individual being measured. For example, height, weight, eye colour, income, country of birth are all types of variables.
- **dataset**: the complete collection of all observations

Types of variables

We can categorise variables into two main types: numeric or categorical.

Numerical variables (also called quantitative variables) comprise data that must be represented by a number, which can be either measured or counted.

Continuous variables can take any value within a defined range.

For example, age, height, weight or blood pressure, are continuous variables because we can make any divisions we want on them, and they can be measured as small as the instrument allows. As an illustration, if two people have the same blood pressure measured to the nearest millimetre of mercury, we may get a difference between them if the blood pressure is measured to the nearest tenth of millimetre. If they are still the same (to the nearest tenth of a millimetre), we can measure them with even finer gradations until we can see a difference.

Discrete variables can only take one of a distinct set of values (usually whole numbers). For discrete variables, observations are based on a quantity where both ordering and magnitude are important, such that numbers represent actual measurable quantities rather than mere labels.

For example, the number of cancer cases in a specified area emerging over a certain period, the number of motorbike accidents in Sydney, the number of times a woman has given birth, the number of beds in a hospital are all discrete variables. Notice that a natural ordering exists among the data points, that is, a hospital with 100 beds has more beds than a hospital with 75 beds. Moreover, a difference between 40 and 50 beds is the same as the difference between 80 and 90 beds.

Categorical variables comprise data that describe a 'quality' or 'characteristic'. Categorical variables, sometimes called qualitative variables, do not have measurable numeric values. Categorical variables can be nominal or ordinal.

A **nominal** variable consists of unordered categories. For example, gender, race, ethnic group, religion, eye colour etc. Both the order and magnitude of a nominal variable are unimportant.

[1] https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/data

If a nominal variable takes on one of two distinct categories, such as black or white then it is called a **binary** or dichotomous variable. Other examples would be smoker or non-smoker; exposed to arsenic or not exposed.

A nominal variable can also have more than two categories, such as blood group, with categories of: Group A, Group B, Group AB and Group O.

Ordinal variables consist of ordered categories where differences between categories are important, such as socioeconomic status (low, medium, high) or student evaluation rating could be classified according to their level of satisfaction: (highly satisfied, satisfied and unsatisfied). Here a natural order exists among the categories.

Note that categorical variables are often stored in data sets using numbers to represent categories. However, this is for convenience only, and these variable must not be analysed as if they were numeric variables.

1.3 Descriptive and inferential statistics

When analysing a set of data, it is important to consider the aims of the analysis and whether these are *descriptive* or *inferential*. Essentially, descriptive statistics summarise data from a single sample or population, and present a "snap-shot" of those data. Inferential statistics use sample data to make statements about larger populations.

Descriptive statistics

Descriptive statistics provide a 'picture' of the characteristics of a population, such as the average age, or the proportion of people born in Australia. Two common examples of descriptive statistics are reports summarising a nation's birth statistics, and death statistics.

Births

The Australian Institute of Health and Welfare produces comprehensive reports on the characteristics of Australia's mothers and babies using the most recent year of data from the National Perinatal Data Collection. The National Perinatal Data Collection comprises *all registered births* in Australia.

The most recent report, published in 2024, summarises Australian births from 2022. (Australian Institute of Health and Welfare (2024)).

One headline from the report is that "More First Nations mothers are accessing antenatal care in the first trimester (up from 51% in 2013 to 71% in 2022)". The report presents further descriptive statistics, such as the average maternal age (31.2 years) and the proportion of women giving birth by caesarean (39%).

Deaths

In another example, consider characteristics of all deaths in Australia in 2023 (Australian Bureau of Statistics (Thu, 10/10/2024 - 11:30)).

> "COVID-19 was the ninth leading cause of death in 2023, after ranking third in 2022."

The report presents the leading causes of death in 2023:

> "The leading cause of death was ischaemic heart disease, accounting for 9.2% of deaths. The gap between ischaemic heart disease and dementia (the second leading cause of death) has continued to narrow over time, with only 237 deaths separating the top two leading causes in 2023."

The top five causes of death are also presented as a graph, enabling a simple comparison of the changes in rates of death between 2014 and 2023.

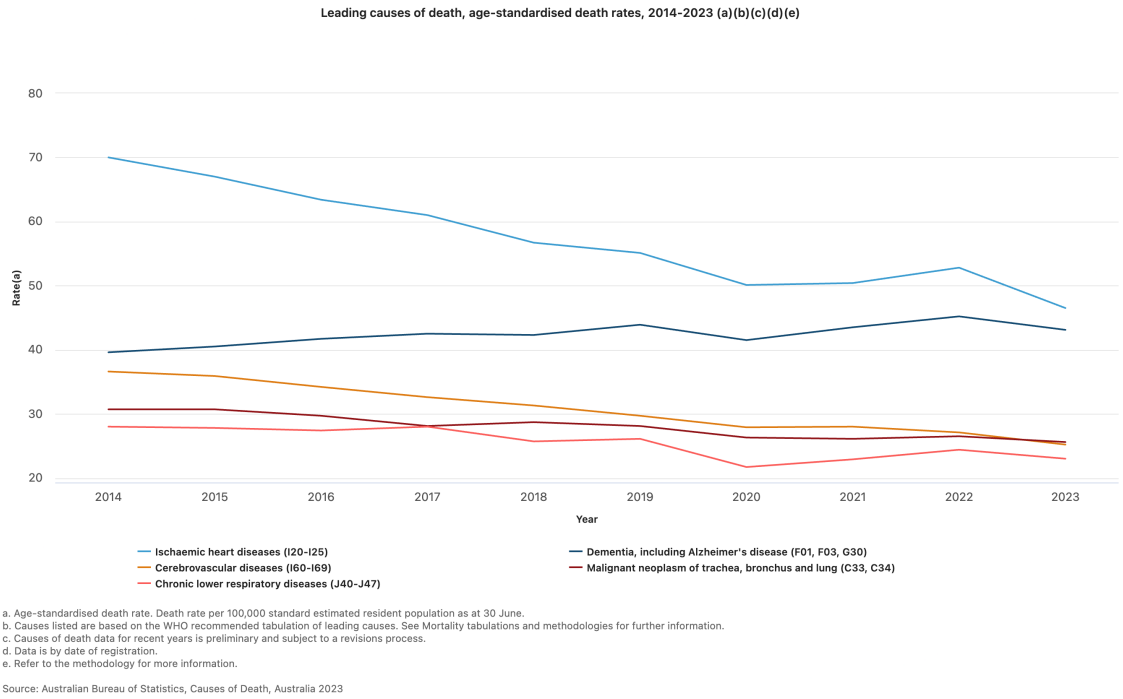


Figure 1.1: Leading causes of death, age-standardised death rates, 2014-2023

Inferential statistics

Inferential statistics use data collected from a sample to make conclusions (inferences) about the whole population from which the sample was drawn. For example, the Australian Institute of Health and Welfare's **Australia's health** reports (eg Australian Institute of Health and Welfare (2025)) use a representative sample to make estimates of the health of the whole of Australia. We will revisit *inferential statistics* in later modules.

1.4 Summarising continuous data

In the first two Modules, we will focus on ways to summarise and present data. We will see that the choice of presentation will depend on the type of variable being summarised. In this Module, we will focus on continuous variables, and will focus on categorical data in Module 2.

Summarising a single continuous variable numerically

When summarising continuous data numerically, there are two things we want to know:

1. What is the average value? And,
2. How variable (or spread out) are the data?

We will use a sample of 35 ages (in whole years) to illustrate how to calculate the average value and measures of variability:

59 41 44 43 31 47 53 59 35 60 54 61 67 52 43 46 39 69 50 64 57 39 54 50 51 31 48 49 70 44 60 51 37 53 34

Measures of central tendency

Mean

The most commonly used measure of the central tendency of the data is the mean, calculated as:

$$\bar{x} = \frac{\sum x}{n}$$

From the age example: $\bar{x}$ = 1745/35 = 49.9. Thus, the mean age of this sample is 49.9 years.

Median

Other measures of central tendency include the median and mode. The median is the middle value of the data, the value at which half of the measurements lie above it and half of the measurements lie below it.

To estimate the median, the data are ordered from the lowest to highest values, and the middle value is used. If the middle value is between two data points (if there are an even number of observations), the median is an average of the two values.

Using our example, we could rank the ages from smallest to largest, and locate the middle value (which has been bolded):

31 31 34 35 37 39 39 41 43 43 44 44 46 47 48 49 50 **50** 51 51 52 53 53 54 54 57 59 59 60 60 61 64 67 69 70

Here, the median age is 50 years.

Note that, in practice, the median is usually calculated by software automatically, and there is no need to rank our data.

Describing the spread of the data

In addition to measuring the centre of the data, we also need an estimate of the variability, or spread, of the data points.

Range

The absolute measure of the spread of the data is the range, that is the difference between the highest and lowest values in the dataset.

Range = highest data value – lowest data value

Using the age example, Range = 70 - 31 = 39 years.

The range is most usefully reported as the actual lowest and highest values e.g. Range: 31 to 70 years.

The range is not always ideal as it only describes the extreme values, without considering how the bulk of the data is distributed between them.

Variance and standard deviation

More useful statistics to describe the spread of the data around a mean value are the variance and standard deviation. These measures of variability depend on the difference between individual observations and the mean value (deviations). If all values are equal to the mean there would be no variability at all, all deviations would be zero; conversely large deviations indicate greater variability.

One way of combining deviations in a single measure is to first square the deviations and then average the squares. Squaring is done because we are equally interested in negative deviations and positive deviations; if we averaged without squaring, negative and positive deviations would ‘cancel out’. This measure is called the variance of the set of observations. It is ‘the average squared deviation from the mean’. Because the variance is in ‘square’ units and not in the units of the measurement, a second measure is derived by taking the square root of the variance. This is the standard deviation (SD), and is the most commonly used measure of variability in practice, as it is a more intuitive interpretation since it is in the same units as the units of measurement.

The formula for the variance of a sample (s^2) is:

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Note that the deviations are first squared before they are summed to remove the negative values; once summed they are divided by the sample size minus 1.

The sample standard deviation is the square root of the of the sample variance:

$$s = \sqrt{s^2}$$

For the age example, we would calculate the sample variance using statistical software. The sample standard deviation is estimated as: $s = 10.47$ years.

Characteristics of the standard deviation:

- It is affected by every measurement
- It is in the same units as the measurements
- It can be converted to measures of precision (standard error and 95% confidence intervals) (Module 3)

Interquartile range

The inter-quartile range (IQR) describes the range of measurements in the central 50% of values lie. This is estimated by calculating the values that cut the data at the bottom 25% and top 25%. The IQR is the preferred measure of spread when the median has been used to describe central tendency.

In the age example, the IQR is estimated as 43 to 58 years.

Population values: mean, variance and standard deviation

The examples above show how the sample mean, range, variance and standard deviation are calculated from the sample of ages from 35 people. If we had information on the age of the *entire* population that the sample was drawn from, we could calculate all the summary statistics described above (for the sample) for the population.

The equation for calculating the population mean is the same as that of sample mean, though now we denote the population mean as μ:

$$\mu = \frac{\sum x}{N}$$

Where $\sum x$ represents the sum of the values in the population, and N represents the total number of measurements in the population.

To calculate the population variance (σ^2) and standard deviation(σ), we use a slightly modified version of the equation for s^2:

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N}$$

with a population standard deviation of: $\sigma = \sqrt{\sigma^2}$.

In practice, we rarely have the information for the entire population to be able to calculate the population mean and standard deviation. Theoretically, however, these statistics are important for two main purposes:

1. the characteristics of the normal distribution (the most important probability distribution discussed in later modules) are defined by the population mean and standard deviation;
2. while calculating sample sizes (discussed in later modules) we need information about the population standard deviation, which is usually obtained from the existing literature.

Summarising a single continuous variable graphically

As well as calculating measures of central tendency and spread to describe the characteristics of the data, a graphical plot can be helpful to better understand the characteristics and distribution of the measurements obtained. *Histograms, density plots* and *box plots* are excellent ways to display continuous data graphically.

Frequency histograms

A frequency histogram is a plot of the number of observations that fall within defined ranges of non-overlapping intervals (called bins). Examples of frequency histograms are given in Figure 1.2.

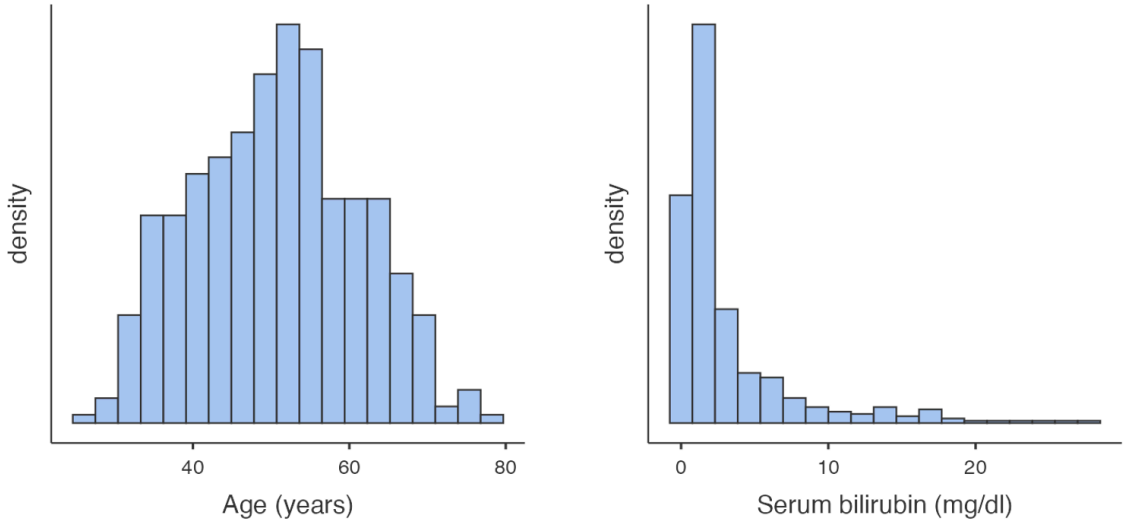


Figure 1.2: Histograms of age (left) and serum bilirubin (right) from PBC data

Some features of a frequency histogram:

- The area under each rectangle is proportional to the frequency
- The rectangles are drawn without gaps between them (that is, the rectangles touch)
- The data are 'binned' into discrete intervals (usually of equal width)

A slight variation on the frequency histogram is the **density histogram**, which plots the density on the y-axis. The density is a technical term, which is similar to the relative frequency, but is scaled so that the sum of the area of the bars is equal to 1.

Both the frequency and density histograms are useful for understanding how the data is distributed across the range of values. Taller bars indicate regions where the data is more densely concentrated, while shorter bars represent areas with fewer data points.

Density plot

A density plot can be thought of as a smoothed version of a density histogram. Like histograms, density plots show areas where there are a lot of observations and areas where there are relatively few observations. Figure 1.3 illustrates example density plots for the same data as plotted in Figure 1.2.

Like histograms, density plots allow you to see the overall shape of a distribution. They are most useful when there are only a small number of observations being plotted. When plotting small datasets, the shape of a histogram can depend on how the bins are defined. This is less of an issue if a density plot is used.

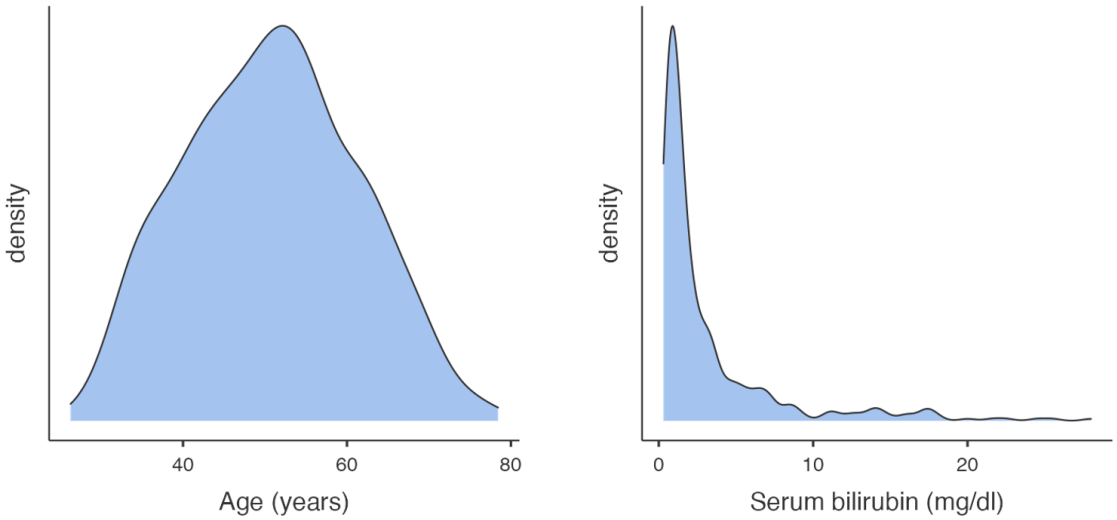


Figure 1.3: Density plots of age (left) and serum bilirubin (right) from a sample of data

Boxplots

Another way to inspect the distribution of data is by using a box plot. In a box plot:

- the line across the box shows the median value
- the limits of the box show the 25-75% range (i.e. the inter-quartile range (IQR) where the middle 50% of the data lie)
- the bars (or whiskers) indicate the most extreme values (highest and lowest) that fall within 1.5 times the interquartile range from each end of the box
  - the upper whisker is the highest value falling within 75th percentile plus 1.5 × IQR
  - the lower whisker is the lowest value falling within 25th percentile minus 1.5 × IQR
- any values in the dataset lying outside the whiskers are plotted individually.

Figure 1.4 presents two example boxplots for age and serum bilirubin.

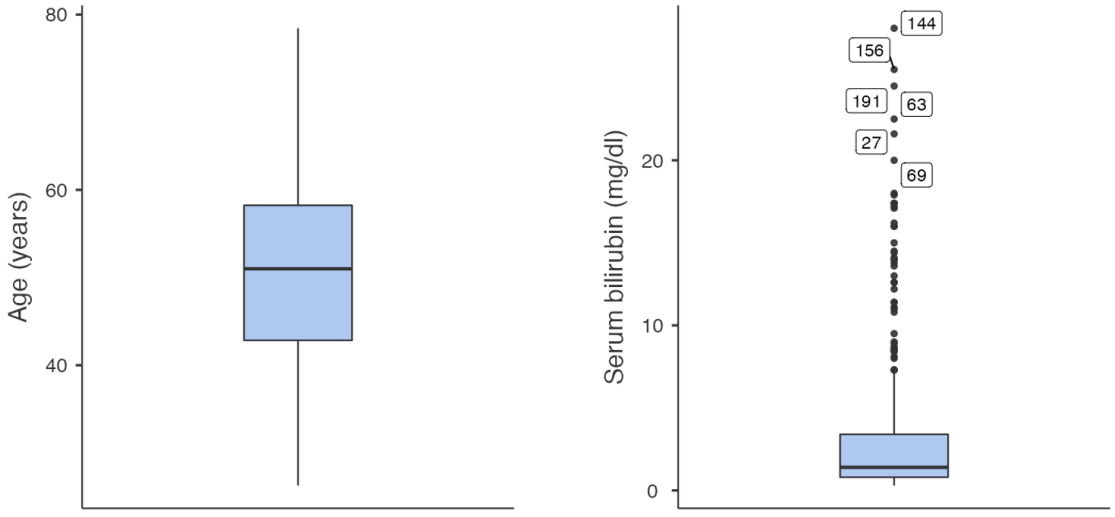


Figure 1.4: Box plot of age (left) and serum bilirubin (right) from PBC study data

The shape of a distribution

Histograms and density plots allow us to consider the shape of a distribution, and in particular, whether a distribution is *symmetric* or *skewed*.

In a histogram, if the rectangles fall in a roughly symmetric shape around a single midpoint, we say that the distribution is symmetric. Similarly, if a density plot looks roughly symmetric around a single point, the distribution is symmetric.

If the histogram or density plot has a longer tail to the right, then the data are said to be positively skewed (or skewed to the right); if the histogram or density plot has an extended tail to the left, then the data are negatively skewed (or skewed to the left).

> The skewness of a distribution is defined by the location of the longer tail in a histogram or density plot, not the location of the peak of the data.

From Figure 1.2 and Figure 1.3, we can see that the distribution for age is roughly symmetric, while the distribution for serum bilirubin is highly positively skewed (or skewed to the right).

While it is technically possible to determine the shape of a distribution using a boxplot, a histogram or density plot gives a more complete illustration of a distribution and would be the preferred method of assessing shape.

Which measure of central tendency to use

We introduced the mean and median in Section 1.4 as measures of central tendency. We need to assess the shape of a distribution to answer which is the more appropriate measure to use.

If a distribution is symmetric, the mean and median will be approximately equal. However, the mean is the preferred measure of central tendency as it makes use of every data point, and has more useful mathematical properties.

The mean is not a good measure of central tendency for skewed distributions, as the calculation will be influenced by the observations in the tail of the distribution. The median is the preferred statistic for describing central tendency in a skewed distribution.

If the data exhibits a symmetric distribution, we use the standard deviation as the measure of spread. Otherwise, the interquartile range is preferred.

1.5 Exploratory data analysis for continuous data

Before conducting any formal analysis, it is good practice to undertake exploratory data analysis. This analysis step gives you a high-level overview of your data: what do your data look like, what are the main features, what shape is the distribution, and are there any unusual values.

Exploring continuous data is best done by examining plots: specifically density plots and boxplots. Density plots are useful in determining the shape of a distribution, to help you decide what summary measures to use, and what type of analysis to conduct. Boxplots can be useful in identifying any unusual points - often called outliers.Outliers can be problematic and the decision to include them or omit them from further analyses can be difficult.

After detecting any outliers or extreme values, **do not automatically exclude them from the analysis**. First, it is important to check the original data collection form or questionnaire to rule out the possibility of a data entry error. If the outlier is not a data entry error, it is then important to decide whether the observation is biologically possible. This step will usually need to be answered by a topic expert. If the outlier is biologically possible, it must be included in the analysis. Only if the outlier is biologically impossible should it be set to missing.

An introduction to jamovi

Learning outcomes

By the end of these notes, you will be able to:

- navigate the jamovi interface
- input and import data into jamovi
- use jamovi menus to summarise data
- perform basic data transformations
- assign variable and value labels
- understand the difference between saving data and saving jamovi output
- copy jamovi output to a standard word processing package

1.6 Introduction

From the jamovi website:

> jamovi is a new "3rd generation" statistical spreadsheet. Designed from the ground up to be easy to use, jamovi is a compelling alternative to costly statistical products such as SPSS and SAS.
>
> jamovi is built on top of the R statistical language, giving you access to the best the statistics community has to offer. Would you like the R code for your analyses? jamovi can provide that too.
>
> jamovi will always be free and open - that's one of our core values - because jamovi is made by the scientific community, for the scientific community.

The notes provided in this course will cover the basics of using jamovi: there is much more to jamovi than we will cover in this course.

1.7 Part 1: An introduction to jamovi

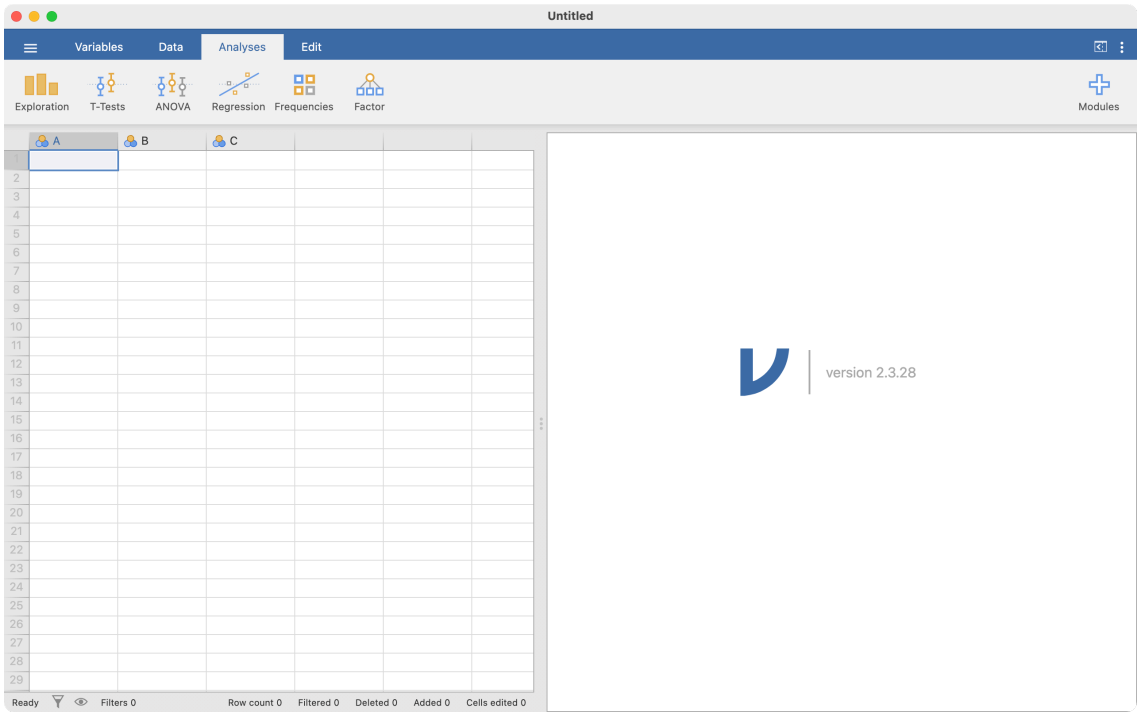
In this very brief section, we will introduce jamovi by calculating the average of six ages. Open the jamovi package in the usual way (note that while jamovi is available for Windows, MacOS and linux, most of the screenshots in these notes will be based on the macOS version.)

1.8 Installing jamovi

jamovi can be downloaded for no cost at https://www.jamovi.org/download.html for Windows, macOS and linux. At the time of writing, Version 2.6.26 solid is the appropriate version to use. Download and install jamovi in the usual way.

1.9 A simple jamovi analysis

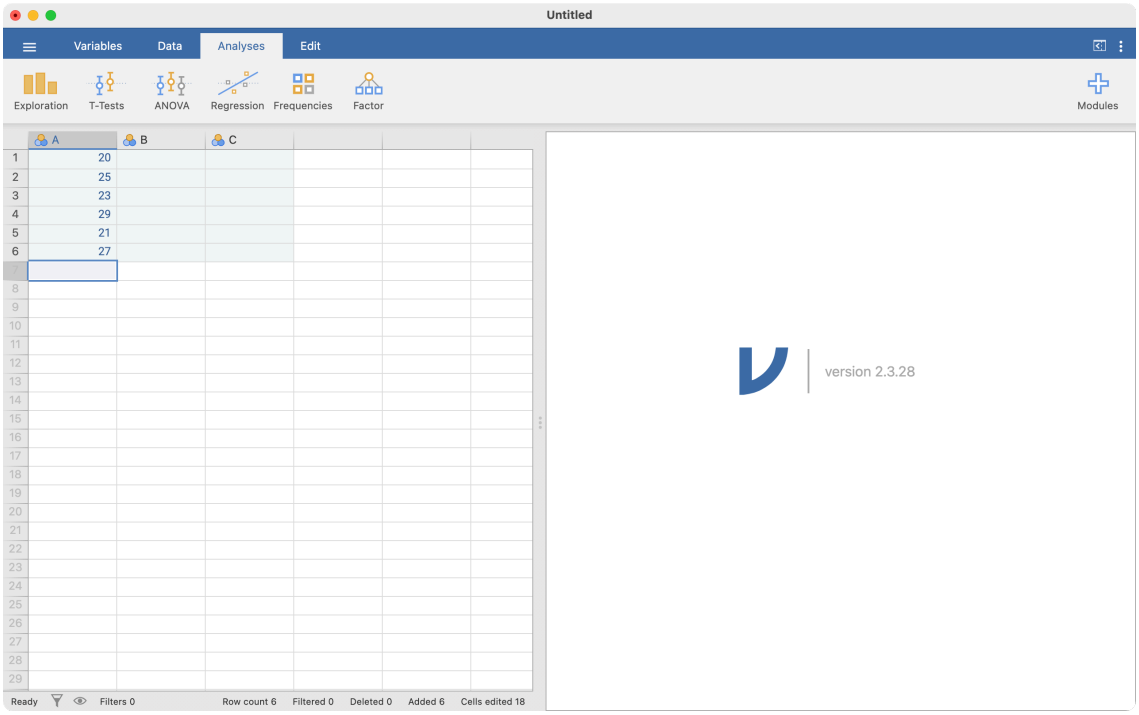
When you first open jamovi, it will look something like the following.



On the left-hand side of the window is the spreadsheet view, and the right is where results of statistical analyses will appear. The **spreadsheet** is where data can be entered or changed.

> TASK
>
> Enter the following six ages into jamovi, starting at the top-left cell, by typing each number and then hitting Enter:
> 20 25 23 29 21 27
> If you make a mistake, simply click the incorrect cell, and enter the correct value.

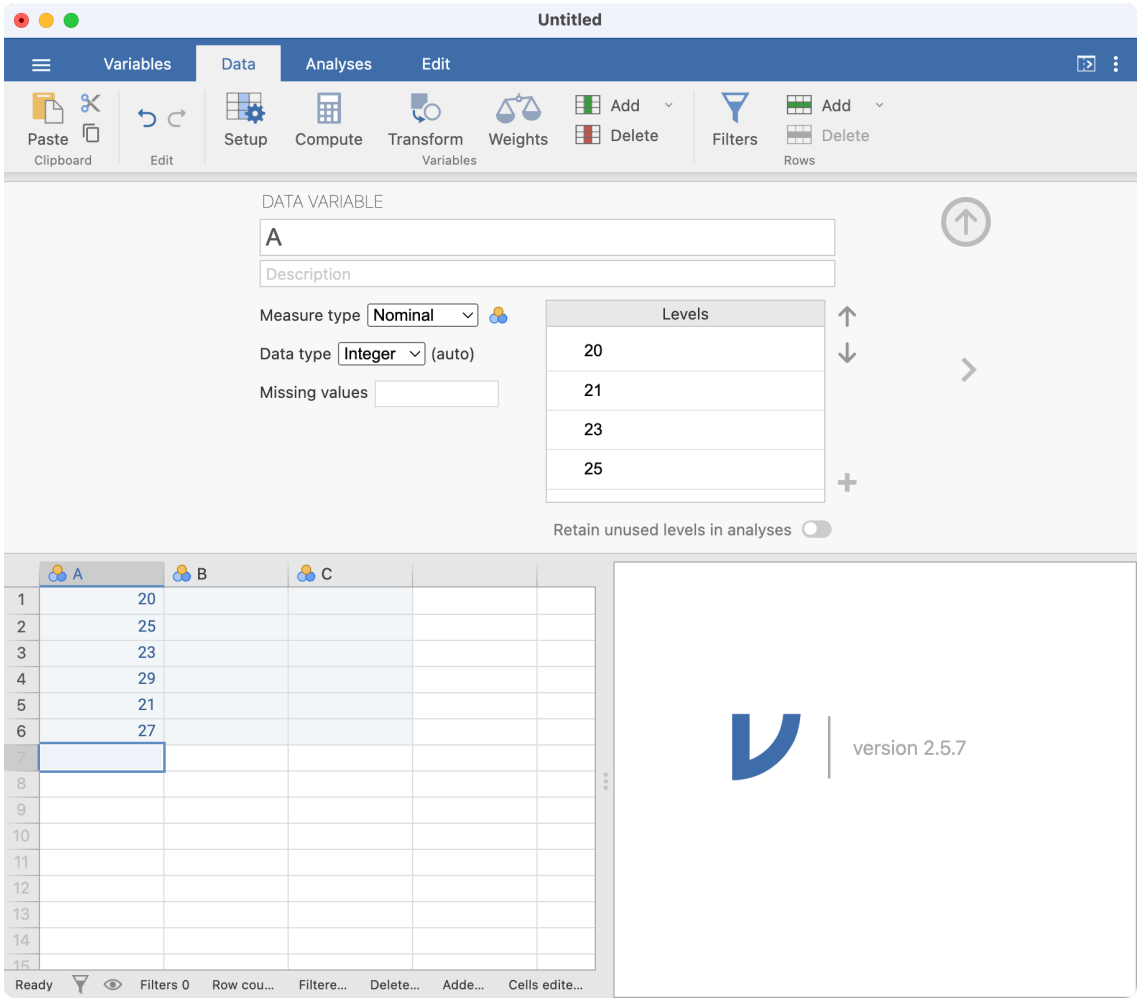
Your screen should look like this:



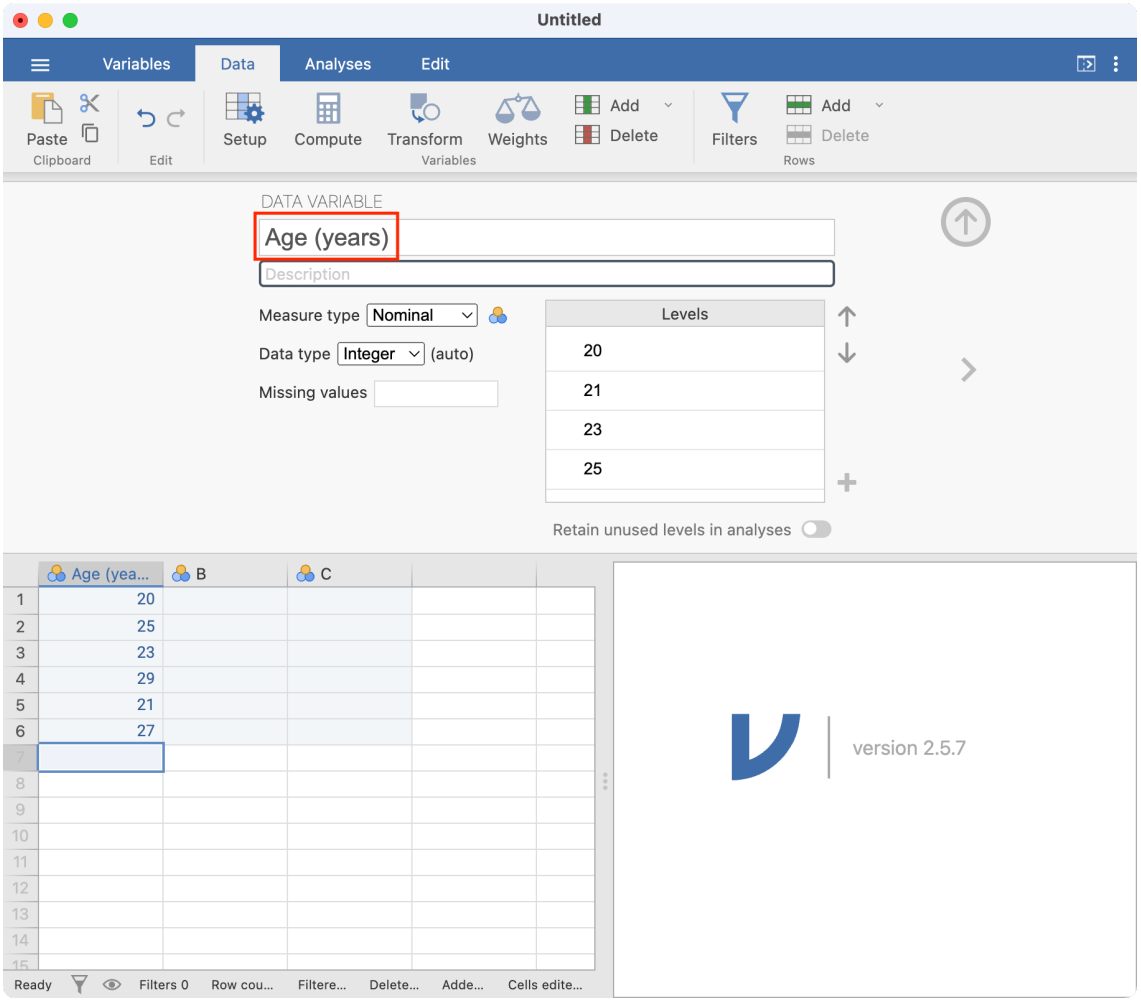
There are two things to note here:

1. Data in jamovi are entered down a column: columns represent variables, and rows represent observations. So our six observations of age are entered in one column.
2. jamovi has given the name of `A` to our column of ages.

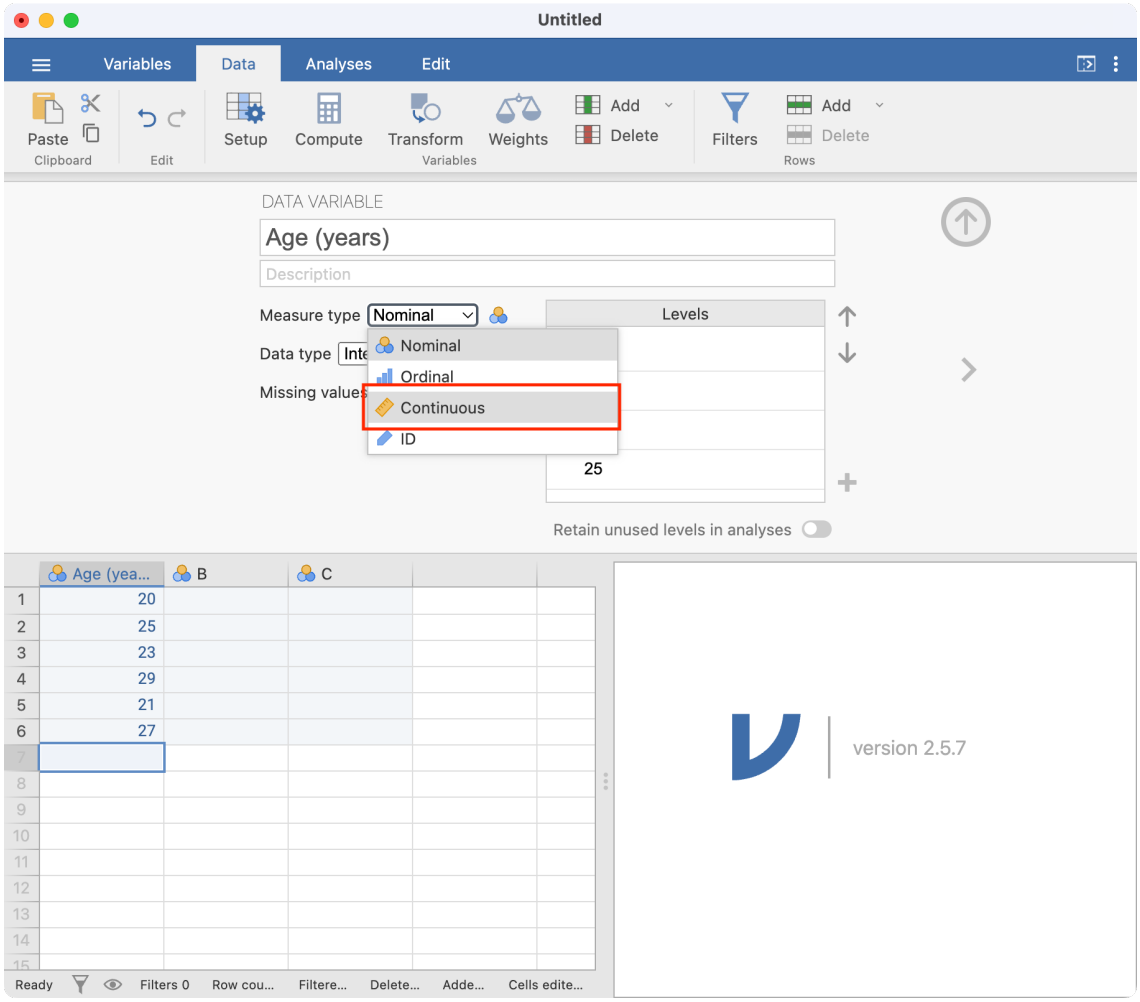
Let's rename our variable from `A` to `Age (years)`. There are a number of ways of doing this: here we will click the **Data** tab which allows us to change aspects of our dataset, and then click **Setup**. By default, the first column is selected - you can choose any column simply by clicking it.



In this window, you can change many variable properties, such as variable names and variable types. To change the variable name, click the name of the variable, currently `A`, at the top of the window. Replace `A` with `Age (years)`:



Note that jamovi has assumed these data are *Nominal* data; this can be changed to *Continuous* by choosing the appropriate item in the **Measure type** drop-down menu:



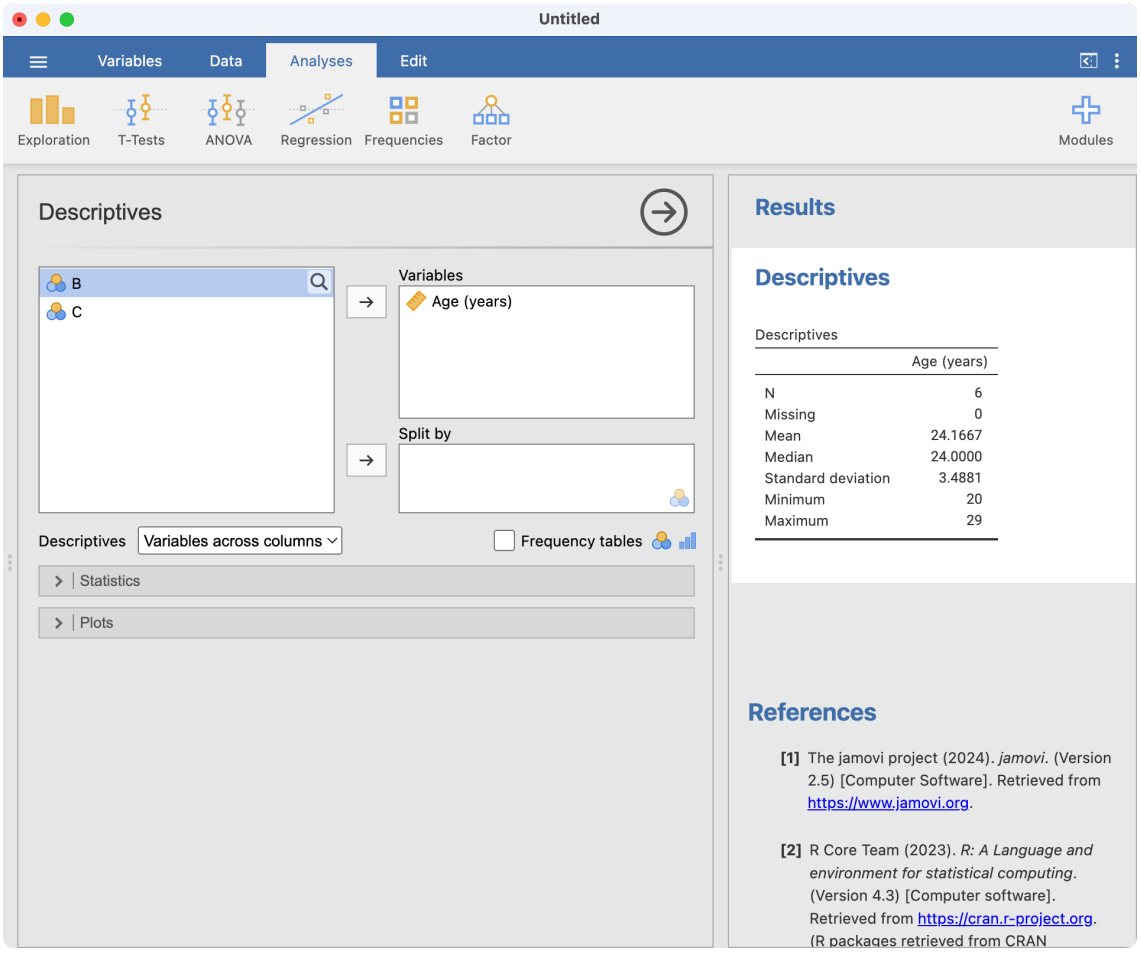
Once you have finished naming all your variables, you can click the **Up** arrow to close the Setup tab to view the spreadsheet again.

jamovi is *very* flexible with its naming convention, and allows variable names that many other statistical packages would not allow. Often, the only way to get publication quality graphs or tables is to name your variables as completely as you can. However, I would recommend the following conventions when choosing variable names in jamovi:

- try to keep your variable names relatively short;
- variable names should start with a letter;
- variable names are case-sensitive (so age, Age and AGE could represent three different variables)

> TASK
>
> Rename the variable `A` with the name `Age (years)`, and define age as a *Continuous* variable.

Now that we have entered our six ages, let's calculate the mean age. Choose **Analyses > Exploration > Descriptives**. The **Descriptives** dialog box will appear. Move the variable `Age` into the **Variables** box by clicking `Age (years)` and then clicking the right-arrow icon. The window should appear as:



You can see the results of the analysis in the right-hand pane.

> TASK
>
> Calculate summary statisitics for `Age (years)` and confirm: there are 6 observations, with a mean age of 24.2 years, a standard deviation of 3.49 years, a minimum of 20 and a maximum of 29 years.

The jamovi environment

Now that we have seen a simple example of how to use jamovi, let's describe the jamovi environment. There are two main views in jamovi: the **Spreadsheet** view, available by clicking the **Data** tab, and the **Analysis** view, available by clicking the **Analyses** tab. We will tend to use these two tabs through the course.

The unique thing about jamovi is that *all analyses are updated whenever the data are changed in the spreadsheet*. While this can be convient, *care must be taken not to make any unintended changes to your data while in the spreadsheet view.*

jamovi also has a way of opening and saving data, using the three lines in the upper-left corner:  . This collection of commands lets you open data, save data and export data and output.

> Tip
>
> The **Special Import** command is hardly ever used. You should use **Open** to open all types of data in jamovi.

1.10 Part 2: Obtaining summary statistics for continuous data

In this exercise (spanning Modules 1 and 2), we will analyse data to complete a descriptive table from a research study. The data come from a study in primary biliary cirrhosis, a condition of the liver, from Therneau and Grambsch (2010), Modeling Survival Data: Extending the Cox Model. By the end of this exercise, we will have completed the following table.

Table 1.1: Summary of 418 participants from the PBC study (Therneau and Grambsch, 2000)

Characteristic		Summary
Age (years)		Mean (SD) or Median [IQR]
Sex	Male	n (%)
	Female	n (%)
AST* (U/ml)		Mean (SD) or Median [IQR]
Serum bilirubin		Mean (SD) or Median [IQR]
Stage	I	n (%)
	II	n (%)
	III	n (%)
	IIIV	n (%)
Vital status at study end	Alive: no transplant	n (%)
	Alive: transplant	n (%)
	Deceased	n (%)

* asparate aminotransferase

TASK

Download the table shell, saved on Moodle as `PBC Table1.docx`, and the information file called `mod01_pbc_info.txt`.

Opening a data file

Typing data directly into jamovi is not common; we usually open data that have been saved as a file. jamovi can open many types of files, including text (txt), comma separated (csv), Microsoft Excel (xlsx), R (rds), Stata (dta) and more. Here, we will open a dataset that has been stored as an R data file (which has the `.rds` suffix).

TASK

Load the sample data set called `mod01_pbc.rds` into jamovi using the following steps:

1. Locate the data set called `mod01_pbc.rds` on Moodle or the PHCM9795 home-page. Click the file to download it, and then save it in a folder you will be able to locate later - for example, your OneDrive folder.
2. In jamovi, click the three-bar icon, then choose **Open**. jamovi usually searches in the most recently used folder, so most times you will need to **click Browse**. Browse to where you stored the dataset and click **Open**.

Confirm that there are 418 rows by examining the **Row count** at the bottom of the screen. Examine the `pbc_info.txt` file for a description of each variable.

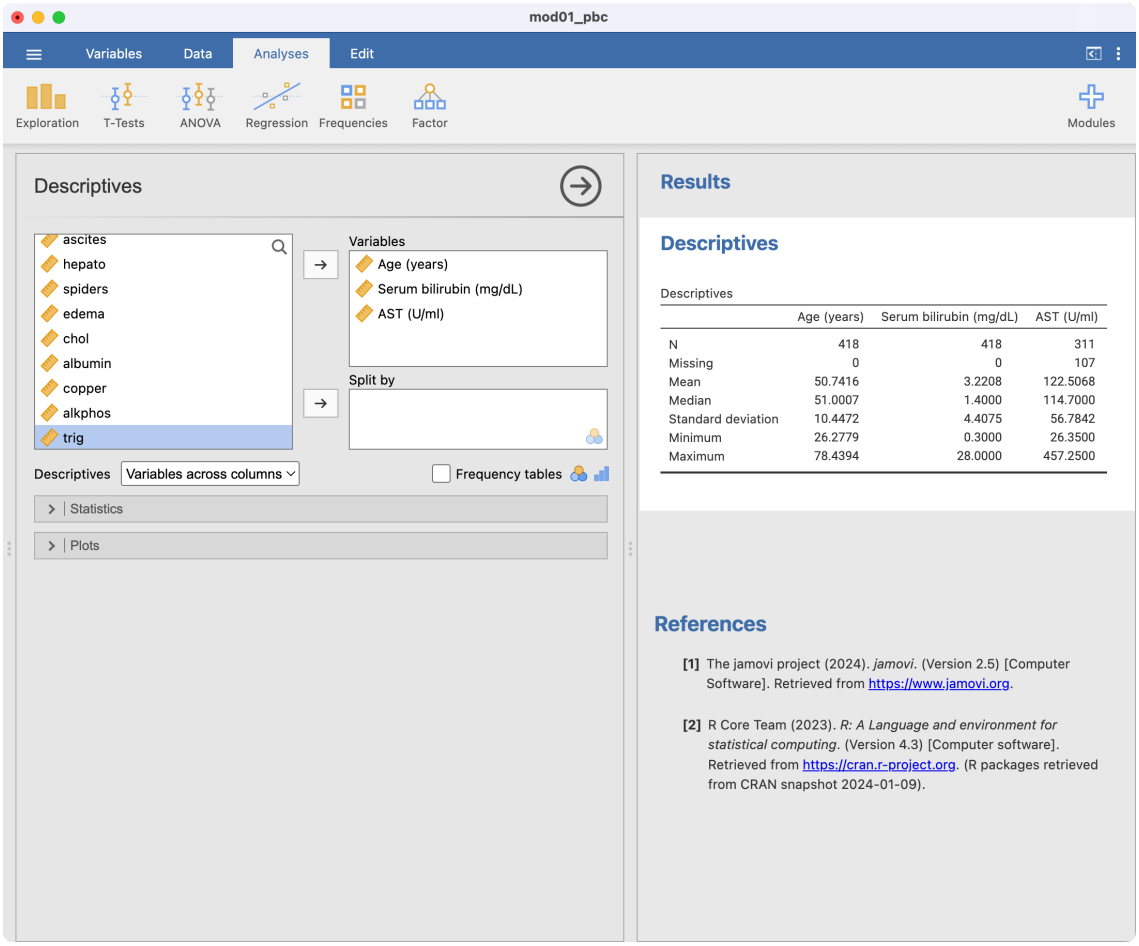
Assigning meaningful variable names

As we saw earlier, jamovi has can allow quite useful variable names, which will appear when creating output. For example, the variable entered as bili could be named `Serum bilirubin (mg/dl)`.

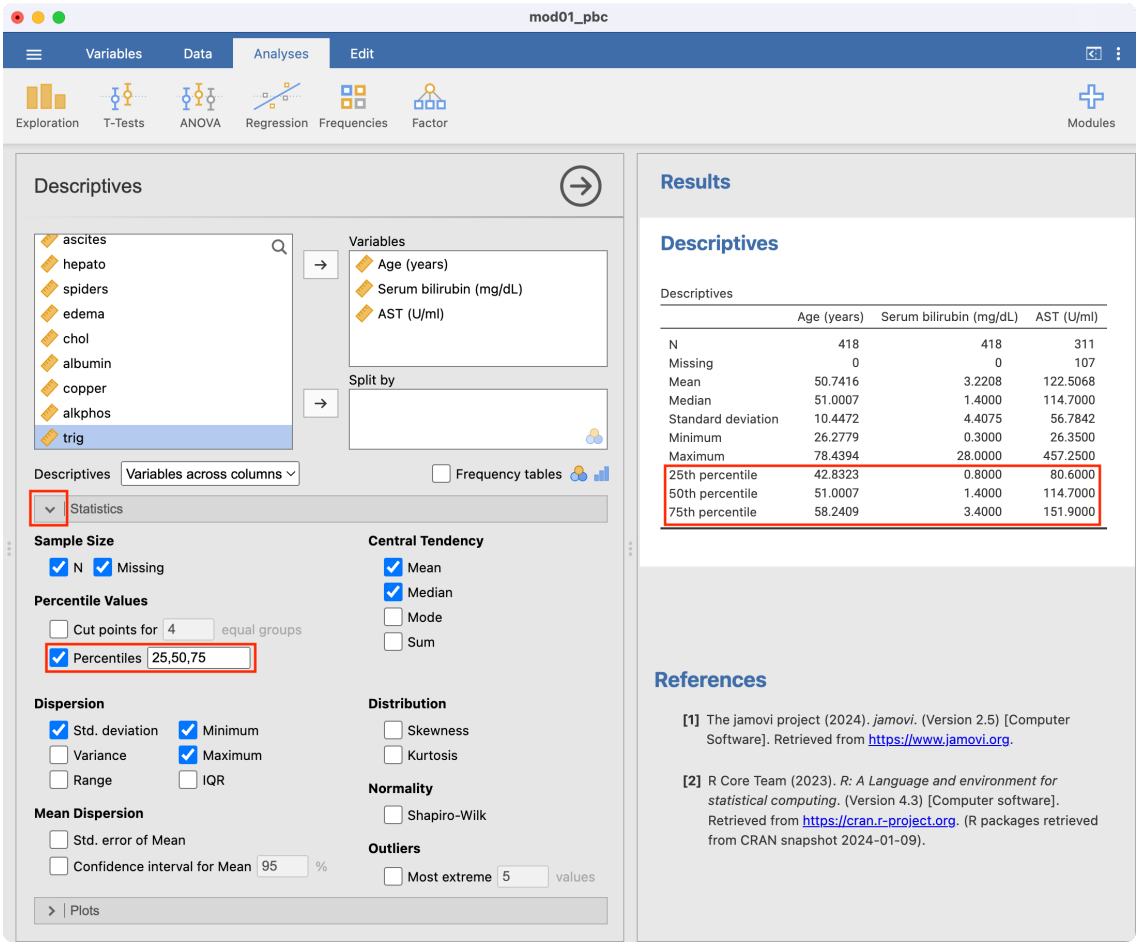
> **TASK**
>
> Assign meaningful variable names to the variables used in Table 1. You should to refer to the file `pbc_info.txt` to determine what each variable represents.

Summarising continuous variables

As we saw in Part 1, continuous variables can be summarised using **Analyses > Exploration**. There are three continuous variables that we would like to summarise: age, AST and serum bilirubin. Each of these can be listed in the **Exploration** dialog box, as shown below. The summaries are calculated automatically:



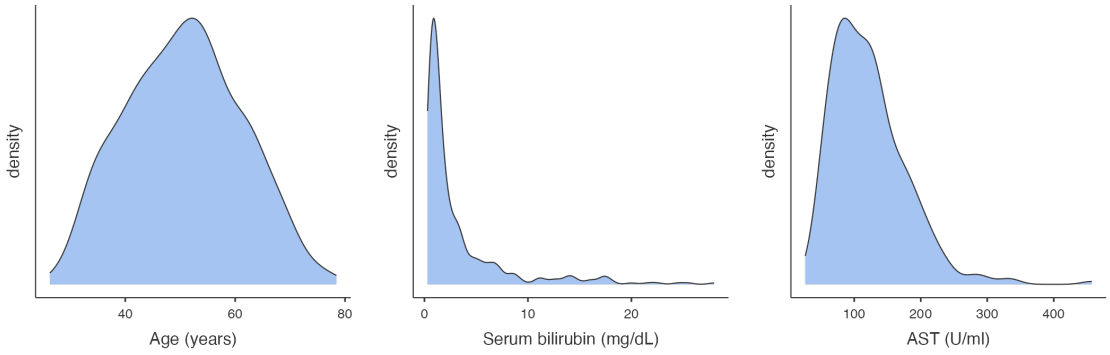
By default, the exploration command presents the number of observations, the number of missing observations, the mean, median, standard deviation, minimum and maximum. We may be interested in obtaining the interquartile range as well, so we select the **Statistics** arrow, and choose **Percentiles**:



For each of our three continuous variables, we need to decide whether to present the mean and standard deviation, or the median and interquartile range. This decision can be made after examining a density plot (and perhaps a boxplot) for each variable.

Producing a density plot

To produce a density plot, click the arrow next to **Plots** and choose **Density**. Plots will be produced for each variable listed in the **Variables** box. The density plots will be produced one after another, but they have been presented horizontally here:



Producing a boxplot

Producing boxplots is done by ticking the **Box plot** box. By default, jamovi labels each of the points that it considers to be an outlier with its row number; this can be turned off if desired.

TASK

Obtain density plots and boxplots for age, AST and bilirubin.
Based on these plots, decide whether the mean or the median is the appropriate summary to use for each variable.

Saving your work from jamovi

Now that you have made some changes to the pbc data and conducted some analyses, it is good practice to save your work. jamovi uses its own file format to save **both data and output**, using files with .omv suffix. All changes to your data will be saved, as well as all existing output. However, work saved by jamovi will only be able to be opened by jamovi - you will not easily be able to share your data or your output with colleagues who do not have jamovi. To save a jamovi session, choose **Save** in the three-lines tab.

If you want to share work with colleagues who do not have jamovi, you can use **Export** to save your data in another file format (recognising that variable and value labels will not be exported), or save your output as a `pdf` or `html` file.

Copying output from jamovi

An easy way to share output between your colleagues is to copy the output into a word processor package (e.g. Microsoft Word). To copy output from jamovi, you can **right-click**[2] the output with your mouse, and choose **Export**. This will copy the output as plain text for pasting into a Word document. If you select a single table for copying, you can also **Copy table** or **Copy table as HTML**. Whichever way you copy output into Word, you will need to make sure your output conforms with all style guides required for your final publication.

TASK

Complete Table 1 for continuous variables using the output generated in this exercise. You should decide on whether to present continuous variables by their means or medians, and present the most appropriate measure of spread. Include footnotes to indicate if any variables contain missing observations.

1.11 Setting a value to missing

As we saw in Section 1.5, it is important to explore our data to identify any unusual observations. If an error is found, the best method for correcting the error is to go back to the original data e.g. the hard copy questionnaire, to obtain the original value, entering the correct value into jamovi. If the original data is not available or the original data is also incorrect, the erroneous value is often excluded from the dataset.

Consider a sample dataset: `mod01_weight_1000.rds`, which contains the weights of 1000 people. A density plot and a boxplot should be examined before we start analysing these data:

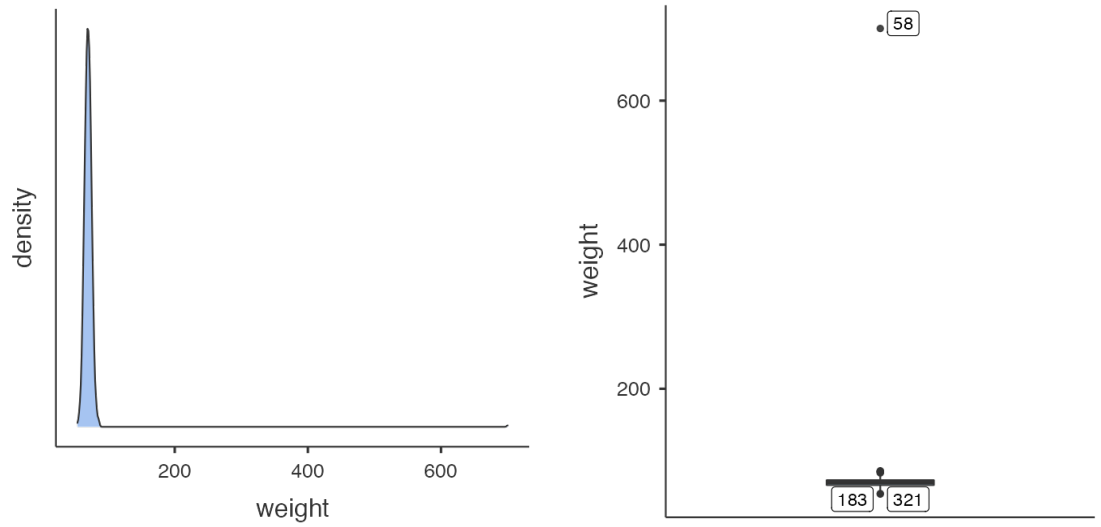
There is a clear outlying point shown in the boxplot. Although not obvious, the same point is shown in the density plot as a small blip around 700kg. Obviously this point is unusual, and we should investigate. You will need to decide if any usual values are a data entry error or are biologically plausible. If an extreme value or "outlier", is biologically plausible, it should be included in all analyses.

[2] **Windows**

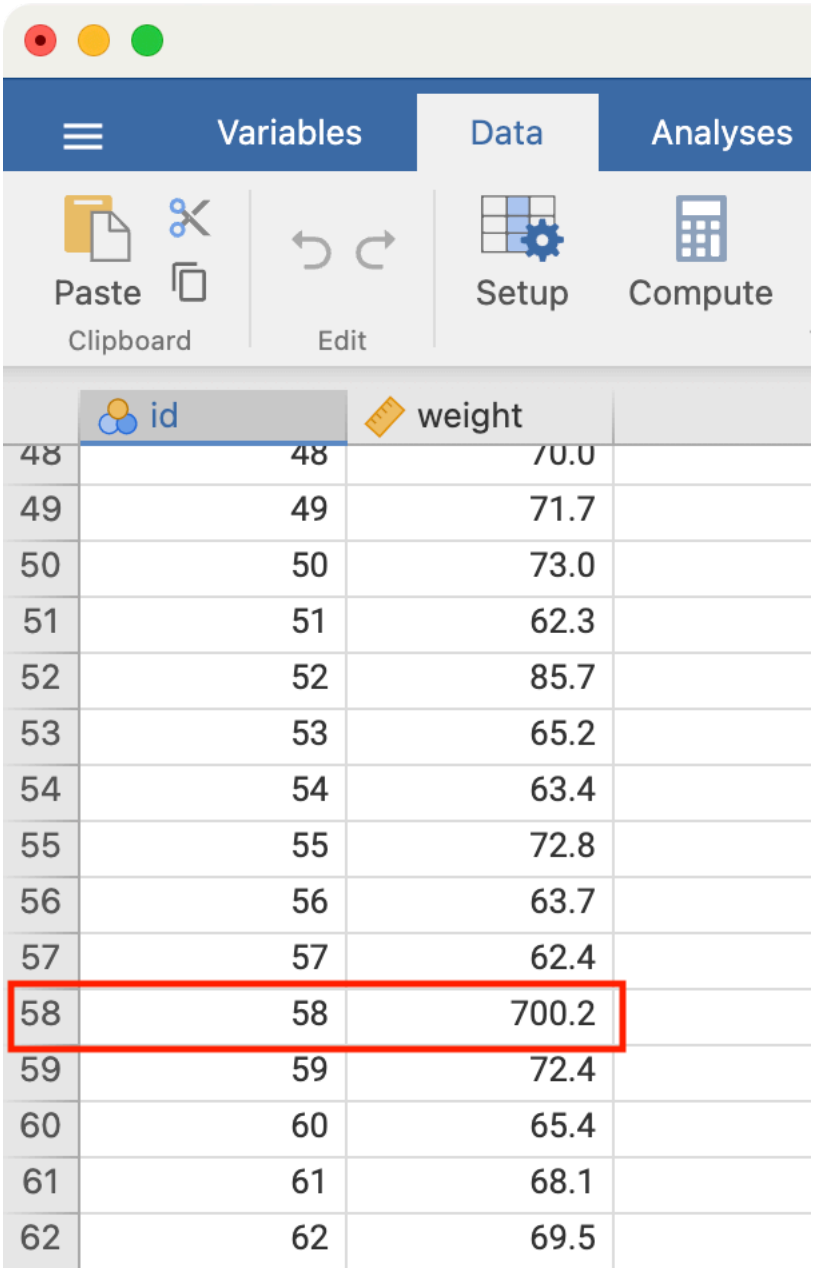
- Click the right mouse button to open a context menu
- Shows options relevant to what you clicked on (copy, export, add note)

Mac

- Two-button mouse: Press right button
- One-button mouse/trackpad: Hold Control (Ctrl) while clicking
- Functions the same as Windows - shows context-specific options



Notice the boxed number in the boxplot: this is the record number in jamovi's spreadsheet. Click **Data** to view the spreadsheet, and scroll to record number 58:



	id	weight
48	48	70.0
49	49	71.7
50	50	73.0
51	51	62.3
52	52	85.7
53	53	65.2
54	54	63.4
55	55	72.8
56	56	63.7
57	57	62.4
58	58	700.2
59	59	72.4
60	60	65.4
61	61	68.1
62	62	69.5

We see that there is a very high value of 700.2kg. A value as high as 700kg is likely to be a data entry error (e.g. error in entering an extra zero) and is not a plausible weight value. Here, **you should check your original data**.

If you do not have access to the original data, it would be safest to set this value as missing. You do change this in jamovi by clicking the datapoint and pressing **Delete** or **Backspace**. This has set this weight to missing.

Note: if an extreme value lies within the range of biological possibility it should not be set to missing.

The same process could be used to replace the incorrect value with the correct value. For example, if you do source the original medical records, you might find that the original weight was recorded in medical records as 70.2kg. We could use the same process to replace 700.2 by 70.2 by entering the correct value in the cell.

Once you have checked your data for errors, you are ready to start analysing your data.

> **!** Important
>
> Whenever you start changing data in jamovi, you should **always keep an original, unedited copy of your data.** You can do this using **Save As** to save your edited work, leaving the original data unchanged.

An introduction to R and RStudio

Learning outcomes

By the end of this Module, you will be able to:

- understand the difference between R and RStudio
- navigate the RStudio interface
- input and import data into R
- use R to summarise data
- perform basic data transformations
- understand the difference between saving R data and saving R output
- copy R output to a standard word processing package

1.12 Part 1: An introduction to R

"R is a language and environment for statistical computing and graphics." Link. It is an open-source programming language, used mainly for statistics (including biostatistics) and data science.

The aim of these notes is to introduce the R language within the RStudio environment, and to introduce the commands and procedures that are directly relevant to this course. There is so much more to R than we can cover in these notes. Relevant information will be provided throughout the course, and we will provide further references that you can explore if you are interested.

R vs RStudio

At its heart, R is a programming language. When you install R on your computer, you are installing the language and its resources, as well as a very basic interface for using R. You can write and run R code using the basic R app, but it's not recommended.

RStudio is an "Integrated Development Environment" that runs R while also providing useful tools to help you write code and analyse data. You can think of R as an engine which does the work, and RStudio as a car that uses the engine, but also provides useful tools like GPS navigation and reversing cameras that help you drive.

Note: even though we recommend that you use RStudio, you still need install R. **RStudio will not run without R installed.**

In summary, we recommend you use RStudio to write R code.

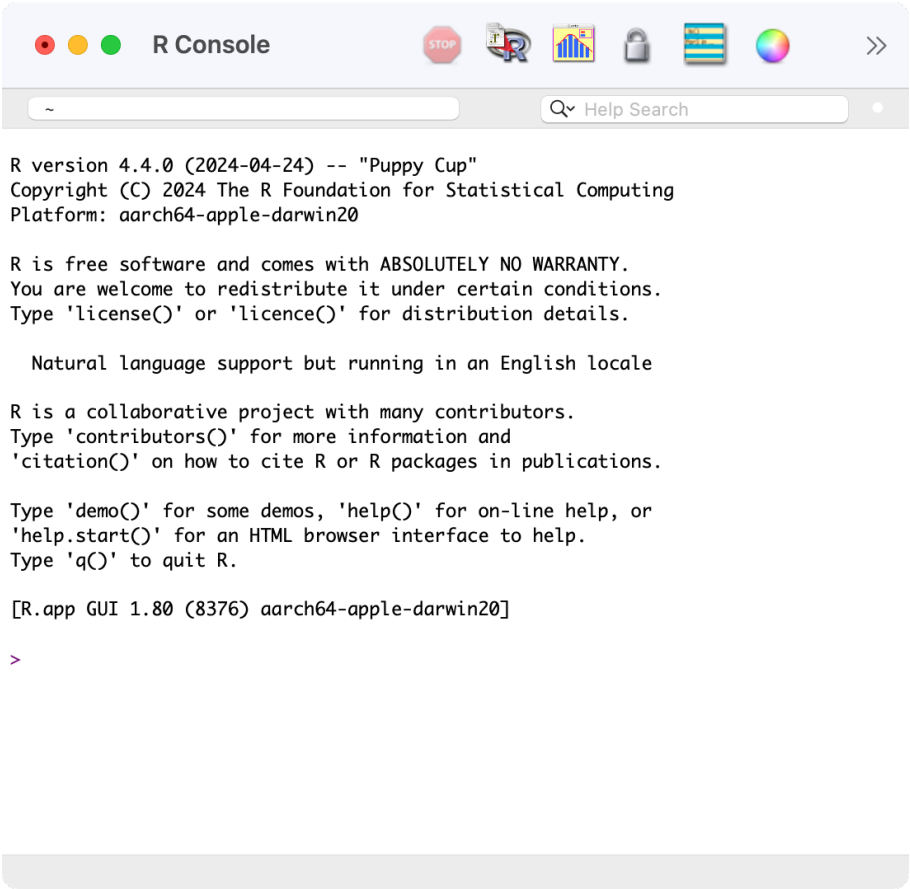
Installing R and RStudio

To install R on your computer

1. Download the R installer from:
    a. for Windows: https://cran.r-project.org/bin/windows/base/
    b. for MacOS: https://cran.r-project.org/bin/macosx/
2. Install R by running the installer and following the installation instructions. The default settings are fine.

- **Note for macOS:** if you are running macOS 10.8 or later, you may need to install an additional application called XQuartz, which is available at https://www.xquartz.org/. Download the latest installer (XQuartz-2.8.5.dmg as of May 2024), and install it in the usual way.

3. Open the R program. You should see a screen similar to below:



Near the bottom of the R screen, you will find the ">" symbol which represents the command line. If you type `1 + 2` into the command line and then hit enter you should get:

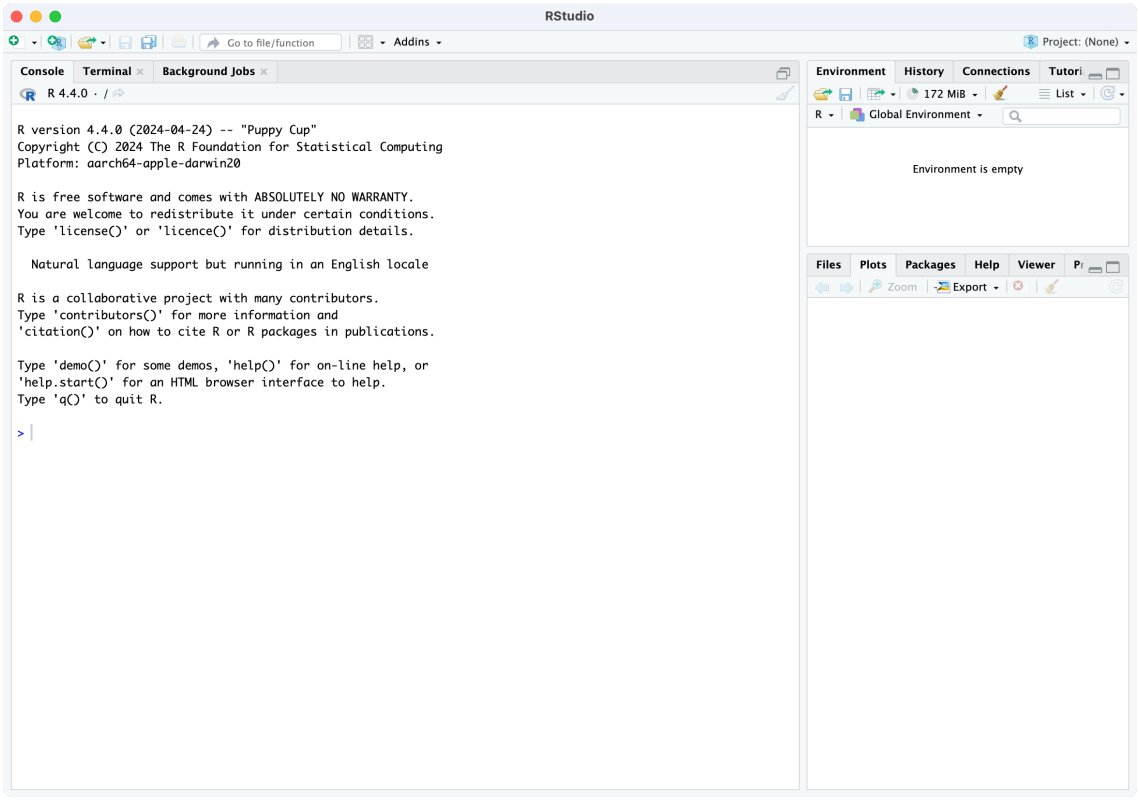
```
[1] 3
```

This is R performing your calculation, with the `[1]` indicating that the solution to `1 + 2` is a single number (the number 3).

At this point, close R - we will not interact with R like this in the future. You can close R by typing `quit()` at the command prompt, followed by the return key, or in the usual way of closing an application in your operating system. There is no need to save anything here if prompted.

To install RStudio on your computer

1. Make sure you have already installed R, and verified that it is working.
2. Download the RStudio desktop installer at: https://posit.co/download/rstudio-desktop/. The website should detect your operating system and link to the appropriate installer for your computer.
3. Install RStudio by running the installer and following the installation instructions. The default settings are fine.
4. Open RStudio, which will appear similar to the screenshot below:



Locate the command line symbol ">" at the bottom of the left-hand panel. Type `1 + 2` into the command line and hit enter, and you will see:

```
[1] 3
```

This confirms that RStudio is running correctly, and can use the R language to correctly calculate the sum between 1 and 2!

RStudio currently comprises three window panes, and we will discuss these later.

> TASK
>
> Install R and RStudio and confirm they are both working correctly.

Recommended setup

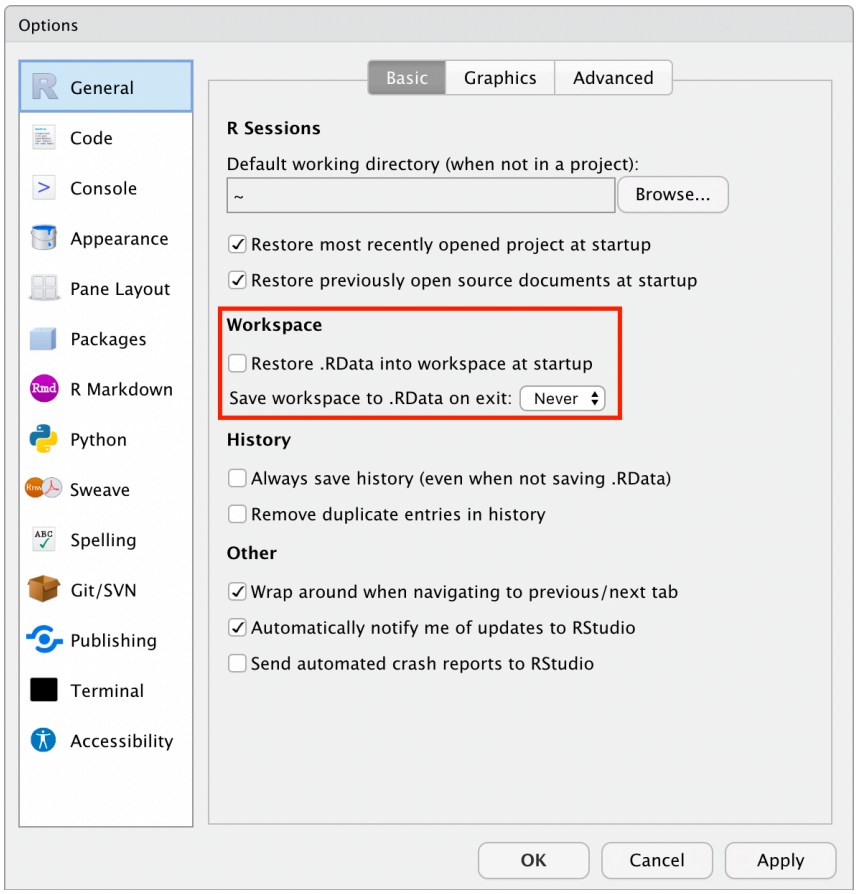
I will provide a recommended setup for R and RStudio in this section. You are free to use alternative workflows and setup, but this setup works well in practice.

RStudio preferences

By default, RStudio will retain data, scripts and other objects when you quit your RStudio session. Relying on this can cause headaches, so I recommend that you set up RStudio so that it does not preserve your workspace between sessions. Open the RStudio options:

- Mac: **Edit > Settings**
- Windows: **Tools > Global Options**

and **deselect "Restore .RData into workplace at startup"**, and choose: "**Save workspace to .RData on exit: Never**".

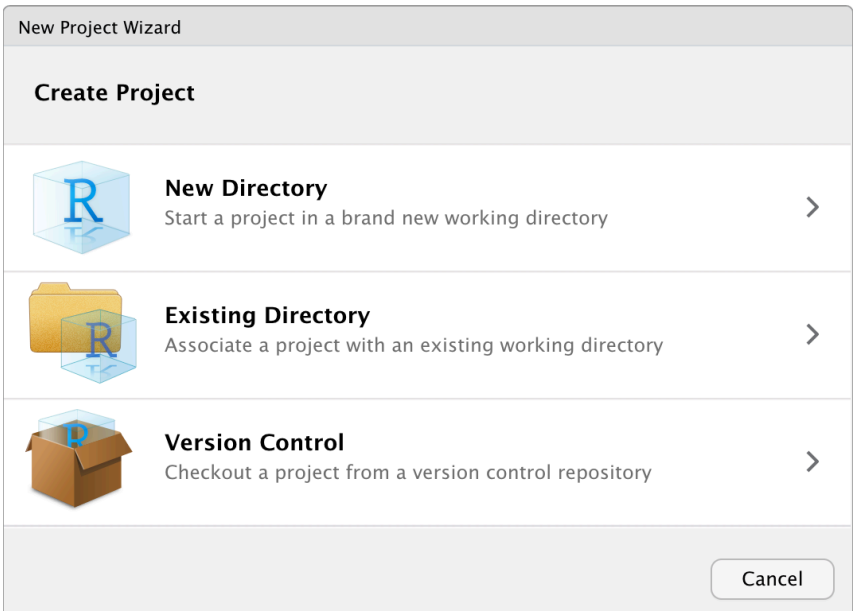


Set up a project

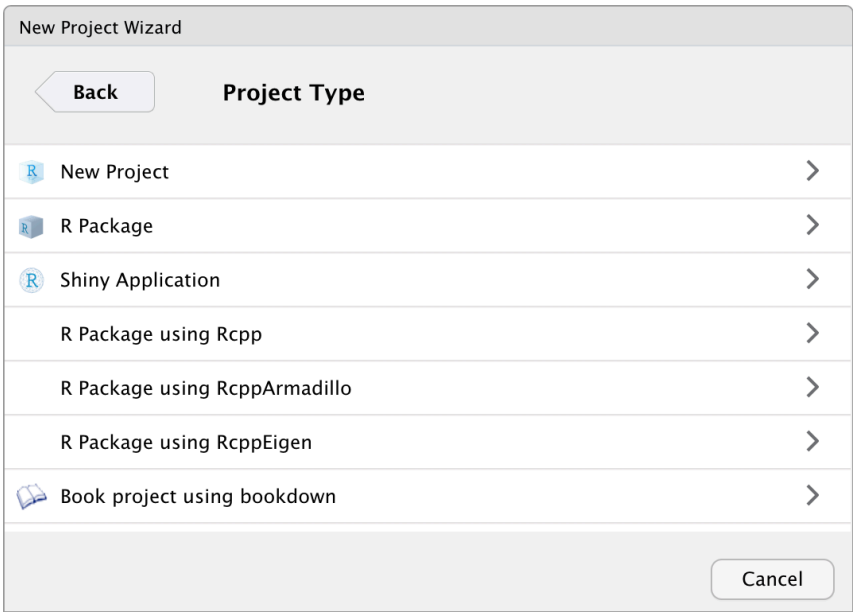
A project in RStudio is a folder that RStudio recognises as a place to store R scripts, data files, figures that are common to an analysis project. Setting up a folder allows much more simple navigation and specification of data files and output. More detail can be found in Chapter 8 of the excellent text: R for Data Science. Using projects is not necessary, but I recommend working with projects from day one.

We will create a project called **PHCM9795** to store all the data you will use and scripts that you will write in this course. First, think about where you want to store your project folder: this could be somewhere in your *Documents* folder.

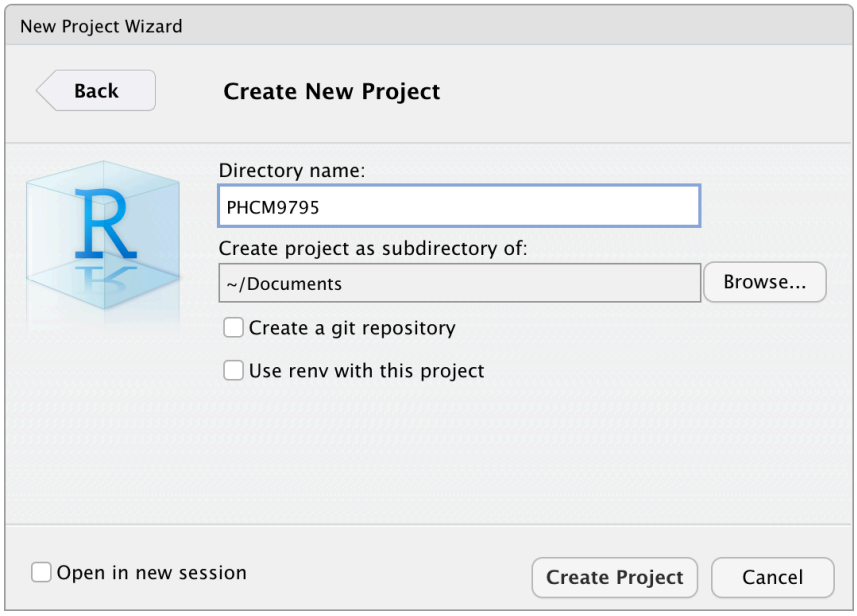
Step 1: Choose **File > New Project...** in RStudio to open the **Create Project** dialog box:



Step 2: Click the first option to create a project in a **New directory**

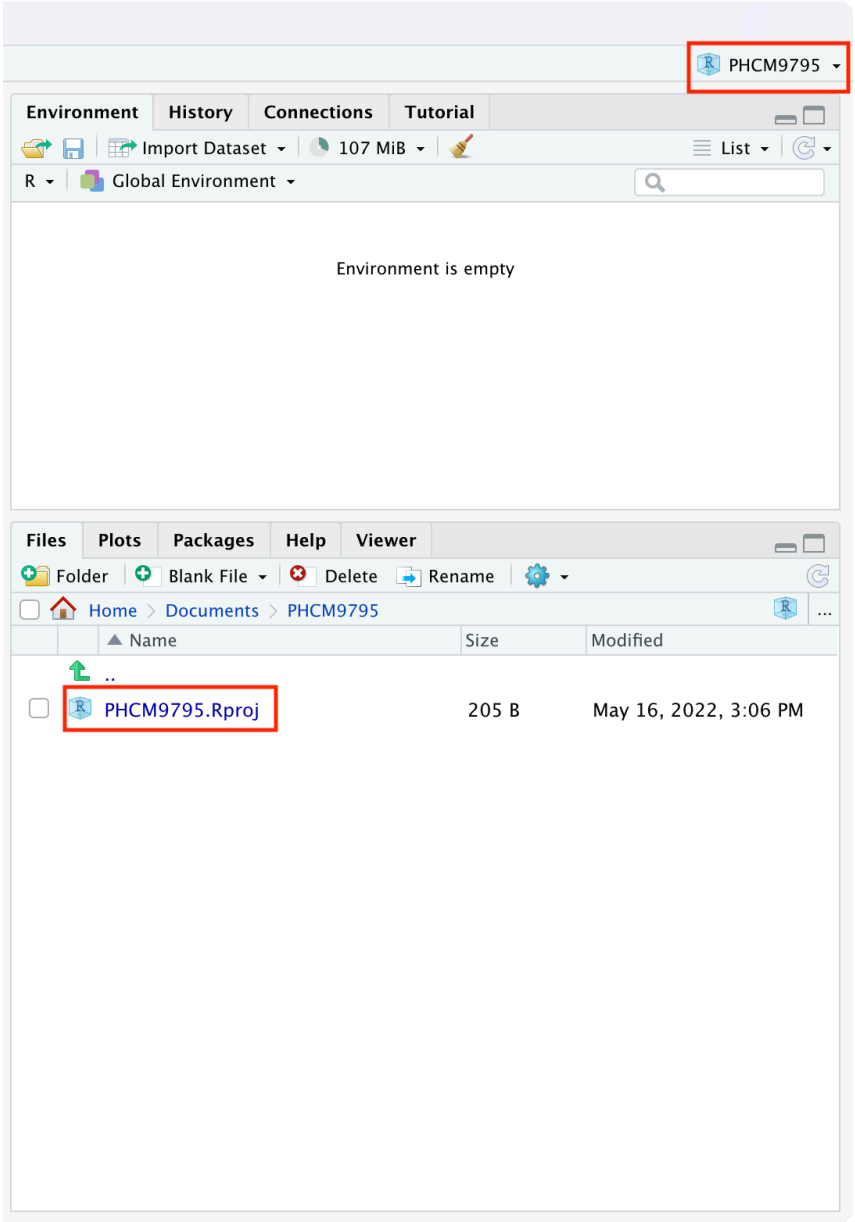


Step 3: Click the first option: **New Project**. Give the project a name, by typing PHCM9795 in the "Directory name", and choose where you want to store the project by clicking the **Browse** button.



Step 4: Click **Create** to create your project.

You will now have a new folder in your directory, which contains only one file: PHCM9795.Rproj, and the two right-hand panes of RStudio will appear as below:



TASK

Create a new project called PHCM9795.

The top-right menu bar is showing that you are working within the PHCM9795 project, and the bottom-right window is showing the contents of that window: the single PHCM9795.Rproj file. We will add some more files to this project later.

A simple R analysis

In this very brief section, we will introduce R by calculating the average of six ages.

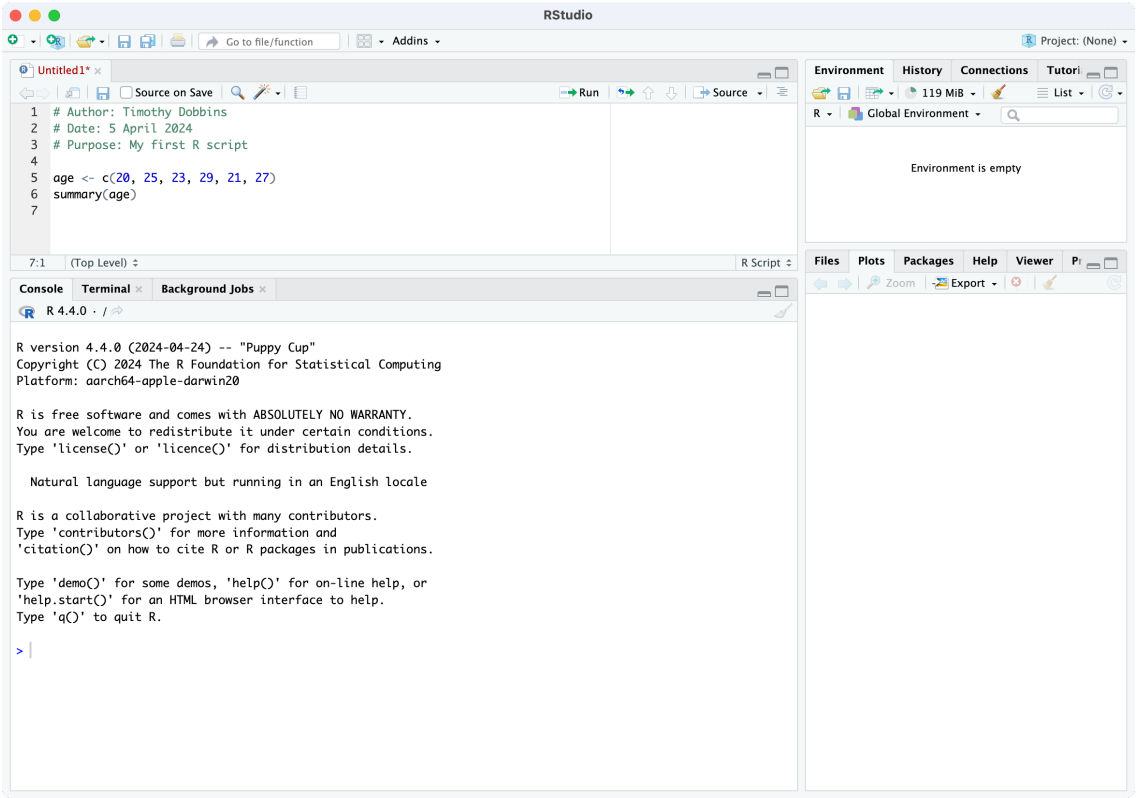
To begin, open a new R Script by choosing **File > New file > R Script** . A script (or a program) is a collection of commands that are sequentially processed by R. You can also type Ctrl+Shift+N in Windows, or Command+Shift+N in MacOS to open a new script in RStudio, or click the **New File** button at the top of the RStudio window.

You should now see four window panes, as below. In the top-left window, type the following (replacing my name with yours, and including today's date):

```
# Author: Timothy Dobbins
# Date: 5 April 2024
# Purpose: My first R script

age <- c(20, 25, 23, 29, 21, 27)
summary(age)
```

Note: R is case-sensitive, so you should enter the text exactly as written in these notes.

Your screen should look something like:



To run your script, choose **Code > Run Region > Run All**. You will see your code appear in the bottom-left window, with the following output:

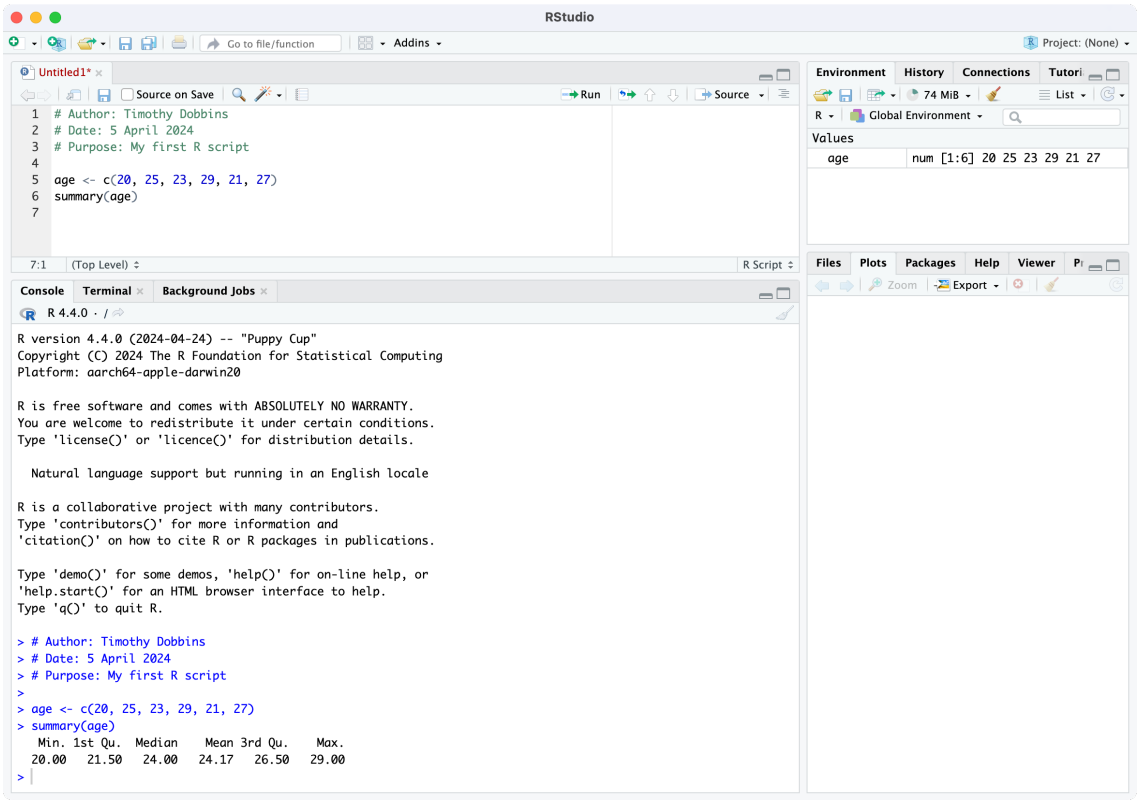
```
> # Author: Timothy Dobbins
> # Date: 5 April 2024
> # Purpose: My first R script
>
> age <- c(20, 25, 23, 29, 21, 27)

> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   21.50   24.00   24.17   26.50   29.00
```

We will explain the key parts of this script later, but for now, you have entered six ages and calculated the mean age (along with five other summary statistics).

> TASK
>
> Type the code above into the top-left window, and run the script.
> Save your script within the PHCM9795 project by using **File > Save As**, using the name `my_first_analysis.R`.

The RStudio environment

Now that we have seen a simple example of how to use R within RStudio, let's describe the RStudio environment. Let's assume that you have just run your first R script, and you have four windows as below:



The top-left window is call the **Source** window, and is where you write and edit your R scripts. Scripts can be saved by clicking **File > Save As** or by clicking on the symbol of a floppy disk at the top of the script. The file will have an extension of .R, for example script.R. Remember to give your script a meaningful title and remember to periodically save as you go.

In RStudio, the name of the script will be black when it has been saved, and will change to red if you have any unsaved changes.

The **Console** window, at the bottom left, contains the command line which is indicated with the symbol >. You can type commands here, but anything executed directly from the console is not saved and therefore is lost when the session ends (when you exit RStudio). You should always run your commands from a script file which you can save and use again later. When you run commands from a script, the output and any notes/errors are shown in the console. The Terminal and Jobs tabs will not be used in this course.

The **Environment** window at the top-right shows a list of objects that have been created during your session. When you close your RStudio session these objects will disappear. We will not use the History or Connections tabs in this course.

The bottom right corner contains some useful tabs, in particular the **Help** tab. When you are troubleshooting errors or learning how to use a function, the Help tab should be the first place you visit. Here you can search the help documents for all the packages you have installed. Whenever you create plots in R, these will be shown in the **Plots** tab. The **Packages** tab contains a list of installed packages and indicates which ones are currently in use (we will learn about packages later). Packages which are loaded, i.e. in use, are indicated with a tick. Some packages are in use by default when you begin a new session. You can access information about a package by clicking on its name. The **Files** tab provides a shortcut to access your files. The Viewer tab will not be used in this course.

Some R basics

While we use R as a statistics package, R is a programming language. In order to use R effectively, we need to define some basics.

Scripts

While R can be run completely from the command line, issuing commands one-by-one, it is most commonly run using **scripts**. A script is simply a list of commands that are processed in order. The simple analysis we conducted earlier is a very simple script. Some things to know about R scripts:

- anything appearing after a # is a comment, and is ignored by R. The first three lines of our script are there for ourselves (either as writers of code, or readers of code). I include comments at the beginning of each of my scripts to describe:
  - who wrote the script (useful if someone else uses your script and wants to ask questions about it);
  - when the script was written;
  - what the script does. This last point may seem odd, but it's useful to describe what this script does, and why it might differ to other scripts being used in the analysis. This is particularly useful if your scripts become long and complex.
- **R is case-sensitive**. So `age`, `AGE` and `Age` could refer to three separate variables (please don't do this!)
- use blank lines and comments to separate sections of your script

Objects

If you do some reading about R, you may learn that R is an "object-oriented programming language". When we enter or import data into R, we are asking R to create **objects** from our data. These objects can be manipulated and transformed by **functions**, to obtain useful insights from our data.

Objects in R are created using the **assignment operator**. The most common form of the assignment operator looks like an arrow: <- and is typed as the < and - symbols. The simplest way of reading <- is as the words "is defined as". Note that it possible to use -> and even = as assignment operators, but their use is less frequent.

Let's see an example:

```
x <- 42
```

This command creates a new object called `x`, which is defined as the number 42 (or in words, "`x` is defined as 42"). Running this command gives no output in the console, but the new object appears in the top-right **Environment** panel. We can view the object in the console by typing its name:

```
# Print the object x
x
```

```
[1] 42
```

Now we see the contents of `x` in the console.

This example is rather trivial, and we rarely assign objects of just one value. In fact, we created an object earlier, called `age`, which comprised six values.

Data structures

There are two main structures we will use to work with data in this course: **vectors** and **data frames**. A **vector** is a combination of data values, all of the same type. For example, our six ages that we entered earlier is a vector. You could think of a vector as a column of data (even though R prints vectors as rows!) And technically, even an object with only one value is a vector, a vector of size 1.

The easiest way of creating a vector in R is by using the `c()` function, where c stands for 'combine'. In our previous Simple Analysis in R (Section 1.12), we wrote the command:

```
age <- c(20, 25, 23, 29, 21, 27)
```

This command created a new object called `age`, and *combined* the six values of age into one vector.

Just as having a vector of size 1 is unusual, having just one column of data to analyse is also pretty unusual. The other structure we will describe here is a **data frame** which is essentially a collection of vectors, each of the same size. You could think of a data frame as being like a spreadsheet, with columns representing variables, and rows representing observations.

There are other structures in R, such as matrices and lists, which we won't discuss in this course. And you may come across the term **tibble**, which is a type of data frame.

Functions

If objects are the nouns of R, functions are the verbs. Essentially, functions transform objects. Functions can transform your data into summary statistics, graphical summaries or analysis results. For example, we used the `summary()` function to display summary statistics for our six ages.

R functions are specified by their arguments (or inputs). The arguments that can be supplied for each function can be inspected by examining the help notes for that function. To obtain help for a function, we can submit `help(summary)` (or equivalently `?summary`) in the console, or we can use the **Help** tab in the bottom-right window of RStudio. For example, the first part of the help notes for `summary` appear as:

summary {base} R Documentation

Object Summaries

Description

`summary` is a generic function used to produce result summaries of the results of various model fitting functions. The function invokes particular `methods` which depend on the `class` of the first argument.

Usage

```
summary(object, ...)

## Default S3 method:
summary(object, ..., digits, quantile.type = 7)
## S3 method for class 'data.frame'
summary(object, maxsum = 7,
        digits = max(3, getOption("digits")-3), ...)

## S3 method for class 'factor'
summary(object, maxsum = 100, ...)
```

The help notes in R can be quite cryptic, but the **Usage** section details what inputs should be specified for the function to run. Here, `summary` requires an object to be specified. In our case, we specified `age`, which is our object defined as the vector of six ages.

Most help pages also include some examples of how you might use the function. These can be found at the very bottom of the help page.

Examples

Run examples

```
summary(attenu, digits = 4) #-> summary.data.frame(...), default precision
summary(attenu $ station, maxsum = 20) #-> summary.factor(...)

lst <- unclass(attenu$station) > 20 # logical with NAs
## summary.default() for logicals -- different from *.factor:
summary(lst)
summary(as.factor(lst))
```

The `summary()` function is quite simple, in that it only requires one input, the object to be summarised. More complex functions might require a number of inputs. For example, the help notes for the `descriptives()` function in the `jmv` package show a large number of inputs can be specified. Instructions for installing the `jmv` package will be provided below, this help-screen is included for illustration only.

descriptives {jmv} R Documentation

Descriptives

Description

Descriptives are an assortment of summarising statistics, and visualizations which allow exploring the shape and distribution of data. It is good practice to explore your data with descriptives before proceeding to more formal tests.

Usage

```
descriptives(data, vars, splitBy = NULL, freq = FALSE,
  desc = "columns", hist = FALSE, dens = FALSE, bar = FALSE,
  barCounts = FALSE, box = FALSE, violin = FALSE, dot = FALSE,
  dotType = "jitter", boxMean = FALSE, boxLabelOutliers = TRUE,
  qq = FALSE, n = TRUE, missing = TRUE, mean = TRUE,
  median = TRUE, mode = FALSE, sum = FALSE, sd = TRUE,
  variance = FALSE, range = FALSE, min = TRUE, max = TRUE,
  se = FALSE, ci = FALSE, ciWidth = 95, iqr = FALSE,
  skew = FALSE, kurt = FALSE, sw = FALSE, pcEqGr = FALSE,
  pcNEqGr = 4, pc = FALSE, pcValues = "25,50,75", formula)
```

There are two things to note here. First, notice that the first two inputs are listed with no = symbol, but all other inputs are listed with = symbols (with values provided after the = symbol). This means that everything apart from `data` and `vars` have **default** values. We are free to not specify values for these inputs if we are happy with the defaults provided. For example, by default the variance is not calculated (as `variance = FALSE`). To obtain the variance as well as the standard deviation, we can change this default to `variance = TRUE`:

```
# Only the standard deviation is provided as the measure of variability
descriptives(data=pbc, vars=age)

# Additionally request the variance to be calculated
descriptives(data=pbc, vars=age, variance=TRUE)
```

Second, for functions with multiple inputs, we can specify the input name and its value, or we can ignore the input name and specify just the input values **in the order listed in the Usage section**. So the following are equivalent:

```
# We can specify that the dataset to be summarised is pbc,
#   and the variable to summarise is age:
descriptives(data=pbc, vars=age)

# We can omit the input name, as long as we keep the inputs in the correct order -
#   that is, dataset first, variable second:
descriptives(pbc, age)
```

```
# We can change the order of the inputs, as long as we specify the input name:
descriptives(vars=age, data=pbc)
```

In this course, we will usually provide all the input names, even when they are not required. As you become more familiar with R, you will start to use the shortcut method.

The curse of inconsistency

As R is an open-source project, many people have contributed to its development. This has led to a frustrating part of R: some functions require a single object to be specified, but some require you to specify a data frame and select variables for analysis. Let's see an example.

The help for `summary()` specifies the usage as: `summary(object, ...)`. This means we need to specify a single object to be summarised. An object could be a single column of data (i.e. a vector), or it could be a data frame. If we have a data frame called `pbc` which contains many variables, the command `summary(pbc)` would summarise every variable in the data frame.

What if we only wanted to summarise the age of the participants in the data frame? To select a single variable from a data frame, we can use the following syntax: `dataframe$variable`. So to summarise just age from this data frame, we would use: `summary(pbc$age)`.

Compare this with the `descriptives()` function in the `jmv` package. We saw earlier that the two required inputs for `descriptives()` are `data` (the data frame to be analysed) and `vars` (the variables to be analysed). So to summarise `age` from the `pbc` data frame, we would specify `descriptives(data=pbc, vars=age)`.

This inconsistency will seem maddening at first, and will continue to be maddening! Reading the **usage** section of the help pages is a useful way to determine whether you should specify an object (like `pbc$age`) or a data frame and a list of variables.

Packages

A **package** is a collection of functions, documentation (and sometimes datasets) that extend the capabilities of R. Packages have been written by R users to be freely distributed and used by others. R packages can be obtained from many sources, but the most common source is CRAN: the Comprehensive R Archive Network.

A useful way of thinking about R is that R is like a smartphone, with packages being like apps which are downloaded from CRAN (similar to an app-store). When you first install R, it comes with a basic set of packages (apps) installed. You can do a lot of things with these basic packages, but sometimes you might want to do things differently, or you may want to perform some analyses that can't be done using the default packages. In these cases, you can install a package.

Like installing an app on a smartphone, you only need to *install* a package once. But each time you want to use the package, you need to *load* the package into R.

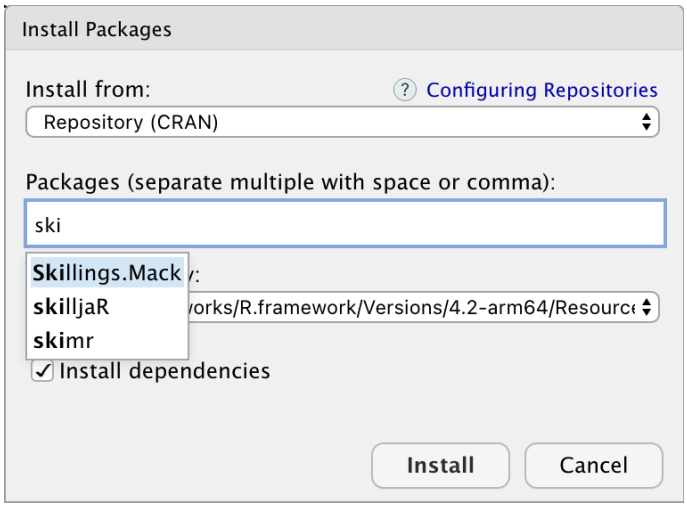
How to install a package

There are a couple of ways to install a package. You can use the `install.packages()` function if you know the exact name of the package. Let's use an example of installing the `skimr` package, which gives a very nice, high-level overview of any data frame. We can install `skimr` by typing the following into the console:

```
install.packages("skimr")
```

Note the use of the quotation marks.

Alternatively, RStudio offers a graphical way of installing packages that can be accessed via **Tools > Install Packages**, or via the **Install** button at the top of the **Packages** tab in the bottom-right window. You can begin typing the name of the package in the dialog box that appears, and RStudio will use predictive text to offer possible packages:



While writing code is usually the recommended way to use R, installing packages is an exception. Using **Tools > Install Packages** is perfectly fine, because you only need to install a package once.

How to load a package

When you begin a new session in RStudio, i.e. when you open RStudio, only certain core packages are automatically loaded. You can use the `library()` function to load a package that has previously been installed. For example, now that we have installed `skimr`, we need to load it before we can use it:

```
library(skimr)
```

Note that quotation marks are not required for the `library()` function (although they can be included if you really like quotation marks!).

> TASK
>
> Install the packages `jmv` and `skimr` using **Tools > Install packages**, or by typing into the console:
> ```
> install.packages("jmv")
> install.packages("skimr")
> ```

Installing vs loading packages

Package installation:

- use the `install.packages()` function (note the 's') or **Tools > Install packages**
- the package name must be surrounded by quotation marks
- only needs to be done once

Package loading

- use the `library()` function
- the package name does not need to be surrounded by quotation marks
- must be done for each R session

What is this thing called the tidyverse?

If you have done much reading about R, you may have come across the tidyverse:

> "The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures." https://www.tidyverse.org/

Packages in the tidyverse have been designed with a goal to make using R more consistent by defining a "grammar" to manipulate data, examine data and draw conclusions from data. While the tidyverse is a common and powerful set of packages, we will not be teaching the tidyverse in this course for two main reasons:

1. The data we provide have been saved in a relatively tidy way, and do not need much manipulation for analyses to be conducted. The cognitive load in learning the tidyverse in this course is greater than the benefit that could be gained.
2. There are many resources (online, in print etc) that are based on `base R`, and do not use the tidyverse. It would be difficult to understand these resources if we taught only tidyverse techniques. In particular, the `dataframe$variable` syntax is an important concept that should be understood before moving into the tidyverse.

In saying all of this, I think the tidyverse is an excellent set of packages, which I frequently use. At the completion of this course, you will be well equipped to teach yourself tidyverse using many excellent resources such as: Tidyverse Skills for Data Science and R for Data Science.

1.13 Part 2: Obtaining summary statistics for continuous data

In this exercise (spanning Modules 1 and 2), we will analyse data to complete a descriptive table from a research study. The data come from a study in primary biliary cirrhosis, a condition of the liver, from Modeling Survival Data: Extending the Cox Model Therneau and Grambsch (2010). By the end of this exercise, we will have completed the following table.

Table 1.2: Summary of 418 participants from the PBC study (Therneau and Grambsch, 2000)

Characteristic		Summary
Age (years)		Mean (SD) or Median [IQR]
Sex	Male	n (%)
	Female	n (%)
AST* (U/ml)		Mean (SD) or Median [IQR]
Serum bilirubin		Mean (SD) or Median [IQR]
Stage	I	n (%)
	II	n (%)
	III	n (%)
	IIIV	n (%)
Vital status at study end	Alive: no transplant	n (%)
	Alive: transplant	n (%)
	Deceased	n (%)

* asparate aminotransferase

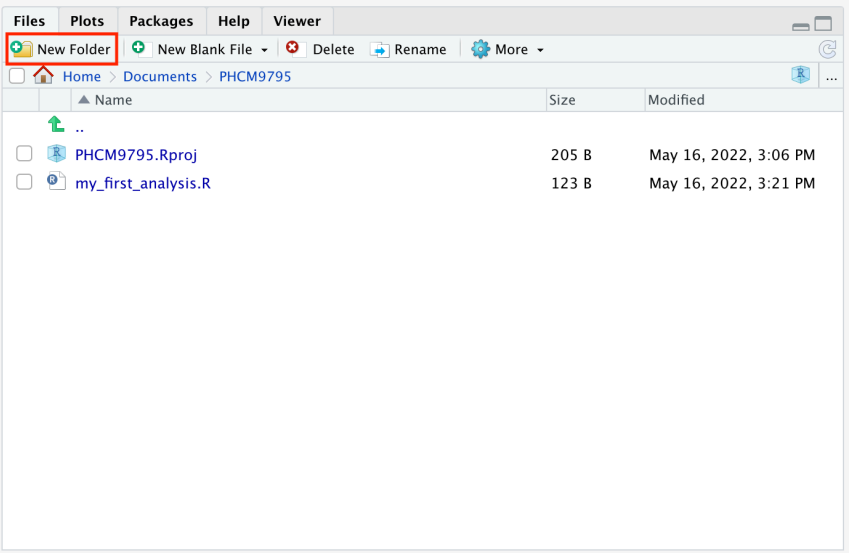
This table is available in Table1.docx, saved on Moodle.

> TASK
>
> Download the table shell, saved on Moodle as `PBC Table1.docx`, and the information file called `mod01_pbc_info.txt`.

Set up your data

We created a project in the previous step. We will now create a folder to store all the data for this course. Storing the data within the project makes life much easier!

Create a new folder by clicking the **New Folder** icon in the **Files** tab at the bottom-right:

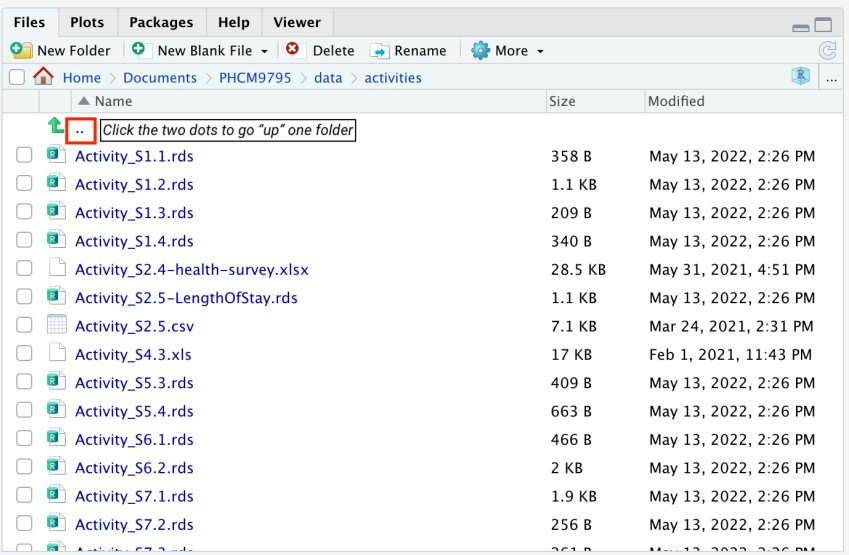


Call the new folder **data**.

Click on this folder to open it, and then create two new folders: **activities** and **examples**.

Download the "Data sets: for learning activities" from Moodle, and use Windows Explorer or MacOS Finder to save these data sets in **activities**. Save the "Data sets: example data from course notes" into the **examples** folder.

Your **activities** folder should look like:



Click the two dots next to the up-arrow at the top of the folder contents to move back up the folder structure. Note that you need to click the dots, and not the up-facing green arrow!

Reading a data file

Typing data directly into R is not common; we usually read data that have been previously saved. In this example, we will read an `.rds` file using the `readRDS()` function, which has only one input: the location of the file.

TASK

1 - Confirm that the `mod01_pdc.rds` file is in the `activities` sub-folder within the `data` folder (as per the previous steps).
2 - Load the `skimr` package, and use the `readRDS()` function to read the file into R, assigning it to a data frame called `pbc`. Because we set up our project, we can locate our data easily by telling R to use the file: `"data/activities/mod01_pdc.rds"`, which translates as: the file `mod01_pdc.rds` which is located in the `activities` sub-folder within the `data` folder.

```
library(skimr)

pbc <- readRDS("data/activities/mod01_pbc.rds")
```

3 - We can now use the `summary()` function to examine the pbc dataset:

```
summary(pbc)
```

```
       id              time          status            trt
 Min.   :  1.0   Min.   :  41   Min.   :0.0000   Min.   :1.000
 1st Qu.:105.2   1st Qu.:1093   1st Qu.:0.0000   1st Qu.:1.000
 Median :209.5   Median :1730   Median :0.0000   Median :1.000
 Mean   :209.5   Mean   :1918   Mean   :0.8301   Mean   :1.494
 3rd Qu.:313.8   3rd Qu.:2614   3rd Qu.:2.0000   3rd Qu.:2.000
 Max.   :418.0   Max.   :4795   Max.   :2.0000   Max.   :2.000
                                                 NA's   :106
      age             sex           ascites           hepato
 Min.   :26.28   Min.   :1.000   Min.   :0.00000   Min.   :0.0000
 1st Qu.:42.83   1st Qu.:2.000   1st Qu.:0.00000   1st Qu.:0.0000
 Median :51.00   Median :2.000   Median :0.00000   Median :1.0000
 Mean   :50.74   Mean   :1.895   Mean   :0.07692   Mean   :0.5128
 3rd Qu.:58.24   3rd Qu.:2.000   3rd Qu.:0.00000   3rd Qu.:1.0000
 Max.   :78.44   Max.   :2.000   Max.   :1.00000   Max.   :1.0000
                                 NA's   :106       NA's   :106
    spiders           edema             bili             chol
 Min.   :0.0000   Min.   :0.0000   Min.   : 0.300   Min.   : 120.0
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.800   1st Qu.: 249.5
 Median :0.0000   Median :0.0000   Median : 1.400   Median : 309.5
 Mean   :0.2885   Mean   :0.1005   Mean   : 3.221   Mean   : 369.5
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 3.400   3rd Qu.: 400.0
 Max.   :1.0000   Max.   :1.0000   Max.   :28.000   Max.   :1775.0
 NA's   :106                                        NA's   :134
    albumin          copper          alkphos            ast
 Min.   :1.960   Min.   :  4.00   Min.   :  289.0   Min.   : 26.35
 1st Qu.:3.243   1st Qu.: 41.25   1st Qu.:  871.5   1st Qu.: 80.60
 Median :3.530   Median : 73.00   Median : 1259.0   Median :114.70
 Mean   :3.497   Mean   : 97.65   Mean   : 1982.7   Mean   :122.56
 3rd Qu.:3.770   3rd Qu.:123.00   3rd Qu.: 1980.0   3rd Qu.:151.90
 Max.   :4.640   Max.   :588.00   Max.   :13862.4   Max.   :457.25
                 NA's   :108      NA's   :106       NA's   :106
     trig           platelet        protime          stage
 Min.   : 33.00   Min.   : 62.0   Min.   : 9.00   Min.   :1.000
 1st Qu.: 84.25   1st Qu.:188.5   1st Qu.:10.00   1st Qu.:2.000
```

```
 Median :108.00   Median :251.0   Median :10.60   Median :3.000
 Mean   :124.70   Mean   :257.0   Mean   :10.73   Mean   :3.024
 3rd Qu.:151.00   3rd Qu.:318.0   3rd Qu.:11.10   3rd Qu.:4.000
 Max.   :598.00   Max.   :721.0   Max.   :18.00   Max.   :4.000
 NA's   :136      NA's   :11      NA's   :2       NA's   :6
```

An alternative to the `summary()` function is the `skim()` function in the `skimr` package, which produces summary statistics as well as rudimentary histograms:

```
skim(pbc)
```

```
── Data Summary ────────────────────────
                           Values
Name                       pbc
Number of rows             418
Number of columns          20
_________________________
Column type frequency:
  numeric                  20
_________________________
Group variables            None

── Variable type: numeric ──────────────────────────────────────────────────────────────────────────────────────────────
   skim_variable n_missing complete_rate      mean        sd     p0     p25     p50     p75     p100 hist
 1 id                    0         1      210.       121.       1     105.    210.    314.     418
 2 time                  0         1     1918.      1105.      41    1093.   1730    2614.    4795
 3 status                0         1        0.830     0.956   0       0       0       2        2
 4 trt                 106         0.746    1.49      0.501   1       1       1       2        2
 5 age                   0         1       50.7      10.4    26.3    42.8    51.0    58.2     78.4
 6 sex                   0         1        1.89      0.307   1       2       2       2        2
 7 ascites             106         0.746    0.0769    0.267   0       0       0       0        1
 8 hepato              106         0.746    0.513     0.501   0       0       1       1        1
 9 spiders             106         0.746    0.288     0.454   0       0       0       1        1
10 edema                 0         1        0.100     0.253   0       0       0       0        1
11 bili                  0         1        3.22      4.41    0.3     0.8     1.4     3.4     28
12 chol                134         0.679  370.      232.    120     250.    310.    400     1775
13 albumin               0         1        3.50      0.425   1.96    3.24    3.53    3.77     4.64
14 copper              108         0.742   97.6      85.6     4      41.2    73      123      588
15 alkphos             106         0.746 1983.     2140.    289     872.   1259    1980    13862.
16 ast                 106         0.746  123.      56.7    26.4    80.6   115.     152.     457.
17 trig                136         0.675  125.      65.1    33      84.2   108      151      598
18 platelet             11         0.974  257.      98.3    62     188.    251      318      721
19 protime               2         0.995   10.7      1.02    9      10      10.6    11.1     18
20 stage                 6         0.986    3.02     0.882   1       2       3       4        4
```

The `summary()` and `skim()` functions are useful to give a quick overview of a dataset: how many variables are included, how variables are coded, which variables contain missing data and a crude histogram showing the distribution of numeric variables.

Summarising continuous variables

One of the most flexible functions for summarising continuous variables is the `descriptives()` function from the `jmv` package. The function is specified as `descriptives(data=, vars=)` where:

- `data` specifies the dataframe to be analysed
- `vars` specifies the variable(s) of interest, with multiple variables combined using the `c()` function

We can summarise the three continuous variables in the pbc data: age, AST and serum bilirubin, as shown below.

```
library(jmv)

descriptives(data=pbc, vars=c(age, ast, bili))
```

```

 DESCRIPTIVES
```

```
 Descriptives

                         age         ast         bili

   N                          418         312          418
   Missing                      0         106            0
   Mean                  50.74155    122.5563     3.220813
   Median                51.00068    114.7000     1.400000
   Standard deviation    10.44721    56.69952     4.407506
   Minimum               26.27789    26.35000    0.3000000
   Maximum               78.43943    457.2500     28.00000
```

By default, the `descriptives` function presents the mean, median, standard deviation, minimum and maximum. We can request additional statistics, such as the quartiles (which are called the percentiles, or pc, in the descriptives function):

```
descriptives(data=pbc, vars=c(age, ast, bili), pc=TRUE)
```

```
 DESCRIPTIVES

 Descriptives

                         age         ast         bili

   N                          418         312          418
   Missing                      0         106            0
   Mean                  50.74155    122.5563     3.220813
   Median                51.00068    114.7000     1.400000
   Standard deviation    10.44721    56.69952     4.407506
   Minimum               26.27789    26.35000    0.3000000
   Maximum               78.43943    457.2500     28.00000
   25th percentile       42.83231    80.60000    0.8000000
   50th percentile       51.00068    114.7000     1.400000
   75th percentile       58.24093    151.9000     3.400000
```

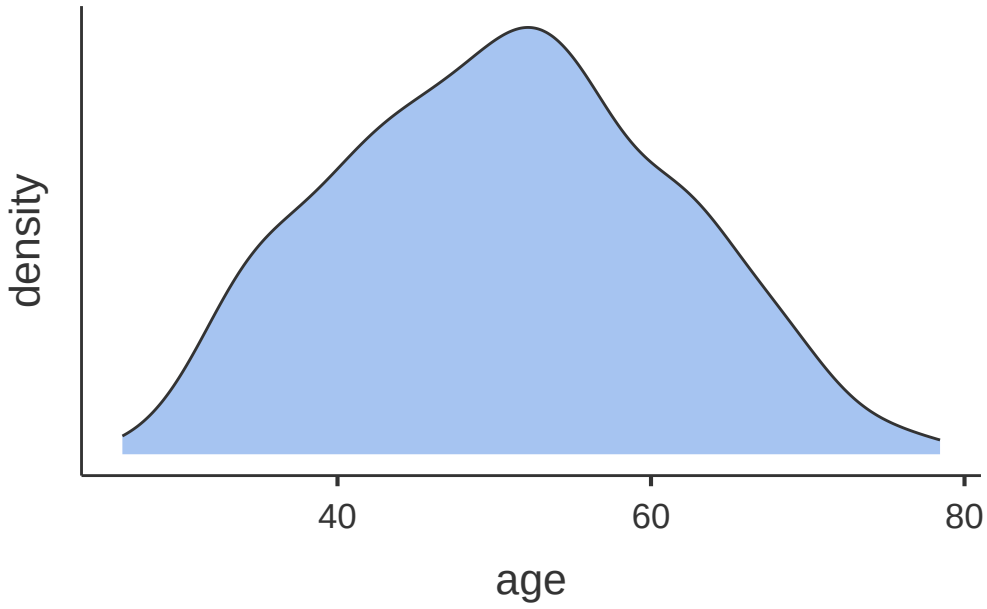
Producing a density plot

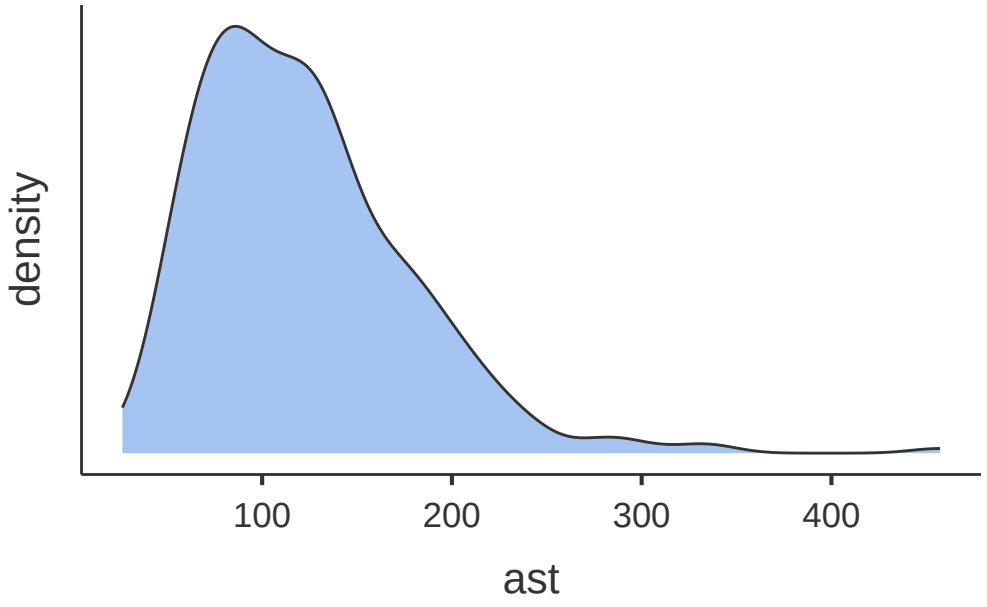
We can add `dens=TRUE` to the `descriptives` function to produce a density plot for each listed variable:

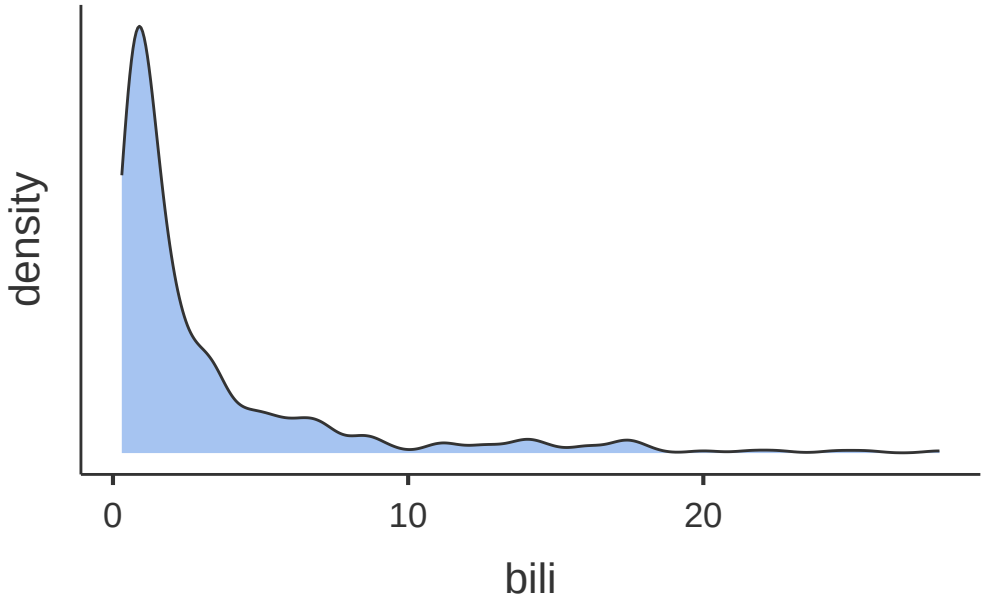
```
descriptives(data=pbc, vars=c(age, ast, bili), pc=TRUE, dens=TRUE)
```

```
 DESCRIPTIVES

 Descriptives

                         age         ast         bili

   N                          418         312          418
   Missing                      0         106            0
   Mean                  50.74155    122.5563     3.220813
   Median                51.00068    114.7000     1.400000
```

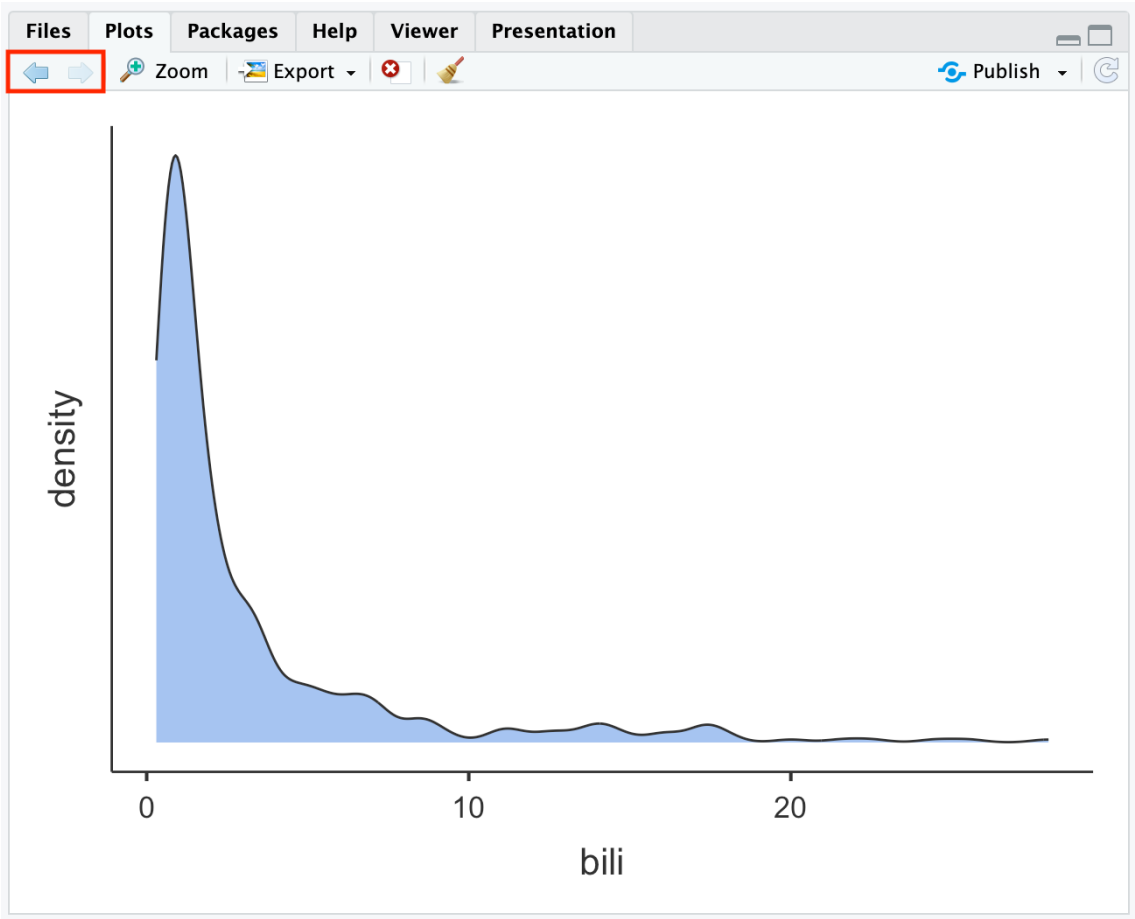
```
Standard deviation     10.44721     56.69952      4.407506
Minimum                26.27789     26.35000     0.3000000
Maximum                78.43943     457.2500      28.00000
25th percentile        42.83231     80.60000     0.8000000
50th percentile        51.00068     114.7000      1.400000
75th percentile        58.24093     151.9000      3.400000
```





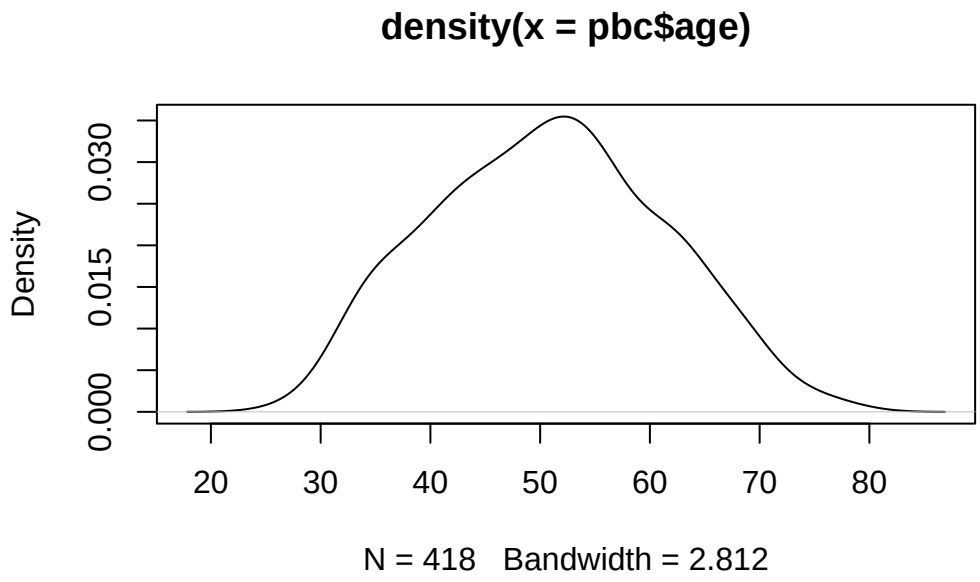


Note that the density plots are plotted separately in the **Plot** window. They can be viewed using the arrows at the top of the **Plot** window:



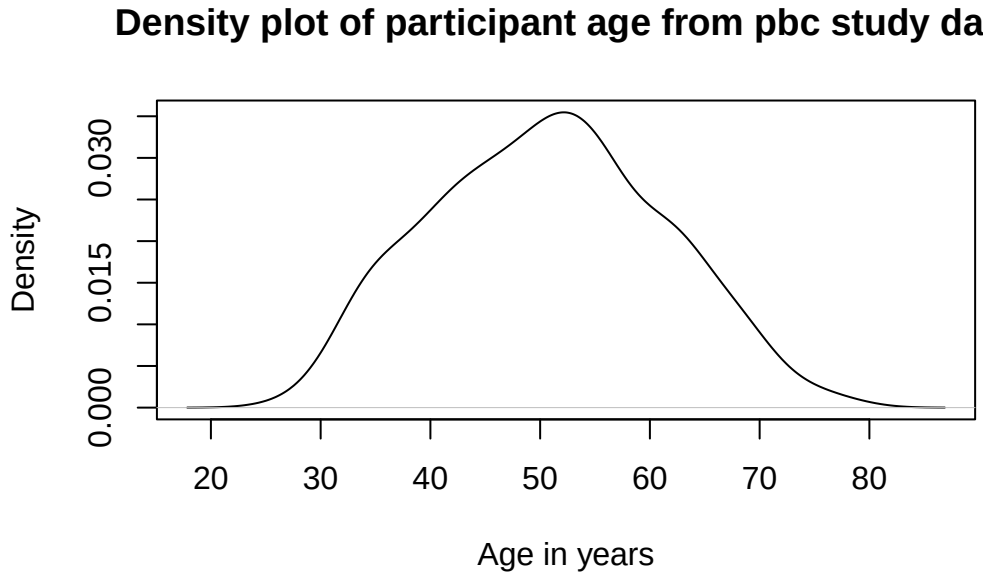
A more flexible way of constructing a density plot is by using the plot() function within R, using the syntax: `plot(density(dataframe$variable))`, which plots the `variable` from the `dataframe`. For example, the default density plot for the `age` column of the `pbc` data:

```
plot(density(pbc$age))
```



This plot can be improved by using `xlab=" "` and `main=" "` to assign labels for the x-axis and overall title respectively:

```
plot(density(pbc$age),
     xlab="Age in years",
     main="Density plot of participant age from pbc study data")
```



Note

The `density()` function requires the analysis variable to contain no missing values, and will give an error if there are any missing values. We can use the option `na.rm=TRUE` to request

that the density function ignore any missing values. For example:

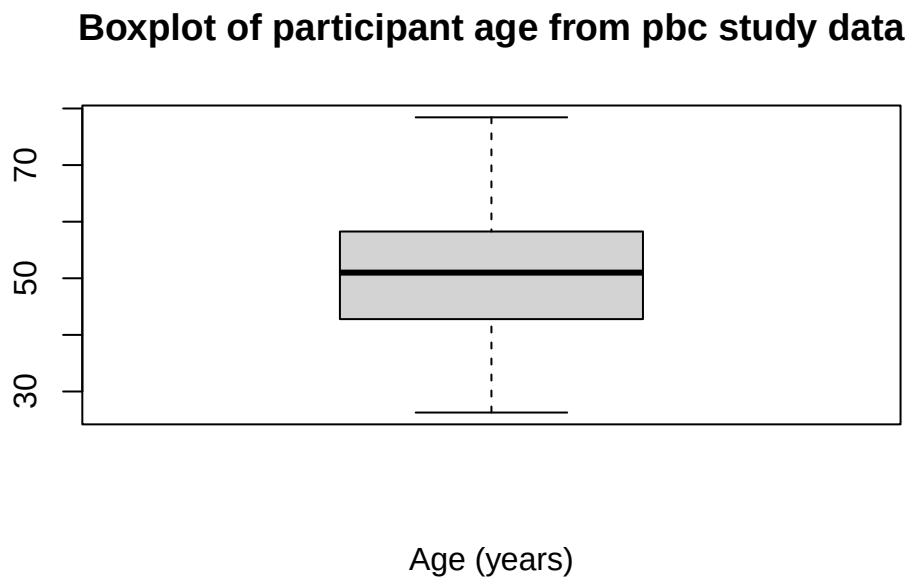
```
plot(density(pbc$ast, na.rm=TRUE),
     xlab="AST (U/mL)",
     main="Density plot of aspartate aminotransferase from pbc study data")
```

Producing a boxplot

Like the density plot, boxplots can be requested in the `descriptives` function by using `box=TRUE`.

The `boxplot` function is an alternative, more flexible function, again specifying the dataframe to use and the variable to be plotted as `dataframe$variable`. Labels can be applied in the same way as the histogram:

```
boxplot(pbc$age, xlab="Age (years)",
   main="Boxplot of participant age from pbc study data")
```



TASK

Obtain density plots and boxplots for age, AST and bilirubin.
Based on these plots, decide whether the mean or the median is the appropriate summary to use for each variable.

Saving data in R

There are many ways to save data from R, depending on the type of file you want to save. The recommendation for this course is to save your data using the `.rds` format, using the `saveRDS()` function, which takes two inputs: `saveRDS(object, file)`. Here, `object` is the R object to be saved (usually a data frame), and `file` is the location for the file to be saved (file name and path, including the `.rds` suffix).

It is not necessary to save our PBC data, as we have made only minor changes to the data that can be replicated by rerunning our script. If you had made major changes and wanted to save your data, you could use:

```
saveRDS(pbc, file="pbc_revised.rds")
```

Copying output from R

It is important to note that saving your data or your script in R will not save your output. The easiest way to retain the output of your analyses is to copy the output from the Console into a word processor package (e.g. Microsoft Word) before closing R.

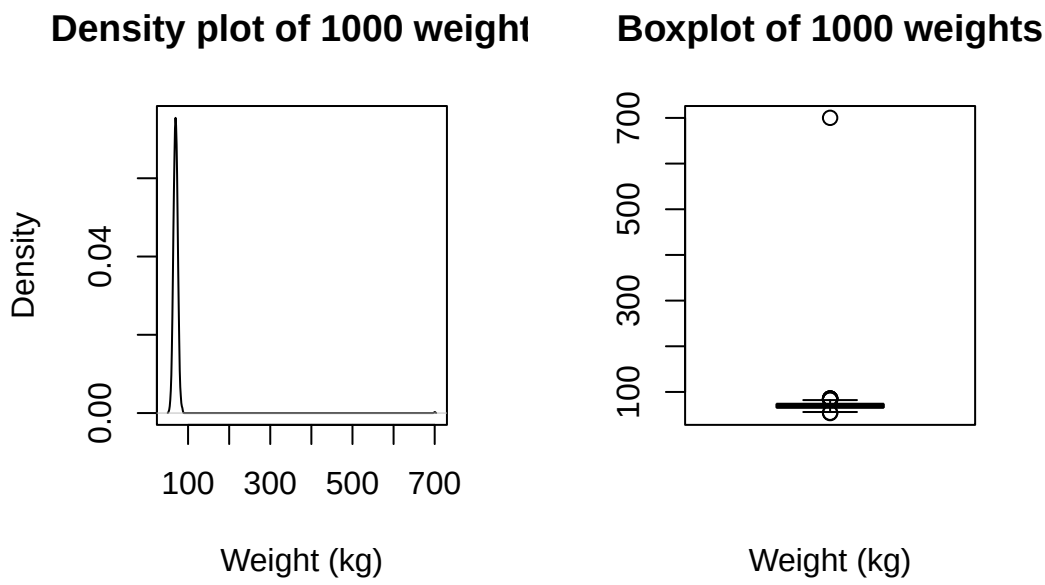
Unfortunately, by default, R is not ideal for creating publication quality tables. There are many packages that will help in this process, such as R Markdown, huxtable, gt and gtsummary, but their use is beyond the scope of this course. R Markdown for Scientists provides an excellent introduction to R Markdown.

> TASK
>
> Complete Table 1 for continuous variables using the output generated in this exercise. You should decide on whether to present continuous variables by their means or medians, and present the most appropriate measure of spread. Include footnotes to indicate if any variables contain missing observations.

1.14 Setting a value to missing

As we saw in Section 1.5, it is important to explore our data to identify any unusual observations. If an error is found, the best method for correcting the error is to go back to the original data e.g. the hard copy questionnaire, to obtain the original value, entering the correct value into R. If the original data is not available or the original data is also incorrect, the erroneous value is often excluded from the dataset.

Consider a sample dataset: `mod01_weight_1000.rds`, which contains the weights of 1000 people. A density plot and a boxplot should be examined before we start analysing these data:



There is a clear outlying point shown in the boxplot. Although not obvious, the same point is shown in the density plot as a small blip around 700kg. Obviously this point is unusual, and we should investigate.

We can view any outlying observations in the dataset using the `subset` function. You will need to decide if these values are a data entry error or are biologically plausible. If an extreme value or "outlier", is biologically plausible, it should be included in all analyses.

For example, to list any observations from the `sample` dataset with a weight larger than 200:

```
subset(sample, weight>200)
```

id	weight
58	700

We see that there is a very high value of 700.2kg. A value as high as 700kg is likely to be a data entry error (e.g. error in entering an extra zero) and is not a plausible weight value. Here, **you should check your original data**.

If you do not have access to the original data, it would be safest to set this value as missing. You do change this in R by using an `ifelse` statement to recode the incorrect weight of 700.2kg into a missing value. **A missing value in R is represented by NA**.

The form of the `ifelse` statement is as follows: `ifelse(test, value_if_true, value_if_false)`

We will write code to:

- create a new column (called `weight_clean`) in the `sample` dataframe (i.e. `sample$weight_clean`)
- test whether `weight` is equal to 700.2
  - if this is true, we will assign `weight_clean` to be NA
  - otherwise `weight_clean` will equal the value of `weight`

Putting it all together:

```
sample$weight_clean = ifelse(sample$weight==700.2, NA, sample$weight)
```

Note: if an extreme value lies within the range of biological possibility it should not be set to missing.

The same syntax could be used to replace the incorrect value with the correct value. For example, if you do source the original medical records, you might find that the original weight was recorded in medical records as 70.2kg. We could use the same syntax to replace 700.2 by 70.2: `sample$weight_clean = ifelse(sample$weight==700.2, NA, sample$weight)`

Once you have checked your data for errors, you are ready to start analysing your data.

What on earth: == ?

In R, the test of equality is denoted by two equal signs: ==. So we would use == to test whether an observation is equal to a certain value. Let's see an example:

```
# Test whether 6 is equal to 6
6 == 6
```

```
[1] TRUE
```

```
# Test whether 6 is equal to 42
6 == 42
```

```
[1] FALSE
```

You can read the == as "is equal to". So the code `sample$weight == 700.2` is read as: "is the value of weight from the data frame sample equal to 700.2?". In our `ifelse` statement above, if this condition is true, we replace `weight` by 70.2; if it is false, we leave weight as is.

Activities

Activity 1.1

25 participants were enrolled in a 3-week weight loss program. The following data present the weight loss (in grams) of the participants.

Table 1.3: Weight loss (g) for 25 participants

255	198	283	312	283
57	85	312	142	113
227	283	255	340	142
113	312	227	85	170
255	198	113	227	255

a) These data have been saved as `Activity_1.1.rds`. Read the data into your software package.
b) What type of data are these?
c) Construct an appropriate graph to display the distribution of participants' weight loss. Provide appropriate labels for the axes and give the graph an appropriate title.

Activity 1.2

Which of the following statements are true? The more dispersed, or spread out, a set of observations are:

a) The smaller the mean value
b) The larger the standard deviation
c) The smaller the variance

Activity 1.3

Estimate the mean, median, standard deviation, range and interquartile range for the data `Activity_1.3.rds`, available on Moodle.

Activity 1.4

Data of diastolic blood pressure (BP) of a sample of study participants are provided in the datasets `Activity_1.4.rds`. Compute the mean, median, range and SD of diastolic BP.

Activity 1.5

The ages of 100 study participants have been saved as `Activity_1.5.rds`. Estimate the:

a. mean and median;
b. standard deviation and interquartile range;
c. range.

Plot the data using a histogram and boxpolot. Is there anything unusual about the ages? What do you think is a possible explanation for this?

A clean version of the data have been saved as `Activity_1.5_clean.rds`. Recalculate the summary statistics and recreate the plots using the clean data.

Based on this exercise, what is your advice on coding unusual or missing values in data?

Module 2

Categorical data, presentation guidelines and probability distributions

Learning objectives

By the end of this module you will be able to:

- Present and report categorical data numerically and graphically
- Describe the concept of probability
- Describe the characteristics of a binomial distribution
- Compute probabilities from a binomial distribution using statistical software

Optional readings

Kirkwood and Sterne (2001); Chapters 3, 14 and 15. [UNSW Library Link]

Bland (2015); Chapters 5 and 6. [UNSW Library Link]

Graphics and statistics for cardiology: designing effective tables for presentation and publication, Boers (2018, UNSW Library Link)

Guidelines for Reporting of Figures and Tables for Clinical Research in Urology, Vickers et al. (2020, UNSW Library Link)

2.1 Introduction

In Module 1, we saw how to summarise continuous data numerically and graphically. In this module, we will discuss summarising categorical data numerically and graphically. We will also introduce the concept of probability which underpins the theoretical basis of statistics, and then introduce the concept of probability distributions. We will present the binomial distribution, which calculates the probability of observing a certain number of events from multiple observations.

2.2 Summarising a single categorical variable numerically

Categorical data are best summarised using a frequency table, where each category is summarised by its frequency: the count of the number of individuals in each category. The **relative frequency** (the frequency expressed as a proportion or percentage of the total frequency) is usually included give further insight.

Table 2.1: Sex of participants in PBC study

Sex	Frequency	Relative frequency (%)
Male	44	10.5
Female	374	89.5

It is sometimes useful to present the cumulative relative frequency, which shows the relative frequency of individuals in a certain category or below (for example, Table 2.2).

Table 2.2: Stage of disease for participants in PBC study

Stage *	Frequency	Relative frequency (%)	Cumulative relative frequency (%)
1	21	5.1	5.1
2	92	22.3	27.4
3	155	37.6	65.0
4	144	35.0	100.0

* Disease stage was missing for 6 participants

From Table 2.2, we can see that 65.0% of participants had Stage 3 disease or lower.

2.3 Summarising a single categorical variable graphically

A categorical variable is best summarised graphically using a **bar chart**. For example, we can present the distribution of Stage of Disease graphically using a bar graph (Figure 2.1). Bar graphs, which are suitable for plotting discrete or categorical variables, are defined by the fact that the bars do not touch.

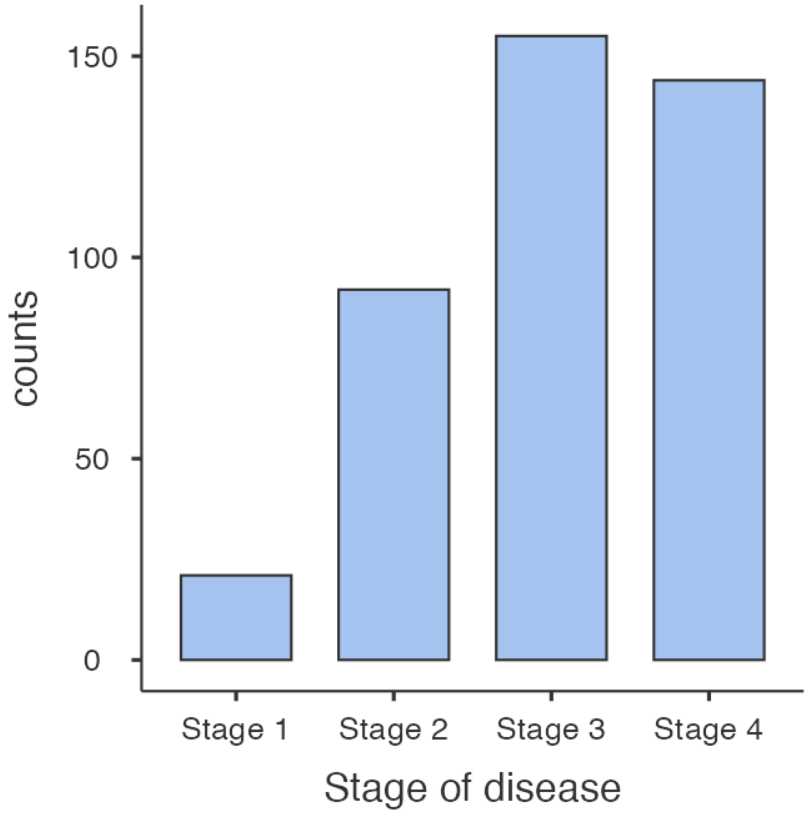


Figure 2.1: Bar graph of stage of disease from PBC study

Pie charts can be an alternative way to summarise a categorical variable graphically, however their use is not recommended for the following reasons:

- Not ideal when there are many categories to compare
- The use of percentages is not appropriate when the sample size is small
- Can be misleading by using different size pies, different rotations and different colours to draw attention to specific groups
- 3D and exploding bar charts further distort the effect of perspective and may confuse the reader

Pie charts will not be discussed further in this course.

2.4 Exploratory data analysis for categorical data

As with continuous data, it is good practice to undertake exploratory data analysis before formally analysing categorical data. You should take a moment and examine a frequency table for categorical variables, to ensure all recorded values are within scope.

2.5 Summarising two categorical variables numerically

So far, we have discussed one-way frequency tables, that is, tables that summarise one variable. We can summarise more than two categorical variables in a table – called a cross tabulation, or a two-way (summarising two variables) table.

Using our PBC data, we can summarise the two categorical variables: sex and stage of disease. The two-way table of frequencies is shown in Table 2.3.

Table 2.3: Frequency of participants by sex and stage of disease*

	Stage				
	1	2	3	4	Total
Sex					
Male	3	8	16	17	44
Female	18	84	139	127	368
Total	21	92	155	144	412

* Disease stage was missing for 6 participants

We can add percentages to two-way tables as either *column* or *row* percents. Using Table 2.3 as an example, column percents represent the relative frequencies of sex within each stage (Table 2.4).

Table 2.4: Frequency of participants by sex and stage of disease*, including column percents

	Stage				
	1	2	3	4	Total
Sex					
Male	3 (14%)	8 (9%)	16 (10%)	17 (12%)	44 (11%)
Female	18 (86%)	84 (91%)	139 (90%)	127 (88%)	368 (89%)
Total	21 (100%)	92 (100%)	155 (100%)	144 (100%)	412 (100%)

* Disease stage was missing for 6 participants

Conversely, row percents represent the relative frequencies of stage within each sex (Table 2.5).

Table 2.5: Frequency of participants by sex and stage of disease, including row percents

	Stage				
	1	2	3	4	Total
Sex					
Male	3 (7%)	8 (18%)	16 (36%)	17 (39%)	44 (100%)
Female	18 (5%)	84 (23%)	139 (38%)	127 (35%)	368 (100%)
Total	21 (5%)	92 (22%)	155 (38%)	144 (35%)	412 (100%)

* Disease stage was missing for 6 participants

Tables containing more than two variables

It is possible to construct multi-way tables that summarise more than two categorical variables in a single table. However, tables can become complex when more than two variables are incorporated, and you may need to present the information as two tables or incorporate additional rows and columns.

In Figure 2.2, characteristics of the sample of prisoners from the NPHDC were presented. This table contains information about sex, age group and Indigenous status from different groups of prisoners; prison entrants, discharges, and prisoners in custody. This type of condensed information is often found in reports and journal articles giving demographic information, by different groups considered in the study.

Table 2.4: Prison entrants (2015), dischargees (2015) and prisoners in custody (2014), by sex, age group and Indigenous status, 2014 and 2015 (per cent)

	Prison entrants[(a)]	Prison dischargees[(a)]	Prisoners in custody[(b)]
Sex			
Male	92	84	92
Female	8	16	8
Age group (years)			
18–24	19	15	18
25–34	42	37	36
35–44	27	30	27
45+	12	17	20
Indigenous status			
Indigenous	24	30	27
Non-Indigenous	75	67	72
Total	**100**	**100**	**100**

(a) Percentage of prison entrants/dischargees (see Note 3) sourced from the 2015 NPHDC.

(b) Percentage of prisoners in custody sourced from ABS 2014e.

Notes

1. Excludes New South Wales which did not provide dischargee data.
2. Percentages may not add exactly to 100, due to unknown demographic information, prisoners in custody aged under 18 and rounding.
3. Prison entrant and prison dischargee data should not be directly compared because they do not relate to the same individuals. See Section 1.4 for details.
4. Totals include 6 entrants and 1 dischargee who identified as transgender, 5 entrants and 4 dischargees of unknown age, and 5 entrants and 14 dischargees of unknown Indigenous status.
5. The proportions for sex and Indigenous status for prison entrants exclude New South Wales because the Inmate Health Survey, from which NSW entrants data are taken, over-sampled females and Indigenous prisoners.

Figure 2.2

We might also consider a table containing further pieces of information. The table presented in Figure 2.3 (from the health of Australia's prisoners 2015 report) compares prison entrants and the general community by three variables: age group, Indigenous status, and highest level of completed education.

Can you see any issues with the presentation of this table?

Table 3.3: Prison entrants and general community, highest level of completed education, 2015 (per cent)

		General community			Prison entrants		
Highest level of educational attainment	Indigenous status	20–24	25–34	35–44	20–24	25–34	35–44
Certificate III or IV	Indigenous	22	26	24	11	7	9
	Non-Indigenous	22	21	20	25	28	26
Year 12 or equivalent	Indigenous	26	14	10	4	2	2
	Non-Indigenous	36	15	13	6	8	11
Year 11 or equivalent	Indigenous	12	11	7	6	3	1
	Non-Indigenous	5	3	4	3	9	10
Year 10 or equivalent	Indigenous	22	20	19	19	10	8
	Non-Indigenous	8	6	11	19	23	25
Below Year 10	Indigenous	13	17	19	19	21	13
	Non-Indigenous	1	2	4	25	24	25

Sources: Entrant form, 2015 NPHDC; ABS 2014b.

Figure 2.3: Highest level of completed education in prison entrants and the general community

Source: Australian Institute of Health and Welfare 2015. The health of Australia's prisoners 2015. Cat. no. PHE 207. Canberra: AIHW.

Some issues in this table:

- The title of the table does not contain full information about the variables in the table;
- It is unclear how the percentages were calculated (which groupings added to 100%);
- The ages are not labelled as such, thus without reading the text in report it is unclear that these are age groupings.

2.6 Summarising two categorical variables graphically

Information from more than one variable can be presented as clustered or multiple bar chart (bars side-by-side) (Figure 2.4). This type of graph is useful when examining changes in the categories separately, but also comparing the grouping variable between the main bar variable. Here we can see that Stage 3 and Stage 4 disease is the most common for both males and females, but there are many more females within each stage of disease.

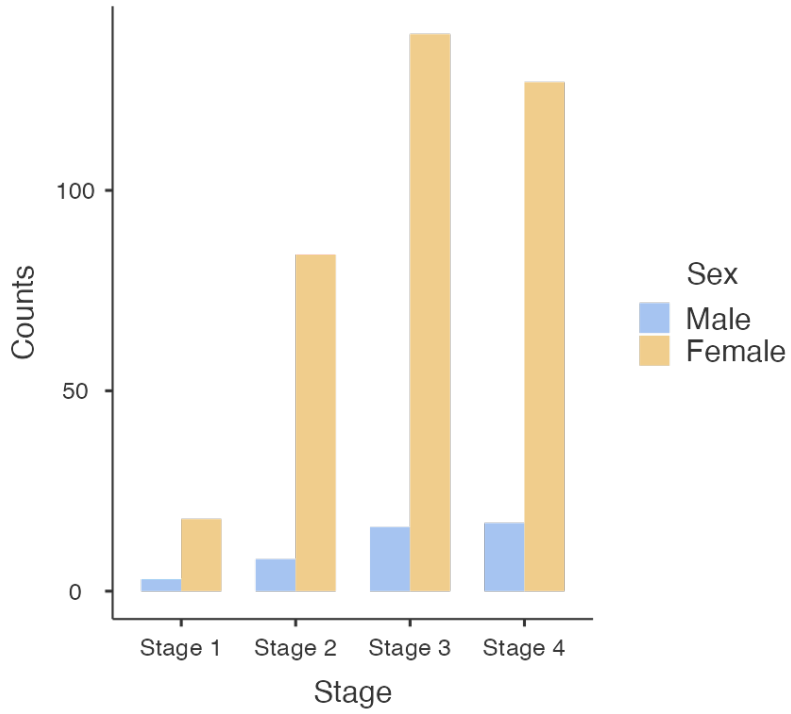


Figure 2.4: Bar graph of stage of disease by sex from PBC study

An alternative bar graph is a stacked or composite bar graph, which retains the overall height for each category, but differentiates the bars by another variable (Figure 2.5).

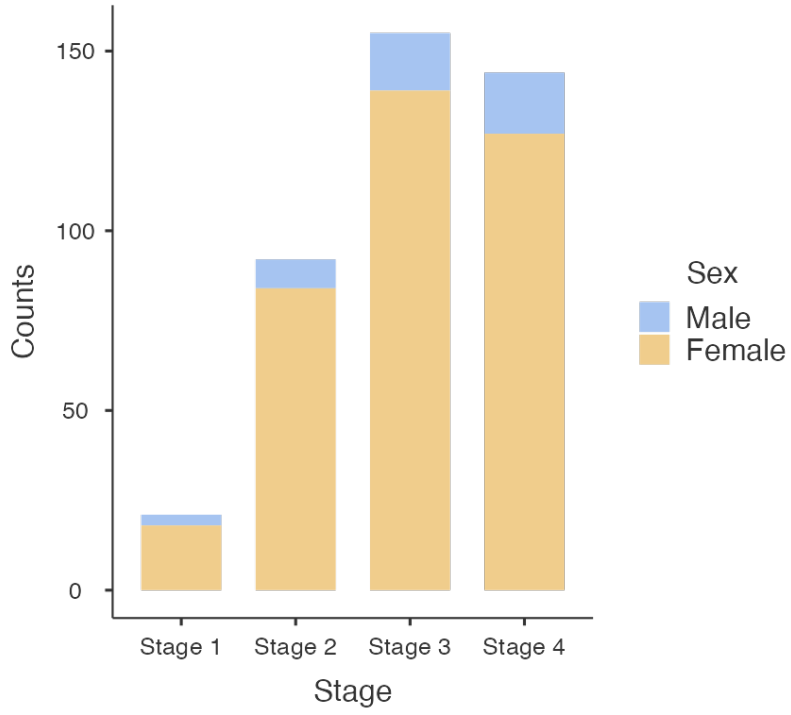


Figure 2.5: Stacked bar graph of stage of disease by sex from PBC study

Finally, a stacked relative bar chart (Figure 2.6) displays the proportion of grouping variable for each bar, where each overall bar represents 100%. These graphs allow the reader to compare the

proportions between categories. We can easily see from Figure 2.6 that the distribution of sex is similar across each stage of disease.

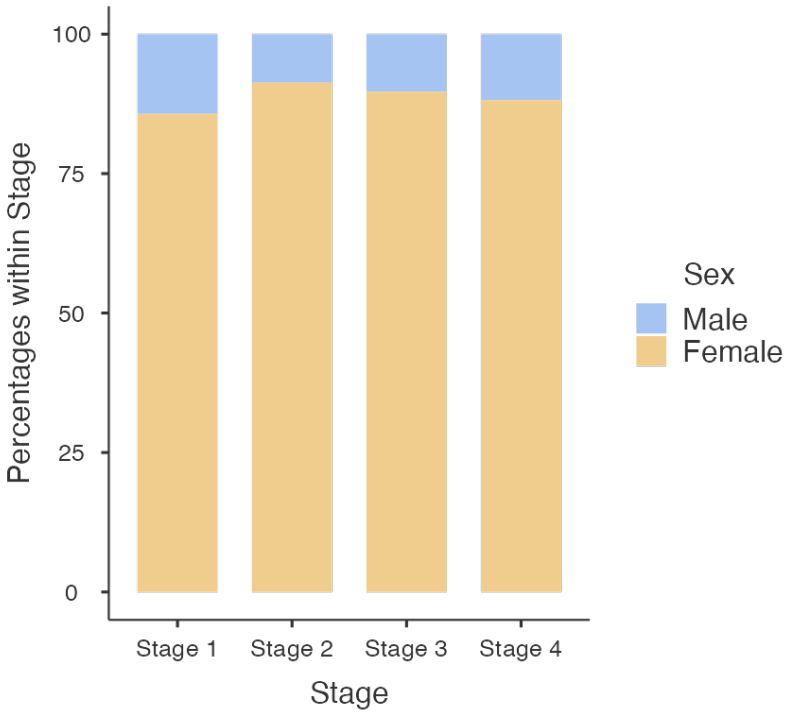


Figure 2.6: Relative frequency of sex within stage of disease from PBC study

2.7 Presentation guidelines

Guidelines for presenting summary statistics

When reporting summary statistics, it is important not to present results with too many decimal places. Doing so implies that your data have a higher level of precision than they do. For example, presenting a mean blood pressure of 100.2487 mmHg implies that blood pressure can be measured accurately to at least three decimal places.

There are a number of guidelines that have been written to help in the presentation of numerical data. Many of these guidelines are based on the number of decimal places, while others are based on the number of significant figures. Briefly, the number of significant figures are "the number of digits from the first non-zero digit to the last meaningful digit, irrespective of the position of the decimal point. Thus, 1.002, 10.02, 100200 (if this number is expressed to the nearest 100) all have four significant digits." Armitage, Berry, and Matthews (2013)

A summary of these guidelines that will be used in this course appear in Table 2.6.

Table 2.6: Guidelines for presentation of statistical results

Summary statistic	Guideline (reference)
Mean	It is usually appropriate to quote the mean to one extra decimal place compared with the raw data. (Altman)
Median, Interquartile range, Range	As medians, interquartile ranges and ranges are based on individual data points, these values should be presented with the same precision as the original data.
Percentage	Percentages do not need to be given with more than one decimal place at most. When the sample size is less than 100, no decimal places should be given. (Altman)

Summary statistic	Guideline (reference)
Probability	It is acceptable to present probabilities to 2 or 3 decimal places. If the probability is presented as a percentage, present the percentage with 0 or 1 decimal place.
Standard deviation	The standard deviation should usually be given to the same accuracy as the mean, or with one extra decimal place. (Altman)
Standard error	As per standard deviation
Confidence interval	Use the same rule as for the corresponding effect size (be it mean, percentage, mean difference, regression coefficient, correlation coefficient or risk ratio) (Cole)
Test statistic	Test statistics should not be presented with more than two decimal places.
P-value	Report P-values to a single significant figure unless the P-value is close to 0.05 (say, 0.01 to 0.2), in which case, report two significant figures. Do not report 'not significant' for P-values of 0.05 or higher. Very low P-values can be reported as $P < 0.001$ or $P < 0.0001$. A P-value can indeed be 1, although some investigators prefer to report this as >0.9. (Based on Assel)
Difference in means	As for the estimated means
Difference in proportions	As for the estimated proportions
Odds ratio / Relative risk	Hazard and odds ratios are normally reported to two decimal places, although this can be avoided for high odds ratios (Assel)
Correlation coefficient	One or two decimal places, or more when very close to ±1 (Cole)
Regression coefficient	Use one more significant figure than the underlying data (adapted from Cole)

Table presentation guidelines

Consider the following guidelines for the appropriate presentation of tables in scientific journals and reports (Woodward, 2013).

1. Each table (and figure) should be self-explanatory, i.e. the reader should be able to understand it without reference to the text in the body of the report.
    - This can be achieved by using complete, meaningful labels for the rows and columns and giving a complete, meaningful title.
    - Footnotes can be used to enhance the explanation.
2. Units of the variables (and if needed, method of calculation or derivation) should be given and missing records should be noted (e.g. in a footnote).
3. A table should be visually uncluttered.
    - Avoid use of vertical lines.
    - Horizontal lines should not be used in every single row, but they can be used to group parts of the table.
    - Sensible use of white space also helps enormously; use equal spacing except where large spaces are left to separate distinct parts of the table.
    - Different typefaces (or fonts) may be used to provide discrimination, e.g. use of bold type and/or italics.
4. The rows and columns of each table should be arranged in a natural order to help interpretation. For instance, when rows are ordered by the size of the numbers they contain for a nominal variable, it is immediately obvious where relatively big and small contributions come from.

5. Tables should have a consistent appearance throughout the report so that the paper is easy to follow (and also for an aesthetic appearance). Conventions for labelling and ordering should be the same (for both tables as well as figures) for ease of comparison of different tables (and figures).
6. Consider if there is a particular table orientation that makes a table easier to read.

Given the different possible formats of tables and their complexity, some further guidelines are given in the following excellent references:

- Graphics and statistics for cardiology: designing effective tables for presentation and publication, Boers (2018)
- Guidelines for Reporting of Figures and Tables for Clinical Research in Urology, Vickers et al. (2020)

Graphical presentation guidelines

Consider the following guidelines for the appropriate presentation of graphs in scientific journals and reports (Woodward, 2013).

- Figures should be self-explanatory and have consistent appearance through the report.
- A title should give complete information. Note that figure titles are usually placed below the figure, whereas for tables titles are given above the table.
- Axes should be labelled appropriately
- Units of the variables should be given in the labelling of the axes. Use footnotes to indicate any calculation or derivation of variables and to indicate missing values
- If the Y-axis has a natural origin, it should be included, or emphasised if it is not included.
- If graphs are being compared, the Y-axis should be the same across the graphs to enable fair comparison
- Columns of bar charts should be separated by a space
- Three dimensional graphs should be avoided unless the third dimension adds additional information

Sources:

Altman (1990)

Cole (2015)

Assel et al. (2019)

2.8 Probability

Probability is defined as:

> the chance of an event occurring, where an event is the result of an observation or experiment, or the description of some potential outcome.

Probabilities range from 0 (where the event will never occur) to 1 (where the event will always occur). For example, tossing a coin is an experiment; one event is the coin landing with head up, while the other event is the coin landing tails up. The set of all possible outcomes in an experiment is called the sample space. For example, by tossing a coin you can get either a head or a tail (called mutually exclusive events); and by rolling a die you can get any of the six sides. Thus, for a die the sampling space is: S = {1, 2, 3, 4, 5, 6}

With a fair (unbiased) die, the probability of each outcome occurring is 1/6 and its probability distribution is simply a probability of 1/6 for each of the six numbers on a die.

Additive law of probability

How do we work out the probability that one roll of a die will turn out to be a 3 or a 6? To do that, we first need to work out whether the events (3 or 6 on the roll of a die) are mutually exclusive. Events are mutually exclusive if they are events which cannot occur at the same time. For example, rolling a die once and getting a 3 and 6 are mutually exclusive events (you can roll one or the other but not both in a single roll).

To obtain the probability of one or the other of two mutually exclusive events occurring, the sum of the probabilities of each is taken. For example, the probability of the roll of a die being a 3 or a 6 is the sum of the probability of the die being 3 (i.e. 1/6) and the probability of the die being 6 (also 1/6). With a fair die:

Probability of a die roll being 3 or 6 = $1/6 + 1/6 = 1/3$

Another way of putting it is:

P(die roll =3 or die roll =6) = P(die roll=3) + P(die roll=6) = $1/6 + 1/6 = 1/3$

Example: Additive law for mutually exclusive events

Consider that blood type can be organised into the ABO system (blood types A, B, AB or O) An individual may only have one blood type.

Using the information from https://www.donateblood.com.au/learn/about-blood let's consider the ABO blood type system. The frequency distribution (prevalence) of the ABO blood type system in the population represents the probability of each of the outcomes. If we consider all possible blood type outcomes, then the total of the probabilities will sum to 1 (100%).

Table 2.7: Frequency of blood types

Blood Type	% of population	Probability
A	38%	0.38
B	10%	0.10
AB	3%	0.03
O	49%	0.49
Total	100%	1.00

In this example we consider: What is the probability that an individual will have either blood group O or A?

Since blood type is mutually exclusive, the probability that either one or the other occurs is the sum of the individual probabilities. These are mutually exclusive events so we can say P(O or A) = P(O) + P(A)

Thus, the answer is: P(Blood type O) + P(Blood type A) = 0.49 + 0.38 = 0.87

Multiplicative law of probability

The additive law of probability lets us consider the probability of different outcomes in a single experiment. The multiplicative law lets us consider the probability of multiple events occurring in a particular order. For example: if I roll a die twice, what is the probability of rolling a 3 and *then* a 6?

These events are independent: the probability of rolling a 6 on the second roll is not affected by the first roll.

The multiplicative law of probability states:

> If A and B are independent, then P(A and B) = P(A) $\times$ P(B).

So, the probability of rolling a 3 and then a 6 is: P(3 and 6) = $1/6 \times 1/6 = 1/36$.

Note here that the order matters – we are considering the probability of rolling a 3 and then a 6, not the probability of rolling a 6 and then a 3.

2.9 Probability distributions

> A probability distribution is a table or a function that provides the probabilities of all possible outcomes for a random event.

For example, the probability distribution for a single coin toss is straightforward: the probability of obtaining a head is 0.5, and the probability of obtaining a tail is 0.5, and this can be summarised in Table 2.8.

Table 2.8: Probability distribution for a single coin toss

Coin face	Probability
Heads	0.5
Tails	0.5

Similarly, the probability distribution for a single roll of a die is straightforward: each face has a probability of 1/6 (Table 2.9).

Table 2.9: Probability distributions for a single roll of a die

Face of a die	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Things become more complicated when we consider multiple coin-tosses, or rolls of a die. These series of events can be summarised by considering the number of times a certain outcome is observed. For example, the probability of obtaining three heads from five coin tosses.

Probability distributions can be used in two main ways:

1. To calculate the probability of an event occurring. This seems trivial for the coin-toss and die-roll examples above. However, we can consider more complex events, as below.
2. To understand the behaviour of a sample statistic. We will see in Modules 3 and 4 that we can assume the mean of a sample follows a probability distribution. We can obtain useful information about the sample mean by using properties of the probability distribution.

2.10 Discrete random variables and their probability distributions

Rather than thinking of random events, we often use the term *random variable* to describe a quantity that can have different values determined by chance.

A *discrete random variable* is a random variable that can take on only countable values (that is, non-negative whole numbers). An example of a discrete random variable is the number of heads observed in a series of coin tosses.

A discrete random variable can be summarised by listing all the possible values that the variable can take. As defined earlier, a table, formula or graph that presents these possible values, and their associated probabilities, is called a probability distribution.

Example: let's consider the number of heads in a series of three coin tosses. We might observe 0 heads, or 1 head, or 2, or 3 heads. If we let X denote the number of heads in a series of three coin tosses, then possible values of X are 0, 1, 2 or 3.

We write the probability of observing x heads as P(X=x). So P(X=0) is the probability that the three tosses has no heads. Similarly, P(X=1) is the probability of observing one head.

The possible combinations for three coin tosses are as follows:

Table 2.10: The number of heads from three coin tosses

Pattern	Number of heads
Tail, Tail, Tail	0
Head, Tail, Tail	
Tail, Head, Tail	1
Tail, Tail, Head	
Head, Head, Tail	
Head, Tail, Head	2
Tail, Head, Head	
Head, Head, Head	3

There are eight possible outcomes from three coin tosses (permutations). If we assume an equal chance of observing a head or a tail, each permutation above is equally likely, and so has a probability of 1/8.

If we consider the possibility of observing just one head out of the three tosses, this can happen in three ways (HTT, THT, TTH). So the probability of observing one head is calculated using the additive law: P(X=1) = $\frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$.

Therefore, the probability distribution for X, the number of heads from three coin tosses, is as follows:

Table 2.11: Probability distribution for the number of heads from three coin tosses

x (number of heads observed)	P(X=x)
0	1/8
1	1/8 + 1/8 + 1/8 = 3/8
2	1/8 + 1/8 + 1/8 = 3/8
3	1/8

Note that the probabilities sum to 1.

The above example was based on a coin toss, where flipping a head or a tail is equally likely (both have probabilities of 0.5). Let's consider a case where the probability of an event is not equal to 0.5: having blood type A.

From Table 2.7, the probability that a person has Type A blood is 0.38, and therefore, the probability that a person does not have Type A blood is 0.62 (1–0.38). If we considered taking a random sample of three people, the probability that all three would have Type A blood is 0.38 × 0.38 × 0.38 (using the multiplicative rule above) – and there is only one way this could happen.

The number of ways two people out of three could have Type A blood is 3, and each permutation is listed in Table 2.12. The probability of observing each of the three patterns is the same, and can be calculated using the multiplicative rule: $0.38 \times 0.38 \times 0.62 = 0.0895$.

Table 2.12: Combinations and probabilities of Type A blood in three people

Person 1	Person 2	Person 3	Probability
A	A	A	$0.38 \times 0.38 \times 0.38 = 0.0549$
A	A	Not A	$0.38 \times 0.38 \times 0.62 = 0.0895$
A	Not A	A	$0.38 \times 0.62 \times 0.38 = 0.0895$
Not A	A	A	$0.62 \times 0.38 \times 0.38 = 0.0895$
A	Not A	Not A	$0.38 \times 0.62 \times 0.62 = 0.1461$
Not A	A	Not A	$0.62 \times 0.38 \times 0.62 = 0.1461$
Not A	Not A	A	$0.62 \times 0.62 \times 0.38 = 0.1461$
Not A	Not A	Not A	$0.62 \times 0.62 \times 0.62 = 0.2383$

Table 2.13 gives the probability of each of the blood type combinations we could observe in three people. The probability of observing a certain number of people (say, k) with Type A blood from a sample of three people can be calculated by summing the combinations:

Table 2.13: Probabilities of observing numbers of people with Type A blood in a sample of three people

Number of people with Type A blood	Probability of each pattern
3	0.0549
2	0.0895 + 0.0895 + 0.0895 = 0.2689
1	0.1461 + 0.1461 + 0.1461 = 0.4382
0	0.2383

2.11 Binomial distribution

The above are examples of the binomial distribution. The binomial distribution is used when we have a collection of random events, where each random event is binary (e.g. Heads vs Tails, Type A blood vs Not Type A blood, Infected vs Not infected). The binomial distribution calculates (in general terms):

- the probability of observing k successes
- from a collection of n trials
- where the probability of a success in one trial is p.

The terms used here can be defined as:

- a success is simply an event of interest from a binary random event. In the coin-toss example, “success” was tossing a Head. In the blood type example, we were only interested in whether someone was Type A or not Type A, so “success” was a blood of Type A. We tend to use the word “success” to mean “an event of interest”, and “failure” as “an event not of interest”.
- the number of trials refers to the number of random events observed. In both examples, we observed three events (three coin tosses, three people).

- the probability of a success (p) simply refers to the probability of the event of interest. In the coin toss example, this was the probability of tossing a Heads (=0.5); for the blood-type example, this was the probability of having Type A blood (0.38).

Putting all this together, we say that we have a binomial experiment. To satisfy the assumptions of a binomial distribution, our experiment must satisfy the following criteria:

1. The experiment consists of fixed number (n) of trials.
2. The result of each trial falls into only one of two categories – the event occurred ("success") or the event did not occur ("failure").
3. The probability, p, of the event occurring remains constant for each trial.
4. Each trial of the experiment is independent of the other trials.

We have shown in the examples above how we can calculate the probabilities for small experiments (n=3). Once n becomes large, constructing such probability distribution tables becomes difficult. The general formula for calculating the probability of observing k successes from n trials, where each trial has a probability of success of p is given by:

$$P(X = k) = \frac{n!}{k!(n-k)!} \times p^k \times (1-p)^{n-k}$$

where $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$.

Note that this formula is almost never calculated by hand. Instructions for calculating binomial probabilities are given in the jamovi and R notes at the end of this Module.

Worked example

A population-based survey conducted by the AIHW (2008) of a random sample of the Australian population estimated that in 2007, 19.8% of the Australian population were current smokers.

a) From a random sample of 6 people from the Australian population in 2007, what is the probability that 3 of them will be smokers?
b) What is the probability that among the six persons, at least 4 will be smokers?
c) What is the probability that at most, 2 will be smokers?

Solution

a) Calculating this single binomial probability is best done using software.

We can use the applet to calculate this, or use the `dbinom` function in R with x=3, size=6, and prob=0.198. This gives an answer of 0.08.

b) In common language, getting "at least 4" smokers means getting 4, 5 or 6 smokers. Since these are mutually exclusive events, we can apply the additive law to find the probability of getting at least 4 smokers:

$$P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6)$$

Using the same binomial probability functions as in the previous question, we could calculate

- P(X=4) = 0.0148
- P(X=5) = 0.00146
- P(X=6) = 0.0000603

Answer: $P(X \geq 4) = 0.0148 + 0.00146 + 0.0000603 = 0.016$

We can use the applet to calculate this, or in R we can use the `pbinom` function with the `lower.tail=FALSE` option.

c) Observing at most two means observing 0, 1 or 2 smokers. Therefore, the probability of observing at most 2 smokers is:

- P(X $\leq$ 2) = P(X=0) + P(X=1) + P(X=2)
- P(X=0) = 0.266
- P(X=1) = 0.394
- P(X=2) = 0.243

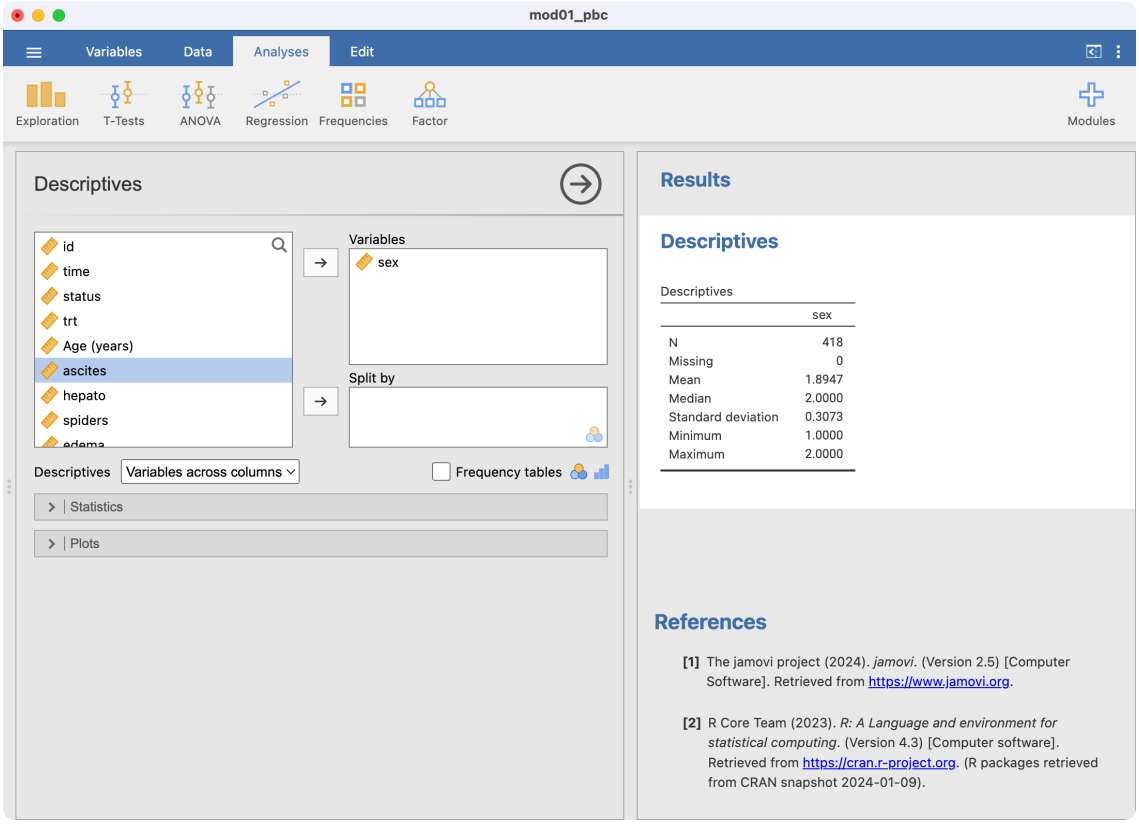
Answer: P(X $\leq$ 2) = 0.266+0.394+0.243=0.903

Again, we can use the applet to calculate this, or use the `pbinom` function in R.

jamovi notes

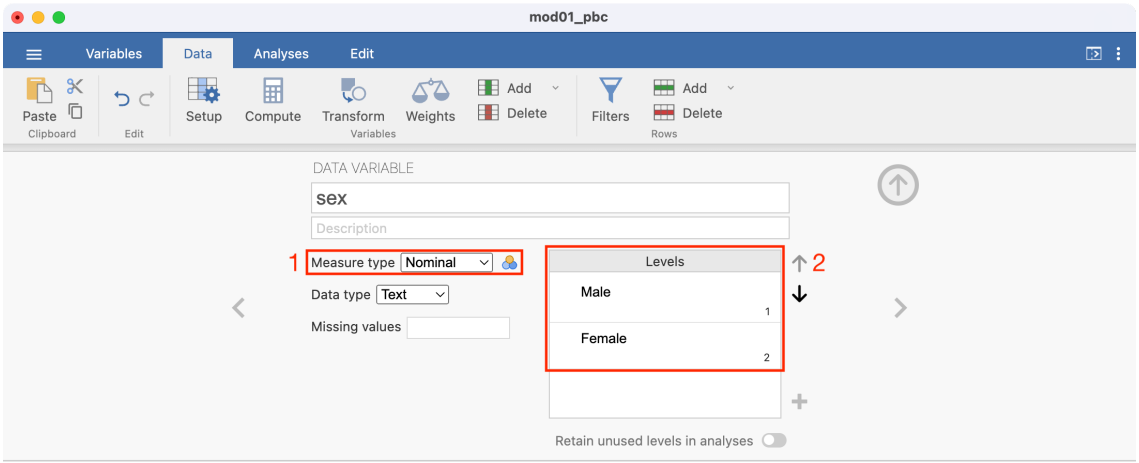
2.12 Producing a one-way frequency table

The simplest way to summarise categorical variables is with a one-way frequency table. These are constructed using **Analyses > Exploration**. We will illustrate this by summarising the variable `sex` from the `pbc.rds` data from the previous module.



jamovi has summarised `sex` here, just as we asked it to, however it has analysed `sex` as if it was a continuous variable. This is incorrect: `sex` is a categorical variable. This can be corrected by defining `sex` to be categorical within **Data > Setup**:

1. Select `sex`, then choose **Nominal** as the Measure type. jamovi now lists the levels of `sex` as 1 and 2. These should be replaced by the categories they represent.
2. From the `mod01_pbc_info.txt` file, we see that 1 represents Male, and 2 represents Female. These labels can be added by typing in the appropriate cell. The completed screen should look like this:



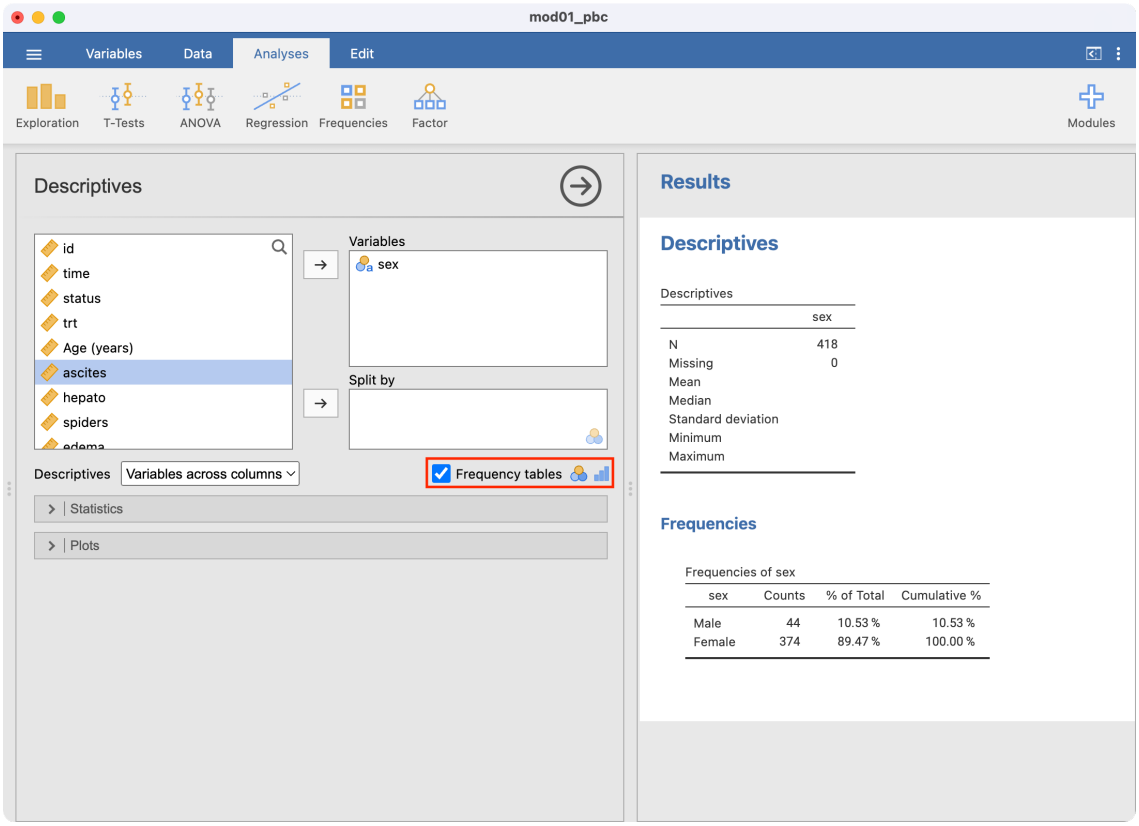
Clicking back to the original summary of sex shows that jamovi is no longer treating `sex` as a continuous variable, but there is little output in the summary:

Descriptives

Descriptives

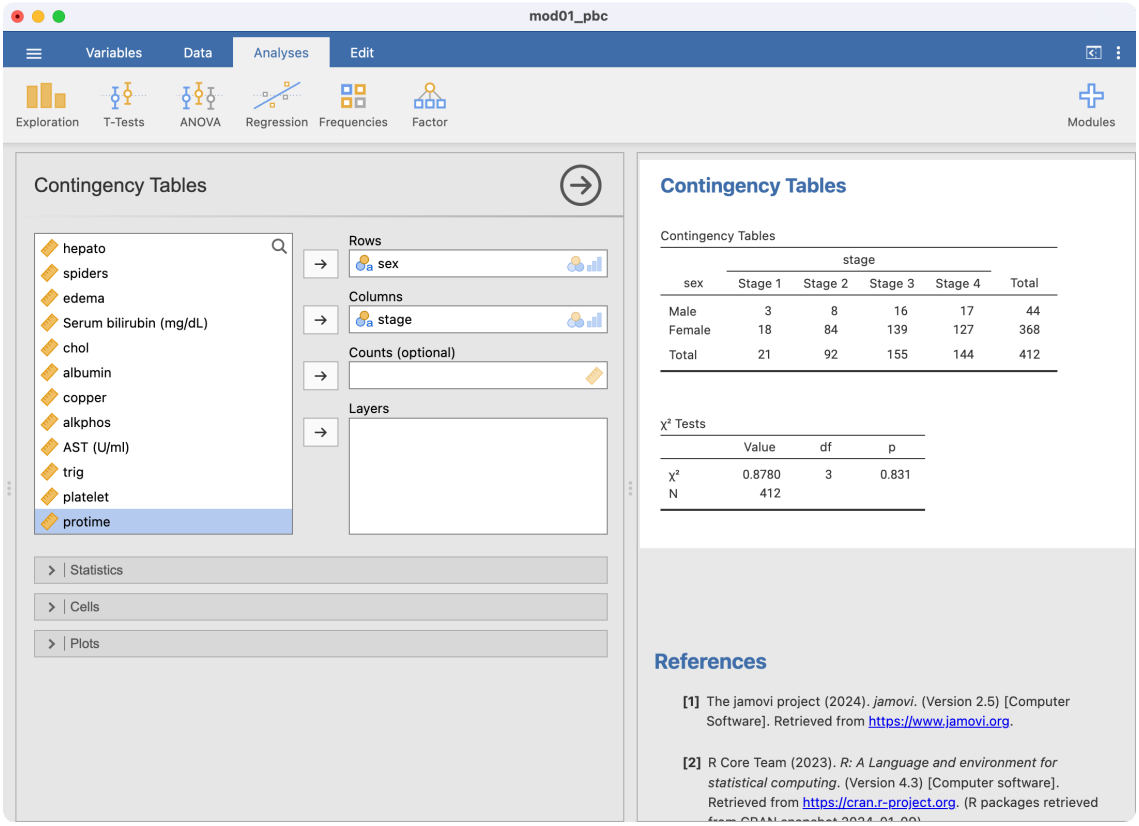
	sex
N	418
Missing	0
Mean	
Median	
Standard deviation	
Minimum	
Maximum	

Click **Frequency tables** in the main Descriptives window to request the one-way frequency table:



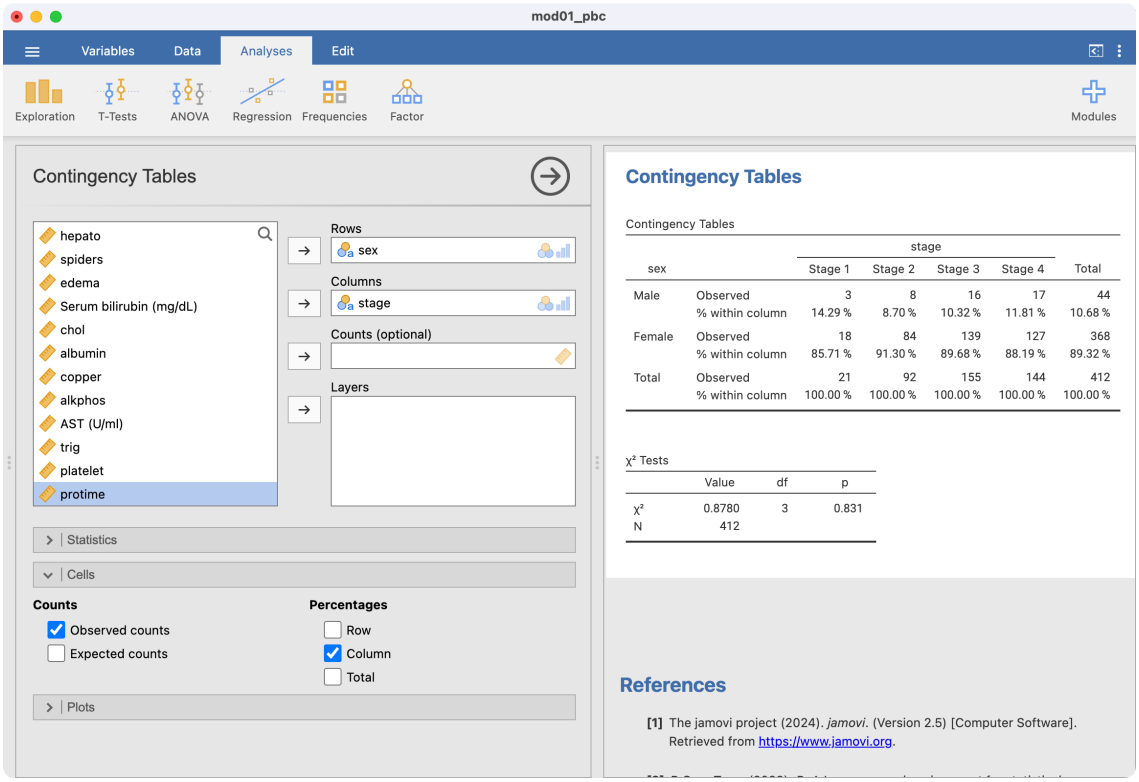
2.13 Producing a two-way frequency table

Two-way tables are constructed using **Analyses > Frequencies > Contingency Tables > Independent Samples**. Note that **both variables must be defined as Nominal variables**. As an example, to produce a two-way table of disease stage by sex:



Ignore the output labelled χ^2 tests for now.

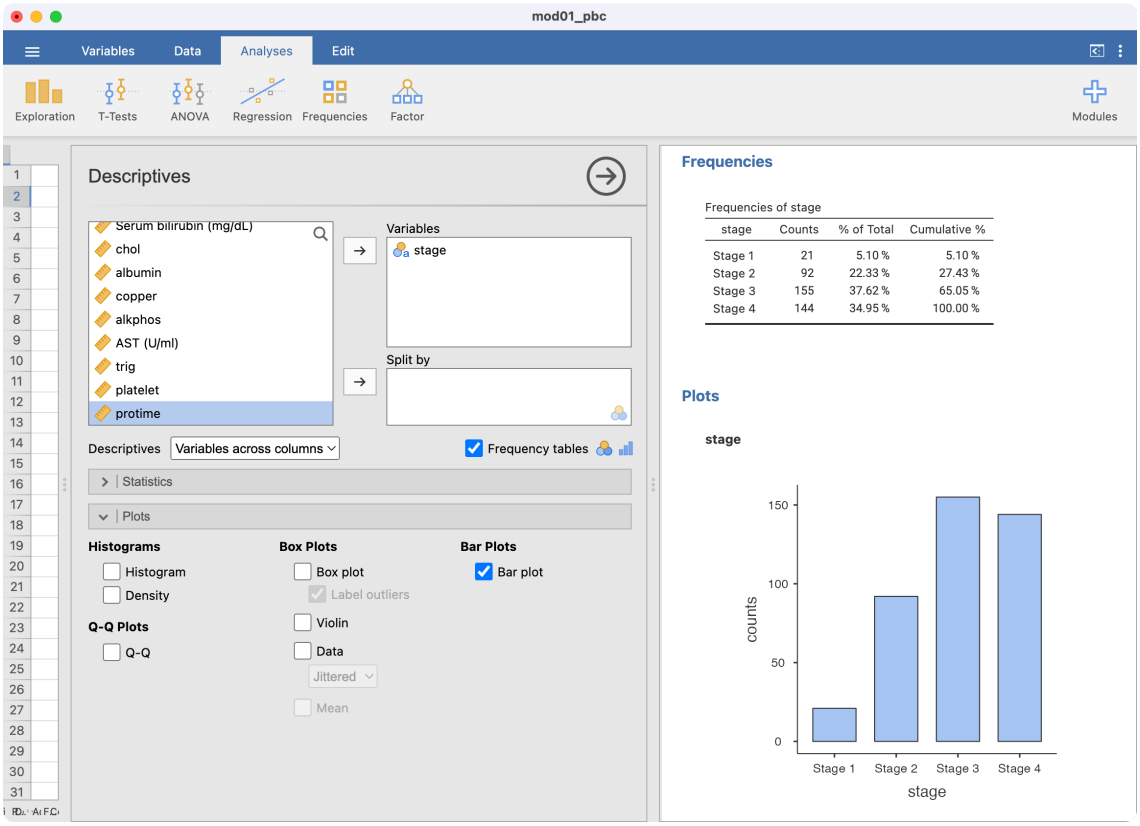
Row or column percents can be requested in the **Cells** section. For example, to calculate the proportion of males within each stage of disease, we would request column percents:



2.14 Creating bar charts for one categorical variable

Here we will create the bar chart shown in Figure 2.1 using the `mod01_pbc.rds` dataset. The x-axis of this graph will be the stage of disease, and the y-axis will show the number of participants in each category.

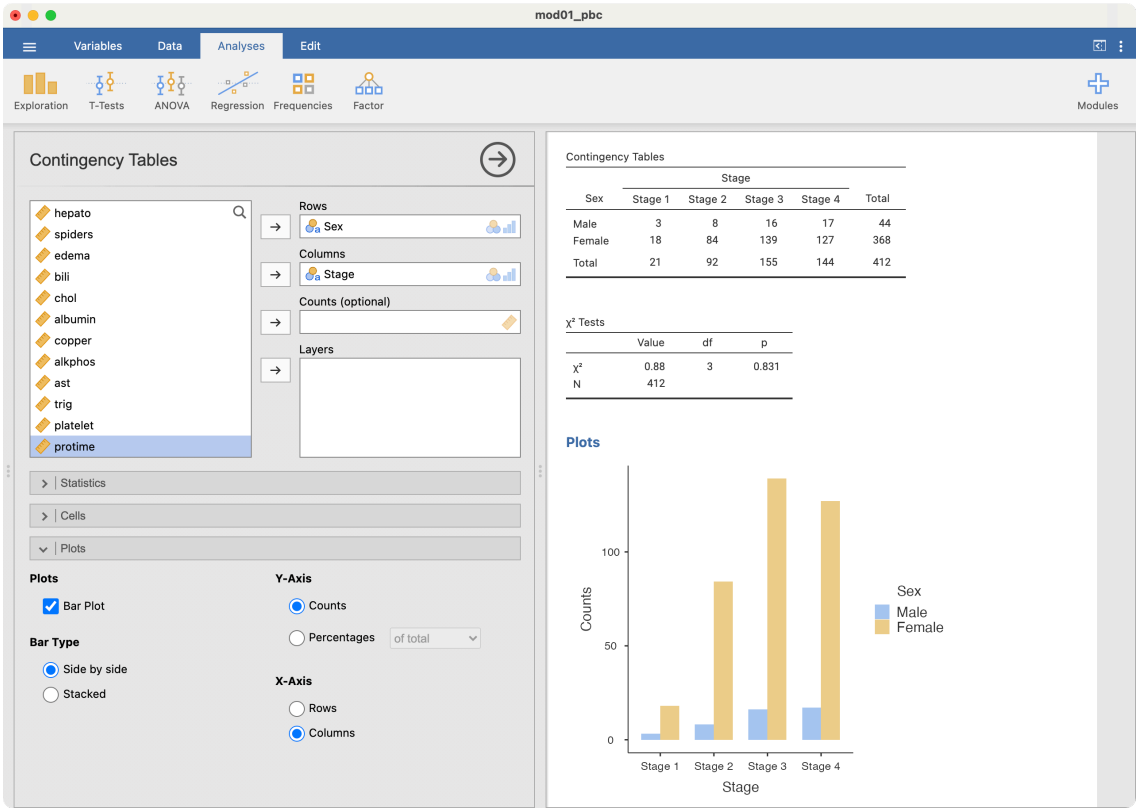
Bar charts are created in the **Exploration** tab. We can summarise `stage`, and request a **Frequency table** in the usual way. To request a bar chart as well, tick **Bar plot** within the **Plots** section:



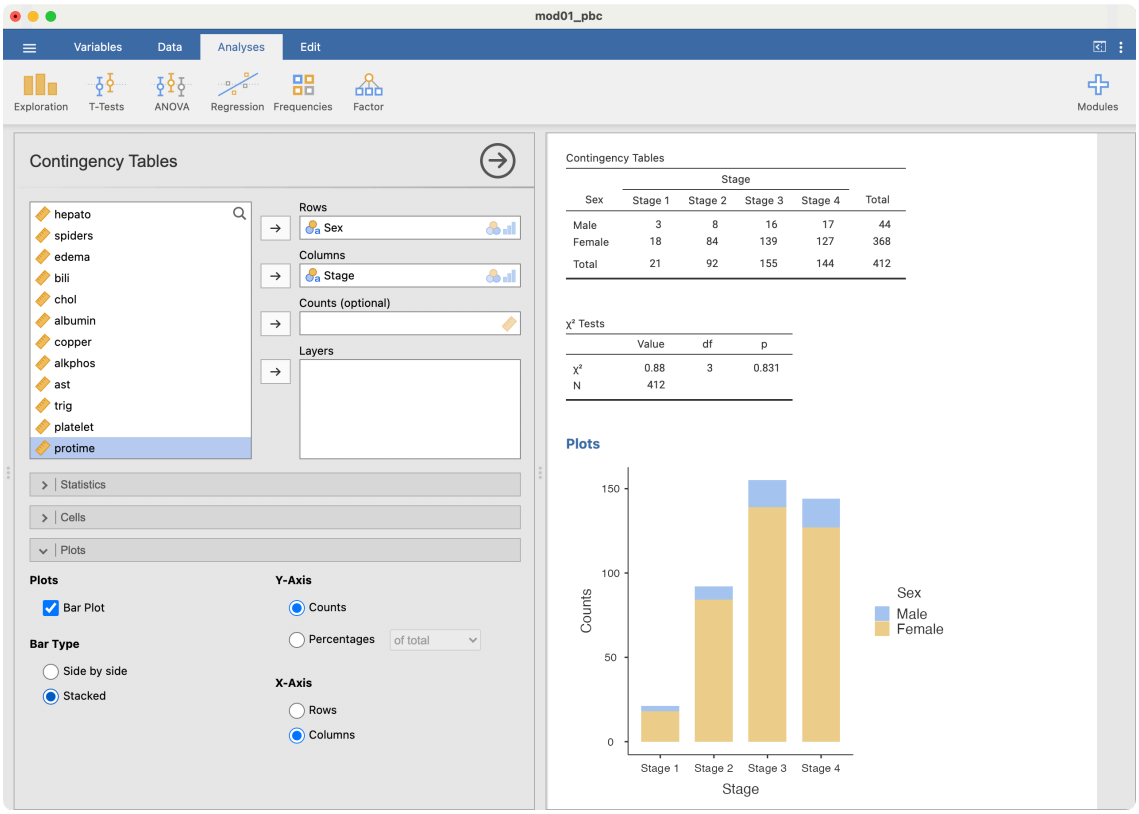
2.15 Creating bar charts for two categorical variables

Option 1: Using Contingency Tables command

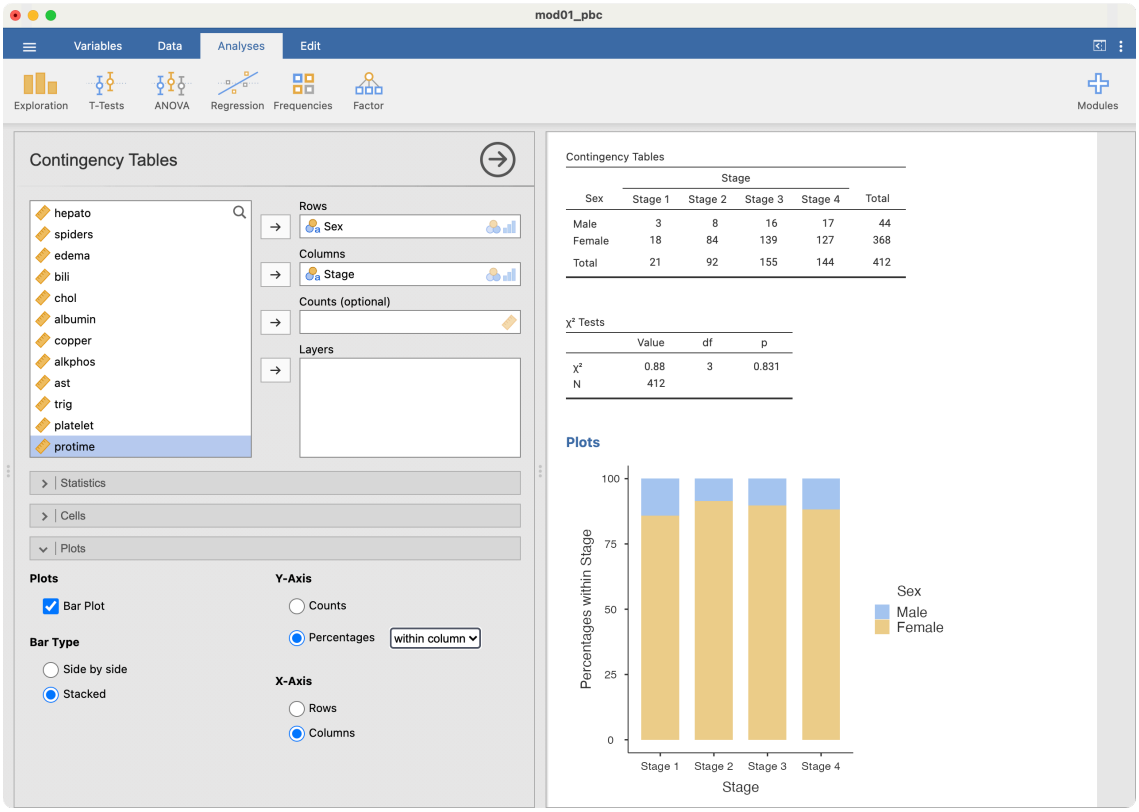
Creating a **clustered bar chart** as shown in Figure 2.4 can be done using **Analyses > Frequencies > Contingency Tables > Independent Samples**. First create the cross-tab of interest, for example, Stage by Sex. Choose **Plots > Bar Plot**, and ensure the Bar Type is selected as **Side by side**. Choose the X-axis as required - here we want to see stage on the x-axis, so we choose the x-axis to be columns (this may need to be adjusted according to how your table has been set up):



To create a **stacked bar chart** (as in Figure 2.5), choose the Bar Type to be **Stacked**:

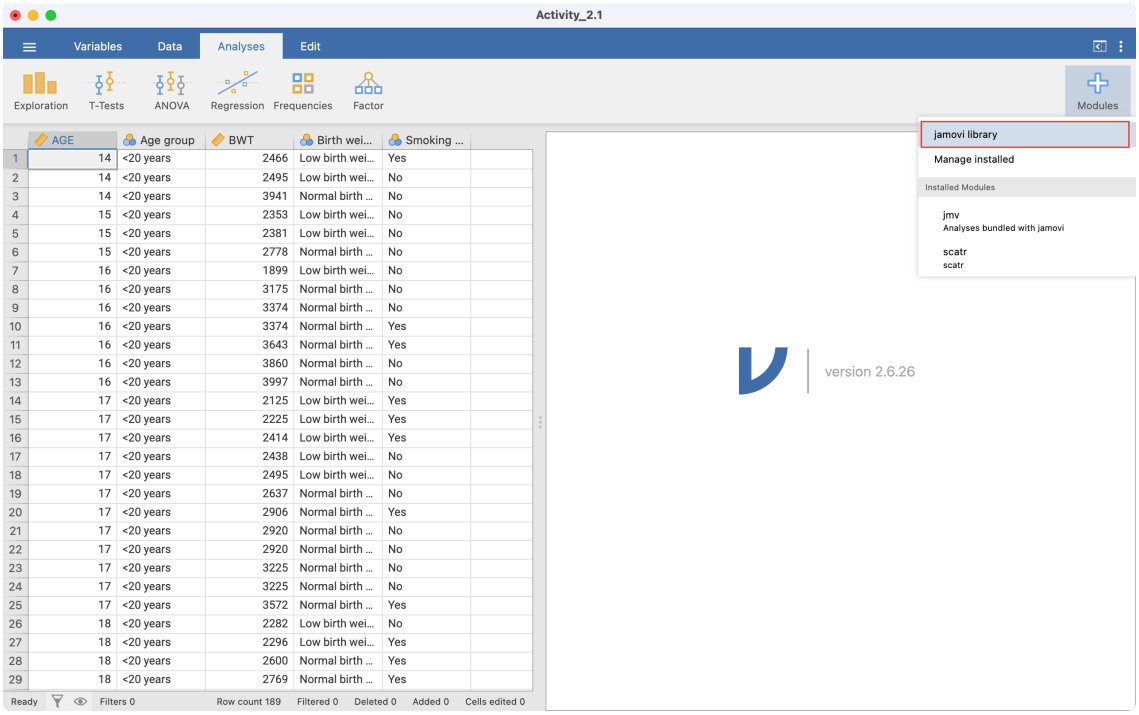


Finally, to create a **stacked relative bar chart** (as in Figure 2.6), choose the Y-axis to be **Percentages within column**:

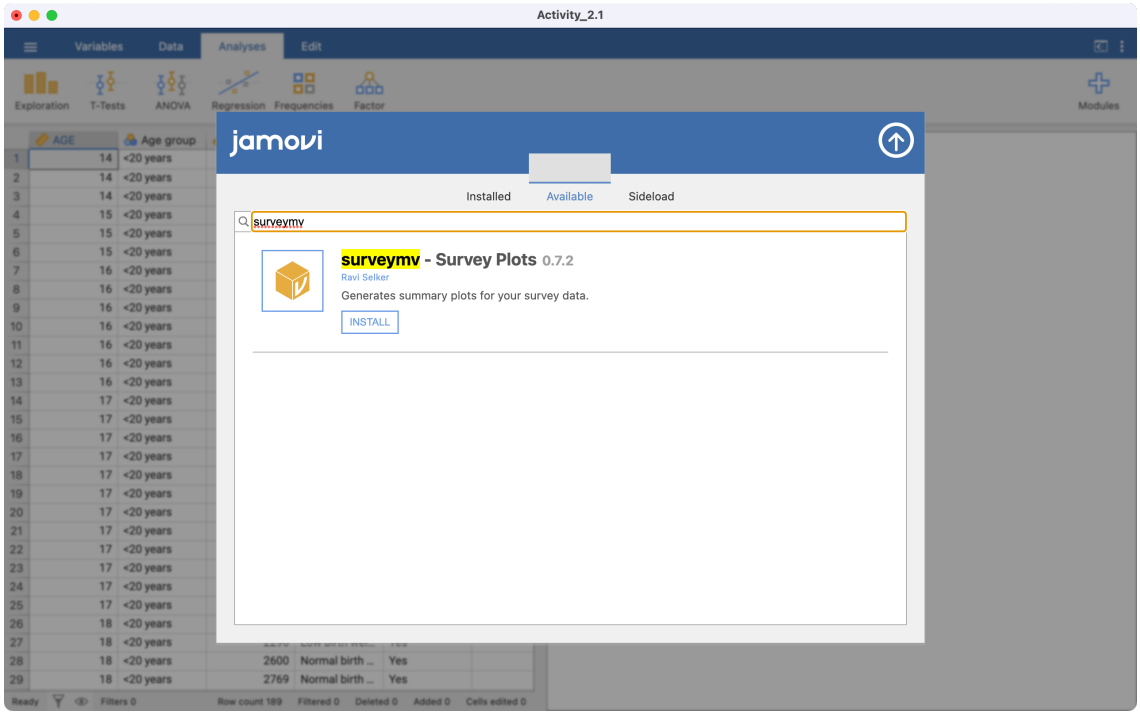


Option 2: Using Survey Plots command

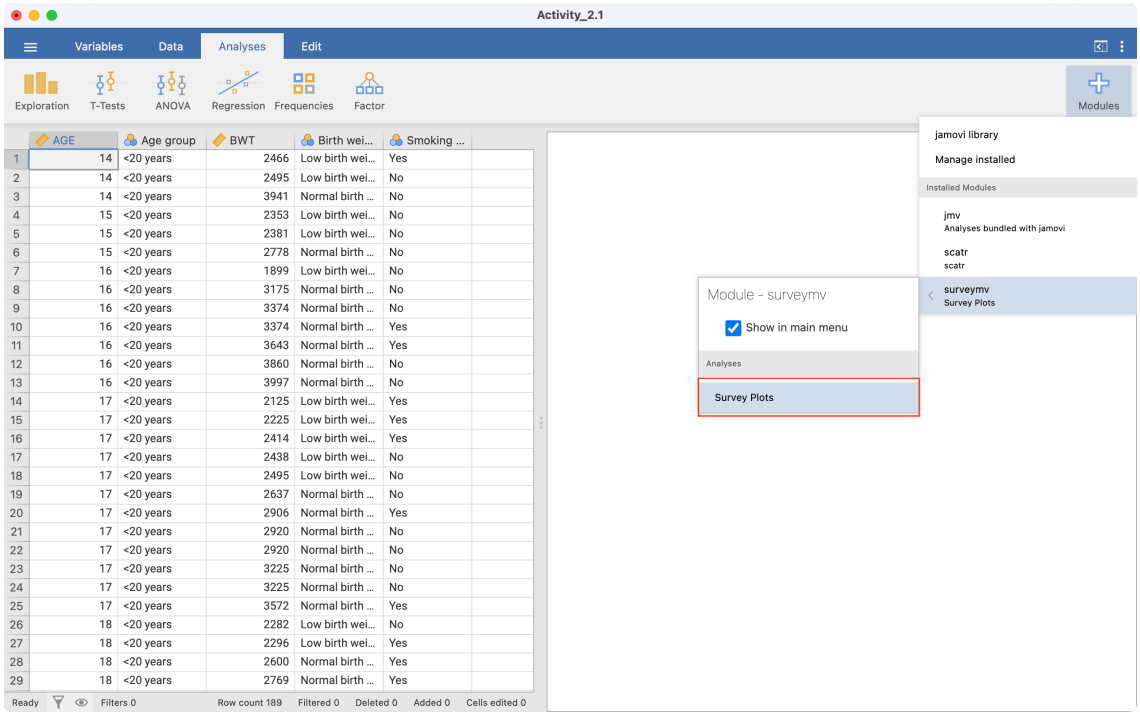
An alternative way of producing barcharts is via a Module called **surveymv** that we add to jamovi. To install the module, click the **Analyses** tab, and click the large **+** at the top-right of the window. Choose **jamovi library**:



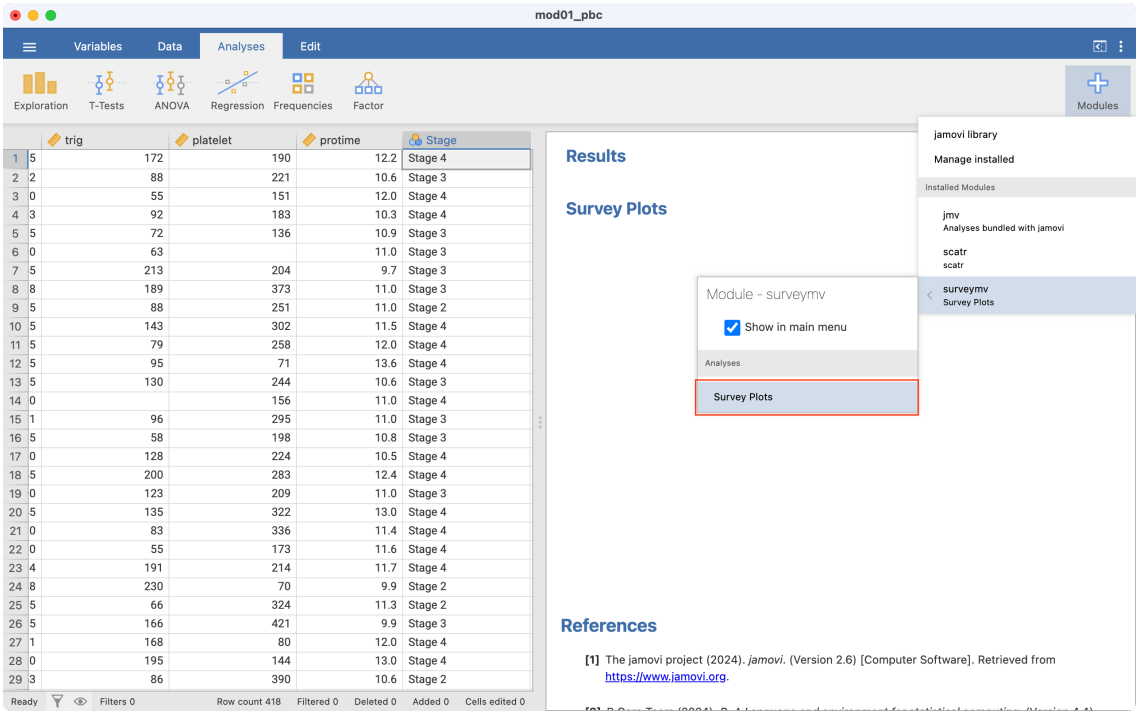
Click **Available** and search for `surveymv`, then click install:



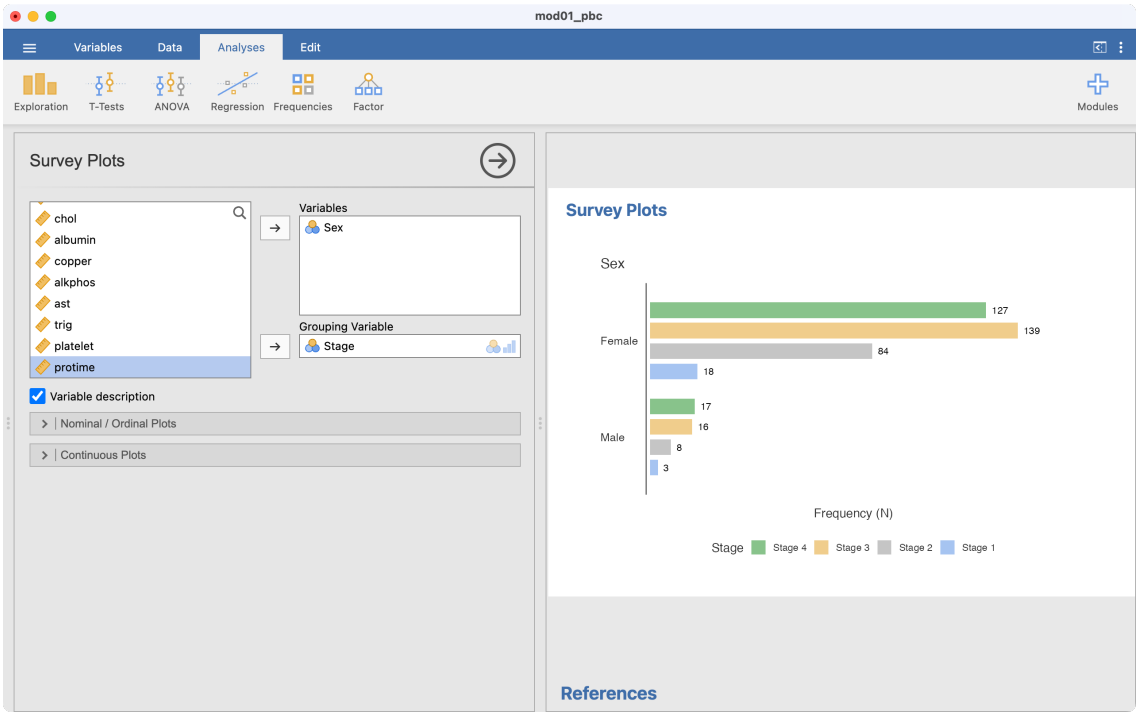
The module has now been installed. To run the module, click the up-arrow to return to the **Analyses** tab, click the large **+** and choose **surveymv > Survey Plots**:



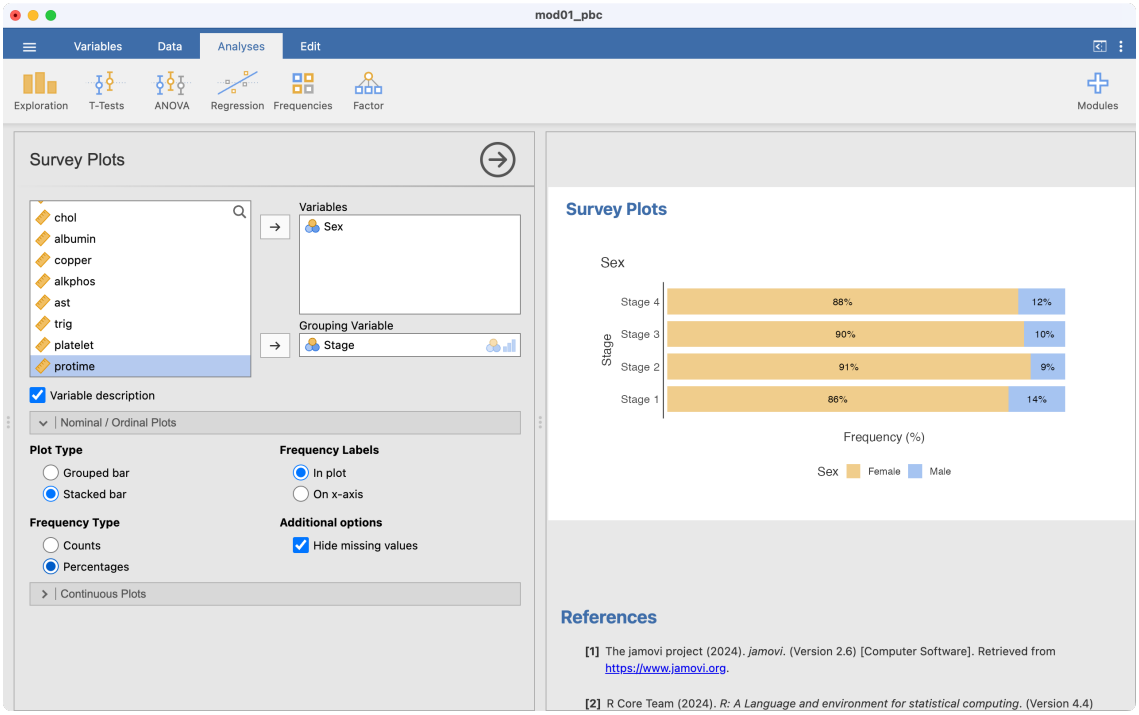
We will demonstrate the module by recreating Figure 2.6, a figure of the relative frequencies of sex within stage of disease. To open the module options, click the large **+** and choose **surveymv > Survey Plots**:



We choose the variable we want to plot, here `Sex`. We want to plot sex within each stage, so `Stage` is entered as a **Grouping Variable**:



This has plotted a clustered bar chart, we want a stacked bar chart, so select **Stacked bar**. Also, we want to plot percentages, not counts, so choose **Percentages**:



Our stacked relative frequency chart has been completed. While this produces a bar chart with horizontal bars, it often performs better with labelling of the groups than the previous method.

2.16 Recoding data

One task that is common in statistical computing is to recode variables. For example, we might want to group some categories of a categorical variable, or to present a continuous variable in a categorical way.

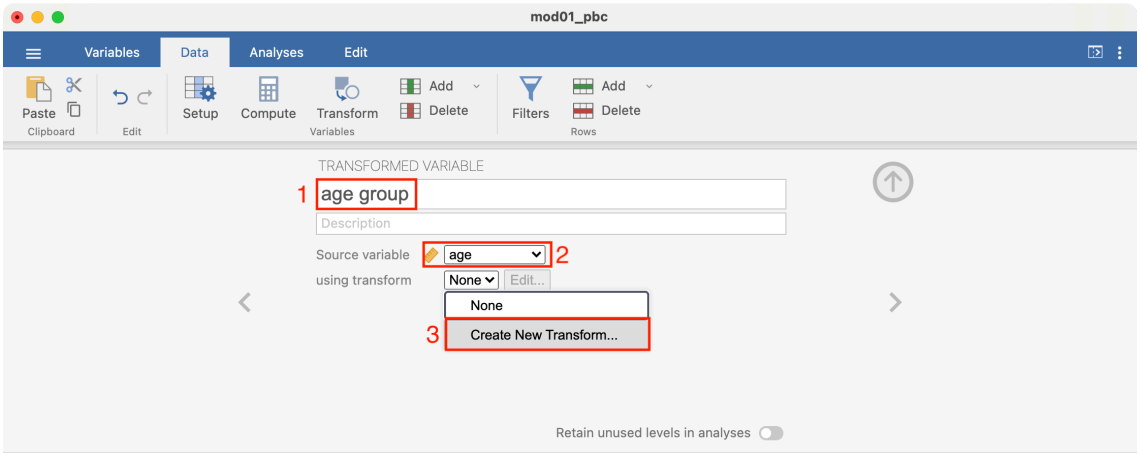
In this example, we can use the `pbc` data and recode age into age groups:

- Less than 30
- 30 to less than 50
- 50 to less than 70
- 70 or older

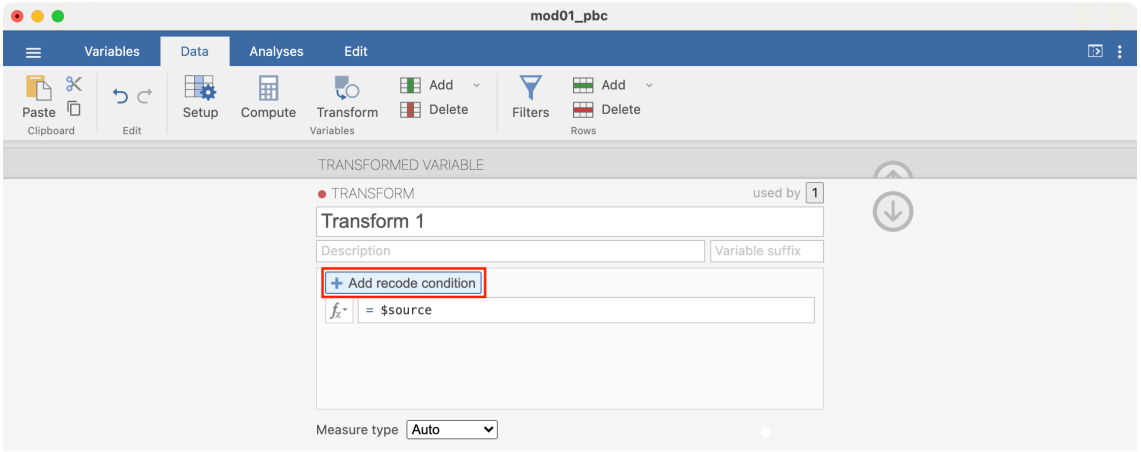
Recoding can be done using **Data > Transform**.

First, click **Data** to view the spreadsheet, then click in an empty column, then click **Setup > NEW TRANSFORMED VARIABLE**. We need to specify three things:

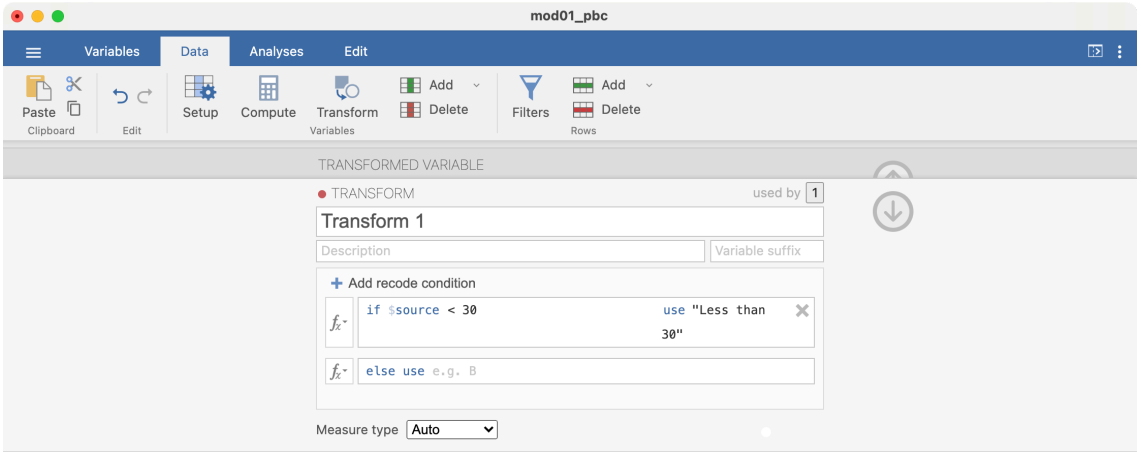
1. The name of the new variable. Here, we will choose `age group`.
2. The source variable. Here, we want to recode **from** `age`, so choose `age`.
3. The `transform`, which is where we define the rules of the recode. Choose **Create New Transform**.



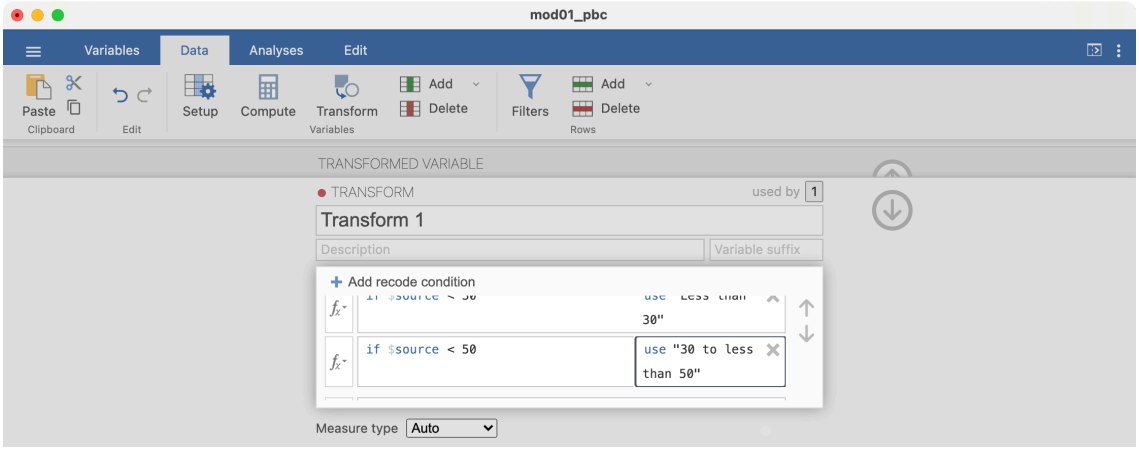
The `transform` is built up by specifying the **recode conditions**. Click **+ Add recode condition** to define the first condition:



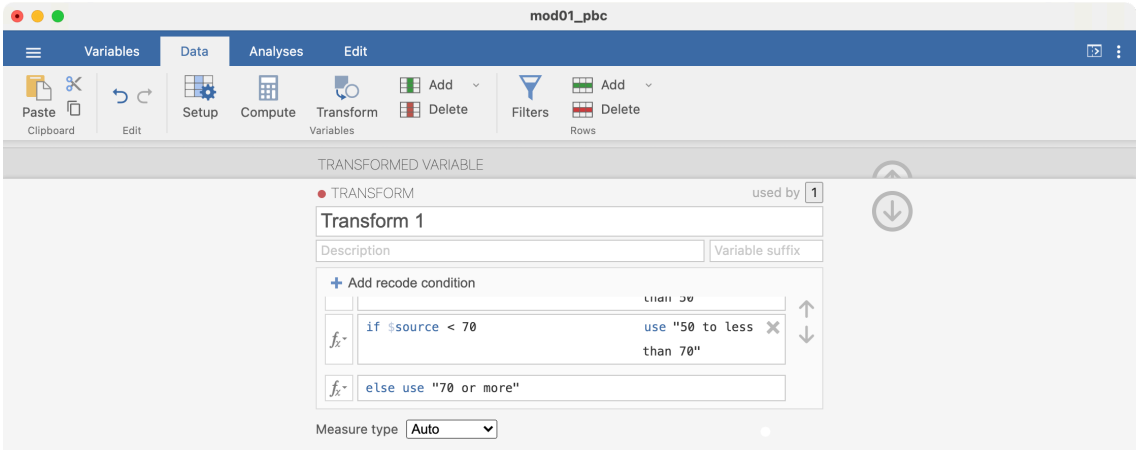
Here we want to define all ages less than 30 as “Less than 30”. Complete the recode condition so it appears as: `if $source < 30 use "Less than 30"`. Note that the quotation marks around “Less than 30” are required:



Add another recode condition, which will be applied if the first condition is not satisfied: `if $source < 50 use "30 to less than 50"`:

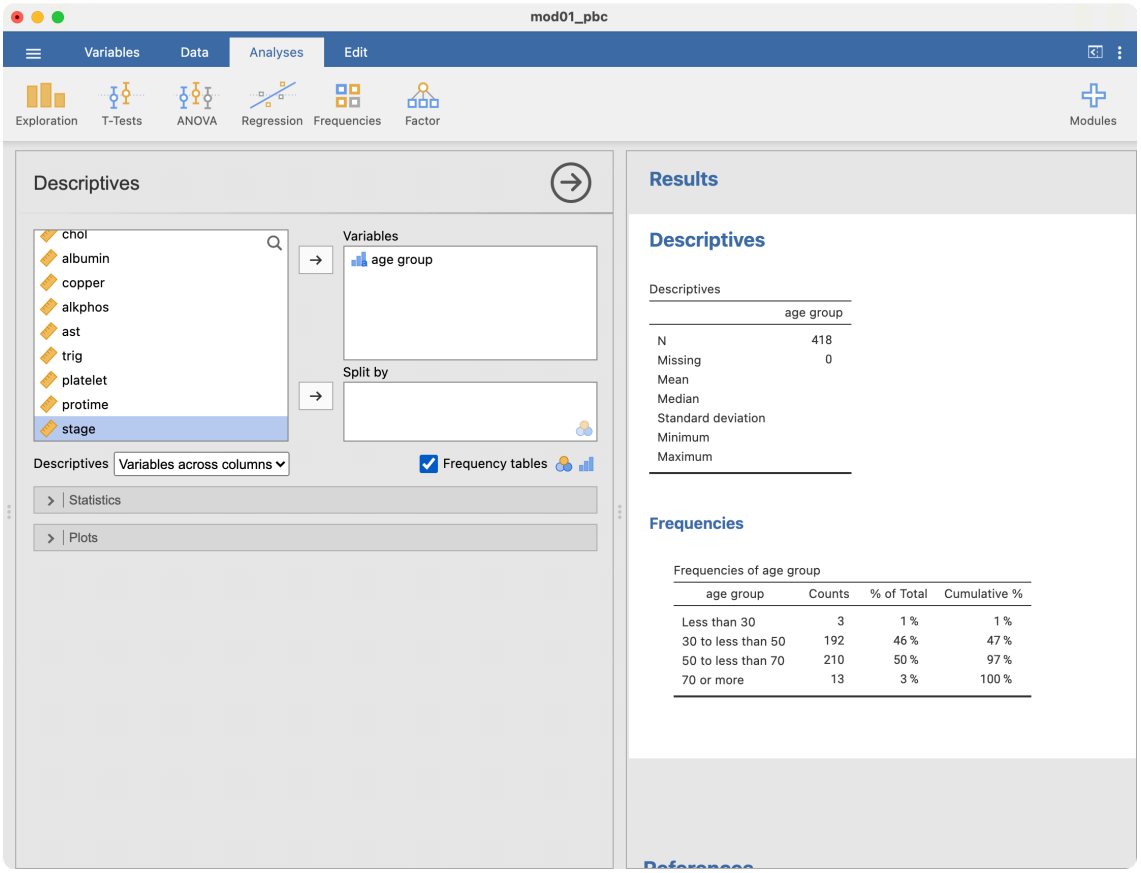


Add another condition: `if $source < 70 use "50 to less than 70"`. There is no need to add a condition for the final condition, simply complete the final line: `else use "70 or more"`:



Finally, click the **down** arrow to dismiss the transform builder, and the **up** arrow to dismiss the transform dialog.

We can examine the new categories by obtaining a frequency table: **Analyses > Exploration > Descriptives**:



2.17 Computing binomial probabilities

jamovi does not have a point-and-click method for computing probabilities from a binomial distribution. Here, instructions are provided for using a third-party applet. This Binomial Distribution Applet has been posted at https://homepage.stat.uiowa.edu/~mbognar/applets/bin.html, and provides a simple and intuitive way to compute probabilities from a binomial distribution.

The applet requires three pieces of information:

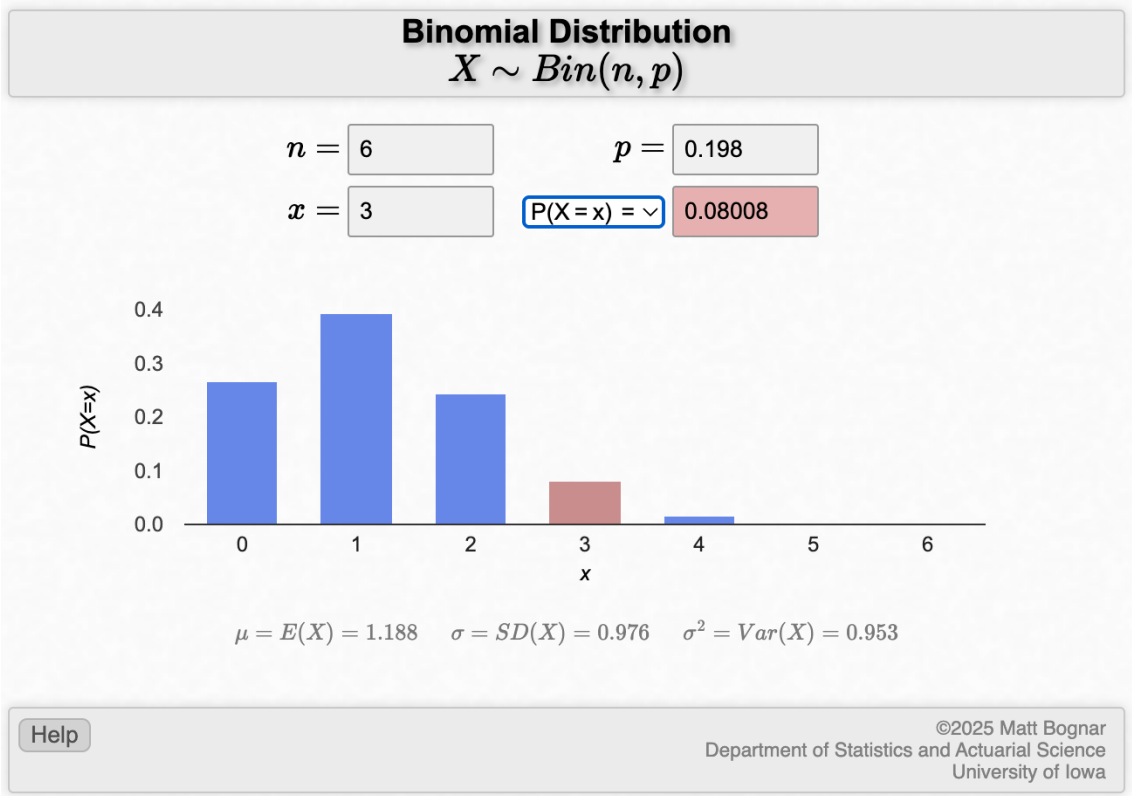
- $/n$: the number of binary trials being considered
- $/p$: the probability of "success" in each trial
- x: the number of success we are interested in

We also need to consider whether we are interested in the probability being equal to, greater than, or less than x.

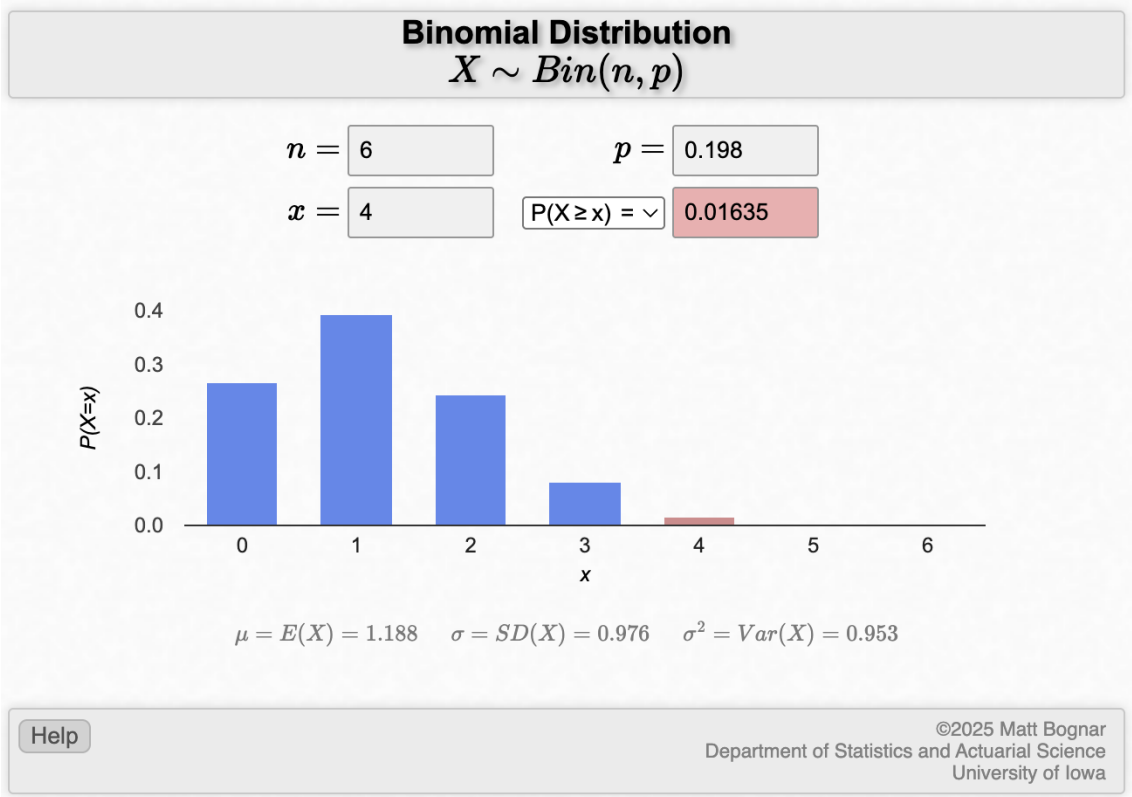
To do the computation for part (a) in Worked Example 2.1:

- *x* is the number of successes, here, the number of smokers (i.e. x=3);
- *n* is the number of trials (i.e. n=6);
- and *p* is probability of drawing a smoker from the population, which is 19.8% (i.e. p=0.198).

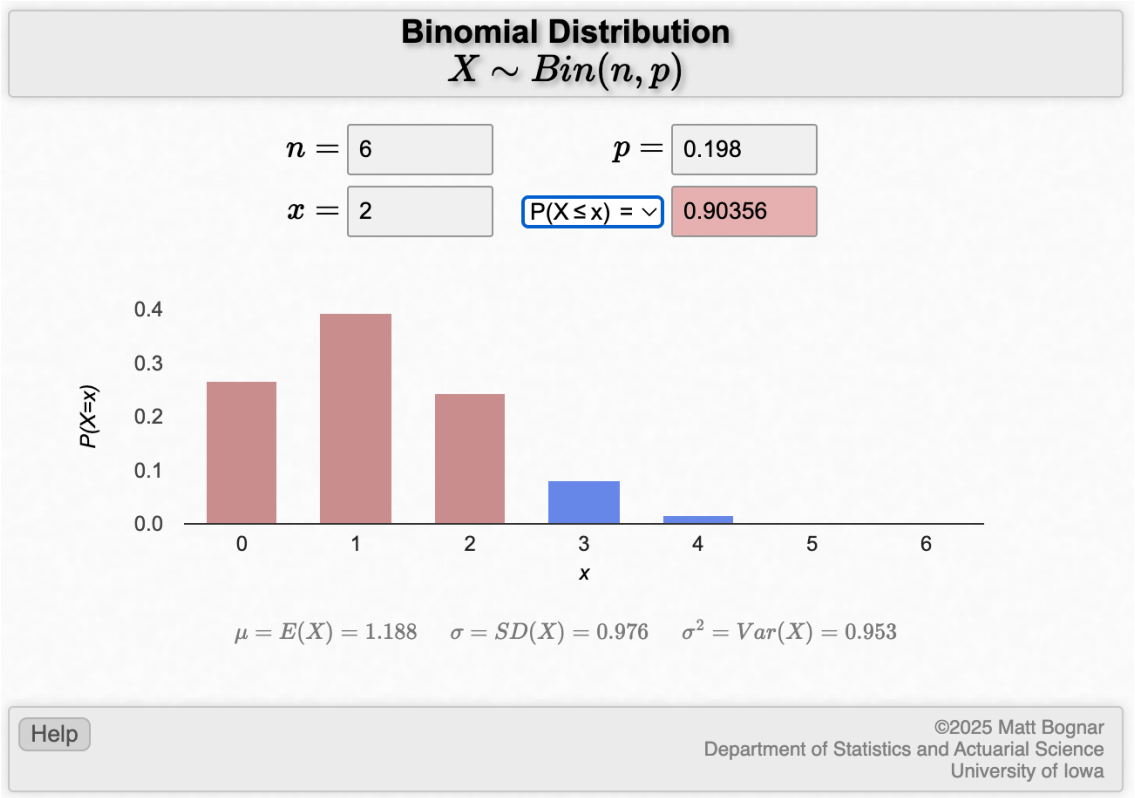
Replace each of these with the appropriate number into the applet:



To calculate the probability of at least 4 smokers in part (b), we change the drop-down to "P(X≥x)", and x to be equal to 4:



To calculate the probability of at most 2 smokers part (c), we we change the drop-down to "P(X≤x)", and x to be equal to 2:


Binomial Distribution
$X \sim Bin(n, p)$
$n =$ 6
$p =$ 0.198
$x =$ 2
P(X ≤ x) =
0.90356
0.4
0.3
0.2
0.1
0.0
P(X=x)
0
1
2
3
4
5
6
x
$\mu = E(X) = 1.188$ $\sigma = SD(X) = 0.976$ $\sigma^2 = Var(X) = 0.953$
Help
©2025 Matt Bognar
Department of Statistics and Actuarial Science
University of Iowa

R notes

Producing a one-way frequency table

We have three categorical variables to summarise in Table 1: sex, stage and vital status. These variables are best summarised using one-way frequency tables, which can be constructed using the `descriptives` function from the `jmv` package, with the `freq = TRUE` option. Before constructing frequency tables however, we *must define the variables as categorical variables*, by converting them to *factors*.

Defining categorical variables as factors

To define a categorical variable as such in R, we define it as a **factor** using the factor function:

```
factor(variable=, levels=, labels=)
```

We specify:

- levels: the values the categorical variable can take
- labels: the labels corresponding to each of the levels (entered in the same order as the levels)

To define our variable `sex` as a factor, we use:

```
pbc <- readRDS("data/activities/mod01_pbc.rds")

pbc$sex <- factor(pbc$sex,
    levels = c(1, 2),
    labels = c("Male", "Female")
)
```

We can then produce a frequency table:

```
descriptives(data = pbc, vars = sex, freq = TRUE)
```

```
 DESCRIPTIVES

 Descriptives

                         sex

   N                     418
   Missing                 0
   Mean
   Median
   Standard deviation
   Minimum
   Maximum
```

```
 FREQUENCIES

 Frequencies of sex

   sex       Counts    % of Total    Cumulative %

   Male          44      10.52632        10.52632
   Female       374      89.47368       100.00000
```

Task: define `stage` and `status` (Vital Status) as factors, and produce one-way frequency tables. Refer to the file `pbc_info.txt` to view the labels for each variable. For example, for Stage:

```
pbc$stage <- factor(pbc$stage,
    levels = c(1, 2, 3, 4),
    labels = c("Stage 1", "Stage 2", "Stage 3", "Stage 4")
)
```

Producing a two-way frequency table

To produce tables summarising two categorical variables, we can use the `contTables()` function within the `jmv` package. The minimal inputs to include are `data`: the name of the data frame to be analysed, `rows`: the variable representing the rows of the table, and `cols`: the name of the columns of the table.

For example, to produce a two-way table showing stage of disease by sex using the pbc data frame, we use:

```
contTables(data = pbc, rows = sex, cols = stage)
```

```
 CONTINGENCY TABLES

 Contingency Tables

   sex       Stage 1    Stage 2    Stage 3    Stage 4    Total

   Male            3          8         16         17       44
   Female         18         84        139        127      368
   Total          21         92        155        144      412


  ²  Tests

         Value        df     p

    ²    0.8779873     3    0.8307365
   N           412
```

[The bottom part of the output, χ^2 Tests, can be ignored for now]

You may notice in the above that the number of observations is now 412. This is because there are missing observations for either sex or stage: which is it, and how would you determine this?

From the cross-tabulation, you can see the individual frequencies of participants in each of the categories in each cell. For example, there are 3 male participants who have Stage 1 disease. You can also read the totals for each row and column. For example, there are 44 males, and 144 participants have Stage 4 disease.

You can also add percentages into your table using `pcCol=TRUE` to include column percents, and `pcRow=TRUE` for row percents. For example, to calculate the relative frequencies (i.e. percentages) of sex within each stage, we would request **column percents** with the option: `pcCol=TRUE`.

```
contTables(data = pbc, rows = sex, cols = stage, pcCol = TRUE)
```

```
 CONTINGENCY TABLES

 Contingency Tables

   sex                        Stage 1      Stage 2      Stage 3      Stage 4      Total

   Male       Observed              3            8           16           17         44
              % within column  14.28571     8.69565     10.32258     11.80556   10.67961

   Female     Observed             18           84          139          127        368
              % within column  85.71429     91.30435    89.67742     88.19444   89.32039

   Total      Observed             21           92          155          144        412
              % within column 100.00000    100.00000    100.00000    100.00000  100.00000


  ² Tests

         Value        df     p

    ²    0.8779873     3    0.8307365
   N           412
```

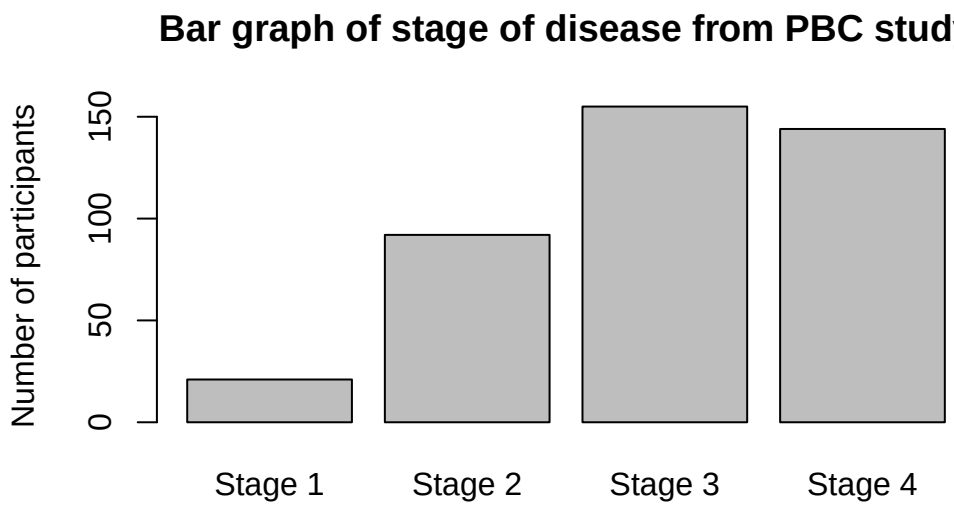
We can see that the 3 male participants with Stage 1 disease represent 14% of those with Stage 1 disease.

2.18 Creating bar charts for one categorical variable

The simplest way to use the `plot()` function is by specifying an object to be plotted. To plot a single variable from a data frame, we must define it using: `dataframe$variable`.

Here we will create the bar chart shown in Figure 2.1 of the statistics notes using the `mod01_pdc.rds` dataset. The x-axis of this graph will be the stage of disease, and the y-axis will show the number of participants in each category.

```
plot(pbc$stage,
    main = "Bar graph of stage of disease from PBC study",
    ylab = "Number of participants"
)
```



Note that stage is a categorical variable, that has been defined as a factor (in Section 2.17). You **must define categorical data as factors** to plot them in a bar graph.

2.19 Creating bar charts for two categorical variables

Option 1: Using the contTables function

Creating a **clustered bar chart** as shown in Figure 2.4 can be done easily using the `contTables` function in the `jmv` package. First create a cross-tab from the variables to be plotted, for example, Stage by Sex:

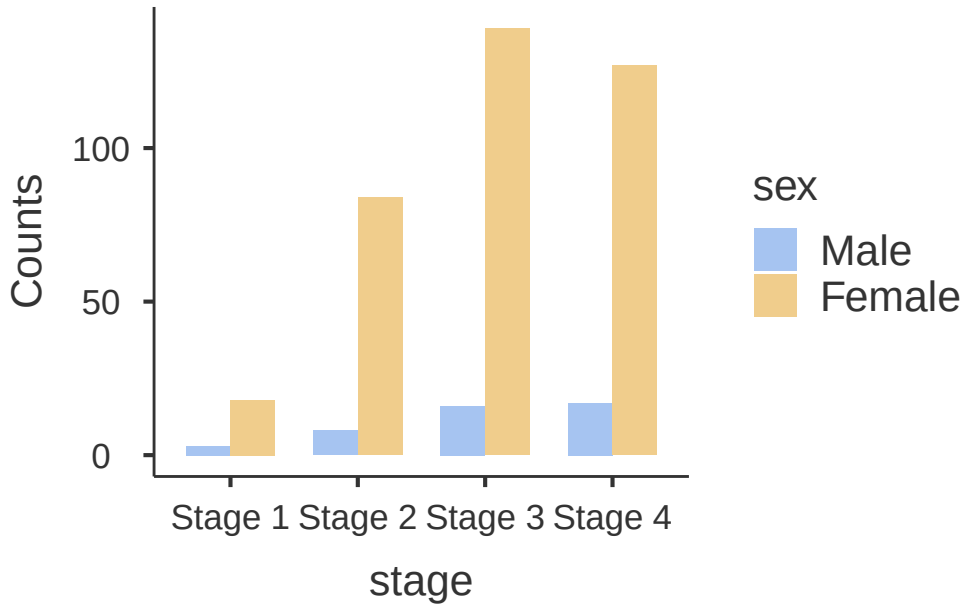
```
contTables(data = pbc, rows = sex, cols = stage)
```

```
 CONTINGENCY TABLES

 Contingency Tables

   sex       Stage 1    Stage 2    Stage 3    Stage 4    Total

   Male            3          8         16         17       44
   Female         18         84        139        127      368
   Total          21         92        155        144      412


 ² Tests

         Value        df     p

   ²     0.8779873     3    0.8307365
   N           412
```

Creating a **clustered bar chart** as shown in Figure 2.4 can be done by requesting a bar chart (`barplot=TRUE`), and the x-axis should be constructed from `stage` - that is, the **column** variable:

```
contTables(pbc,
    rows = sex, cols = stage,
    barplot = TRUE, xaxis = "xcols"
)
```

```
 CONTINGENCY TABLES

 Contingency Tables

   sex       Stage 1    Stage 2    Stage 3    Stage 4    Total

   Male            3          8         16         17       44
   Female         18         84        139        127      368
   Total          21         92        155        144      412


  ² Tests

        Value        df    p

   ²    0.8779873     3    0.8307365
   N           412
```

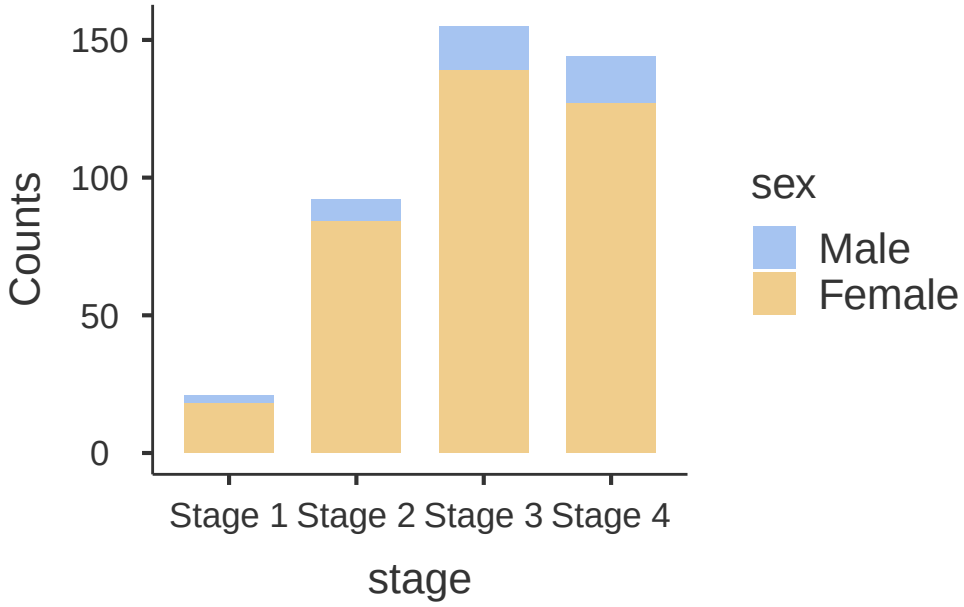


If you want the x-axis to be constructed from the row variable, you would use `xaxis = "xrows"`.

To create a **stacked bar chart** (as in Figure 2.5), specify `bartype` to be `stack`:

```
contTables(pbc,
    rows = sex, cols = stage,
    barplot = TRUE, xaxis = "xcols", bartype = "stack"
)
```

```
CONTINGENCY TABLES

Contingency Tables

  sex       Stage 1    Stage 2    Stage 3    Stage 4    Total

  Male            3          8         16         17       44
  Female         18         84        139        127      368
  Total          21         92        155        144      412


  χ² Tests

        Value        df     p

    χ²    0.8779873     3    0.8307365
  N           412
```



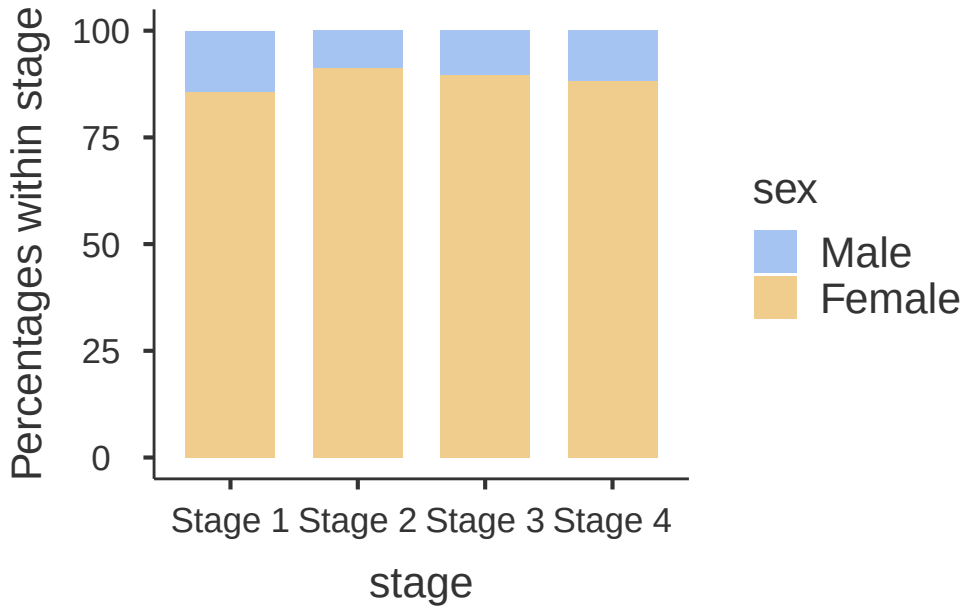
Finally, to create a **stacked relative bar chart** (as in Figure 2.6), specify the y-axis to be a percent (`yaxis="ypc"`), and the percentage be calculated from the columns of the frequency table (`yaxisPc = "column_pc"`):

```
contTables(pbc,
    rows = sex, cols = stage,
    barplot = TRUE, bartype = "stack", xaxis = "xcols",
    yaxis = "ypc", yaxisPc = "column_pc"
)
```

```
CONTINGENCY TABLES

Contingency Tables

  sex       Stage 1    Stage 2    Stage 3    Stage 4    Total
```

```
Male            3          8         16         17       44
Female         18         84        139        127      368
Total          21         92        155        144      412
```

```
² Tests

      Value        df     p

 ²    0.8779873     3    0.8307365
N           412
```



Option 2: Using surveyPlot function

An alternative way of producing barcharts is via a package called `surveymv`. Unfortunately, surveymv is not hosted on the standard package repository, we need to install the package from github.com. This is a straight-forward process, which involves installing a package called `devtools` that allows packages to be installed from alternative locations:

```
install.packages("devtools")
library(devtools)
install_github("raviselker/surveymv")
```

These commands have installed the `surveymv` package. We load the packing using the standard `library()` command:

```
library(surveymv)
```

`surveymv` has only one function: `surveyPlot`, with the following syntax:

```
surveyPlot(
  data = babies,
```

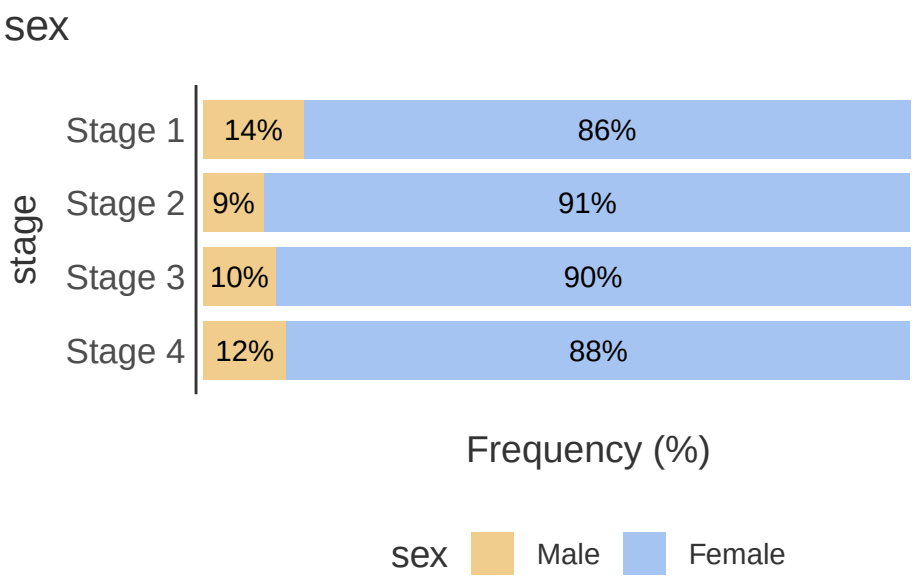
```
  vars = "Birth weight",
  group = "Age group",
  type = "stacked",
  freq = "perc")
```

We specify our data (`data=`), and the main variable to be plotted (`vars=`). If we have a grouping variable, we specify a `group=` variable. We define the chart to be either a stacked (`type = "stacked"`) or clustered (`type = "grouped"`) bar chart, and specify whether to plot frequencies (`freq = "count"`) or percentages (`freq = "perc"`).

To demonstrate, we will recreate Figure 2.6, a figure of the relative frequencies of sex within stage of disease:

```
library(surveymv)

surveyPlot(
  data = pbc,
  vars = "sex",
  group = "stage",
  type = "stacked",
  freq = "perc")
```

```
 SURVEY PLOTS
```



While this produces a bar chart with horizontal bars, it often performs better with labelling of the groups than the previous method.

2.20 Importing data into R

We have described previously how to import data that have been saved as R .rds files. It is quite common to have data saved in other file types, such as Microsoft Excel, or plain text files. In this section, we will demonstrate how to import data from other packages into R.

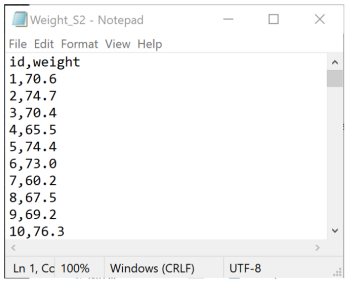
There are two useful packages for importing data into R: `haven` (for data that have been saved by jamovi, SAS or SPSS) and `readxl` (for data saved by Microsoft Excel). Additionally, the `labelled` package is useful in working with data that have been labelled in jamovi.

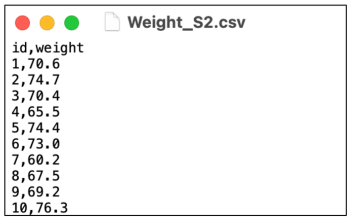
Importing plain text data into R

A `csv` file, or a "comma separated variables" file is commonly used to store data. These files have a very simple structure: they are plain text files, where data are separated by commas. csv files have the advantage that, as they are plain text files, they can be opened by a large number of programs (such as Notepad in Windows, TextEdit in MacOS, Microsoft Excel - even Microsoft Word). While they can be opened by Microsoft Excel, they can be opened by many other programs: the csv file can be thought of as the lingua-franca of data.

In this demonstration, we will use data on the weight of 1000 people entered in a csv file called `mod02_weight_1000.csv` available on Moodle.

To confirm that the file is readable by any text editor, here are the first ten lines of the file, opened in Notepad on Microsoft Windows, and TextEdit on MacOS.





We can use the `read.csv` function:

```
sample <- read.csv("data/examples/mod02_weight_1000.csv")
```

Here, the `read.csv` function has the default that the first row of the dataset contains the variable names. If your data do not have column names, you can use `header=FALSE` in the function.

Note: there is an alternative function `read_csv` which is part of the `readr` package (a component of the `tidyverse`). Some would argue that the `read_csv` function is more appropriate to use because of an issue known as `strings.as.factors`. The `strings.as.factors` default was removed in R Version 4.0.0, so it is less important which of the two functions you use to import a `.csv` file. More information about this issue can be found here and here.

2.21 Recoding data

One task that is common in statistical computing is to recode variables. For example, we might want to group some categories of a categorical variable, or to present a continuous variable in a categorical way.

In this example, we can use the `pbc` data and recode age into age groups:

- Less than 30
- 30 to less than 50
- 50 to less than 70
- 70 or older

The quickest way to recode a continuous variable into categories is to use the `cut` command which takes a continuous variable, and "cuts" it into groups based on the specified "cutpoints"

```
pbc$agegroup <- cut(pbc$age,
    breaks = c(0, 30, 50, 70, 100)
)
```

Notice that some numbers need to be defined for the lowest (age=0) and highest (age=100) bounds: both a lower and upper limit must be defined for each group.

If we examine the new `agegroup` variable:

```
summary(pbc$agegroup)
```

```
  (0,30]  (30,50]  (50,70] (70,100]
       3      192      210       13
```

we see that each group has been labelled in the form of (a, b]. This notation is equivalent to: greater than a, and less than or equal to b. The `cut` function excludes the lower limit, but includes the upper limit. Our age groups have been defined to include the lower limit, and exclude the upper limit (for example, greater than or equal to 30 and less than 50).

We can specify this recoding using the `right=FALSE` option:

```
pbc$agegroup <- cut(pbc$age,
    breaks = c(0, 30, 50, 70, 100),
    right = FALSE
)

summary(pbc$agegroup)
```

```
  [0,30)  [30,50)  [50,70) [70,100)
       3      192      210       13
```

Finally, we can specify labels for the groups using the `labels` option:

```
pbc$agegroup <- cut(pbc$age,
    breaks = c(0, 30, 50, 70, 100),
    right = FALSE,
    labels = c(
        "Less than 30", "30 to less than 50",
        "50 to less than 70", "70 or more"
    )
)

summary(pbc$agegroup)
```

```
      Less than 30 30 to less than 50 50 to less than 70         70 or more
                 3                192                210                 13
```

2.22 Computing binomial probabilities using R

There are two R functions that we can use to calculate probabilities based on the binomial distribution: `dbinom` and `pbinom`:

- `dbinom(x, size, prob)` gives the probability of obtaining `x` successes from `size` trials when the probability of a success on one trial is `prob`;
- `pbinom(q, size, prob)` gives the probability of obtaining `q` **or fewer** successes from `size` trials when the probability of a success on one trial is `prob`;
- `pbinom(q, size, prob, lower.tail=FALSE)` gives the probability of obtaining **more than** `q`successes from `size` trials when the probability of a success on one trial is `prob`.

To do the computation for part (a) in Worked Example 2.1, we will use the `dbinom` function with:

- *x* is the number of successes, here, the number of smokers (i.e. k=3);
- *size* is the number of trials (i.e. n=6);

- and *prob* is probability of drawing a smoker from the population, which is 19.8% (i.e. p=0.198).

Replace each of these with the appropriate number into the formula:

```
dbinom(x = 3, size = 6, prob = 0.198)
```

```
[1] 0.08008454
```

To calculate the upper tail of probability in part (b), we use the `pbinom(lower.tail=FALSE)` function. Note that the `pbinom(lower.tail=FALSE)` function **does not include** `q`, so to obtain 4 or more successes, we need to enter `q=3`:

```
pbinom(q = 3, size = 6, prob = 0.198, lower.tail = FALSE)
```

```
[1] 0.01635325
```

For the lower tail for part (c), we use the `pbinom` function:

```
pbinom(q = 2, size = 6, prob = 0.198)
```

```
[1] 0.9035622
```

Activities

Activity 2.1

Researchers at a maternity hospital in the 1970s conducted a study of low birth weight babies. Low birth weight is classified as a weight of 2500g or less at birth. Data were collected on age and smoking status of mothers and the birth weight of their babies. The file `Activity_2.1.rds` contain data on the participants in the study. The file is located on Moodle in the Learning Activities section.

Create a 2 by 2 table to show the proportions of low birth weight babies born to mothers who smoked during pregnancy and those that did not smoke during pregnancy. Answer the following questions:

a) What was the total number of mothers who smoked during pregnancy?
b) What proportion of mothers who smoked gave birth to low birth weight babies? What proportion of non-smoking mothers gave birth to low birth weight babies?
c) Construct a stacked bar chart of the data to examine if there a difference in the proportion of babies born with a low birth weight in relation to the age group of the mother? Provide appropriate labels for the axes and give the graph an appropriate title. [Hint: plot the data using the `AgeGrp` variable]
d) Using your answers to the question a) and b), write a brief conclusion about the relationship of low birth weight and mother's age and smoking status.

Activity 2.2

In a Randomised Controlled Trial, the preference of a new drug was tested against an established drug by giving both drugs to each of 90 people. Assume that the two drugs are equally preferred, that is, the probability that a patient prefers either of the drugs is equal (50%). Use either the web applet, or one of the binomial functions in R to compute the probability that 60 or more patients would prefer the new drug. In completing this question, determine:

a) The number of trials (`n` for the web applet, `size` for R)
b) The number of successes we are interested in (`x` for web applet, `x` or `q` for R)
c) The probability of success for each trial (`p` for the web applet, `prob` for R)
d) The form of the binomial function
    - for the web applet: `P(X=x)`, `P(X x)` or `P(X x)`;
    - for R: `dbinom`, `pbinom` or `pbinom(lower.tail=FALSE)`
e) The final probability.

Activity 2.3

A case of Schistosomiasis is identified by the detection of schistosome ova in a faecal sample. In patients with a low level of infection, a field technique of faecal examination has a probability of 0.35 of detecting ova in any one faecal sample. If five samples are routinely examined for each patient, use the web applet or R to compute the probability that a patient with a low level of infection:

a) Will not be identified?

b) Will be identified in two of the samples?
c) Will be identified in all the samples?
d) Will be identified in at most 3 of the samples?

Activity 2.4

A health survey was conducted, and an extract of data has been provided in `Activity_2.4-health-survey.csv`. Categorise height into 20cm intervals, and present the height-groups appropriately.

Activity 2.5

The data in the file `Activity_2.5-LengthOfStay.rds` (available on Moodle) has information about **birth weight** and **length of stay** collected from 117 babies admitted consecutively to a hospital for surgery. For each variable:

a. Create a histogram, density plot and boxplot to inspect the distribution of birth weight and length of stay;

b. Complete the following summary statistics for each variable:

   - mean and median;
   - standard deviation and interquartile range.

Make a decision about whether each variable is symmetric or not, and which measure of central tendency and variability should be reported.

Module 3

Continuous probability distributions, sampling and precision

Learning objectives

By the end of this module you will be able to:

- Describe the characteristics of a Normal distribution
- Compute probabilities from a Normal distribution using statistical software
- Briefly outline other types of distributions
- Explain the purpose of sampling, different sampling methods and their implications for data analysis
- Distinguish between standard deviation of a sample and standard error of a mean
- Calculate and interpret confidence intervals for a mean

Optional readings

Kirkwood and Sterne (2001); Chapters 4, 5 and 6. [UNSW Library Link]

Bland (2015); Sections 3.3 and 3.4, Chapter 7, Sections 8.1 to 8.3. [UNSW Library Link]

3.1 Introduction

In this module, we will continue our introduction to probability by considering probability distributions for continuous data. We will introduce one of the most important distributions in statistics: the Normal distribution.

To describe the characteristics of a population we can gather data about the entire population (as is undertaken in a national census) or we can gather data from a sample of the population. When undertaking a research study, taking a sample from a population is far more cost-effective and less time consuming than collecting information from the entire population. When a sample of a population is selected, summary statistics that describe the sample are used to make inferences about the total population from which the sample was drawn. These are referred to as inferential statistics.

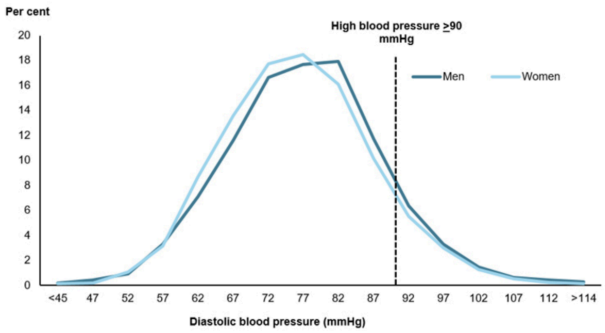
However, for the inferences about the population to be valid, a random sample of the population must be obtained. The goal of using random sampling methods is to obtain a sample that is representative of the target population. In other words, apart from random error, the information derived from the sample is expected to be much the same as the information collected from a complete population census as long as the sample is large enough.

3.2 Probability for continuous variables

Calculating the probability for a categorical random variable is relatively straightforward, as there are only a finite number of possible events. However, there are an infinite number of possible values for a continuous variable, and we calculate the probability that the continuous variable lies in a range of values.

3.3 Normal distribution

The frequency plot for many biological and clinical measurements (for example blood pressure and height) follow a bell shape where the curve is symmetrical about the mean value and has tails at either end. Figure 3.1 [1] and Figure 3.2 [2] demonstrate this type of distribution.



Note: Measured high blood pressure excludes self-reported hypertension prevalence rates. In 2017-18, 31.6% of respondents aged 18 years and over did not have their blood pressure measured. For these respondents, imputation was used to obtain blood pressure. For more information see Appendix 2: Physical measurements in the National Health Survey.

Source: AIHW analysis of ABS 2019. (see Table S3 for footnotes).

Figure 3.1: Distribution of diastolic blood pressure, 2017–18 Australian Bureau of Statistics National Health Survey

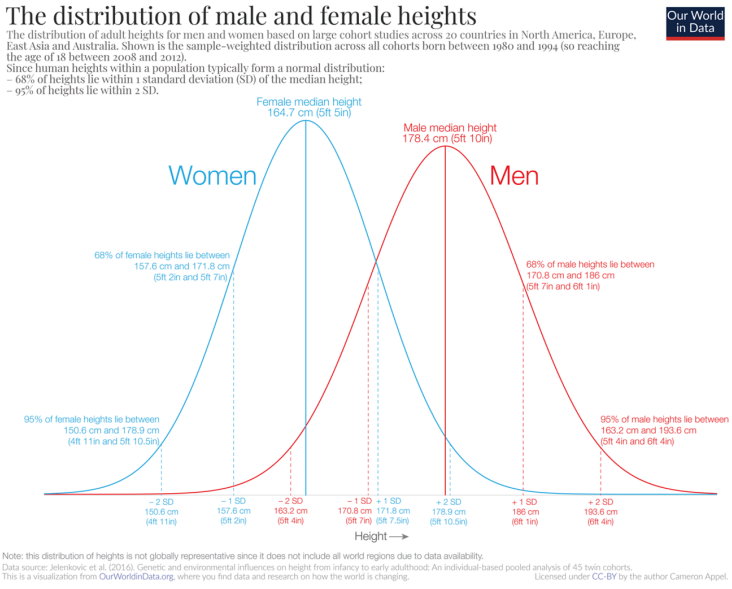


Figure 3.2: Distribution of male and female heights

The Normal distribution, also called the Gaussian distribution (named after Johann Carl Friedrich Gauss, 1777–1855), has been shown to fit the frequency distribution of many naturally occurring variables. It is characterised by its bell-shaped, symmetric curve and its tails that approach zero on either side.

[1] Source: https://www.aihw.gov.au/reports/risk-factors/high-blood-pressure/contents/high-blood-pressure (accessed March 2021)

[2] Source: https://ourworldindata.org/human-height (accessed March 2021)

There are two reasons for the importance of the Normal distribution in biostatistics (Kirkwood and Sterne, 2003). The first is that many variables can be modelled reasonably well using the Normal distribution. Even if the observed data were not Normally distributed, it can often be made reasonably Normal after applying some transformation of the data. The second (and possible most important) reason, is based on the central limit theorem and will be discussed later in this module.

The Normal distribution is characterised by two parameters: the mean (μ) and the standard deviation (σ). The mean defines where the middle of the Normal distribution is located, and the standard deviation defines how wide the tails of the distribution are.

For a Normal distribution, about 68% of the observations lie between $-\sigma$ and σ of the mean; 95% of the observations lie between $-1.96 \times \sigma$ and $1.96 \times \sigma$ from the mean; and almost all the observations (99.7%) lie between $-3 \times \sigma$ and $3 \times \sigma$ (Figure 3.3). Also note that the mean is the same as the median, as the curve is symmetric about its mean.

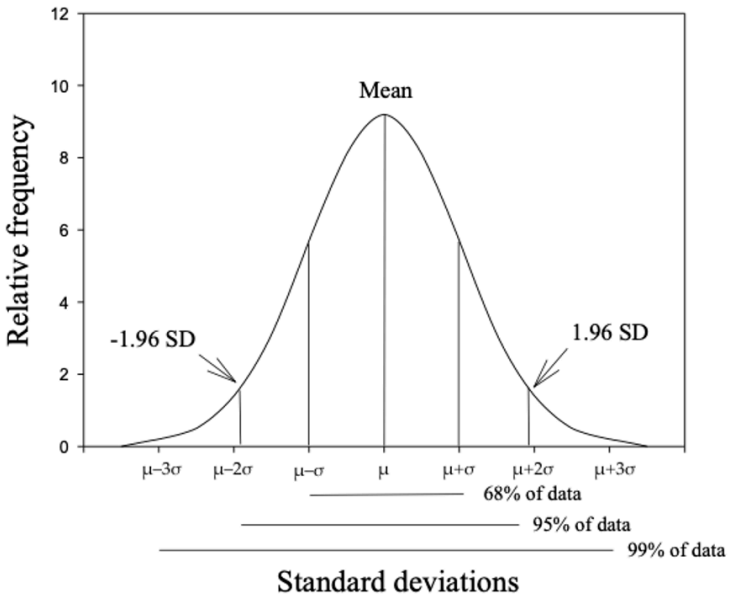


Figure 3.3: Characteristics of a Normal distribution

3.4 The Standard Normal distribution

As each Normal distribution is defined by its mean and standard deviation, there are an infinite number of possible Normal distributions. However, every Normal distribution can be transformed to what we call the Standard Normal distribution, which has a mean of zero ($\mu = 0$) and a standard deviation of one ($\sigma = 1$). The Standard Normal distribution is so important that it has been assigned its own symbol: Z.

Every observation from a Normal distribution X with a mean μ and a standard deviation σ can be transformed to a z-score (also called a Standard Normal deviate) by the formula:

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

The z-score is simply how far an observation lies from the population mean value, scaled by the population standard deviation.

We can use z-scores to estimate probabilities, as shown in Worked Example 2.2.

Worked Example

This example extends the example of diastolic blood pressure shown in Figure 3.1. Assume that the mean diastolic blood pressure for men is 77.9 mmHg, with a standard deviation of 11. What

is the probability that a man selected at random will have high blood pressure (i.e. diastolic blood pressure ≥ 90)?

To estimate the probability that diastolic blood pressure ≥ 90 (i.e. the upper tail probability), we first need to calculate the z-score that corresponds to 90 mmHg.

Using the z-score formula, with x=90, μ=77.9 and σ=11:

$$z = \frac{90 - 77.9}{11} = 1.1$$

Thus, a blood pressure of 90 mmHg corresponds to a z-score of 1.1, or a value 1.1 $\times\sigma$ above the mean weight of the population.

Figure 3.4 shows the probability of a diastolic blood pressure of 90 mmHg or more in the population for a z-score of greater than 1.1 on a Standard Normal distribution.

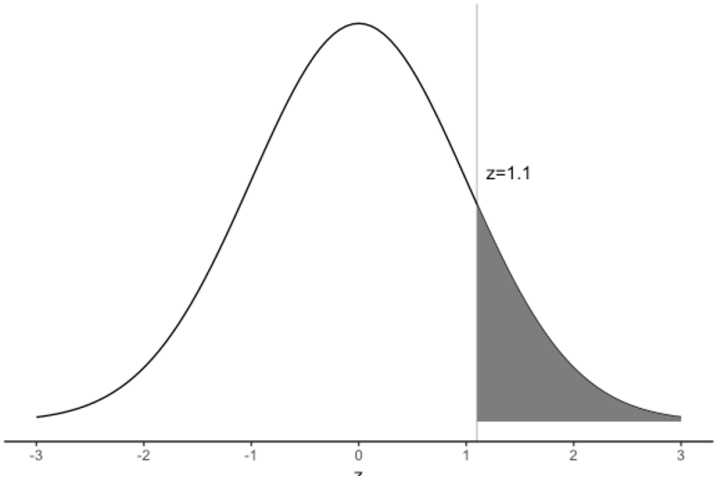


Figure 3.4: Area under the Standard Normal curve (as probability) for Z > 1.1

Using software, we find the probability that a person has a diastolic blood pressure of 90 mmHg or more as P(Z ≥ 1.1) = 0.136.

Apart from calculating probabilities, z-scores are most useful for comparing measurements taken from a sample to a known population distribution. It allows measurements to be compared to one another despite being on different scales or having different predicted values.

For example, if we take a sample of children and measure their weights, it is useful to describe those weights as z-scores from the population weight distribution for each age and gender. Such distributions from large population samples are widely available. This allows us to describe a child's weight in terms of how much it is above or below the population average. For example, if mean weights were compared, children aged 5 years would be on average heavier than the children aged 3 years simply because they are older and therefore larger. To make a valid comparison, we could use the Z-scores to say that children aged 3 years tend to be more overweight than children aged 5 years because they have a higher mean z-score for weight.

3.5 Assessing Normality

There are several ways to assess whether a continuous variable is Normally distributed. The best way to assess whether a variable is Normally distributed is to plot its distribution, using a density plot for example. If the density plot looks appoximately bell-shaped and approximately symmetrical, assuming Normality would be reasonable.

It may be useful to examine a boxplot of a variable in conjunction with a density plot. However a boxplot in isolation is not as useful as a density plot, as a boxplot only indicates whether a variable is distributed symmetrically (indicated by equal "whiskers"). A boxplot cannot give an indication of whether the distribution is bell-shaped, or flat.

For your information: There are formal tests that test for Normality. These tests are beyond the scope of this course and are not recommended.

We can construct a density plot for `age` in the `pbc` data introduced in Module 1. We can see that the density plot is approximately approximately bell-shaped and roughly symmetrical. The mean (50.7 years) and median (51 years) are similar, as would be expected for a Normal distribution. Thus, it would be reasonable to assume that age is Normally distributed in this set of data.

Figure 3.5: Density plot of participant age from PBC study data

3.6 Non-Normally distributed measurements

Not all measurements are Normally distributed, and the symmetry of the bell shape may be distorted by the presence of some very small or very large values. Non-Normal distributions such as this are called skewed distributions.

When there are some very large values, the distribution is said to be positively skewed.This often occurs when measuring variables related to time, such as days of hospital stay, where most patients have short stays (say 1 - 5 days) but a few patients with serious medical conditions have very long lengths of hospital stay (say 20 - 100 days).

In practice, most parametric summary statistics are quite robust to minor deviations from Normality and non-parametric statistical methods are only required when the sample size is small and/or the data are obviously skewed with some influential outliers.

When the data are markedly skewed, density plots are not all bell-shaped. For example, serum bilirubin measured from participants in the PBC study are presented in Figure 3.6.

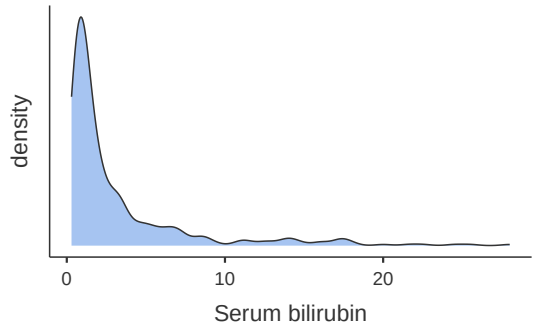


Figure 3.6: Density plot of serum bilirubin from PBC study data

In the plot of Figure 3.6, there is a tail of values to the right, so we would conclude that the distribution is skewed to the right. The mean (3.2 mg/dL) is much larger than the median (1.4 mg/dL), as expected from a skewed distribution.

3.7 Parametric and non-parametric statistical methods

Many statistical methods are based on assumptions about the distribution of the variable – these methods are known as parametric statistical methods. Many methods of statistical inferences based on theoretical sampling properties that are derived from a Normal distribution with the characteristics described above. Thus, it is important that measurements approximate to a Normal distribution before these parametric methods are used. The methods are called 'parametric' because they are based on the parameters – the mean and standard deviation - that underlie a Normal distribution. Statistics which do not assume a particular distribution are called distribution-free statistics, or 'non-parametric statistics'.

In this course, you will learn about both parametric and non-parametric statistical methods. Parametric summary statistical methods include those based on the mean, standard deviation and range (Module 1), and standard error and 95% confidence interval (Module 3). Parametric statistical tests also include t-tests which will be covered in Modules 4 and 5, and correlation and regression described in Module 8.

Non-parametric summary statistical methods are often based on ranks, and may use such statistics as the median, mode and inter-quartile range (Module 1). Non-parametric statistical tests that use ranking are described in Module 9.

3.8 Other types of probability distributions

In this module we have considered a Normal probability distribution and how to use it to measure the precision of continuously distributed measurements. Data also follow other types of distributions which are briefly described below. In other modules in this course, we will be looking at a range of methods to analyse health data and will refer back to these different distributions.

Normal approximation of binomial: When the sample size becomes large, it becomes cumbersome to calculate the exact probability of an event using the binomial distribution. Conveniently, with large sample sizes, the binomial distribution approximates a Normal distribution. The mean and SD of a binomial distribution can be used to calculate the probability of the event as though it was from a Normal distribution.

Poisson distribution: is another distribution which is often used in health research for modelling count data. The Poisson distribution is followed when a number of events happen in a fixed time interval. This distribution is useful for describing data such as deaths in the population in a time period. For example, the number of deaths from breast cancer in one year in women over 50 years old will be an observation from a Poisson distribution. We can also use this to make comparisons of mortality rates between populations.

Many other probability distributions can be derived for functions which arise in statistical analyses but the chi-squared, t and F distributions are the three distributions that are most widely used. These have many applications, some of which are described in later modules.

The chi-squared distribution is a skewed distribution which allows us to determine the probability of a deviation between a count that we observe and a count that we expect for categorical data. One use of this is in conducting statistical tests for categorical data. See Module 7.

A t-distribution is used when the population standard deviation is not known. The t-distribution is appropriate for small samples (<30) and its distribution is bell shaped similar to a Normal distribution but slightly flatter. The t-distribution is useful for comparing mean values. See Module 4 and Module 5.

3.9 Sampling methods

Methods have been designed to select participants from a population such that each person in the target population has an equal probability of being chosen. Methods that use this approach are called random sampling methods. Examples include simple random sampling and stratified random sampling.

In simple random sampling, every person in the population from which the sample is drawn has the same random chance of being selected into the sample. To implement this method, every person in the population is allocated an ID number and then a random sample of the ID numbers is selected. Software packages can be used to generate a list of random numbers to select the random sample.

In stratified sampling, the population is divided into distinct non-overlapping subgroups (strata) according to an important characteristic (e.g. age or sex) and then a random sample is selected from each of the strata. This method is used to ensure that sufficient numbers of people are sampled from each stratum and therefore each subgroup of interest is adequately represented in the sample.

The purpose of using random sampling is to minimise selection bias to ensure that the sample enrolled in a study is representative of the population being studied. This is important because the summary statistics that are obtained can then be regarded as valid in that they can be applied (generalised) back to the population.

A non-representative sample might occur when random sampling is used, simply by chance. However, non-random sampling methods, such as using a study population that does not represent the whole population, will often result in a non-representative sample being selected so that the summary statistics from the sample cannot be generalised back to the population from which the participants were drawn. The effects of non-random error are much more serious than the effects of random error. Concepts such as non-random error (i.e. systematic bias), selection bias, validity and generalisability are discussed in more detail in PHCM9794: Foundations of Epidemiology.

3.10 Standard error and precision

Module 1 introduced the mean, variance and standard deviation as measures of central tendency and spread for continuous measurements from a sample or a population. As described in Module 1, we rarely have data on the entire population but we *infer* information *about* the population from a *sample*. For example, we use the sample mean $\bar{x}$ as an *estimate* of the true population mean μ.

However, a sample taken from a population is usually a small proportion of the total population. If we were to take multiple samples of data and calculate the sample mean for each sample, we would not expect them to be identical. If our samples were very small, we would not be surprised if our estimated sample means were somewhat different from each other. However, if our samples were large, we would expect the sample means to be less variable, i.e. the estimated sample means would be more close to each other, and hopefully, to the true population mean.

The standard error of the mean

A point estimate is a single best guess of the true value in the population - taken from our sample of data. Different samples will provide slightly different point estimates. The standard error is a measure of variability of the point estimate.

In particular, the *standard error of the mean* measures the extent to which we expect the means from different samples to vary because of chance due to the sampling process. This statistic is directly proportional to the standard deviation of the variable, and inversely proportional to the size of the sample. The standard error of the mean for a continuously distributed measurement for which the SD is an accurate measure of spread is computed as follows:

$$\text{SE}(\bar{x}) = \frac{\text{SD}}{\sqrt{n}}$$

Take for example, a set of weights of students attending a university gym in a particular hour. The thirty weights are given below:

Table 3.1: Weight of 30 gym attendees

65.0	70.0	70.0	67.5	65.0	80.0
70.0	72.5	67.5	62.5	67.5	72.5
60.0	65.0	72.5	77.5	75.0	75.0
75.0	70.0	67.5	77.5	67.5	62.5
75.0	62.5	70.0	75.0	72.5	70.0

We can calculate the mean (70.0kg) and the standard deviation (5.04kg). Hence, the standard error of the mean is estimated as:

$$\text{SE}(\bar{x}) = \frac{5.04}{\sqrt{30}} = 0.92$$

Because the calculation uses the sample size (n) (i.e. the number of study participants) in the denominator, the SE will become smaller when the sample size becomes larger. A smaller SE indicates that the estimated mean value is more precise.

The standard error is an important statistic that is related to sampling variation. When a random sample of a population is selected, it is likely to differ in some characteristic compared with another random sample selected from the same population. Also, when a sample of a population is taken, the true population mean is an unknown value.

Just as the standard deviation measures the spread of the data around the population mean, the standard error of the mean measures the spread of the sample means. Note that we do not have different samples, only one. It is a theoretical concept which enables us to conduct various other statistical analyses.

3.11 Central limit theorem

Even though we now have an estimate of the mean and its standard error, we might like to know what the mean from a different random sample of the same size might be. To do this, we need to know how sample means are distributed. In determining the form of the probability distribution of the sample mean ($\bar{x}$), we consider two cases:

When the population distribution is unknown:

The central limit theorem for this situation states:

> In selecting random samples of size n from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean $\bar{x}$ approaches a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size becomes large.

The sample size n = 30 and above is a rule of thumb for the central limit theorem to be used. However, larger sample sizes may be needed if the distribution is highly skewed.

When the population is assumed to be normal:

> In this case the sampling distribution of $\bar{x}$ is normal for any sample size.

3.12 95% confidence interval of the mean

Earlier, we showed that the characteristics of a Standard Normal Distribution are that 95% of the data lie within 1.96 standard deviations from the mean (Figure 3.2). Because the central limit theorem states that the sampling distribution of the mean is approximately Normal in large enough samples, we expect that 95% of the mean values would fall within 1.96 × SE units above and below the measured mean population value.

For example, if we repeated the study on weight 100 times using 100 different random samples from the population and calculated the mean weight for each of the 100 samples, approximately 95% of the values for the mean weight calculated for each of the 100 samples would fall within 1.96 × SE of the population mean weight.

This interpretation of the SE is translated into the concept of precision as a 95% confidence interval (CI). A 95% CI is a range of values within which we have 95% confidence that the true population mean lies. If an experiment was conducted a very large number of times, and a 95%CI was calculated for each experiment, 95% of the confidence intervals would contain the true population mean.

The calculation of the 95% CI for a mean is as follows:

$$\bar{x} \pm 1.96 \times \text{SE}(\bar{x})$$

This is the generic formula for calculating 95% CI for any summary statistic. In general, the mean value can be replaced by the point estimate of a rate or a proportion and the same formula applies for computing 95% CIs, i.e.

$$95\% \text{ CI} = \text{point estimate} \pm 1.96 \times \text{SE}(\text{point estimate})$$

The main difference in the methods used to calculate the 95% CI for different point estimates is the way the SE is calculated. The methods for calculating 95% CI around proportions and other ratio measures will be discussed in Module 6.

The use of 1.96 as a general critical value to compute the 95% CI is determined by sampling theory. For the confidence interval of the mean, the critical value (1.96) is based on normal distribution (true when the population SD is known). However, in practice, statistical packages will provide slightly different confidence intervals because they use a critical value obtained from the t-distribution. The t-distribution approaches a normal distribution when the sample size approaches infinity, and is close to a normal distribution when the sample size is ≥30.The critical values obtained from the t-distribution are always larger than the corresponding critical value from the normal distribution. The difference gets smaller as the sample size becomes larger. For example, when the sample size n=10, the critical value from the t-distribution is 2.26 (rather than 1.96); when n= 30, the value is 2.05; when n=100, the value is 1.98; and when n=1000, the critical value is 1.96.

The critical value multiplied by SE (for normal distribution, 1.96 × SE) is called the maximum likely error for 95% confidence.

The t-distribution and when should I use it?

The population standard deviation (σ) is required for calculation of the standard error. Usually, σ is not known and the sample standard deviation (s) is used to estimate it. It is known, however, that the sample standard deviation of a normally distributed variable underestimates the true value of σ, particularly when the sample size is small.

Someone by the pseudonym of Student came up with the Student's t distribution with $(n-1)$ degrees of freedom to account for this underestimation. It looks very much like the standardised normal distribution, only that it has fatter tails (Figure 3.7). As the degrees of freedom increase (i.e. as n increases), the t-distribution gradually approaches the standard normal distribution. With a sufficiently large sample size, the Student's t-distribution closely approximates the standardised normal distribution.

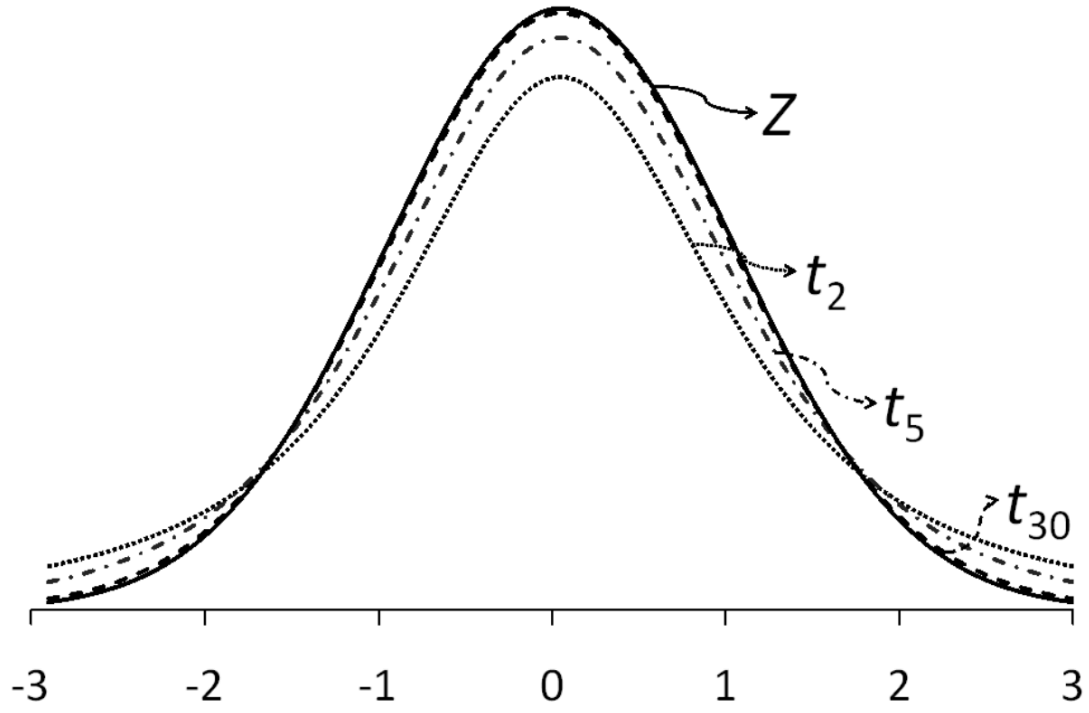


Figure 3.7: The normal (Z) and the student's t-distribution with 2, 5 and 30 degrees of freedom

If a variable X is normally distributed and the population standard deviation σ is known, using the normal distribution is appropriate. However, if σ is not known then one should use the student t-distribution with $(n-1)$ degrees of freedom.

Worked Example 3.1: 95% CI of a mean using individual data

The diastolic blood pressure of 733 female Pima indigenous Americans was measured, and a density plot showed that the data were approximately normally distributed. The mean diastolic blood pressure in the sample was 72.4 mmHg with a standard deviation of 12.38 mmHg. These data are saved as `mod03_blood_pressure.csv`.

Use Jamovi or R, we can calculate the mean, its Standard Error, and the 95% confidence interval:

Table 3.2: Summary of blood pressure from female Pima indigenous Americans

n	Mean	Standard deviation	Standard error of the mean	95% confidence interval of the mean
733	72.4	12.38	0.46	71.5 to 73.3

We can interpret this confidence interval as: we are 95% confident that the true mean of female Pima indigenous Americans lies between 71.5 and 73.3 mmHg.

Worked Example 3.2: 95% CI of a mean using summarised data

The publication of a study using a sample of 242 participants reported a sample mean systolic blood pressure of 128.4 mmHg and a sample standard deviation of 19.56 mmHg. Find the 95% confidence interval for the mean systolic blood pressure.

Using jamovi or R, we obtain a 95% confidence interval from 125.9232 to 130.8768.

We are 95% confident that the true mean systolic blood pressure of the population from which the sample was drawn lies between 125.9 kg and 130.9 mmHg.

Jamovi notes

3.13 Generating new variables

We commonly need to create new variables based on existing variables in our data. For example, body size is often summarised using the **Body Mass Index (BMI)**. BMI is calculated as: $\frac{\text{weight (kg)}}{\text{height (m)}^2}$.

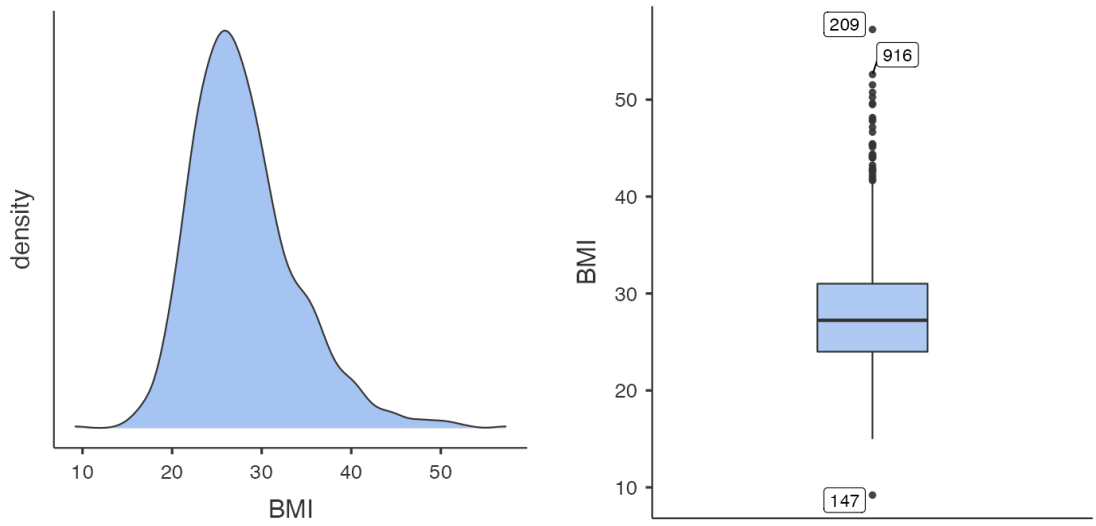
In this demonstration, we will import a selection of records from a large health survey, stored in the file `mod03_health_survey.xlsx`. The health survey data contains 1140 records, comprising:

- sex: 1 = respondent identifies as male; 2 = respondent identifies as female
- height: height in meters
- weight: weight in kilograms

To generate a new variable, we use **Data > Compute**:

1. click **Data** to open the spreadsheet, and click into an empty column
2. click **Setup** then **NEW COMPUTED VARIABLE**
3. enter the name of the new variable, here `BMI`
4. in the formula (*f_x) box enter: `weight / height^2` (note: ^2 represents "to the power of 2", or "squared")

You should check the construction of any new variable buy examining a density plot and/or a boxplot:

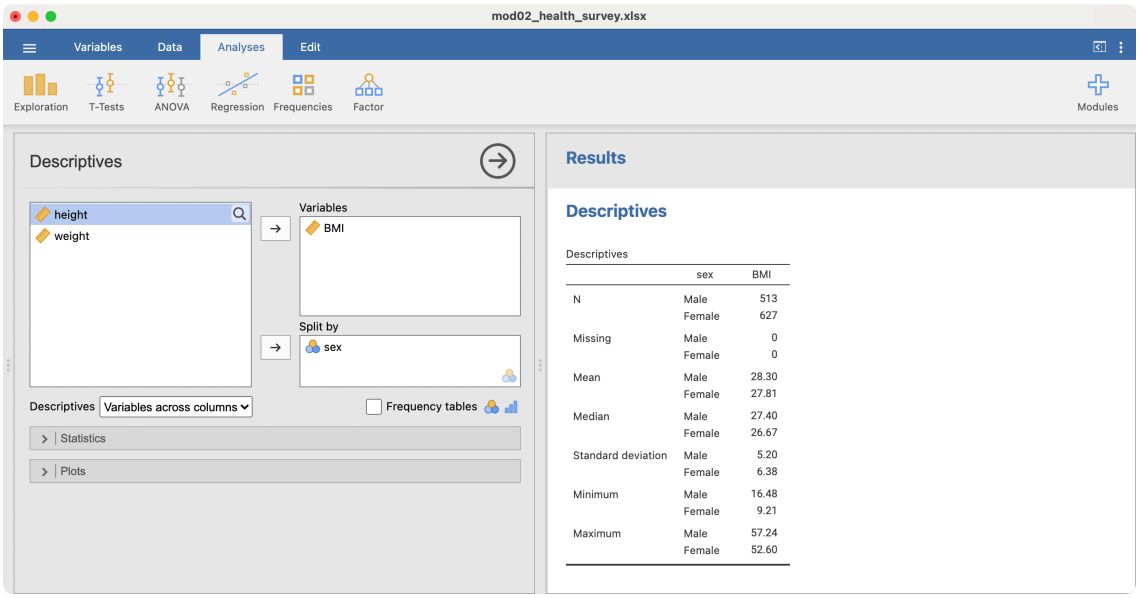


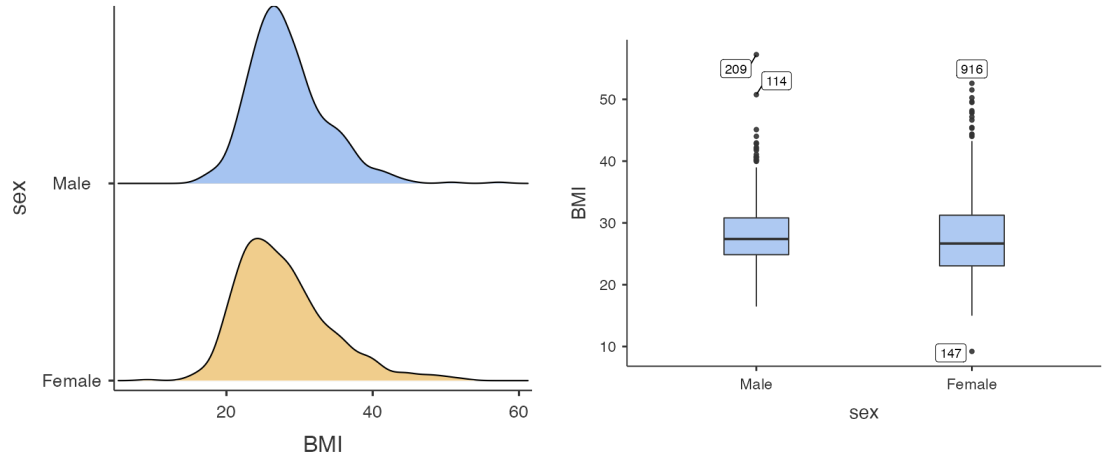
In the general population, BMI ranges between about 15 to 30. It appears that BMI has been correctly generated in this example, perhaps with some unusual values that might require investigation.

3.14 Summarising data by another variable

We will often want to calculate the same summary statistics by another variable. For example, we might want to calculate summary statistics for BMI for males and females separately. We can do this in jamovi by defining sex as a *by-variable*.

This can be done easily in jamovi by defining a **Split by** variable. For example, to summarise BMI for males and females separately, we use `sex` as the Split by variable:





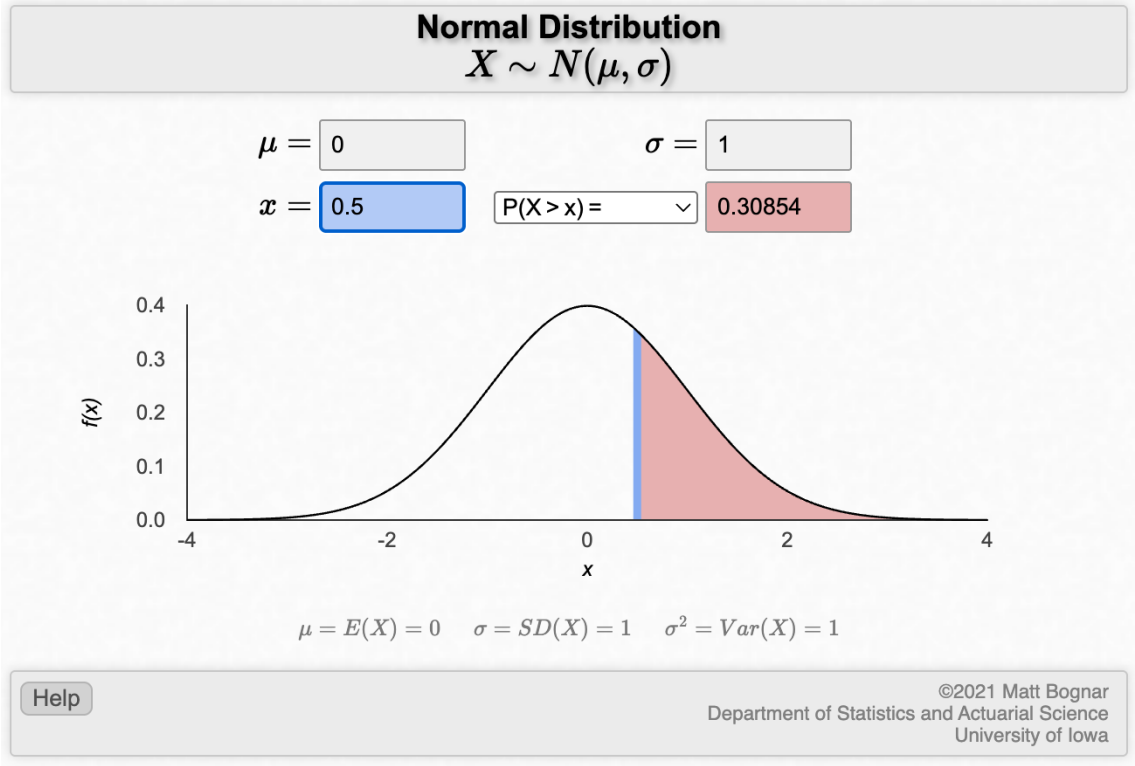
3.15 Computing probabilities from a Normal distribution

jamovi does not have a point-and-click method for computing probabilities from a Normal distribution. Here, instructions are provided for using a third-party applet. This Normal Distribution Applet has been posted at https://homepage.stat.uiowa.edu/~mbognar/applets/normal.html, and provides a simple and intuitive way to compute probabilities from a Normal distribution. The applet requires three pieces of information:

- μ: the mean of the Normal distribution being considered
- σ: the standard deviation of the Normal distribution being considered
- x: the value being considered

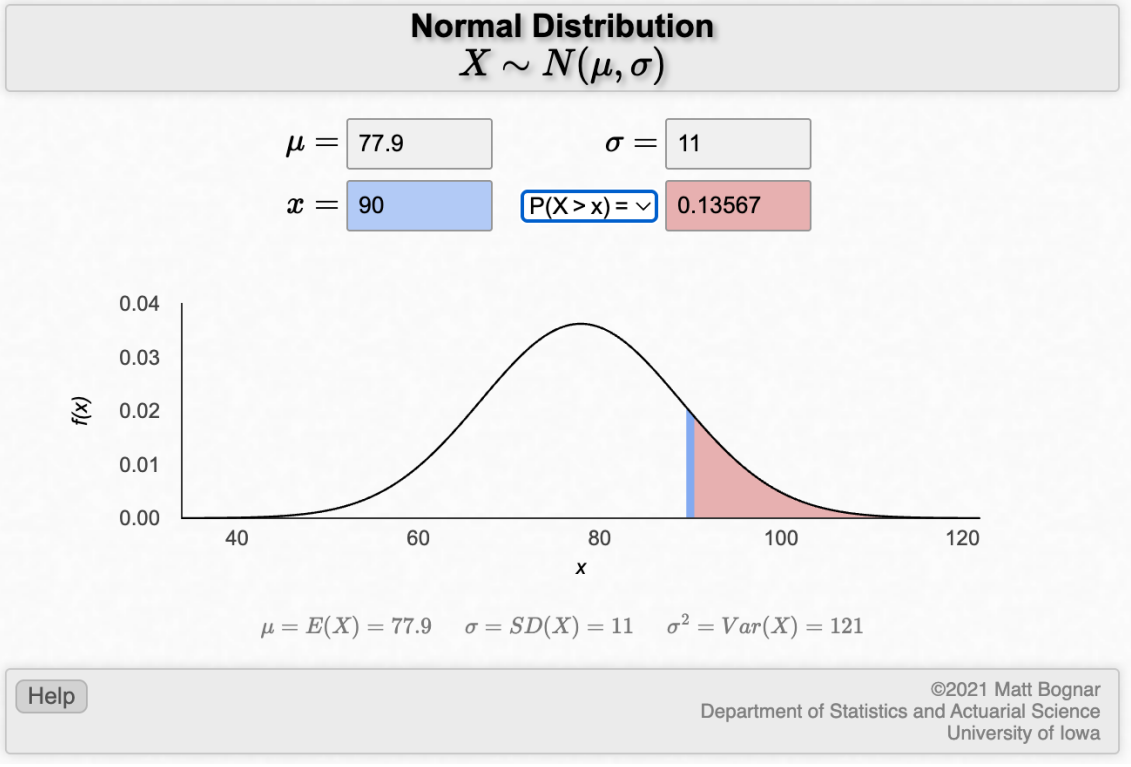
We also need to consider whether we are interested in the probability being greater than x, or less than x.

For example, to obtain the probability of obtaining 0.5 or greater from a standard normal (i.e. $/mu$=0, $/sigma$=1) distribution:



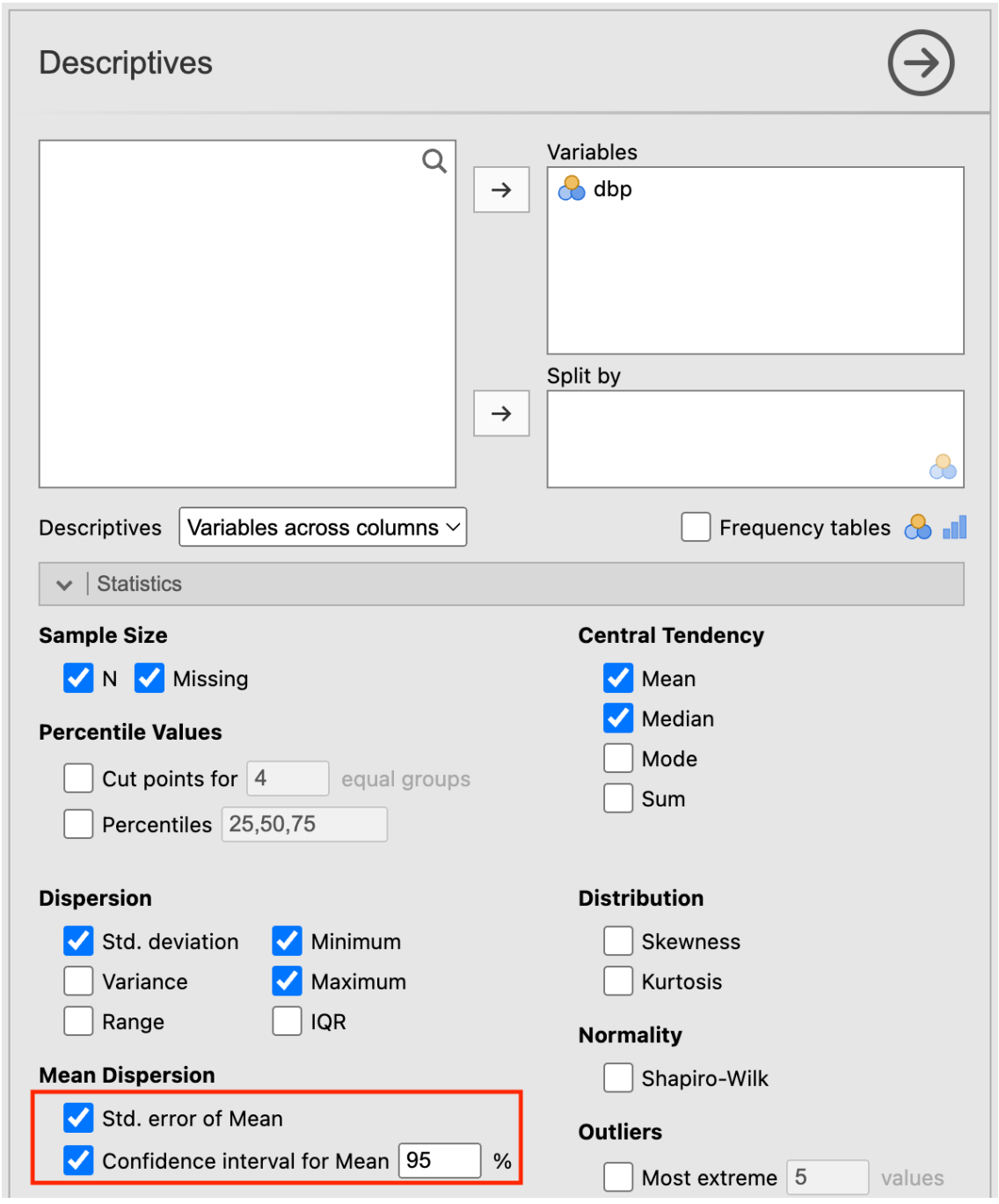
The Normal curve of interest is shaded, and the probability is provided as 0.30854.

To calculate the worked example: Assume that the mean diastolic blood pressure for men is 77.9 mmHg, with a standard deviation of 11. What is the probability that a man selected at random will have high blood pressure (i.e. diastolic blood pressure greater than or equal to 90)?



3.16 Calculating a 95% confidence interval of a mean: Individual data

To demonstrate the computation of the 95% confidence interval of a mean, we can use the data from `mod03_blood_pressure.csv`. We can use **Exploration > Descriptives** to calculate the mean, its standard error and the 95% confidence interval for the mean. Choose **dbp** as the analysis variable, and select **Std. error of Mean** and **Confidence interval for Mean** in the **Statistics** section:



The descriptives output appears:

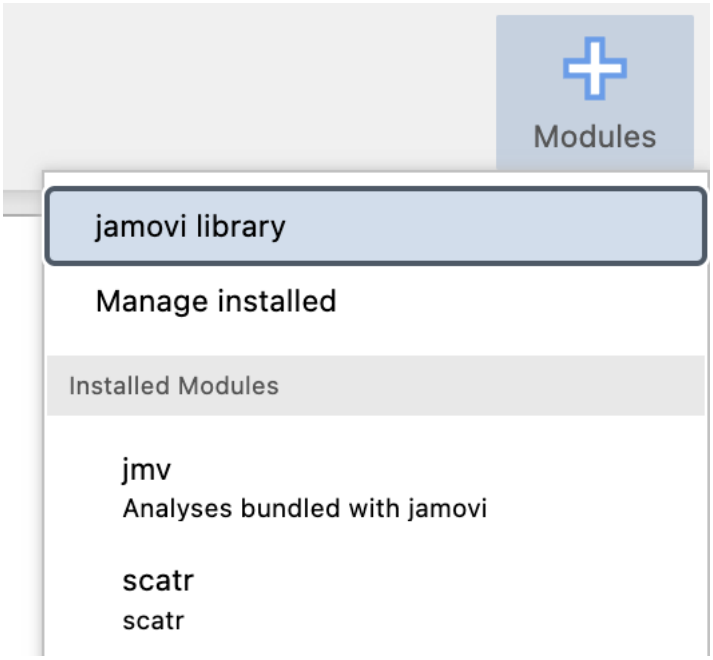
Descriptives

	dbp
N	733
Missing	35
Mean	72.41
Std. error mean	0.46
95% CI mean lower bound	71.51
95% CI mean upper bound	73.30
Median	72
Standard deviation	12.38
Minimum	24
Maximum	122

Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom

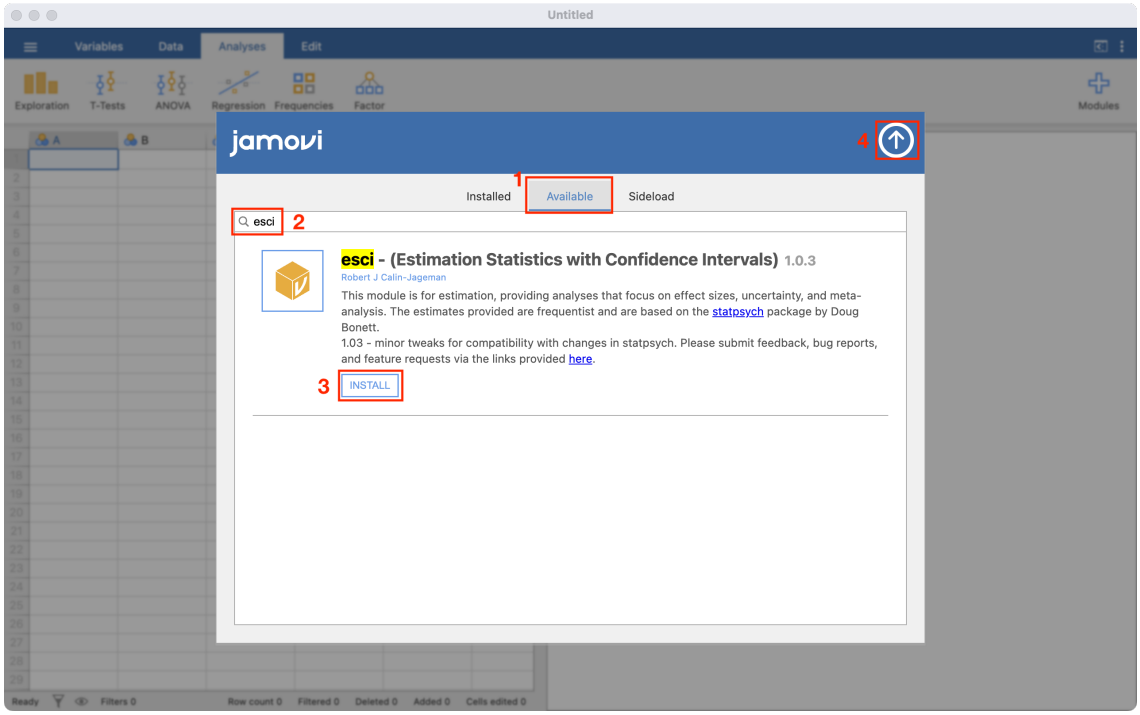
3.17 Calculating a 95% confidence interval of a mean: Summarised data

For Worked Example 3.2 where we are given the sample mean, sample standard deviation and sample size, we need to install a new Jamovi module, called **esci**. To install a new module, click the large **+ Modules** button on the right-hand side of the Jamovi window, and then choose **jamovi library**:

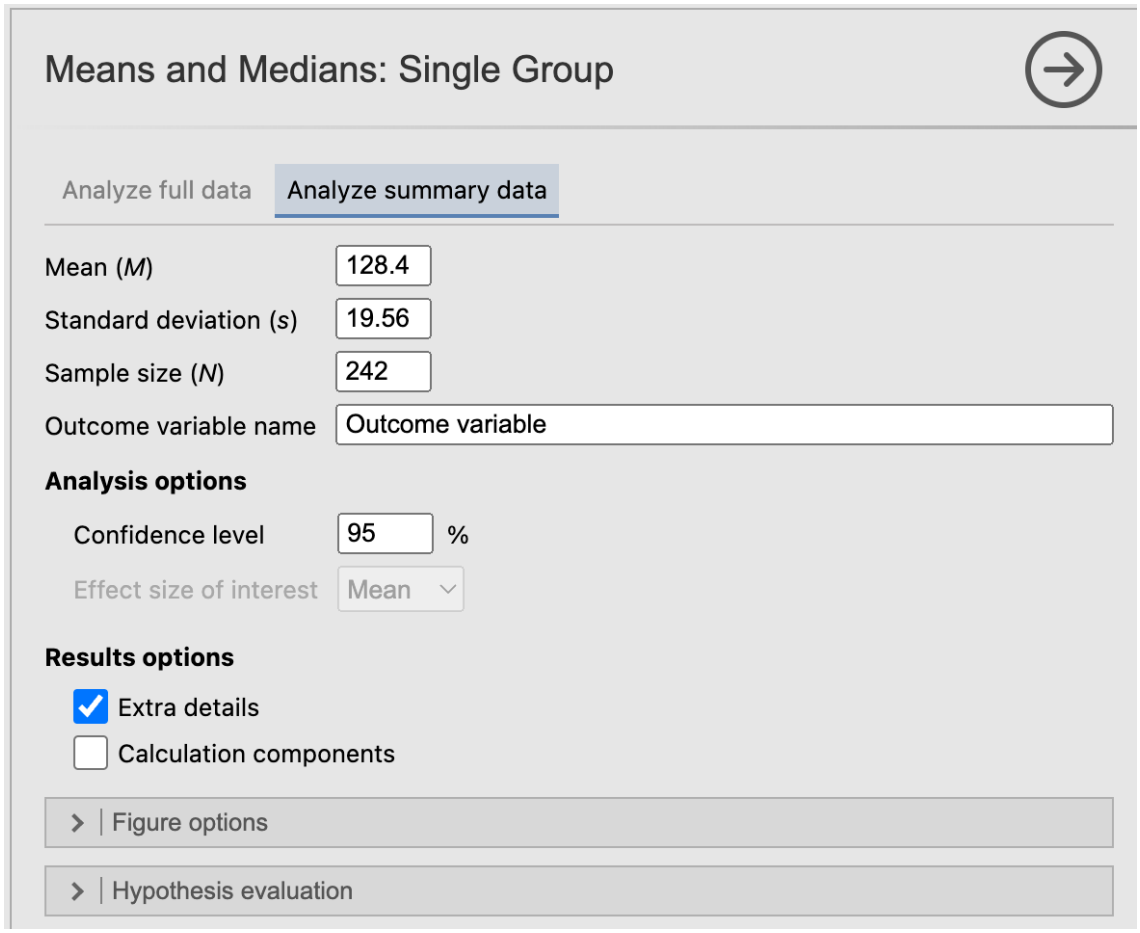


To install a new module:

1 - Ensure that the middle tab, **Available** is selected; 2 - Type **esci** in the search bar. The **esci** module will appear; 3 - Click **INSTALL** to install the module 4 - Click the up-arrow to exit from the Install Module window



To calculate the 95% confidence interval, choose **esci > Means and Medians > Single Group**. Select the **Analyze summary data** tab, and enter the known information: here `128.4` as the **Mean**, `19.56` as the **Standard deviation** and `242` as the **Sample size**. Choose **Extra details** to obtain the Standard Error of the mean:



The 95% confidence interval is listed as the lower limit (LL) and the upper limit (UL):

Means and Medians: Single Group

Overview

Outcome variable	*M*	95% CI		*MoE*	SE_{Mean}	*s*	*N*	*df*
		LL	UL					
Outcome variable	128.40	125.92	130.88	2.48	1.26	19.56	242	241

R notes

3.18 Importing Excel data into R

Another common type of file that data are stored in is a Microsoft Excel file (.xls or .xlsx). In this demonstration, we will import a selection of records from a large health survey, stored in the file `mod03_health_survey.xlsx`.

The health survey data contains 1140 records, comprising:

- sex: 1 = respondent identifies as male; 2 = respondent identifies as female
- height: height in meters
- weight: weight in kilograms

To import data from Microsoft Excel, we can use the `read_excel()` function in the `readxl` package.

```
library(readxl)

survey <- read_excel("data/examples/mod03_health_survey.xlsx")
summary(survey)
```

```
      sex           height          weight      
 Min.   :1.00   Min.   :1.220   Min.   : 22.70  
 1st Qu.:1.00   1st Qu.:1.630   1st Qu.: 68.00  
 Median :2.00   Median :1.700   Median : 79.40  
 Mean   :1.55   Mean   :1.698   Mean   : 81.19  
 3rd Qu.:2.00   3rd Qu.:1.780   3rd Qu.: 90.70  
 Max.   :2.00   Max.   :2.010   Max.   :213.20  
```

We can see that sex has been entered as a numeric variable. We should transform it into a factor so that we can assign labels to each category:

```
survey$sex <- factor(survey$sex, level=c(1,2), labels=c("Male", "Female"))

summary(survey$sex)
```

```
  Male Female 
   513    627 
```

We also note that height looks like it has been entered as meters, and weight as kilograms.

3.19 Generating new variables

Our health survey data contains information on height and weight. We often summarise body size using BMI: body mass index which is calculated as: $\frac{\text{weight (kg)}}{(\text{height (m)})^2}$

We can create a new column in our data frame in many ways, such as using the following approach:

```
dataframe$new_column <- <formula>
```

For example:

```
survey$bmi <- survey$weight / (survey$height^2)
```

We should check the construction of the new variable by examining some records. The `head()` and `tail()` functions list the first and last 6 records in any dataset.

```
head(survey)
```

```
# A tibble: 6 x 4
  sex    height weight   bmi
  <fct>   <dbl>  <dbl> <dbl>
1 Male     1.63   81.7  30.8
2 Male     1.63   68    25.6
3 Male     1.85   97.1  28.4
4 Male     1.78   89.8  28.3
5 Male     1.73   70.3  23.5
6 Female   1.57   85.7  34.8
```

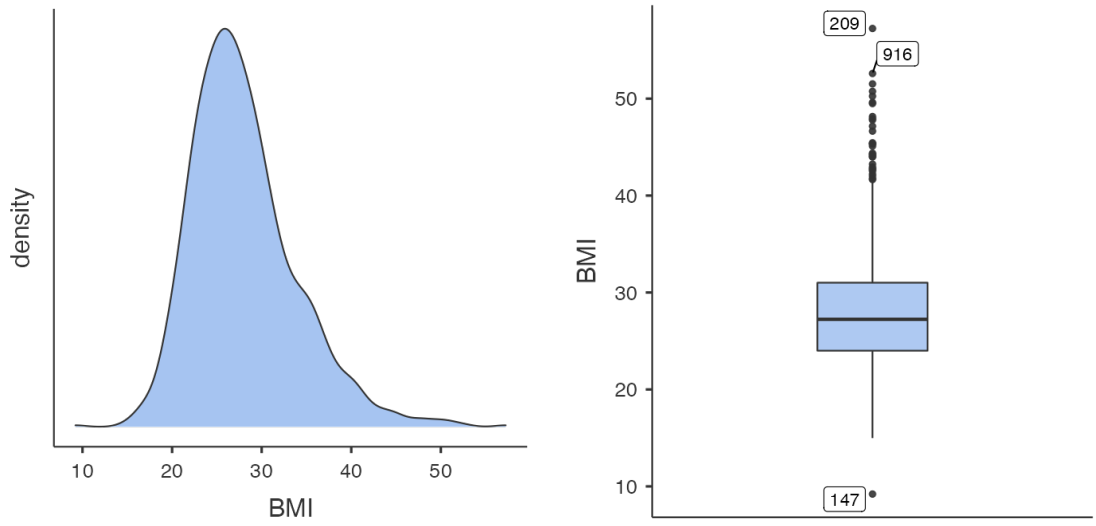
```
tail(survey)
```

```
# A tibble: 6 x 4
  sex    height weight   bmi
  <fct>   <dbl>  <dbl> <dbl>
1 Female   1.65   95.7  35.2
2 Male     1.8    79.4  24.5
3 Female   1.73   83    27.7
4 Female   1.57   61.2  24.8
5 Male     1.7    73    25.3
6 Female   1.55   91.2  38.0
```

We should also check the construction of any new variable buy examining a density plot and/or a boxplot:

```
descriptives(data=survey, vars=bmi, dens=TRUE, box=TRUE)
```

In the general population, BMI ranges between about 15 to 30. It appears that BMI has been correctly generated in this example. We should investigate the very low and some of the very high values of BMI, but this will be left for another time.

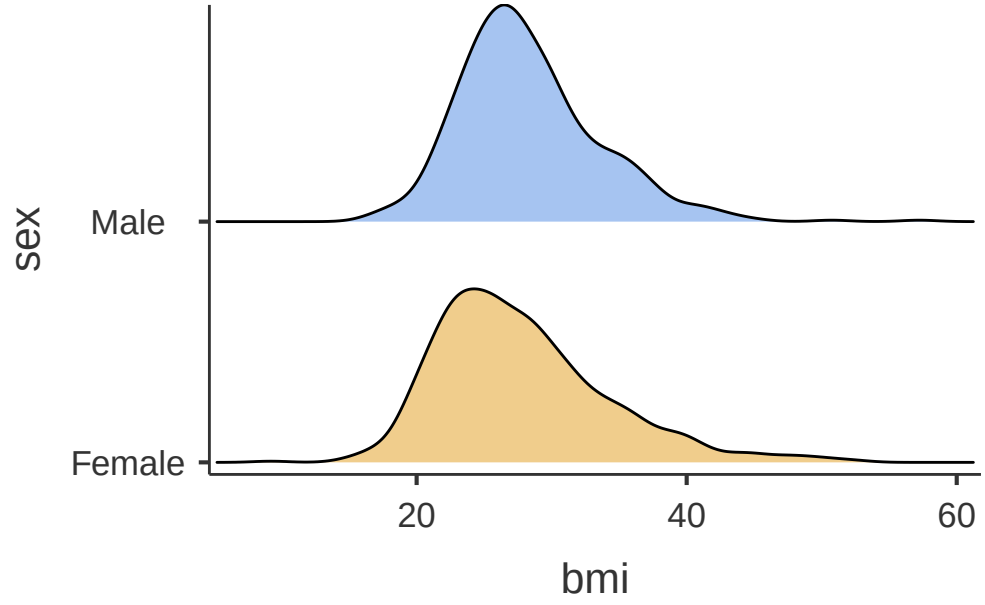


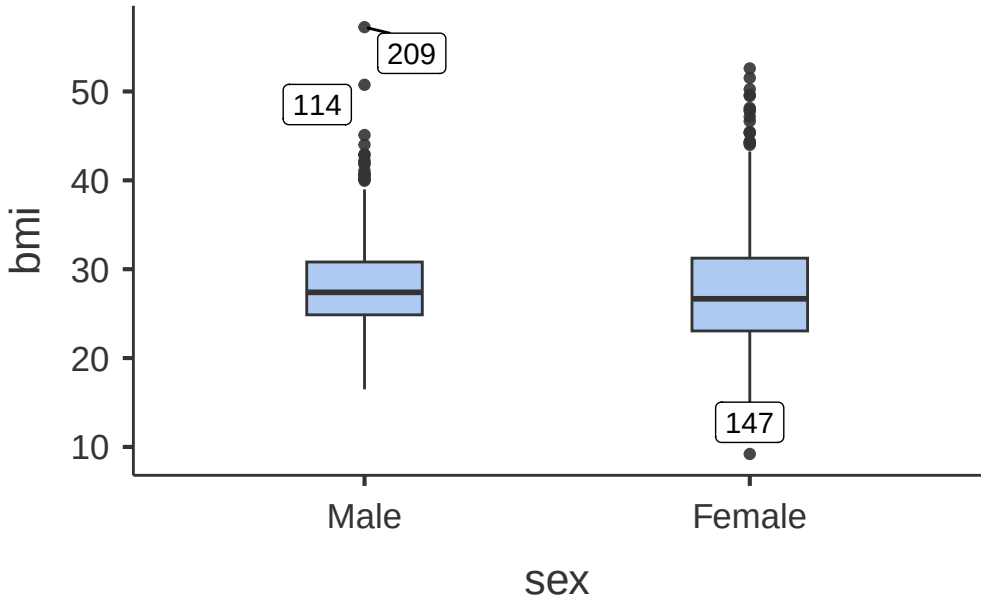
3.20 Summarising data by another variable

We will often want to calculate the same summary statistics by another variable. For example, we might want to calculate summary statistics for BMI for males and females separately. We can do this in in the `descriptives` function by defining sex as a `splitBy` variable:

```
library(jmv)
descriptives(data=survey, vars=bmi, splitBy = sex, dens=TRUE, box=TRUE)
```

```
 DESCRIPTIVES

 Descriptives

                          sex       bmi

   N                      Male          513
                          Female        627
   Missing                Male            0
                          Female          0
   Mean                   Male      28.29561
                          Female    27.81434
   Median                 Male      27.39592
                          Female    26.66667
   Standard deviation     Male      5.204975
                          Female    6.380523
   Minimum                Male      16.47519
                          Female    9.209299
   Maximum                Male      57.23644
                          Female    52.59516
```





3.21 Summarising a single column of data

In Module 1, we started with a very simple analysis: reading in six ages, and them using `summary()` to calculate descriptive statistics. We then went on to use the `decriptives()` function in the `jmv` package as more flexible way of calculating descriptive statistics. Let's revisit this analysis:

```
# Author: Timothy Dobbins
# Date: 5 April 2022
# Purpose: My first R script
library(jmv)

age <- c(20, 25, 23, 29, 21, 27)
```

```
# Use "summary" to obtain descriptive statistics
summary(age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   21.50   24.00   24.17   26.50   29.00
```

```
# Use the "descriptives" function from jmv to obtain descriptive statistics
descriptives(age)
```

```
Error: Argument 'data' must be a data frame
```

The `summary()` function has worked correctly, but the `descriptives()` function has given an error: `Error: Argument 'data' must be a data frame`. What on earth is going on here?

The error gives us a clue here - the `descriptives()` function requires a data frame for analysis. We have provided the object `age`: a **vector**. As we saw in Section 1.12, a vector is a single column of data, while a data frame is a collection of columns.

In order to summarise a vector using the `descriptives()` function, we must first convert the vector into a data frame using `as.data.frame()`. For example:

```
# Author: Timothy Dobbins
# Date: 5 April 2024
# Purpose: My first R script
library(jmv)

age <- c(20, 25, 23, 29, 21, 27)

# Use "summary" to obtain descriptive statistics
summary(age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   21.50   24.00   24.17   26.50   29.00
```

```
# Create a new data frame from the vector age:

age_df <- as.data.frame(age)

# Use "descriptives" to obtain descriptive statistics for age_df
descriptives(age_df)
```

```
 DESCRIPTIVES

 Descriptives

                        age

   N                      6
   Missing                0
   Mean            24.16667
   Median          24.00000
   Standard deviation  3.488075
   Minimum         20.00000
   Maximum         29.00000
```

3.22 Computing probabilities from a Normal distribution

We can use the `pnorm` function to calculate probabilities from a Normal distribution:

- `pnorm(q, mean, sd)` calculates the probability of observing a value of `q` or less, from a Normal distribution with a mean of `mean` and a standard deviation of `sd`. Note that if `mean` and `sd` are not entered, they are assumed to be 0 and 1 respectively (i.e. a standard normal distribution.)
- `pnorm(q, mean, sd, lower.tail=FALSE)` calculates the probability of observing a value of more than `q`, from a Normal distribution with a mean of `mean` and a standard deviation of `sd`.

To obtain the probability of obtaining 0.5 or greater from a standard normal distribution:

```
pnorm(0.5, lower.tail = FALSE)
```

```
[1] 0.3085375
```

To calculate the worked example: Assume that the mean diastolic blood pressure for men is 77.9 mmHg, with a standard deviation of 11. What is the probability that a man selected at random will have high blood pressure (i.e. diastolic blood pressure greater than or equal to 90)?

```
pnorm(90, mean = 77.9, sd = 11, lower.tail = FALSE)
```

```
[1] 0.1356661
```

3.23 Calculating a 95% confidence interval of a mean: individual data

To demonstrate the computation of the 95% confidence interval of a mean, we can use the data from `mod03_blood_pressure.csv`:

```
pima <- read.csv("data/examples/mod03_blood_pressure.csv")
```

We can examine the data set using the `summary` command:

```
summary(pima)
```

```
      dbp
 Min.   : 24.00
 1st Qu.: 64.00
 Median : 72.00
 Mean   : 72.41
 3rd Qu.: 80.00
 Max.   :122.00
```

The mean and its 95% confidence interval can be obtained many ways in R. We will use the `descriptives()` function within the `jmv` package to calculate the standard error of the mean, and a confidence interval, by including `se = TRUE` and `ci = TRUE`:

```
library(jmv)
descriptives(data=pima, vars=dbp, se=TRUE, ci=TRUE)
```

```
 DESCRIPTIVES

 Descriptives

                              dbp

   N                              733
   Missing                          0
   Mean                      72.40518
   Std. error mean           0.4573454
   95% CI mean lower bound   71.50732
   95% CI mean upper bound   73.30305
   Median                         72
   Standard deviation        12.38216
   Minimum                        24
   Maximum                       122

   Note. The CI of the mean assumes
   sample means follow a
   t-distribution with N - 1 degrees
   of freedom
```

3.24 Calculating a 95% confidence interval of a mean: summarised data

For Worked Example 3.2 where we are given the sample mean, sample standard deviation and sample size. R does not have a built-in function to calculate a confidence interval from summarised data, but we can write our own.

Note: writing your own functions is beyond the scope of this course. You should copy and paste the code provided to do this.

```
### Copy this section
ci_mean <- function(n, mean, sd, width=0.95, digits=3){
  lcl <- mean - qt(p=(1 - (1-width)/2), df=n-1) * sd/sqrt(n)
  ucl <- mean + qt(p=(1 - (1-width)/2), df=n-1) * sd/sqrt(n)

  print(paste0(width*100, "%", " CI: ", format(round(lcl, digits=digits), nsmall = digits),
               " to ", format(round(ucl, digits=digits), nsmall = digits) ))

}
### End of copy

ci_mean(n=242, mean=128.4, sd=19.56, width=0.95)


[1] "95% CI: 125.923 to 130.877"

ci_mean(n=242, mean=128.4, sd=19.56, width=0.99)


[1] "99% CI: 125.135 to 131.665"
```

Activities

Activity 3.1

An investigator wishes to study people living with agoraphobia (fear of open spaces). The investigator places an advertisement in a newspaper asking for volunteer participants. A total of 100 replies are received of which the investigator randomly selects 30. However, only 15 volunteers turn up for their interview.

1. Which of the following statements is true?

a) The final 15 participants are likely to be a representative sample of the population available to the investigator
b) The final 15 participants are likely to be a representative sample of the population of people with agoraphobia
c) The randomly selected 30 participants are likely to be a representative sample of people with agoraphobia who replied to the newspaper advertisement
d) None of the above

2. The basic problem confronted by the investigator is that:

a) The accessible population might be different from the target population
b) The sample has been chosen using an unethical method
c) The sample size was too small
d) It is difficult to obtain a sample of people with agoraphobia in a scientific way

Activity 3.2

A dental epidemiologist wishes to estimate the mean weekly consumption of sweets among children of a given age in her area. After devising a method which enables her to determine the weekly consumption of sweets by a child, she conducted a pilot survey and found that the standard deviation of sweet consumption by the children per week is 85 gm (assuming this is the population standard deviation, σ). She considers taking a random sample for the main survey of:

- 25 children, or
- 100 children, or
- 625 children or
- 3,000 children.

a) Estimate the standard error of the sample mean for each of these four sample sizes.
b) What happens to the standard error as the sample size increases? What can you say about the precision of the sample mean as the sample size increases?

Activity 3.3

The dataset for this activity is the same as the one used in Activity 1.4 in Module 1. The file is `Activity_1.4.rds` on Moodle.

a) Plot a histogram of diastolic BP and describe the distribution.
b) Use jamovi or R to obtain an estimate of the mean, standard error of the mean and the 95% confidence interval for the mean diastolic blood pressure.
c) Interpret the 95% confidence interval for the mean diastolic blood pressure.

Activity 3.4

Suppose that a random sample of 81 newborn babies delivered in a hospital located in a poor neighbourhood during the last year had a mean birth weight of 2.7 kg and a standard deviation of 0.9 kg. Calculate the 95% confidence interval for the unknown population mean. Interpret the 95% confidence interval.

Activity 3.5

Using the health survey data (`Activity_3.5.xlsx`) described in the computing notes of this module, create a new variable, BMI, which is equal to a person's weight (in kg) divided by their height (in metres) squared (i.e. BMI $= \frac{\text{weight (kg)}}{[\text{height (m)}]^2}$. Categorise BMI using the categories:

- Underweight: BMI < 18.5
- Normal weight: 18.5 $\leq$ BMI < 25
- Overweight: 25 $\leq$ BMI < 30
- Obese: BMI $\geq$ 30

Note: BMI does not necessarily reflect body fat distribution or describe the same degree of fatness in different individuals. However, at a population level, BMI is a practical and useful measure for identifying overweight or obesity. [3]

Create a two-way table to display the distribution of BMI categories by sex (sex: 1 = respondent identifies as male; 2 = respondent identifies as female). Does there appear to be a difference in categorised BMI between males and females?

Activity 3.6

The data set of hospital stay data for 1323 hypothetical patients is available on Moodle in csv format (`Activity_3.6.csv`). Import this dataset into jamovi or R. There are two variables in this dataset:

- female: female=1; male=0
- los: length of stay in days

a) Use jamovi or R to examine the distribution of length of stay: overall; and separately for females and males. Comment on the distributions.
b) Use jamovi or R to calculate measures of central tendency for hospital stay to obtain information about the average duration of hospital stay. Which summary statistics should you report and why? Report the appropriate statistics of the spread and measure of central tendency chosen.
c) Calculate the measures of central tendency for hospital duration separately for males and females. What can you conclude from comparing these measures for males and females?

Activity 3.7

If weights of men are Normally distributed with a population mean μ = 87, and a population standard deviation, σ = 8 kg:

a) What is the probability that a man will weigh 95 kg or more? Draw a Normal curve of the area represented by this probability in the population (i.e. with μ = 87 kg and σ = 8 kg).

[3] https://www.aihw.gov.au/reports/overweight-obesity/overweight-and-obesity/contents/measuring-overweight-and-obesity

b) What is the probability that a man will weigh more than 75 kg but less than 95 kg? Draw the area represented by this probability on a Normal curve.

Module 4

An introduction to hypothesis testing

Learning objectives

By the end of this module you will be able to:

- Formulate a research question as a hypothesis;
- Understand the concepts of a hypothesis test;
- Consider the difference between statistical significance and clinical importance;
- Use 95% confidence intervals to conduct an informal hypothesis test;
- Perform and interpret a one-sample t-test;
- Explain the concept of one and two tailed statistical tests.

Optional readings

Kirkwood and Sterne (2001); Chapter 8. [UNSW Library Link]

Bland (2015); Sections 9.1 to 9.7; Sections 10.1 and 10.2. [UNSW Library Link]

4.1 Introduction

In earlier modules, we examined sampling and how summary statistics can be used to make inferences about a population from which a sample is drawn. In this module, we introduce hypothesis testing as the basis of the statistical tests that are important for reporting results from research and surveillance studies, and that you will be learning in the remainder of this course.

We use hypothesis testing to answer questions such as whether two groups have different health outcomes or whether there is an association between a treatment and a health outcome. For example, we may want to know:

- whether a safety program has been effective in reducing injuries in a factory, i.e. whether the frequency of injuries in the group who attended a safety program is lower than in the group who did not receive the safety program;
- whether a new drug is more effective in reducing blood pressure than a conventional drug, i.e. whether the mean blood pressure in the group receiving the new drug is lower than the mean blood pressure in the group receiving the conventional medication;
- whether an environmental exposure increases the risk of a disease, i.e. whether the frequency of disease is higher in the group who have been exposed to an environmental factor than in the non-exposed group.

We may also want to know something about a single group. For example, whether the mean blood pressure of a sample is the same as the general population.

These questions can be answered by setting up a null hypothesis and an alternative hypothesis, and performing a hypothesis test (also known as a significance test).

4.2 Hypothesis testing

Hypothesis testing is a statistical technique that is used to quantify the evidence against a null hypothesis. A null hypothesis (H_0) is a statement that there is no difference in a summary statistic between groups. For example, a null hypothesis may be stated as follows:

H_0 = there is no difference in mean systolic blood pressure between a group taking a conventional drug and a group taking a newly developed drug

We also have an alternative hypothesis that is opposite or contrasting to the null hypothesis. In our example above, the alternative hypothesis for the above null hypothesis is that there is a difference between groups. The alternative hypothesis is usually of most interest to the researcher but in practice, formal statistical tests are used to test the null hypothesis (not the alternative hypothesis). The hypotheses are always in reference to the population from which the sample is drawn, not the sample itself.

After setting up our null and alternative hypotheses, we use the data to generate a test statistic. The particular test statistic differs depending on the type of data being analysed (e.g. continuous or categorical), the study design (e.g. paired or independent) and the question being asked.

The test statistic is then compared to a known distribution to calculate the probability of observing a test statistic which is as large or larger than the observed test statistic, if the null hypothesis was true. The probability is known as the P-value.

Informally, the P-value can be interpreted as the probability of observing data like ours, or more extreme, if the null hypothesis was true.

If the P-value is small, it is unlikely that we would observe data like ours or more extreme if the null hypothesis was true. In other words, our data are not consistent with the null hypothesis, and we conclude that we have evidence against the null hypothesis. If the P-value is not small, the probability of observing data like ours or more extreme is not unlikely. We therefore have little or no evidence against the null hypothesis. In hypothesis testing, the null hypothesis cannot be proven or accepted; we can only find evidence to refute the null hypothesis.

To summarise:

- a small P-value gives us evidence against the null hypothesis;
- a P-value that is not small provides little or no evidence against null hypothesis;
- the smaller the P-value, the stronger the evidence against the null hypothesis.

Historically, a value of 0.05 has been used as a cut-point for finding evidence against the null hypothesis. A P-value less than 0.05 would be interpreted as “statistically significant”, and would allow us to “reject the null hypothesis”. A P-value greater than 0.05 would be interpreted as “not significant”, and we would “fail to reject the null hypothesis”. This arbitrary dichotomy is overly simplistic, and a more nuanced view is now recommended. Recommended interpretations for P-values are given in Table 4.1.

Table 4.1: Interpretation of P-values

P-value	Strength of evidence
<0.001	Very strong evidence
0.001 to <0.01	Strong evidence
0.01 to <0.05	Evidence
0.05 to <0.1	Weak evidence
≥0.1	Little or no evidence

P-values are usually generated using statistical software although other methods such as statistical tables or Excel functions can be used to generate test statistics and determine the P-value. In traditional statistics, the probability level was described as a lower-case p but in many journals today, probability is commonly described by upper case P. Both have the same meaning.

4.3 Effect size

In hypothesis testing, P-values convey only part of the information about the hypothesis and need to be accompanied by an estimation of the effect size, that is, a description of the magnitude of the difference between the study groups. The effect size is a summary statistic that conveys the size of the difference between two groups. For continuous variables, it is usually calculated as the difference between two mean values.

If the variable is binary, the effect size can be expressed as the absolute difference between two proportions (attributable risk), or as an odds ratio or relative risk.

Reporting the effect size enables clinicians and other researchers to judge whether a statistically significant result is also a clinically important finding. The size of the difference or the risk statistic provides information to help health professionals decide whether the observed effect is large and important enough to warrant a change in current health care practice, is equivocal and suggests a need for further research, or is small and clinically unimportant.

4.4 Statistical significance and clinical importance

When applying statistical methods in health and medical research, we need to make an informed decision about whether the effect size that led to a statistically significant finding is also clinically important (see Figure 4.1)). The decision about whether a statistically significant result is also clinically important depends on expert knowledge and is best made by practitioners with experience in the field.

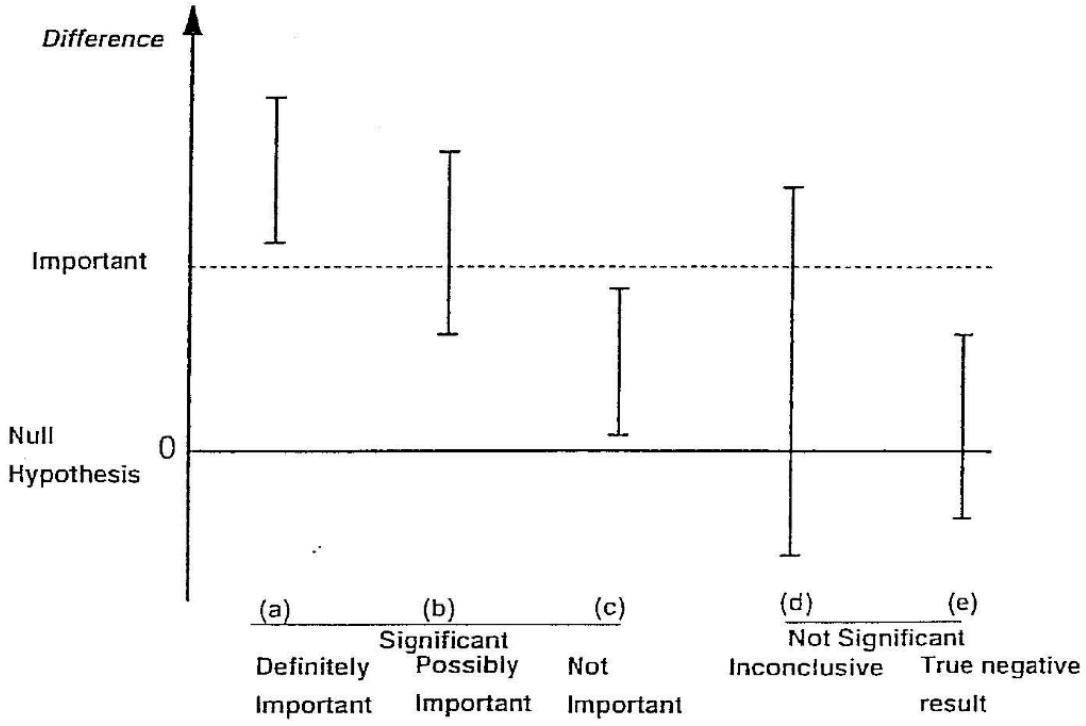


Figure 4.1: Statistical significance vs. clinical importance (Source: Armitage P, Berry G, Matthews JNS. (2001)

It is possible when conducting significance tests, particularly in very large studies, that a small effect is found to be statistically significant. For example, say in a large study of over 1000 patients, a new medication was found to lower blood pressure on average by 1 mmHg more than

a currently accepted drug and this was statistically significant (P < 0.05). However, such a small decrease in blood pressure would probably not be considered clinically important. The cost and side effects of prescribing the new medication would need to be weighed against the very small average benefit that would be expected. In this case, although the null hypothesis would be rejected (i.e. the result is statistically significant), the result would not be clinically important. This is the situation described in scenario (c) of Figure 4.1.

Conversely, it is possible to obtain a large, clinically important difference between groups, but a P value that does not demonstrate a statistically significant difference.

For example, consider a study to measure the rate of hospital admissions. We may find that 80% of children who present to the Emergency Department are admitted before an intervention is introduced compared to only 65% of children after the intervention. However, the P value may be calculated as 0.11 and is non-significant. This is because only 60 children were surveyed in each period. Here, the reduction in the admission rate by 15% represents a clinically important difference, but not statistically significant. This situation is represented in scenario (d) of Figure 4.1.

The important thing to remember is that statistical significance does not always correspond to practical importance. A statistically significant result may be practically unimportant, and a statistically non-significant results may be practically important.

4.5 Errors in significance testing

There are two conclusions we can draw when conducting a hypothesis test: if the P-value is small, there is strong evidence against the null hypothesis and we reject the null hypothesis. If the P-value is not small, there is little evidence against the null hypothesis and we fail to reject the null hypothesis. As discussed above, the “small” cut-point for the P-value is often taken as 0.05. We refer to this value as α (alpha).

We can conduct a thought experiment and compare our hypothesis test conclusion to reality. In reality, either the null hypothesis is true, or it is false. Of course, if we knew what reality was, we would not need to conduct a hypothesis test. But we can compare our possible hypothesis test conclusions to the true (unobserved) reality.

If the null hypothesis was true in reality, our hypothesis test can fail to reject the null hypothesis – this would be a correct conclusion. However, the hypothesis test could lead us to rejecting the null hypothesis – this would be an incorrect conclusion. We call this scenario a Type I error, and it has a probability of α.

The other situation is where, in reality, the null hypothesis is false. A correct conclusion would be where our hypothesis test rejects the null hypothesis. However, if our hypothesis test fails to reject the null hypothesis, we have made a Type II error. The probability of making a Type II error is denoted β (beta). We will see in Module 10 that β is determined by the size of the study.

The error in falsely rejecting the null hypothesis when it is true (type I error), or in falsely accepting the null hypothesis when it is not true (type II error) is summarised in Table 4.2. We will return to these concepts in Module 10, when discussing how to determine the appropriate sample size of a study.

Table 4.2: Comparison of study result with the truth

	Effect	No effect
Evidence	Correct	α
No evidence	β	Correct

4.6 Confidence intervals in hypothesis testing

In Module 3, the 95% confidence interval around a mean value was calculated to show the precision of the summary statistic. The 95% confidence intervals around other summary statistics can also be calculated.

For example, if we were comparing the means of two groups, we would want to test the null hypothesis that the difference in means is zero, that there is no true difference between the groups.

From the data from the two groups, we could estimate the difference in means, the standard error of the difference in means and the 95% confidence interval around the difference. To estimate the 95% confidence interval, we use the formula given in Module 3, that is:

$$95\% \text{ CI} = \text{Difference in means} \pm 1.96 \times \text{SE}(\text{Difference in means})$$

It is important to remember that the 95% CI is estimated from the standard error, and that the standard error has a direct relationship to the sample size. For small sample sizes, the standard error is large and the 95% CI becomes wider. Conversely, the larger the sample size, the smaller the standard error and the narrower the 95% CI becomes indicating a more precise estimate of the mean difference.

The 95% CI tells us the region in which we are 95% confident that the true difference between the groups in the population lies. If this region contains the null value of no difference, we can say that we are 95% confident that there is no true difference between the groups and therefore we would not reject the null hypothesis. This is shown in the top two estimates in Figure 4.2. If the zero value lies outside the 95% confidence interval, we can conclude that there is evidence of a difference between the groups because we are 95% confident that the difference does not encompass a zero value (as shown in the lower two estimates in Figure 4.2.

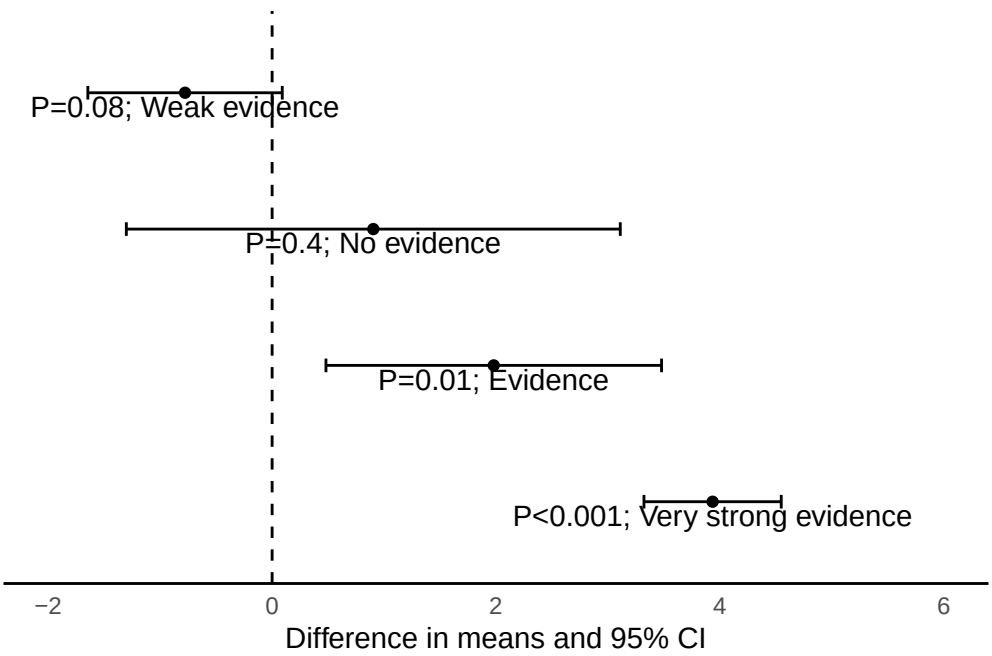


Figure 4.2: Using confidence intervals as informal hypothesis tests

For relative risk and odds ratio measures, when the 95% CI includes the value of 1 it indicates that we can be 95% confident that the true RR or OR of the association between the study factor and outcome factor includes 1.0 in the source population. This indicates little evidence of an association between the study factor and the outcome factor, e.g. if the results of a study were reported as RR = 1.10 (95% CI 0.95 to 1.25). The P-value can be calculated to assess this (discussed in Module 7).

Table 4.3: Values indicating no effect

Type of outcome	Measure of effect	Null value
Continuous	Difference in means	0
Binary	Difference in proportions	0
Binary	Relative risk	1
Binary	Odds ratio	1

4.7 One-sample t-test

A one-sample t-test tests whether a sample mean is different to a hypothesised value. The t-distribution and its relation to normal distribution has been discussed in detailed in Module 3.

In a one-sample t-test, a t-value is computed as the sample mean divided by the standard error of the mean. The significance of the t-value is then computed using software, or can be obtained from a statistical table.

The principles of this test can be used for applications such as testing whether the mean of a sample is different from a known population mean, for example testing whether the IQ of a group of children is different from the population mean of 100 IQ points or testing whether the number of average hours worked in an adult sample is different from the population mean of 38 hours.

Worked Example

The mean diastolic blood pressure (BP) of the general US population is known to be 71 mm Hg. The diastolic blood pressure of 733 female Pima indigenous Americans was measured and a density plot showed that the data were approximately normally distributed. The mean diastolic blood pressure in the sample was 72.4 mm Hg with a standard deviation of 12.38 mm Hg.

We can use jamovi or R to conduct a one sample t-test using the data available on Moodle (`mod04_blood_pressure.csv`). The results from this test are summarised below.

Table 4.4: Summary of blood pressure from female Pima indigenous Americans

n	Mean	Standard deviation	Standard error	95% confidence interval of the mean
733	72.4	12.38	0.46	71.5 to 73.3

The test statistic for the one-sample t-test is calculated as t_{732}=3.07, with a P-value of 0.002.

The mean diastolic blood pressure of females from Pima is estimated as 72.4 mmHg (95% CI: 71.5 to 73.3 mmHg), which is higher than that of the general US population. Note that this interval does not contain the mean of the general US population (71 mm Hg), providing some indication that the mean diastolic blood pressure of female Pima people is higher than that of the general US population.

The result from the formal hypothesis test gives strong evidence that the mean diastolic BP of the female Pima people is higher than that of the general US population (t_{732}=3.07, P=0.002).

4.8 One and two tailed tests

Most statistical tests are two tailed tests, that is, we conduct a test that allows for the summary statistic in the group of interest to be either higher or lower than in the comparison group. For a t-test, this requires that we obtain a two-tailed P value which gives us the probability of the t-value being in either one of the two tails of the t-distribution as shown in Figure 4.3. The shaded regions show the t values that indicate a P value less than 0.05.

Occasionally, one tailed tests are conducted in which the summary statistic in the group of interest can only be higher or lower than the comparison group, i.e. a difference is specified to

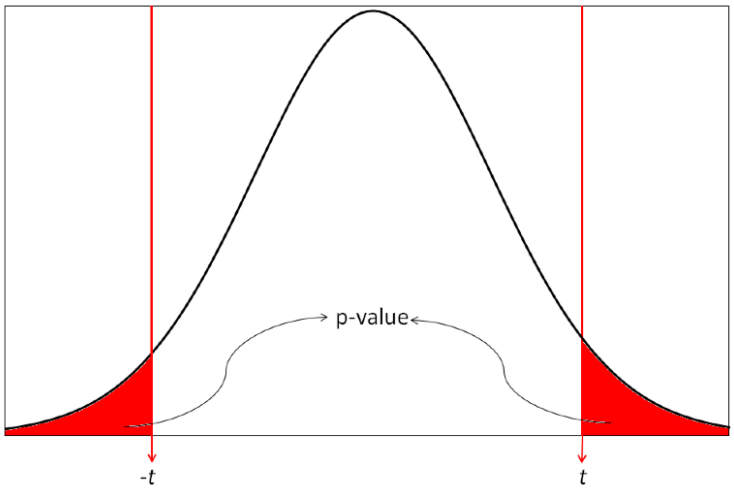


Figure 4.3: P-value for a 2-tailed test

occur in one direction only. In most studies, two tailed tests of significance are used to allow for the possibility that the effect size could occur in either direction. In clinical trials, this would mean allowing for a result that can indicate a benefit or an adverse effect in response to a new treatment. In epidemiological studies, two tailed tests are used to allow for the fact that exposure to a factor of interest may be adverse or may be beneficial. This conservative approach is usually adopted to prevent missing important effects that occur in the opposite direction to that expected by the researchers.

4.9 A note on P-values displayed by software

You will often see P-values generated by statistical software presented as 0.000 or 0.0000. As P-values can never be equal to zero, any P-value displayed in this way should be converted to <0.001 or <0.0001 respectively (i.e. replace the last 0 with a 1, and use the less-than symbol).

R can display P-values in a very cryptic way: 6.478546e-05 for example. This is translated as:

$$\begin{aligned} 6.478546e-05 &= 6.478546 \times 10^{-5} \\ &= 6.478546 \times 0.00001 \\ &= 0.00006478546 \end{aligned}$$

Such a P-value should be presented as P<0.0001.

4.10 Decision Tree

In the following modules in this course, several formal statistical tests will be described to analyse different types of data sets that have been collected to test set null hypotheses. It is important that the correct statistical test is selected to generate P-values and estimate effect size. If an incorrect statistical test is used, the assumptions of the test may be violated, the effect size may be biased and the P value generated may be incorrect.

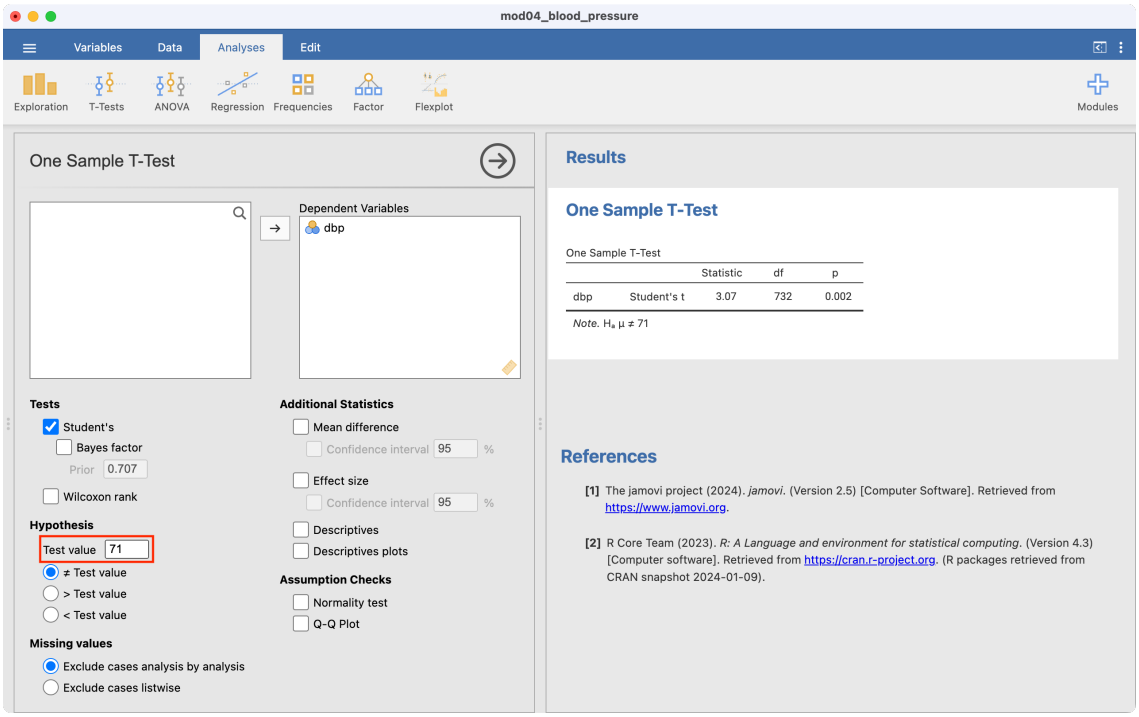
Selecting the correct test to use in each situation depends on the study design and the nature of the variables collected. Figure 1 in the Appendix shows a decision tree which enables you to decide the type of test to select based on the nature of the data.

Jamovi notes

4.11 One sample t-test

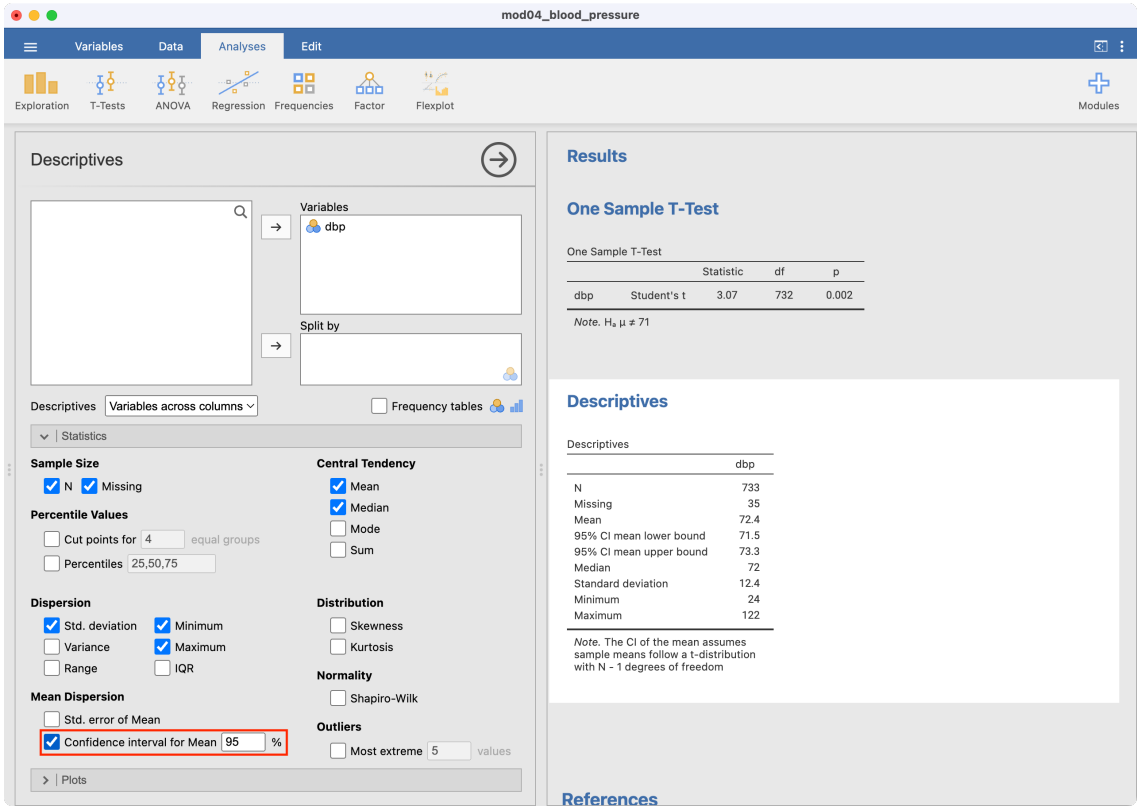
We will use data from `mod04_blood_pressure.csv` to demonstrate how a one-sample t-test is conducted in Jamovi. To perform the test, go to **Analyses > T-Tests > One Sample T-Test**.

Select `dbp` as the **Dependent Variable**. Enter the hypothesised mean as the **Test value** (`71` in this example) as shown below.



The test statistic (i.e. t and its degrees of freedom) are provided. By default, Jamovi will calculate a P-value for the two-sided test (H_a: mean $\neq$ 71), which appropriate for this example.

Note that the one-sample t-test output does not include the mean or the 95% confidence interval of the mean of your variable of interest. This should be obtained using **Exploration > Descriptives**:


mod04_blood_pressure
Variables
Data
Analyses
Edit
Exploration
T-Tests
ANOVA
Regression
Frequencies
Factor
Flexplot
Modules
Descriptives
Variables
dbp
Split by
Descriptives
Variables across columns
Frequency tables
Statistics
Sample Size
N
Missing
Percentile Values
Cut points for 4 equal groups
Percentiles 25,50,75
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
IQR
Mean Dispersion
Std. error of Mean
Confidence interval for Mean 95 %
Central Tendency
Mean
Median
Mode
Sum
Distribution
Skewness
Kurtosis
Normality
Shapiro-Wilk
Outliers
Most extreme 5 values
Plots
Results
One Sample T-Test
One Sample T-Test
Statistic df p
dbp Student's t 3.07 732 0.002
Note. Hₐ μ ≠ 71
Descriptives
Descriptives
dbp
N 733
Missing 35
Mean 72.4
95% CI mean lower bound 71.5
95% CI mean upper bound 73.3
Median 72
Standard deviation 12.4
Minimum 24
Maximum 122
Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom
References

R notes

4.12 One sample t-test

We will use data from `mod04_blood_pressure.csv` to demonstrate how a one-sample t-test is conducted in R. To test whether the mean diastolic blood pressure of the population from which the sample was drawn is equal to 71, we can use the t.test command:

```
pima <- read.csv("data/examples/mod04_blood_pressure.csv")
t.test(pima$dbp, mu=71)
```

```
    One Sample t-test

data:  pima$dbp
t = 3.0725, df = 732, p-value = 0.002202
alternative hypothesis: true mean is not equal to 71
95 percent confidence interval:
 71.50732 73.30305
sample estimates:
mean of x
 72.40518
```

The output provides:

- a test statistic (t=3.07);
- degrees of freedom for the test statistic (df = 732);
- a P-value from the two-sided test (P=0.002);
- the mean of the sample (72.4);
- and the 95% confidence interval of the mean (71.5 to 73.3).

Activities

Activity 4.1

In each of the following situations, what decision should be made about the null hypothesis if the researcher indicates that:

a) $P < 0.01$
b) $P > 0.05$
c) “ns”
d) “significant differences exist”

Activity 4.2

For the following hypothetical situations, formulate the null hypothesis and alternative hypothesis and write a conclusion about the study results:

a) A study was conducted to investigate whether the mean systolic blood pressure of males aged 40 to 60 years was different to the mean systolic blood pressure of females aged 40 to 60 years. The result of the study was that the mean systolic blood pressure was higher in males by 5.1 mmHg (95% CI 2.4 to 7.6; P = 0.008).
b) A case-control study was conducted to investigate the association between obesity and breast cancer. The researchers found an OR of 3.21 (95% CI 1.15 to 8.47; P = 0.03).
c) A cohort study investigated the relationship between eating a healthy diet and the incidence of influenza infection among adults aged 20 to 60 years. The results were RR = 0.88 (95% CI 0.65 to 1.50; P = 0.2).

Activity 4.3

A pilot study was conducted to compare the mean daily energy intake of women aged 25 to 30 years with the recommended intake of 7750 kJ/day. In this study, the average daily energy intake over 10 days was recorded for 12 healthy women of that age group. The data are in the the Excel file Activity_4.3.xlsx. Import the file into jamovi or R for this activity.

a) State the research question
b) Formulate the null hypothesis
c) Formulate the alternative hypothesis
d) Analyse the data in jamovi or R and report your conclusions

Activity 4.4

Which procedure gives the researcher the better chance of rejecting a null hypothesis?

a) comparing the data-based p-value with the level of significance at 5%
b) comparing the 95% CI with a nominated value
c) neither procedure

Activity 4.5

Setting the significance level at P < 0.10 instead of the more usual P < 0.05 increases the likelihood of:

a) a Type I error
b) a Type II error
c) rejecting the null hypothesis
d) Not rejecting the null hypothesis

Activity 4.6

For a fixed sample size setting the significance level at a very extreme cutoff such as P < 0.001 increases the chances of: a) obtaining a significant result b) rejecting the null hypothesis c) a Type I error d) a Type II error

Module 5

Comparing the means of two groups

Learning objectives

By the end of this module you will be able to:

- Decide whether to use an independent samples t-test or a paired t-test to compare two the means of two groups;
- Conduct and interpret the results from an independent samples t-test;
- Describe the assumptions of an independent samples t-test;
- Conduct and interpret the results from a paired t-test;
- Describe the assumptions of a paired t-test;
- Conduct an independent samples t-test and a paired t-test using software;
- Report results and provide a concise summary of the findings of statistical analyses.

Optional readings

Kirkwood and Sterne (2001); Sections 7.1 to 7.5. [UNSW Library Link]

Bland (2015); Section 10.3. [UNSW Library Link]

5.1 Introduction

In Module 4, a one-sample t-test was used for comparing a single mean to a hypothesised value. In health research, we often want to compare the means between two groups. For example, in an observational study, we may want to compare cholesterol levels in people who exercise regularly to the levels in people who do not exercise regularly. In a clinical trial, we may want to compare cholesterol levels in people who have been randomised to a dietary modification or to usual care. In this module, we show how to compare the means of two groups where the analysis variable is normally distributed.

From the decision tree presented in the Appendix, we can see that if we have a continuous outcome measure and two categorical groups that are not related, i.e. a binary exposure measurement, the test for such data is an independent samples t-test. The test is also sometimes called a 2-sample t-test.

In research, data are often 'paired' or 'matched', that is the two data points are related to one another. This occurs when measurements are taken:

- From each participant on two occasions, e.g. at baseline and follow-up in an experimental study or in a longitudinal cohort study;
- From related people, e.g. a mother and daughter or a child and their sibling;
- From related sites in the same person, e.g. from both limbs, eyes or kidneys;

- From matched participants e.g. in a matched case-control study;
- In cross-over clinical trials where the patient receives both drugs, often in random order.

An independent samples t-test cannot be used for analysing paired or matched data because the assumption that the two groups are independent is violated. Treating paired or matched measurements as independent samples would artificially inflate the sample size and lead to inaccurate P values. When the data are related in a paired or matched way and the outcome is continuous, a paired t-test is the appropriate statistic to use if the data are normally distributed.

5.2 Independent samples t-test

An independent samples t-test is a parametric test that is used to assess whether the mean values of two groups are different from one another. Thus, the test is used to assess whether two mean values are similar enough to have come from the same population or whether the difference between them is so large that the two groups can be considered to have come from separate populations with different characteristics.

The null hypothesis is that the mean values of the two groups are not different, that is:

H_0: $(\mu_1 - \mu_2) = 0$

Rejecting the null hypothesis using an independent samples t-test indicates that the difference between the means of the two groups is large in relation to the variability in the samples and is unlikely to be due to chance or to sampling variation.

Assumptions for an independent samples t-test

The assumptions that must be met before an independent samples t-test can be used are:

- The two groups are independent
- The measurements are independent
- The analysis variable must be continuous and must be normally distributed in each group

The first two assumptions are determined by the study design. The two samples must be independent, i.e. if a person is in one group then they cannot be included in the other group, and the measurements within a sample must be independent, i.e. each person must be included in their group once only.

The third assumption of normality is important although t-tests are robust to some degree of non-normality as long as there are no influential outliers and, more importantly, if the sample size is large. We examined how to assess normality in Module 2. If the data are not normally distributed, it may be possible to transform them using a mathematical function such as a logarithmic transformation. If not, then we may need to use non-parametric tests. This is examined in Module 9.

Traditionally, the variance of the analysis variable in each group was assumed to be equal. However, this assumption can be relaxed by using Welch's variation of the t-test. It has been recommended that this unequal-variances t-test be used in most, if not all situations (West 2021; Delacre, Lakens, and Leys 2017; Ruxton 2006).

Worked Example 5.1

In an observational study of a random sample of 100 full term babies from the community, birth weight and gender were measured. There were 44 male babies and 56 female babies in the sample. The research question asked whether there was a difference in birth weights between boys and girls. The two groups are independent of each other and therefore an independent samples t-test can be used to test the null hypothesis that there is no difference in weight between the genders.

Exploratory data analysis of the variable of interest in each group should always be obtained before a t-test is undertaken to ensure that the assumptions are met. In particular, the

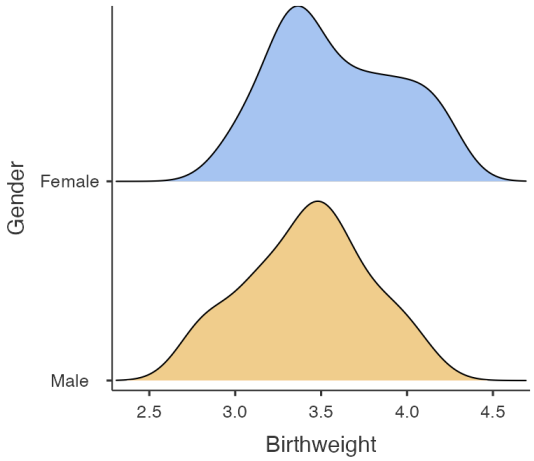


Figure 5.1: Distribution of birth weight by gender

distribution of the analysis variable should be examined for each group, as shown in Figure 5.1. The `mod05_birthweight.rds` dataset is available on Moodle.

The plots show that the data are approximately normally distributed: the density curves are relatively bell shaped and symmetric, and there are no outliers.

We can also describe the data using summary statistics:

Table 5.1: Summary of birthweight by gender

Characteristic	Female	Male
Birthweight		
Number	56	44
Mean (SD)	3.59 (0.36)	3.42 (0.35)
Median (Q1, Q3)	3.53 (3.32, 3.88)	3.43 (3.15, 3.63)
Range	2.95 to 4.25	2.75 to 4.10

The table shows that girls have a mean weight of 3.59 kg (SD 0.36) and boys have a mean weight of 3.42 kg (SD 0.35) with females being heavier than males. The variabilities of birth weight (i.e. the standard deviation) are similar between the two groups.

Conducting and interpreting an independent samples t-test

An independent samples t-test provides us with a t statistic from which we can compute a P value. The computation of the t statistic is as follows:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{SE(\overline{x}_1 - \overline{x}_2)}$$

with the standard error and degrees of freedom calculated from software. Note that by using Welch's t-test, the degrees of freedom will usually not be a whole number, and will appear with decimals.

Looking at the formula for the t-statistic, we can see that the t is an estimate of how different the mean values are compared to the variability of the difference in means. So t will become larger as the difference in means increases with respect to the variability.

Statistical software will calculate both the t and P values. If the t-value is large, the P value will be small, providing evidence against the null hypothesis of no difference between the groups.

Table 5.2 summarises the results of an independent samples t-test using `mod05_birthweight.rds`. The process of conducting the t-test is summarised for jamovi and R in the following sections.

Table 5.2: Birthweight (kg) by sex

Sex	n	Mean (SE)	95% Confidence Interval
Female	56	3.59 (0.049)	3.49 to 3.68
Male	44	3.42 (0.053)	3.31 to 3.53
Difference		0.17 (0.072)	0.02 to 0.31

Here we see that girls are heavier than boys, and the mean difference in weights between the genders is 0.17 kg (95% CI 0.02, 0.31). We are 95% confident that the true mean difference of weight between girls and boys lies between 0.02 and 0.31 kg. Note that this interval does not contain the null value of 0.

Here we are testing the null hypothesis of no difference in mean birthweights between females and males: a two-sided test. The t-value is calculated as 2.30 with 93.5 degrees of freedom, and yields a two-sided P value of 0.023, providing evidence of a difference in mean birthweight between sex.

5.3 Paired t-tests

If the outcome of interest is the difference in the continuously outcome measurement between each pair of observations, a paired t-test is used. In effect, a paired t-test is used to assess whether the mean of the differences between the two related measurements is significantly different from zero. In this sense, a paired t-test is very closely aligned with a one sample t-test.

When using a paired t-test, the variation *between the pairs* of measurements is the most important statistic. The variation between the participants is of little interest.

For related measurements, the data for each pair of values must be entered on the same row of the spreadsheet. Thus, the number of rows in the data sheet is the number of pairs of observations. Thus, the effective sample size is the total number of pairs and not the total number of measurements.

Assumptions for a paired t-test

The assumptions for a paired t-test are:

- the outcome variable is continuous
- the differences between the pair of the measurements are normally distributed

For a paired samples t-test, it is important to test whether the *differences* between the two measurements are normally distributed. If the assumptions for a paired t-test cannot be met, a non-parametric equivalent is a more appropriate test to use (Module 9).

Computing a paired t-test

The null hypothesis for using a paired t-test is as follows:

H_0: Mean (Measurement1 – Measurement2) = 0

To compute a t-value, the size of the mean difference between the two measurements is compared to the standard error of the paired differences, i.e.

$$t = \frac{\overline{d}}{SE(\overline{d})}$$

with $n-1$ degrees of freedom, where n is the number of pairs.

Because the standard error becomes smaller as the sample size becomes larger, the t-value increases as the sample size increases for the same mean difference.

Worked Example 5.2

A total of 107 people were recruited into a study to assess whether ankle blood pressure measured in two different sites would be the same. For each person, systolic blood pressure (SBP) was measured in two sites: dorsalis pedis and tibialis posterior.

The dataset `mod05_ankle_bp.xlsx` is available on Moodle. First, we need to compute the pairwise difference between SBP measured in the two sites. The distribution of the difference between SBP measured in dorsalis pedis and tibialis posterior is shown in Figure 5.2. The differences approximate a normal distribution and therefore a paired t-test can be used.

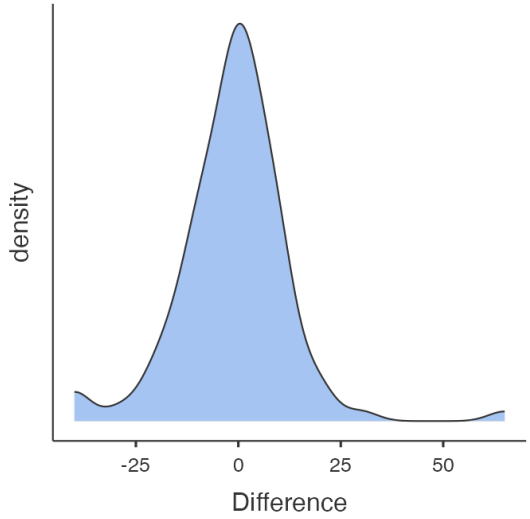


Figure 5.2: Distribution of differences in ankle SBP between two sites of 107 participants

The paired t-test can be performed using statistical software, with a summary of the results presented in Table 5.3. We can see that the mean SBP is very similar in the two sites.

Table 5.3: Systolic blood pressure (mmHg) measured at two sites on the ankle

Site	n	Mean (SE)	95% Confidence Interval
Dorsalis pedis	107	116.7 (3.46)	(109.9 to 123.6)
Tibialis posterior	107	118.0 (3.43)	(111.2 to 124.8)
Difference	107	-1.3 (1.31)	(-3.9 to 1.3)

The t-value is calculated as −0.96 with 106 degrees of freedom, providing a two-sided P-value of 0.34. Thus these data provide no evidence of a difference in systolic blood pressure between the two sites.

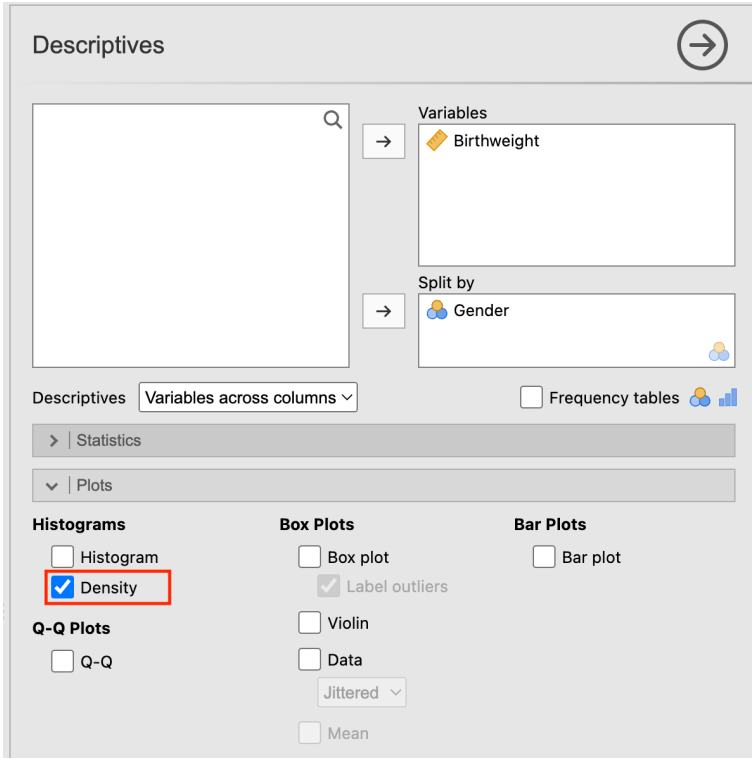
Jamovi notes

5.4 Checking data for the independent samples t-test

Examining variable distributions by a second variable

We can use **Analyses > Exploration > Descriptives** to obtain the distribution plots in Figure 5.1. Choose `birthweight` to appear in the **Variables** box, and `gender` as the **Split by** variable. Choose **Density** in the **Plots** section:

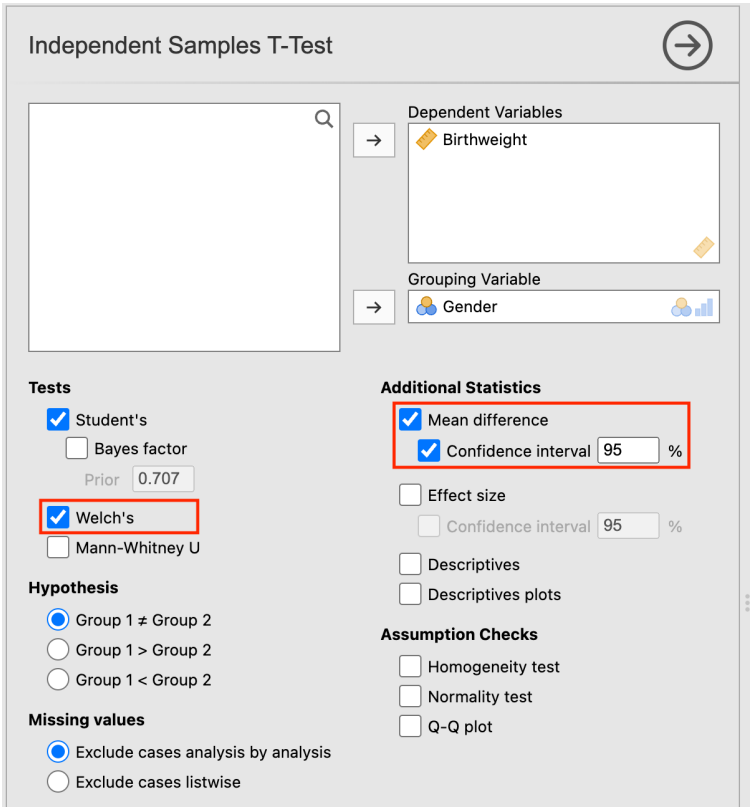
Jamovi also produces summary statistics for each level of the **Split by** variable, and we can select the statistics of interest in the **Statistics** section as necessary.



5.5 Independent samples t-test

To carry out an independent sample t-test, go to **Analyses > T-Tests > Independent Samples T-Test**. Move `birthweight` into **Dependent variables** and `gender` as the **Grouping Variable**. Because we don't assume equal variances of `birthweight` for males and females, we tick the **Welch's** box.

In order to obtain an estimate of the difference in means, with its 95% Confidence Interval, tick the relevant boxes in **Additional Statistics**:



5.6 Checking the assumptions for a Paired t-test

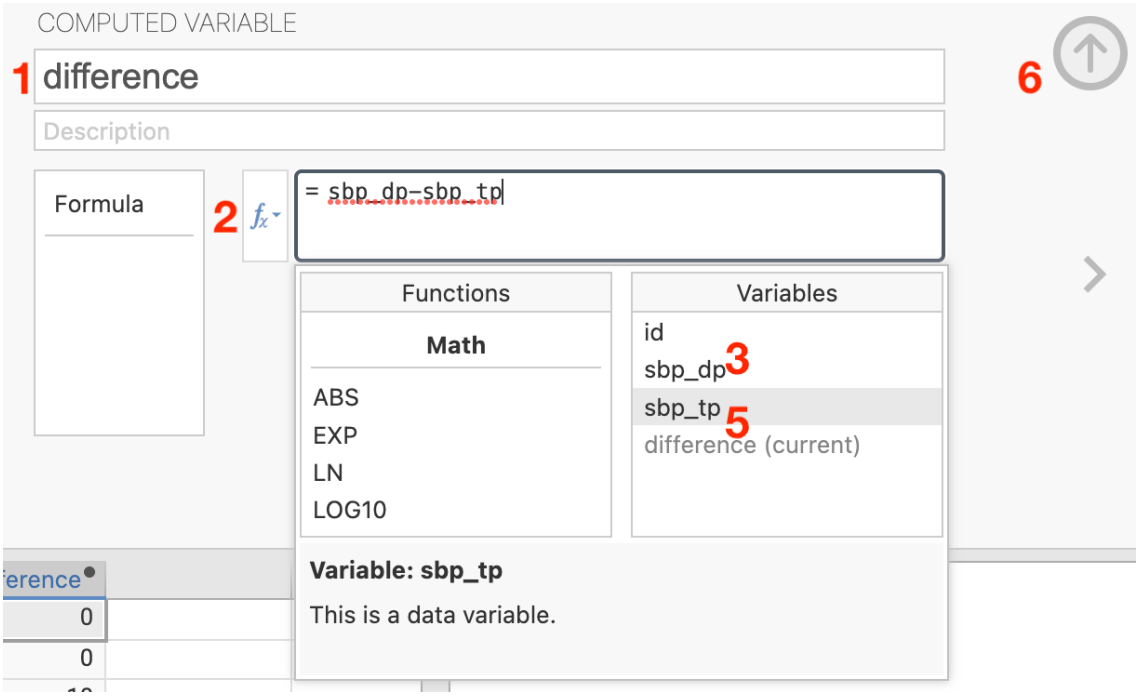
Before performing a paired t-test, you must check that the assumptions for the test have been met. Using the dataset `mod05_ankle_bp.xlsx` to show that the difference between the pair of measurements between the sites is normally distributed, we first need to compute a new variable of the differences.

To create a new column at the end of your dataset, click a cell in the first empty column, then choose **Data > Compute**. We want to compute `difference` as `sbp_dp - sbp_tp`, so we enter this in the formula box as below:

1. Specify the name of the variable to be created: here `difference`
2. Click the f_x button to display a list of variable names in your dataset
3. Double-click the variable `sbp_dp` to bring it into the formula box
4. Type – to represent "minus"
5. Double-click the variable `sbp_tp` to bring it into the formula box (do not worry if your formula is underlined in red, this is simply a spell-check)
6. Click the up arrow to close the **Compute** dialog box

(Note that steps 3 to 5 could also be completed by typing the formula.)

You will see a new column called `difference` which represents the difference between the two blood pressures.

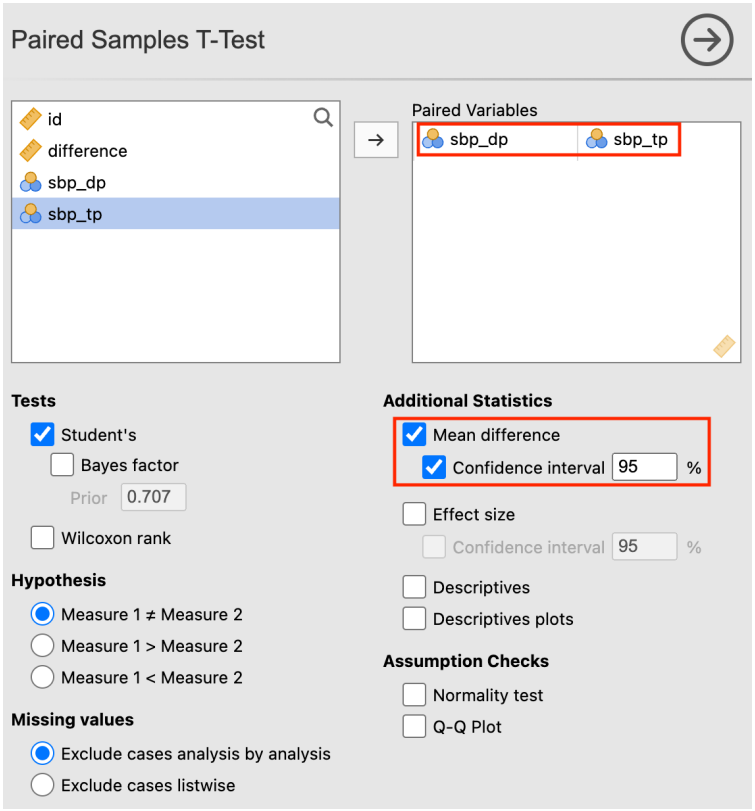


A density plot of the differences can be constructed in the usual way.

5.7 Paired t-Test

Using the same blood pressure data as previously, choose **Analyses > T-Tests > Paired Samples T-Test**. Select `sbp_dp` and `sbp_tp` as the **Paired Variables**.

To obtain more informative output, select **Mean difference** and **Confidence interval** as additional statistics. The dialog box will look like:



With the following output:

Paired Samples T-Test

Paired Samples T-Test

								95% Confidence Interval	
			statistic	df	p	Mean difference	SE difference	Lower	Upper
sbp_dp	sbp_tp	Student's t	−0.96	106.00	0.338	−1.26	1.31	−3.86	1.34

Note. H_a $\mu_{\text{Measure 1 - Measure 2}} \neq 0$

R notes

5.8 Checking data for the independent samples t-test

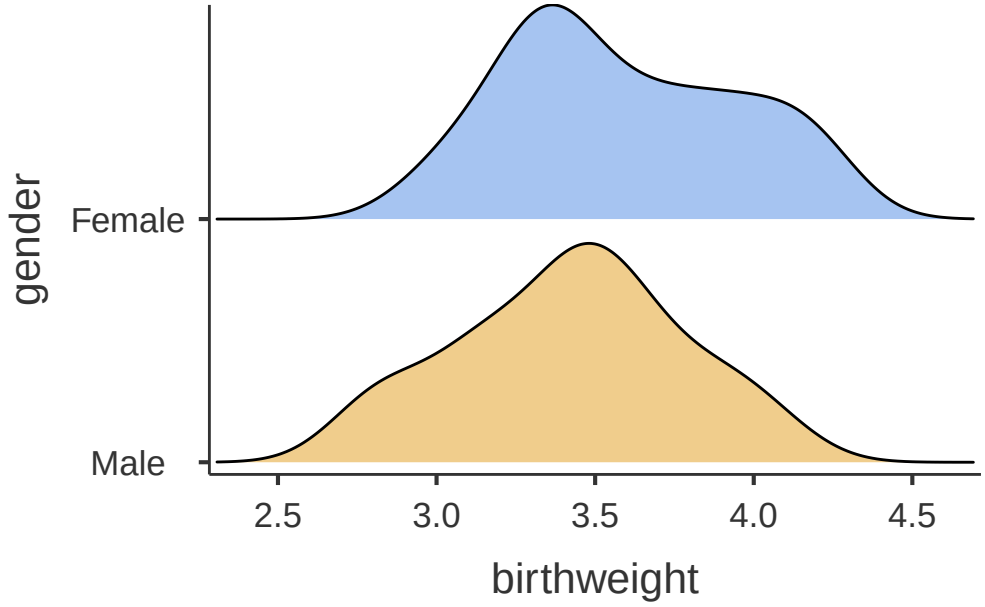
Examining variable distributions by a second variable

We can use the a `splitBy` variable in the `descriptives` function in the `jmv` package to obtain summary statistics for each level of a grouping variable. Further, specifying `dens = TRUE` will produce density plots for the analysis variable for each level of the grouping variable.

For example, to create the distribution plots in Figure 5.1, we can use

```
library(jmv)

bwt <- readRDS("data/examples/mod05_birthweight.rds")

descriptives(data = bwt, vars = birthweight, splitBy = gender, dens = TRUE)
```

```
 DESCRIPTIVES

 Descriptives

                        gender    birthweight

   N                    Female            56
                        Male              44
   Missing              Female             0
                        Male               0
   Mean                 Female      3.587411
                        Male        3.421364
   Median               Female      3.530000
                        Male        3.430000
   Standard deviation   Female     0.3629788
                        Male       0.3536165
   Minimum              Female      2.950000
                        Male        2.750000
   Maximum              Female      4.250000
                        Male        4.100000
```



5.9 Independent samples t-test

We can use the `ttestIS()` (t-test, independent samples) function from the `jmv` package to perform the independent samples t-test. We include the `meanDiff=TRUE` and `ci=TRUE` options to obtain the difference in means, with its 95% confidence interval. We can request a Welch's test (which does not assume equal variances) by the `welchs=TRUE` option:

```
ttestIS(data = bwt, vars = birthweight, group = gender, meanDiff = TRUE, ci = TRUE, welchs = TRUE)
```

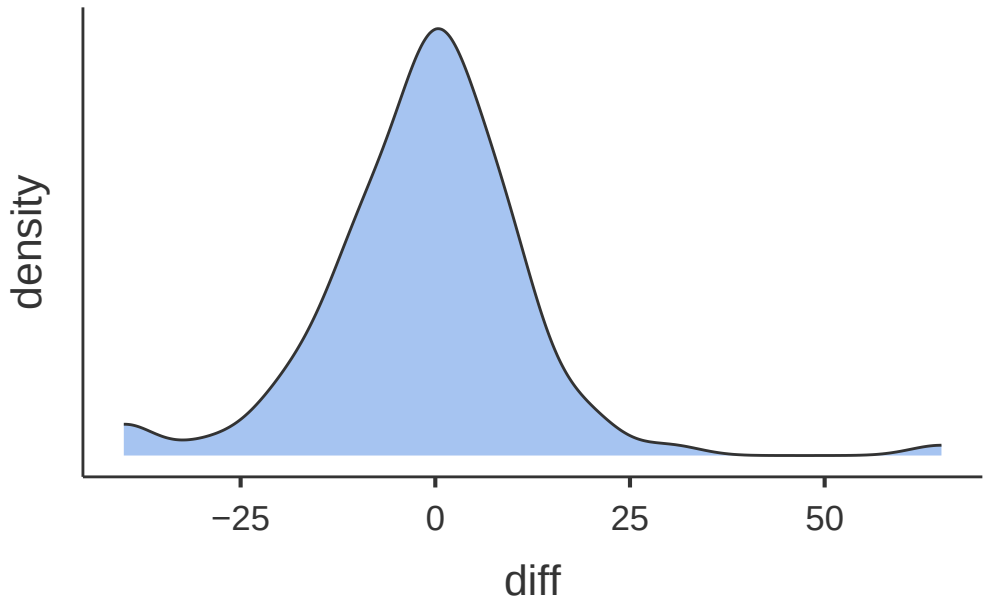
```
 INDEPENDENT SAMPLES T-TEST

 Independent Samples T-Test

                                Statistic    df          p            Mean difference    SE differe

   birthweight    Student's t     2.296556    98.00000    0.0237731          0.1660471       0.07230
                  Welch's t       2.303840    93.54377    0.0234458          0.1660471       0.07207

   Note. H   <sub>Female</sub>    <sub>Male</sub>
```

5.10 Checking the assumptions for a Paired t-test

Before performing a paired t-test, you must check that the assumptions for the test have been met. Using the dataset `mod05_ankle_bp.xlsx` to show that the difference between the pair of measurements between the sites is normally distributed, we first need to compute a new variable of the differences and examine its distribution.

```
library(readxl)

sbp <- read_excel("data/examples/mod05_ankle_bp.xlsx")
sbp$diff <- sbp$sbp_dp - sbp$sbp_tp
descriptives(data = sbp, vars = diff, dens = TRUE)
```

```
DESCRIPTIVES

Descriptives

                        diff

  N                          107
  Missing                      0
  Mean                 -1.261682
  Median                0.000000
  Standard deviation    13.56489
  Minimum              -40.00000
  Maximum               65.00000
```



While there is a large difference in blood pressure (around 60 mmHg) that warrants further checking, the curve is roughly symmetric with an approximately Normal distribution.

5.11 Paired t-Test

To perform a paired t-test we will use the dataset `mod05_ankle_bp.xlsx`. We can perform a paired t-test using the `ttestPS()` function within the `jmv` package, where we defined the paired observations as: 'pairs=list(list(i1 = 'variable1', i2 = 'variable2'))

```
ttestPS(data = sbp, pairs = list(list(i1 = "sbp_dp", i2 = "sbp_tp")), meanDiff = TRUE, ci = TRU
```

```
PAIRED SAMPLES T-TEST

Paired Samples T-Test

                                        statistic     df          p            Mean difference

  sbp_dp     sbp_tp     Student's t     -0.9621117    106.0000    0.3381832          -1.261682

  Note. H   <sub>Measure 1 - Measure 2</sub>   0
```

The syntax of the ttestPS function is a little cumbersome. The `t.test` function can be used as an alternative:

```
t.test(sbp$sbp_dp, sbp$sbp_tp, paired = TRUE)
```

```
    Paired t-test

data:  sbp$sbp_dp and sbp$sbp_tp
t = -0.96211, df = 106, p-value = 0.3382
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -3.861596  1.338232
sample estimates:
mean difference
      -1.261682
```

Activities

Activity 5.1

Indicate what type of t-test could be used to analyse the data from the following studies and provide reasons:

a) A total of 60 university students are randomly assigned to undergo either behaviour therapy or Gestalt therapy. After twenty therapeutic sessions, each student earns a score on a mental health questionnaire.
b) A researcher wishes to determine whether attendance at a day care centre increases the scores of three year old twins on a motor skills test. Random assignment is used to decide which member from each of 30 pairs of twins attends the day care centre and which member stays at home.
c) A child psychologist assigns aggression scores to each of 10 children during two 60 minute observation periods separated by an intervening exposure to a series of violent TV cartoons.
d) A marketing researcher measures 100 doctors' reports of the number of their patients asking them about a particular drug during the month before and the month after a major advertising campaign.

Activity 5.2

A study was conducted to compare haemoglobin levels in the blood of children with and without cystic fibrosis. It is known that haemoglobin levels are normally distributed in children. The data are stored in Activity_5.2.csv as:

- cf: cystic fibrosis: 0=no, 1=yes
- haem: haemoglobin (g/dL)

a) State the appropriate null hypothesis and alternate hypothesis
b) Use jamovi or R to conduct an appropriate statistical test to evaluate the null hypothesis. Are the assumptions for the test met for this analysis to be valid?

Activity 5.3

A randomised controlled trial (RCT) was carried out to investigate the effect of a new tablet supplement in increasing the hematocrit (%) value in anaemic participants. In the study, hematocrit was measured as the proportion of blood that is made up of red blood cells. Hematocrit levels are often lower in anaemic people who do not have sufficient healthy red blood cells. In the RCT, 33 people in the intervention group received the new supplement and 31 people in the control group received standard care (i.e. the usual supplement was given). After 4 weeks, hematocrit values were measured as entered in the file `Activity_5.3.rds`. In the community, hematocrit levels are normally distributed.

a) State the research question and formulate a null hypothesis.
b) Use jamovi or R to conduct an appropriate statistical test to answer the research question. Before using the test, check the data to see if the assumptions required for the test are met.
c) Run your statistical test.
d) Construct a table to show how you would report your results and write a conclusion.

Activity 5.4

A total of 41 babies aged 6 months to 2 years with haemangioma (birth mark) were enrolled in a study to test the effect of a new topical medication in reducing the volume of their haemangioma. Parents were asked to apply the medication twice daily. The volume (in mm3) of the haemangioma was measured at enrolment and again after 12 weeks of using the medication. The data are saved in `Activity_5.4.rds`.

Analyse these data fully to answer the research question. Are there any limitations of this study?

Supplementary Activity 5.5

A study was conducted to investigate cardiovascular health of Australians. In this study heart rates were recorded by a heart rate monitor on each participant following 30 minutes of intense aerobic exercise. The researchers are interested in whether there is a difference in the mean post-exercise heart rates of females aged 20 – 24 years compared to females aged 25 – 30 years.

A dataset containing the study information is provided in the file `Activity_5.5_heartrate.csv`. There are 149 observations in the dataset with the following variables:

- id: id number
- agegroup: age range of the females (1 = 20 – 24 years; 2 = 25 – 30 years)
- heartrate: heart rate (beats/minute)

Analyse these data to answer the research question. Write a brief report summarising your results and state your conclusion.

Activity 5.6

A study was conducted to assess the effectiveness of a health promotion program to improve the fitness of school children. The fitness of 58 children from an inner-city school was assessed by measuring the total distance each child could run in a 10-minute period.

The following data are stored in `Activity_5.6_fitness.csv`:

- id: participant ID
- before: 10-minute running distance before the program (metres)
- after: 10-minute running distance after the program (metres)

Analyse these data to assess whether there has been a change in running distance after the health promotion campaign. Write a brief report summarising your results and state your conclusion.

Module 6

Summary statistics for binary data

Learning objectives

By the end of this module you will be able to:

- Compute and interpret 95% confidence intervals for proportions;
- Conduct and interpret a significance test for a one-sample proportion;
- Use statistical software to compute 95% confidence intervals for a difference in proportions, a relative risk and an odds ratio.

Optional readings

Kirkwood and Sterne (2001); Chapter 16 [UNSW Library Link]

Bland (2015); Section 8.6, Section 13.7 [UNSW Library Link]

6.1 Introduction

In Modules 4 and 5, we discussed methods used to analyse continuous data. In Modules 6 and 7, we will focus on analysing categorical data.

In health research, we often collect information that can be put into two categories, e.g. male and female, disease present or disease absent etc. Binary categorical variables such as these are summarised using proportions.

6.2 Calculating proportions and 95% confidence intervals

Calculating a proportion

We need two pieces of information to calculate a proportion: n, the number of trials, and k, the number of 'successes'. Note that we use the term 'success' to describe the outcome of interest, recognising that a success may be a adverse outcome such as death or disease.

The following formula is used to calculate the proportion, p:

$$p = k/n$$

The proportion, p, is a number that lies between 0 and 1. Proportions and their confidence intervals can easily be converted to percentages by multiplying by 100 once computed.

As for all summary statistics, it is useful to compute the precision of the estimate as a 95% confidence interval (CI) to indicate the range of values in which are 95% confident that the true population value lies. In this module, we present two methods for computing a 95% confidence interval around a proportion.

Calculating the 95% confidence interval of a proportion (Wald method)

The Wald method for calculating the 95% confidence interval is based on assuming that the proportion, p, is Normally distributed. This assumption is reasonable if the sample is sufficiently large (for example, if $n > 30$) and if $n \times (1-p)$ and $n \times p$ are both larger than 5.

The Wald method for calculating a 95% confidence interval is given by:

$$\text{95\% CI} = p \pm (1.96 \times \text{SE}(p))$$

where the standard error of a proportion is computed as:

$$\text{SE}(p) = \sqrt{\frac{p \times (1-p)}{n}}$$

Worked Example 6.1

In a cross-sectional study of children living in a rural village, 47 children from a random sample of 215 children were found to have scabies. Here $n = 215$ and $k = 47$, so the proportion of children with scabies is estimated as:

$$p = \frac{47}{215} = 0.2186$$

Given the large sample size and the number of children with the rarer outcome is larger than 5, the Wald method is used to calculate the standard error of the proportion as:

$$\text{SE}\,(p) = \sqrt{\frac{0.2186 \times (1-0.2186)}{215}} = 0.02819$$

Then, the 95% confidence interval is estimated as:

$$\text{95\% CI} = 0.2186 \pm 1.96 \times 0.02819$$

$$= 0.1634 \text{ to } 0.2739$$

The prevalence of scabies among children in the village is 21.9% (95% CI 16.3%, 27.4%). These values tell us that we are 95% confident that the true prevalence of scabies among children in the village is between 16.3% and 27.4%.

Calculating the 95% confidence interval of a proportion (Wilson method)

Another method to calculate the confidence interval of a proportion is the Wilson (sometimes also called the 'score') method. We can use it in situations where it is not appropriate to use the normal approximation to the binomial distribution as described above i.e. if the sample size is small ($n < 30$) or the number of subjects with the rarer outcome is 5 or fewer. This method much more difficult to implement by hand than the standard confidence interval, and so we will not discuss the hand calculation using the mathematical equation in this course. Instead, we use statistical software to do this (see the jamovi or R notes for detail).

When using software, our worked example provides a 95% confidence interval of the prevalence of scabies of 16.9% to 27.9%.

Wald vs Wilson methods

The Wald method, which assumes that the underlying proportion follows a Normal distribution, is easy to calculate and follows the form of other confidence intervals. The Wilson method, which is difficult to calculate by hand, has nicer mathematical properties. There are also a number of other methods for calculating confidence intervals for proportions, but we do not discuss these in this course.

A paper by Brown, Cai and DasGupta (Brown, Cai, and DasGupta (2001)) has compared the properties of the Wald and Wilson methods (among others) and concluded that the Wilson method is preferred over the Wald method. **Therefore, we recommend the Wilson method be used to calculate 95% confidence intervals for a proportion.** Note that it is not possible to compute a Wilson confidence interval using jamovi. The interval calculated by jamovi is the Clopper-Pearson interval.

6.3 Hypothesis testing for one sample proportion

We can carry out a hypothesis test to compare a sample proportion to a hypothesised proportion. In much the same way as a one sample t-test was used in Module 5 to test a sample mean against a hypothesised mean, we can perform a one-sample test to test a sample proportion against a hypothesised proportion. The significance test will provide a P-value to assess the evidence against the null hypothesis, while the 95% confidence interval will provide the range in which we are 95% confident that the true proportion lies.

For example, we can test the following null hypothesis:

H_0: sample proportion is not different from the hypothesised proportion

Much like constructing a 95% confidence interval, there are two main options when performing a hypothesis test on a single proportion: the first assumes that the proportion follows a Normal distribution, while the second relaxes this assumption.

z-test for testing one sample proportion

The first step in the z-test is to calculate a z-statistic, which is then used to calculate a P-value. The z-statistic is calculated as the difference between the population proportion and the sample proportion divided by the standard error of the population proportion, i.e.

$$z = \frac{(p_{sample} - p_{population})}{\text{SE}(p_{population})}$$

This z-statistic is then compared to the standard Normal distribution to calculate the P-value.

Worked Example 6.2

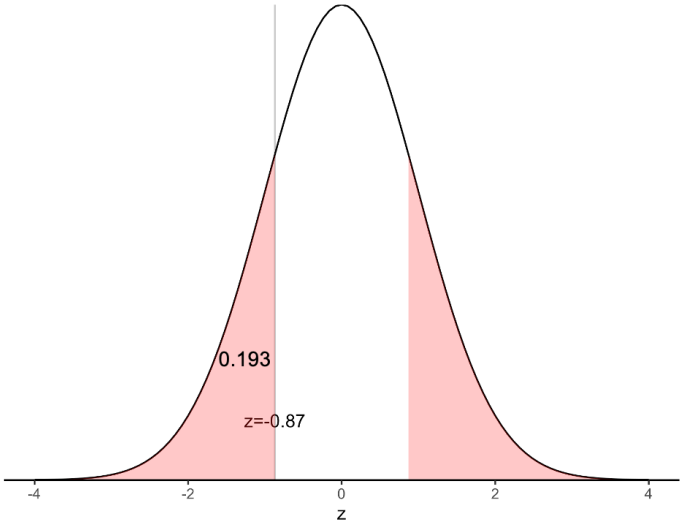
A national census in a country shows that 20% of the population are smokers. A survey of a community within the country that has received a public health anti-smoking intervention shows that 54 of 300 people sampled are smokers (18%). We can calculate a 95% confidence interval around this proportion using the Wilson method, which is calculated as 14.1% to 22.7%.

The researchers are interested in whether the proportion of smoking in this community is the same as the population prevalence of smoking of 20%. The null hypothesis can be written as: H_0: the proportion of smokers in the community is 20% (the same as in the national census).

We can test this by calculating a z-statistic:

$$\begin{aligned} z &= \frac{(0.18 - 0.20)}{\sqrt{\frac{0.20 \times (1-0.20)}{300}}} \\ &= -0.87 \end{aligned}$$

The P-value for the test above can be obtained from a Normal distribution table as $P = 2 \times 0.192 = 0.38$ (using statistical software). This indicates that there is insufficient evidence to conclude that there is a difference between the proportion of smokers in the community and the country. This is consistent with our 95% confidence interval which crosses the null value of 20%.



Binomial test for testing one sample proportion

We can use the binomial distribution to obtain an exact P-value for testing a single proportion. Historically, this was a time consuming process with much hand calculation. These days, statistical software performs the calculations quickly and efficiently, and is the preferred method.

Worked example 6.3

The file `mod06_smoking_status.rds` contains the data for this example. In the data file, smokers are coded as 1 and non-smokers are coded as 0. In jamovi, we can perform the binomial test, while in R, we can use the `prop.test` function to perform a z-test, or the `binom.test` function to perform the binomial test.

The z-test provides a two-sided P-value of 0.39, while the binomial test gives a two-sided P-value of 0.43. Both tests provide little evidence against the hypothesis that the prevalence of smoking in the community is 20%.

6.4 Contingency tables

As introduced in PHCM9794: Foundations of Epidemiology, 2-by-2 contingency tables can be used to examine associations between two binary variables, most commonly an exposure and an outcome. The traditional form of a 2-by-2 contigency table is given in Table 6.1.

Table 6.1: Traditional format for presenting a contingency table

	Outcome present	Outcome absent	Total
Exposure present	a	b	a+b
Exposure absent	c	d	c+d
Total	a+c	b+d	N

When using a statistics program, it is recommended that the outcome and exposure variables are coded by assigning 'absent' as 0 and 'present' as 1, for example 'No' = 0 and 'Yes' = 1. This coding ensures that measures of association, such as the odds ratio or relative risk, are computed correctly. While R does not strictly require this coding to be followed, it is good practice nonetheless.

6.5 A brief summary of epidemiological study types

In this section, we wil present a very brief summary of three study types commonly used in population health research. This topic is covered in much more detail in **PHCM9794: Foundations of Epidemiology**, and more detail can be found in Chapter 4 of Essential Epidemiology (3rd or 4th edition) Webb, Bain and Page (Webb, Bain, and Page (2016)).

Randomised controlled trial

A randomised controlled trial addresses the research question: what is the effect of an intervention on an outcome. In the simplest form of a randomised controlled trial, a group of participants is randomly allocated to a group that receives the treatment of interest or to a control group that does not receive the treatment of interest. Participants are followed up over time, and the outcome is measured at the conclusion of the study.

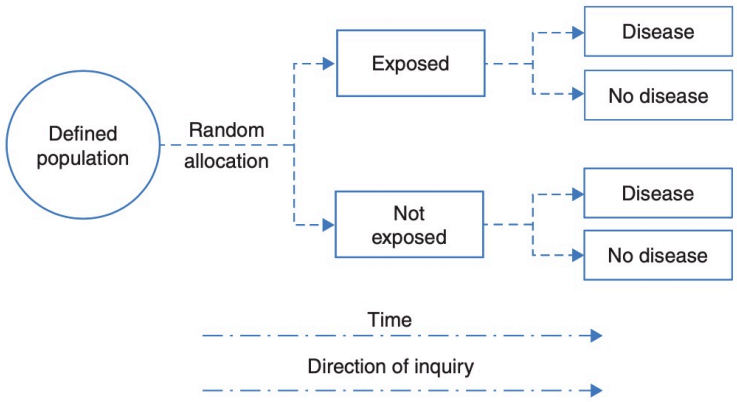


Figure 6.1: The design of a randomised controlled trial [Figure 4.1, Essential Epidemiology]

Cohort study

A cohort study is an *observational study* that addresses the research question: what is the effect of an exposure on an outcome. This research question is similar to that studied in a randomised controlled trial, but the exposure is defined by the participants' circumstances, and not manipulated by the researchers. In a cohort study, participants without the outcome of interest are enrolled, followed over time, and information on their exposure to a factor is measured (either at baseline or over time). At the conclusion of the study, information on the outcome is measured to identify new (incident) cases.

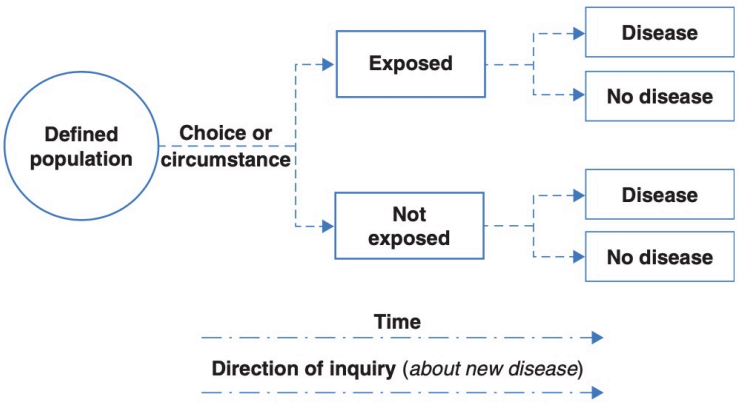


Figure 6.2: The design of a cohort study [Figure 4.2, Essential Epidemiology]

Case control study

While the randomised controlled trial and cohort study begin with a population without the outcome, a case-control study begins by assembling a group with the outcome of interest (cases), and a group without the outcome of interest (controls). The researchers then ask the cases and controls about their previous exposures.

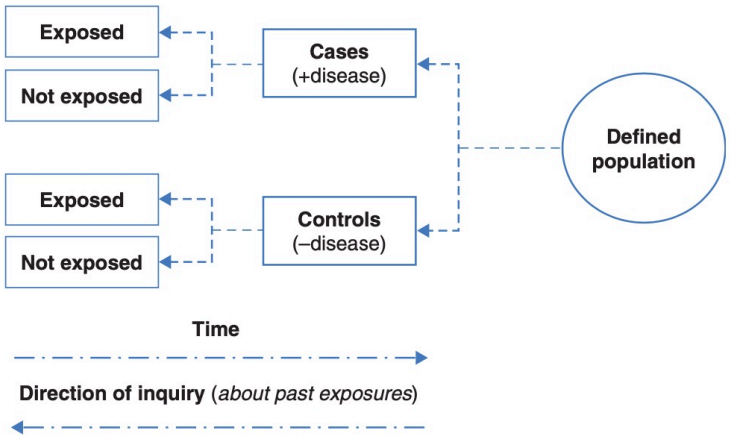


Figure 6.3: The design of a case-control trial [Figure 4.3, Essential Epidemiology]

Cross-sectional study

In a cross-sectional study, the exposure and the outcome are measured at the same time. While this results in a study that is relatively quick to conduct, it does not allow for any temporal relationships to be assessed.

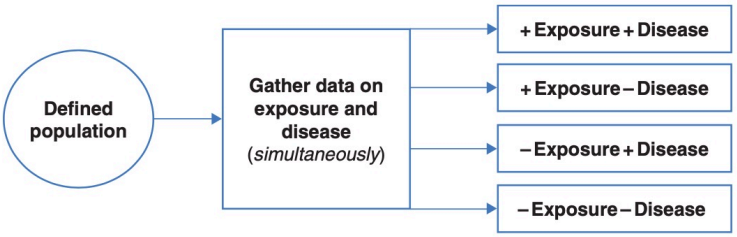


Figure 6.4: The design of a cross-sectional study [Figure 4.4, Essential Epidemiology]

6.6 Measures of effect for epidemiological studies

We can calculate a **relative** measure of association between an exposure and an outcome as either a relative risk or odds ratio. The relative risk is a direct comparison of the risk in the

exposed group with the risk in the non-exposed group, and can only be calculated for a cohort study (including a randomised controlled trial) or a cross-sectional study (where it is also called a *prevalence ratio*).

For cohort studies, randomised controlled trials and cross-section studies, we can calculate an **absolute** measure of association between an exposure and an outcome as a difference in proportions (also known as an *attributable risk*).

For case-control studies, as we sample participants based on their outcome, we can not estimate the risk of the outcome. Hence, calculating a relative risk or risk difference is inappropriate. Instead of calculating risks in a case-control study, we instead calculate *odds*, where the odds of an event are calculated as the number with the event divided by the number without the event.

Table 6.2: Contingency table for a case-control study

	Cases	Controls	Total
Exposure present	a	b	a+b
Exposure absent	c	d	c+d
Total	a+c	b+d	N

In the example in Table Table 6.2, we can calculate the odds of being exposed in the cases as $a \div c$. Similarly, we can calculate the odds of being exposed in the controls as $b \div d$. We can the calculate the *odds ratio* as:

$$\begin{aligned} \text{Odds ratio} &= (a \div c) \div (b \div d) \\ &= \frac{a \times d}{b \times c} \\ &= \frac{ad}{bc} \end{aligned}$$

Note that some authors say we should think of the odds ratio being based on the odds of being a case in the exposed group compared to the odds of being a case in the unexposed group. Here, the exposed group comprises cells “a” and “b”, so the odds of being a case in the exposed group is (a/b). Similarly, for the unexposed group, the odds of being exposed is (c/d). So our odds ratio becomes (a/b) / (c/d). If we rearrange this, we get the same odds ratio as above: (ad)/(bc).

The interpretation of an odds ratio is discussed in detail in PHCM9794: Foundations of Epidemiology, and an excerpt is presented here: The meaning of the calculated odds ratio as a measure of association between exposure and outcome is the same as for the rate ratio (relative risk) where:

- An odds ratio >1 indicates that exposure is positively associated with disease (i.e. the exposure may be a cause of disease);
- An odds ratio < 1 indicates that exposure is negatively associated with disease (i.e. the exposure may be protective against disease); and
- An odds ratio = 1 indicates no association between the exposure and the outcome.

In some situations, related to how well controls are recruited into this study, the odds ratio is a close approximation of the relative risk. Therefore, you may see in some published papers of case control studies the OR interpreted as you would interpret a RR. This should be avoided in this course.

More information about the problems of interpreting odds-ratios as relative risks has been presented by Deeks (1998) and Schmidt and Kohlmann (2008).

Worked Example 6.4

A randomised controlled trial was conducted among a group of patients to estimate the side effects of a drug. Fifty patients were randomly allocated to receive the active drug and 50 patients were allocated to receive a placebo drug. The outcome measured was the experience of nausea. The data is given in the file `mod06_nausea.rds`.

A summary table can be constructed as in Table 6.3.

Table 6.3: Nausea status by drug exposure

	Nausea	No nausea	Total
Active drug	15	35	50
Placebo	4	46	50
Total	19	81	100

We can use jamovi or R to calculate the relative risk (RR=3.75) and its 95% confidence interval (1.34 to 10.51). This tells us that nausea is 3.75 times more likely to occur in the active drug group compared with the placebo group. Because this is a randomised controlled trial, the relative risk would be an appropriate measure of association.

We can confirm the estimated relative risk:

$$\begin{aligned} \text{RR} &= \frac{a/(a+b)}{c/(c+d)} \\ &= \frac{15/(15+35)}{4/(4+46)} \\ &= \frac{0.3}{0.08} \\ &= 3.75 \end{aligned}$$

Worked Example 6.5

A case-control study investigated the association between human papillomavirus and oropharyngeal cancer (D'Souza, et al. NEJM 2007), and the results appear in Table 6.4.

Table 6.4: Association between human papillomavirus and oropharyngeal cancer

	Cases	Controls	Total
HPV Positive	57	14	71
HPV Negative	43	186	229
Total	100	200	300

The odds ratio is the odds of being HPV positive in cases (those with oropharyngeal cancer) compared to the odds of being HPV positive in the controls (those without oropharyngeal cancer):

$$\begin{aligned}\text{OR} &= \frac{a/c}{b/d} \\ &= \frac{57/43}{14/186} \\ &= 17.6\end{aligned}$$

We can use jamovi or R to estimate the odds ratio and its 95% confidence interval. The odds ratio is estimated as 17.6, and its 95% confidence interval is estimated as 9.0 to 34.5.

The interpretation of the confidence intervals for both the relative risk and the odds ratio is the same as for the confidence intervals around other summary measures in that it shows the region in which we are 95% confident that the true population estimate lies.

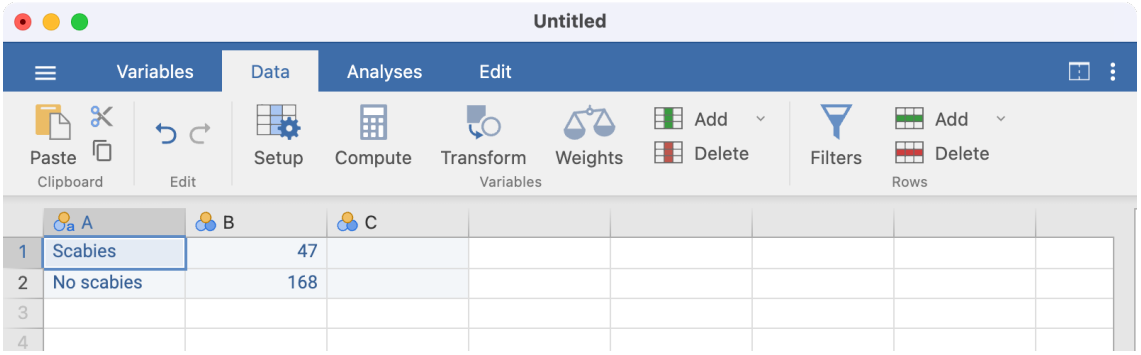
jamovi notes

6.7 95% confidence intervals for proportions

To analyse proportions in jamovi, we use **Frequencies > One Sample Proportion Tests > 2 Outcomes | Binomial Test**. The procedure is slightly different if we are using individual level vs summary data. Here, the procedure will be illustrated as if we have summary data.

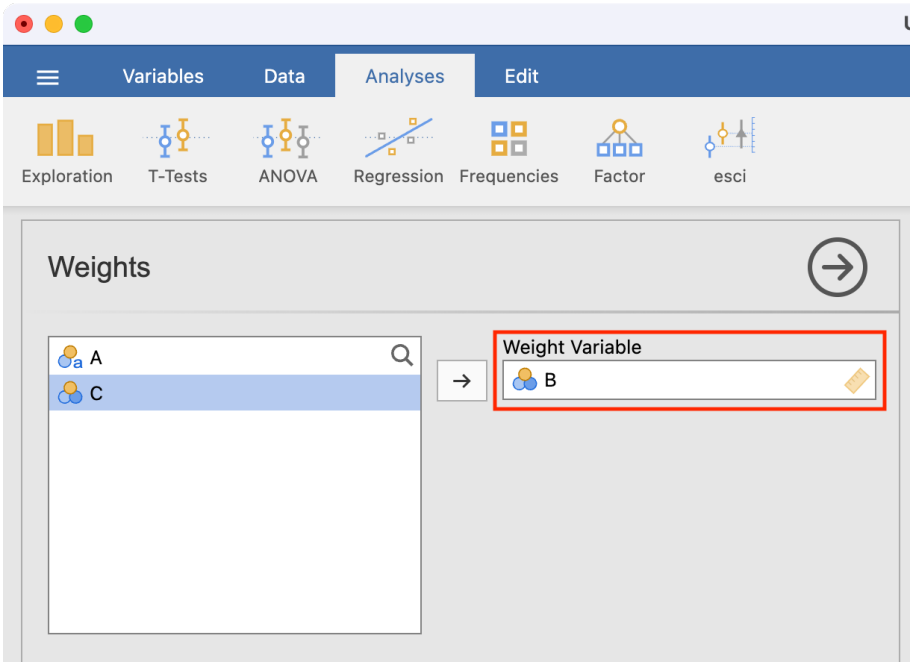
In Worked Example 6.1, 47 children were found to have scabies and 168 (i.e. 215 - 47) were found not to have scabies. We need to enter two columns of data into jamovi: the first indicating whether a count is for scabies or no scabies, and the second representing the number in each category.

Our data are entered as follows:

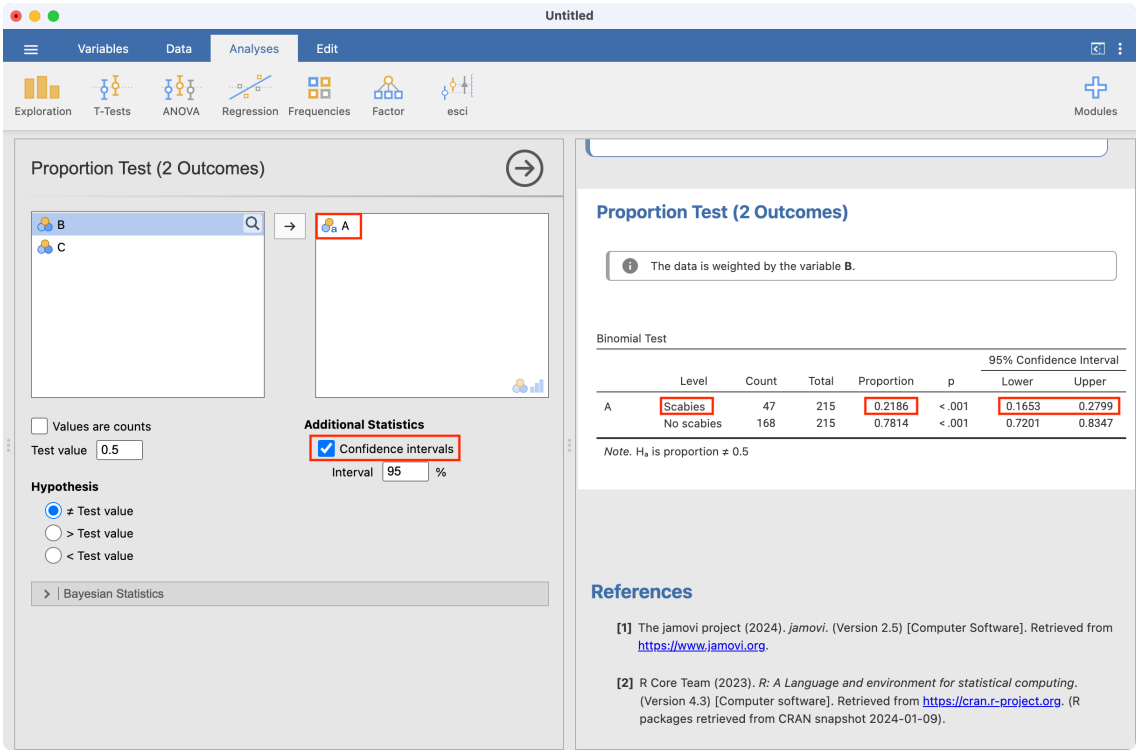


Note that there is no need to name the columns here; using `A` and `B` is fine.

Before we analyse these data, we need to tell jamovi that column `B` represents the count of each category. We do this by using **Data > Weights** and defining `B` as the weight variable. This essentially says that the first row represents 47 observations, and the second row represents 168 observations:



To estimate the proportion with scabies, we use **Frequencies > One Sample Proportion Tests > 2 Outcomes | Binomial Test**, defining Column `A` as the analysis variable and requesting confidence intervals:

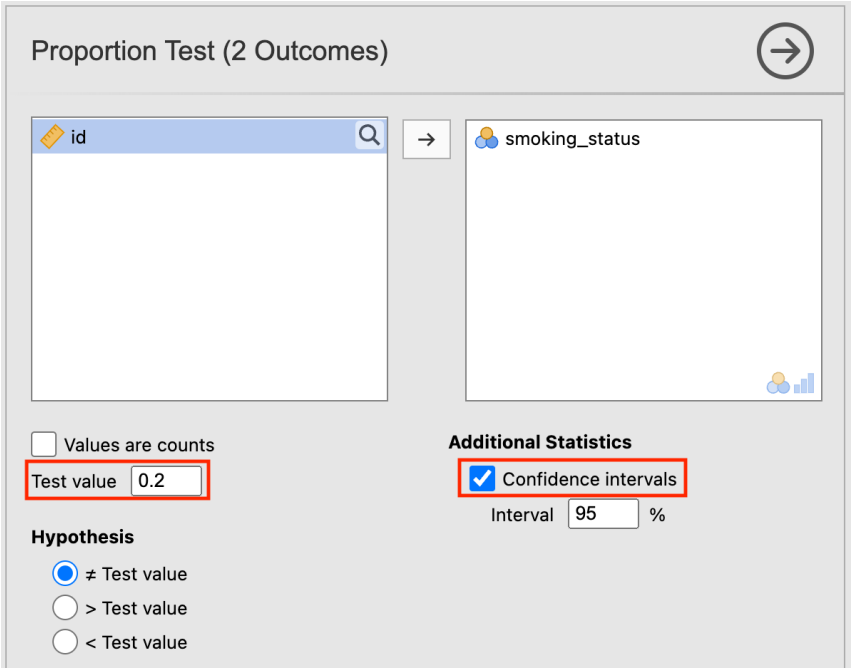


Note that the jamovi provides the proportion with scabies as well as the proportion without scabies.

Binomial test for testing one sample proportion

A binomial test can be performed using a similar approach. Here we consider Worked example 6.3, testing whether a sample is consistent with a true smoking proportion of 20%. `mod06_smoking_status.rds` contains individual level data, so we do not need to use weighting.

After opening the data, click **Analyses > Frequencies > One Sample Proportion Tests > 2 Outcomes | Binomial Test**. Set the **Test value** as 0.2, and tick the **Confidence intervals** box:



The P-value for testing whether the true proportion of smokers is 20% is provided as P=0.43:

Proportion Test (2 Outcomes)

Binomial Test

	Level	Count	Total	Proportion	p	95% Confidence Interval Lower	95% Confidence Interval Upper
smoking_status	Non-smokers	246	300	0.8200	< .001	0.7718	0.8618
	Smokers	54	300	0.1800	0.427	0.1382	0.2282

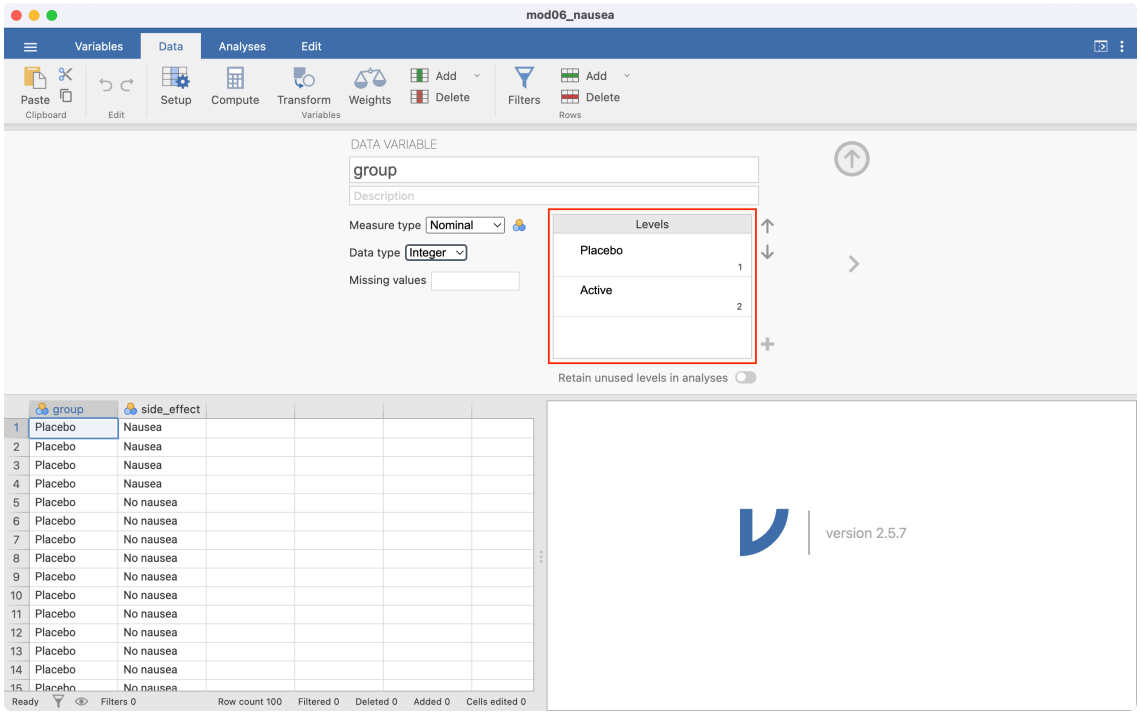
Note. H_a is proportion ≠ 0.2

6.8 Computing a relative risk and its 95% confidence interval

To calculate relative risks, odds ratios and risk differences correctly, **we must define the positive exposure and positive outcome to be the first level of a factor**. When defining an exposure for example, we should define the active treatment or the positive exposure as the first category. When defining an outcome, we should define the category of interest (e.g. disease, or side effect) as the first category.

We will use Worked Example 6.4 to demonstrate calculating a relative risk and its 95% CI, by opening `mod06_nausea.rds`.

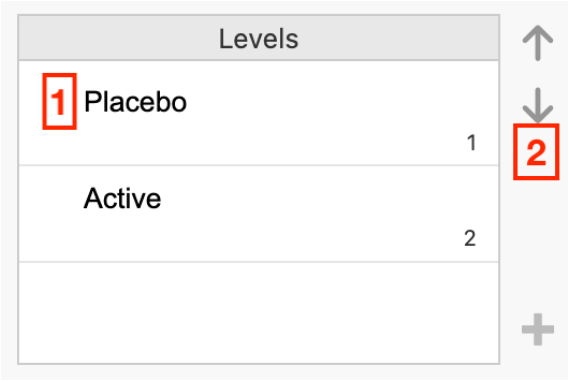
Before analysing these data, we should check that the exposure (`group`) and outcome (`side_effect`) variables have been set up correctly, with the correct level chosen to be of interest. In this example, we will define `Active` as the first level in the `group` factor, and `Nausea` to be the first level of the `side_effect` factor. Let's consider the exposure variable first. Click **Data > Setup** - this will open the following window:



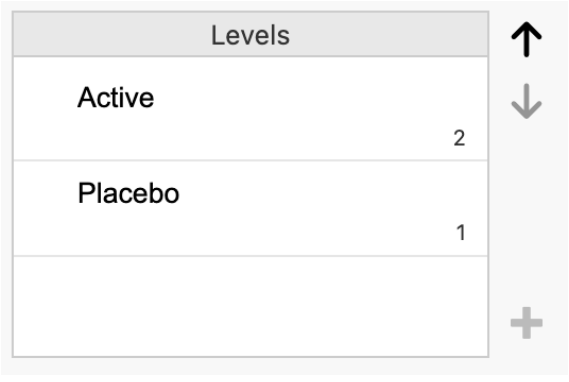
Notice the **Levels** section, highlighted in red. This means that that the `group` variable has been entered in jamovi with *Placebo* as the first category, and *Active* as the second. This ordering means that jamovi will incorrectly consider *Placebo* as "Exposed", and *Active* as "Unexposed". We need to change the ordering, so that *Active* is the first level, and *Placebo* is second.

To re-order the levels:

1. click the *Placebo* cell in the **Levels** box, then
2. click the down arrow to move *Placebo* to be the second level:

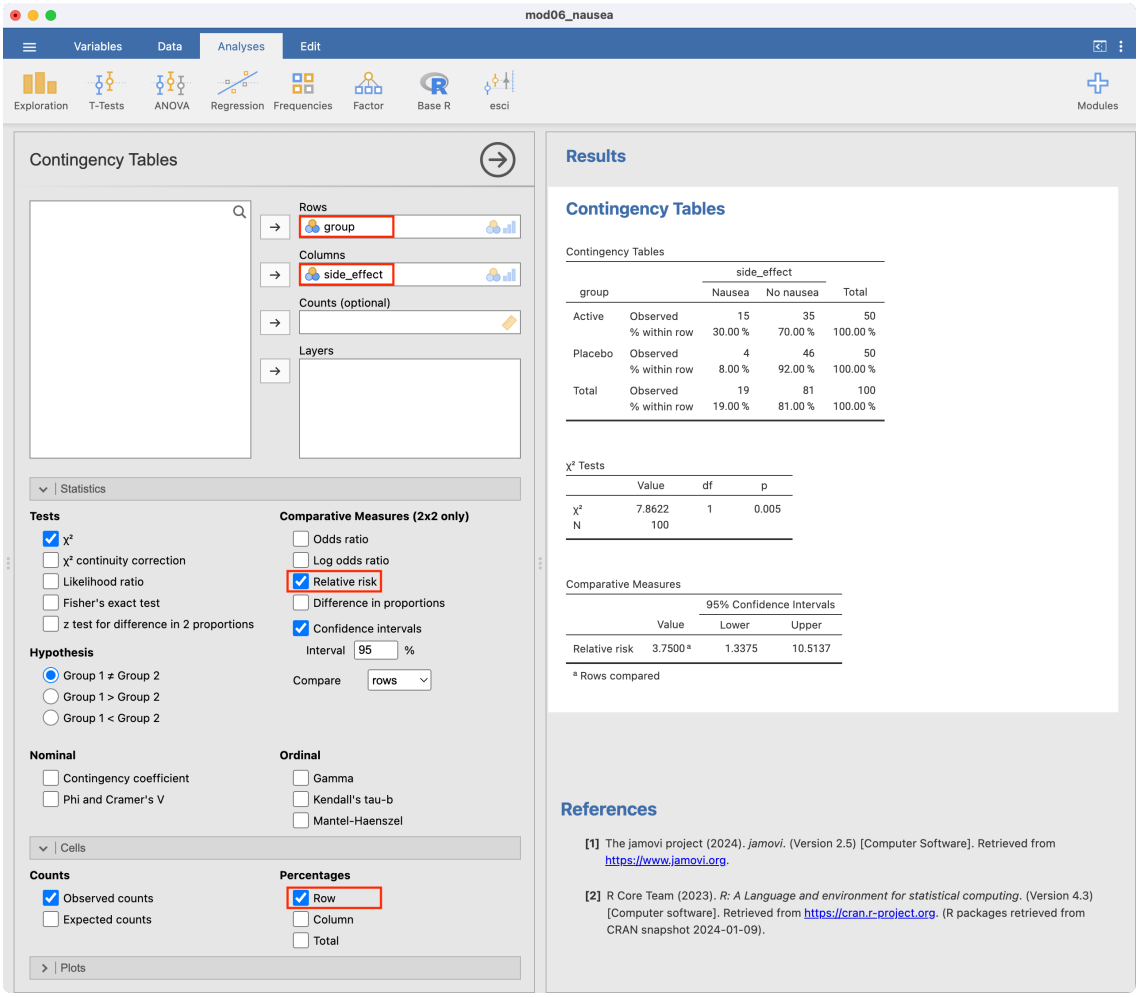


The **Levels** section should appear as below:



Click the `side_effect` column to investigate the ordering of the outcome variable. Repeat the process to set *Nausea* as the first level, and *No nausea* as the second variable.

To construct the 2-by-2 table and calculate a relative risk, we use **Analyses > Frequencies > Independent samples**. Define **Rows** as the exposure variable (`group`), and **Columns** as the outcome (`side_effect`). We can request the row-percents by ticking **Row** in the **Cells** section, and request the relative risk and confidence interval by ticking **Relative risk** in the **Statistics** section:



6.9 Computing other measures of effect

An **Odds ratio** or **Difference in proportions** can be requested in the **Statistics** section.

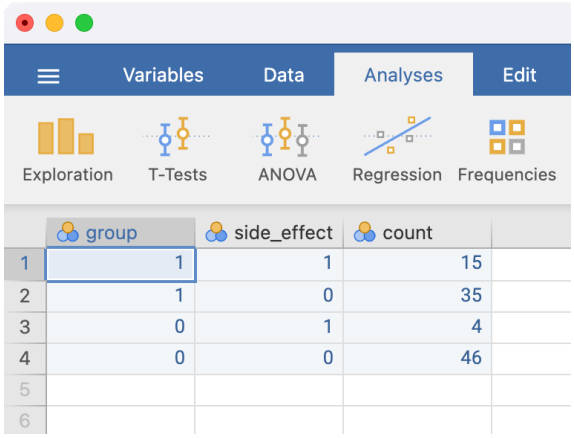
6.10 Working with summarised data

If you only have the cross-tabulated data (i.e. the summarised or aggregated data), you will need to enter your data into a new spreadsheet. For example, to recreate the above analyses, we could re-write the 2-by-2 table as follows:

Group	Side effect	Count
Active	Nausea	15
Active	No nausea	35
Placebo	Nausea	4
Placebo	No nausea	46

We can enter these data in a new spreadsheet, entering the exposure and outcome using the values of 0 or 1. By convention, we use 1 to represent the exposed category, and 0 to represent

the unexposed category. Similarly, we use 1 to represent the outcome category of interest, and 0 to represent the outcome category not of interest. Our entered data would look as follows:

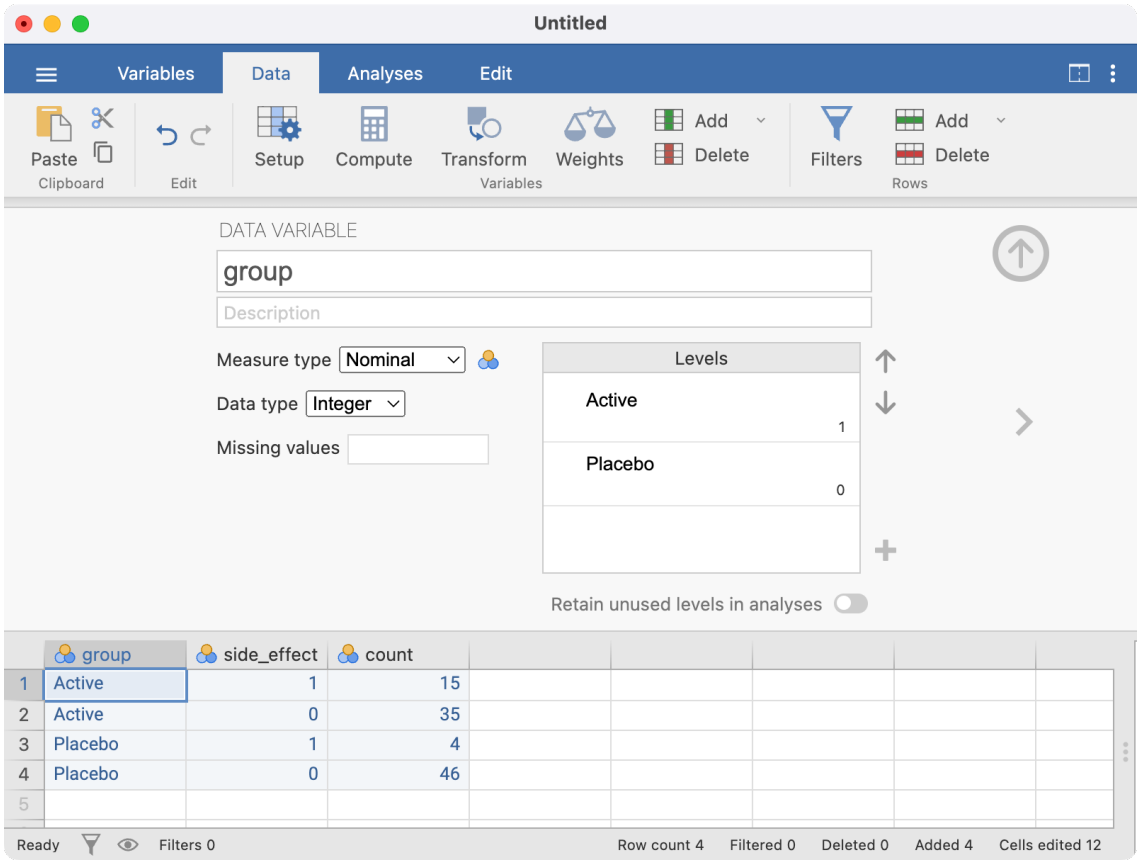


The variable names can be changed in the variables tab in the usual way. It is good practice to label the **levels** of the exposure and outcome variables - this can be done in **Data > Setup**. Click the variable to be defined, and type the labels of each level in the **Levels** section.

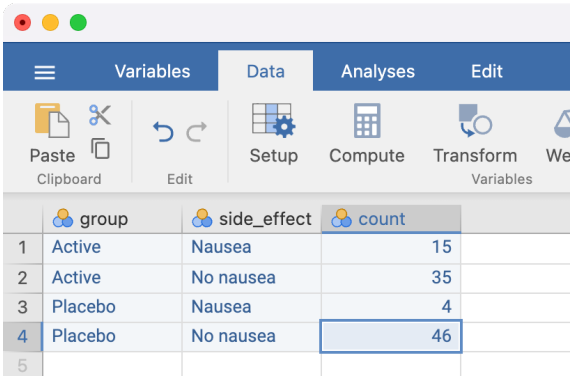
Here, the variable **group** is defined as:

- 1 represents Active
- 0 represents Placebo

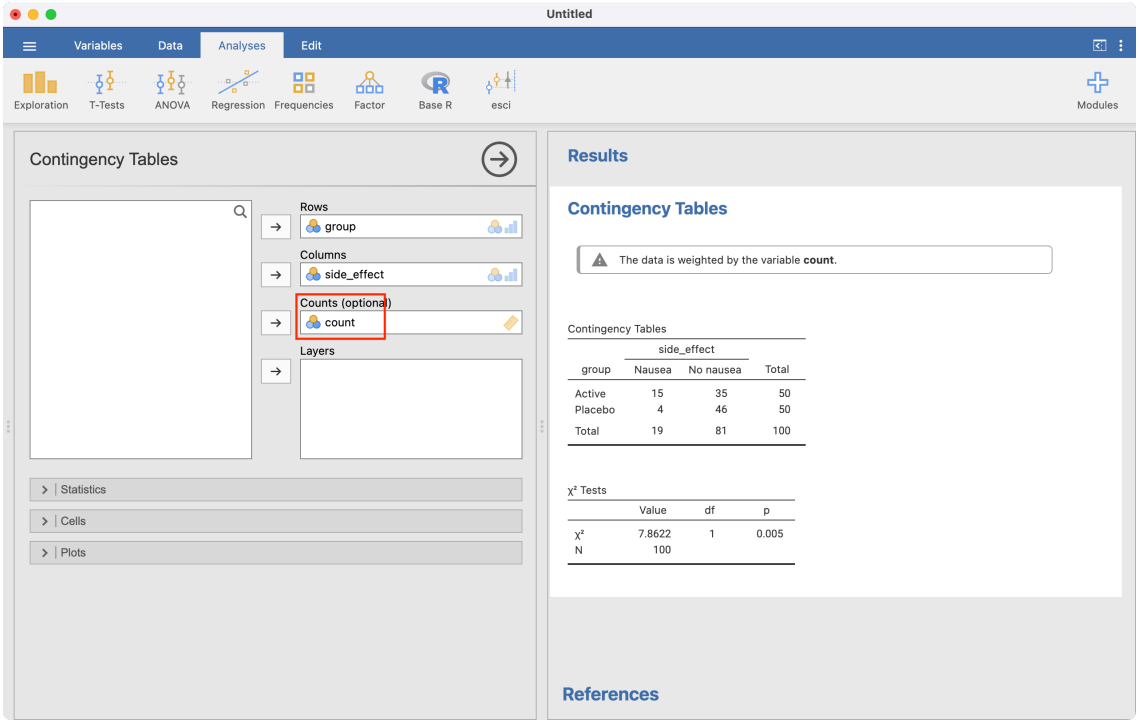
The **Setup** screen is completed as follows:



The `side_effect` variable should be set up using a similar approach, with the final spreadsheet looking like:



The analysis is conducted in the same way as for individual data, but we must now specify the **Counts** field:



R notes

6.11 95% confidence intervals for proportions

We can use the `BinomCI(x=, n=, method=)` function within the `DescTools` package to compute 95% confidence intervals for proportions. Here we specify `x`: the number of successes, `n`: the sample size, and optionally, the `method` (which defaults to Wilson's method).

```
library(DescTools)

BinomCI(x=47, n=215, method='wald')
```

```
           est    lwr.ci    upr.ci
[1,] 0.2186047 0.1633595 0.2738498
```

```
BinomCI(x=47, n=215, method='wilson')
```

```
           est    lwr.ci    upr.ci
[1,] 0.2186047 0.1685637 0.2785246
```

6.12 Significance test for single proportion

We can use the `binom.test` function to perform a significance test for a single proportion: `binom.test(x=, n=, p=)`. Here we specify `x`: the number of successes, `n`: the sample size, and `p`: the hypothesised proportion (which defaults to 0.5 if nothing is entered).

```
binom.test(x=54, n=300, p=0.2)
```

```

    Exact binomial test

data:  54 and 300
number of successes = 54, number of trials = 300, p-value = 0.4273
alternative hypothesis: true probability of success is not equal to 0.2
95 percent confidence interval:
 0.1382104 0.2282394
sample estimates:
probability of success
                  0.18
```

Note that the `binom.test` function also produces a 95% confidence interval around the estimated proportion. This confidence interval is based on the inferior Wald method: *the confidence interval derived from the Wilson method is preferred*.

We can also conduct a z-test for a single proportion:

```
prop.test(x=54, n=300, p=0.2, correct=FALSE)
```

```
    1-sample proportions test without continuity correction

data:  54 out of 300, null probability 0.2
X-squared = 0.75, df = 1, p-value = 0.3865
alternative hypothesis: true p is not equal to 0.2
95 percent confidence interval:
 0.1406583 0.2274332
sample estimates:
   p
0.18
```

6.13 Computing a relative risk and its 95% confidence interval

We will use Worked Example 6.4 to demonstrate calculating a relative risk and its 95% CI:

```
library(jmv)

drug <- readRDS("data/examples/mod06_nausea.rds")

summary(drug)
```

```
     group       side_effect
 Placebo:50   No nausea:81
 Active :50   Nausea   :19
```

By using the `head()` function to view the first six lines of data, we see that both `group` and `side_effect` have been entered as factors. Notice the order in which the factor levels are presented: `group` has the `Placebo` level defined as the first level, and the `Active` level defined as the second; `side_effect` has `No nausea` defined as the first level, and the `Nausea` level defined as the second.

We will use `jmv` to calculate relative risks, odds ratios and risk differences. To calculate these estimates correctly, **we must define the positive exposure and positive outcome to be the first level of a factor**. When defining an exposure for example, we should define the active treatment or the positive exposure as the first category. When defining an outcome, we should define the category of interest (e.g. disease, or side effect) as the first category.

In this example, we will define `Active` as the first level in the `group` factor, and `Nausea` to be the first level of the `side_effect` factor.

We can do this using the `relevel()` function, which re-orders the levels of a factor so that the level specified is defined as the first level, and the others are moved down:

```
# Define "Active" as the first level of group:
drug$group <- relevel(drug$group, ref="Active")


# Define "Nausea" as the first level of side_effect:
drug$side_effect <- relevel(drug$side_effect, ref="Nausea")
```

Upon re-leveling the factors, we can check that the levels of interest have been defined as the first levels:

```
summary(drug)
```

```
     group       side_effect
 Active :50   Nausea   :19
 Placebo:50   No nausea:81
```

To construct the 2-by-2 table and calculate a relative risk, we use the `contTables()` function in `jmv`. We request the row-percents using `pcRow = TRUE` and the relative risk and confidence interval using `relRisk = TRUE`:

```
contTables(data=drug,
           rows=group, cols=side_effect,
           pcRow=TRUE, relRisk = TRUE)
```

```
 CONTINGENCY TABLES

 Contingency Tables

   group                      Nausea       No nausea    Total

   Active     Observed             15           35          50
              % within row   30.00000     70.00000    100.00000

   Placebo    Observed              4           46          50
              % within row    8.00000     92.00000    100.00000

   Total      Observed             19           81         100
              % within row   19.00000     81.00000    100.00000


  ² Tests

        Value       df     p

    ²   7.862248     1    0.0050478
   N          100


 Comparative Measures

                    Value         Lower       Upper

   Relative risk    3.750000      1.337540    10.51370

     Rows compared
```

If you only have the cross-tabulated data (i.e. aggregated), you will need to enter your data into a new data frame. For example, to recreate the above analyses, we can re-write the 2-by-2 table as follows:

Group	side_effect	Number
Active	Nausea	15
Active	No nausea	35
Placebo	Nausea	4
Placebo	No nausea	46

We can enter these data in a dataframe, comprising three vectors, as follows:

```
drug_aggregated <- data.frame(
  group = c("Active", "Active", "Placebo", "Placebo"),
  side_effect = c("Nausea", "No nausea", "Nausea", "No nausea"),
  n = c(15, 35, 4, 46)
)
```

We need to define group and side_effect as factors. Here we must define the `levels` **in the order we want the categories to appear in the table**. Note that as group and side_effect are entered as text variables, we can omit `labels` command when defining the factors, and the factor will be labelled using the text entry:

```
drug_aggregated$group <- factor(drug_aggregated$group,
                                levels=c("Active", "Placebo"))

drug_aggregated$side_effect <- factor(drug_aggregated$side_effect,
                                      levels=c("Nausea", "No nausea"))
```

We can calculate the relative risk using the summarised data in the same was done previously. However, we need to include the number of observations in each cell using the `counts` command:

```
contTables(data=drug_aggregated,
           rows=group, cols=side_effect, count=n,
           pcRow=TRUE, relRisk = TRUE)
```

```
CONTINGENCY TABLES

 Contingency Tables

   group                      Nausea       No nausea    Total

   Active     Observed        15.000000     35.00000     50.00000
              % within row     30.00000     70.00000    100.00000

   Placebo    Observed         4.000000     46.00000     50.00000
              % within row      8.00000     92.00000    100.00000

   Total      Observed        19.000000     81.00000    100.00000
              % within row     19.00000     81.00000    100.00000



  [2] Tests

         Value       df    p
```

```
   ²     7.862248      1     0.0050478
  N           100


 Comparative Measures

                    Value          Lower        Upper

   Relative risk    3.750000       1.337540     10.51370

     Rows compared
```

6.14 Computing a difference in proportions and its 95% confidence interval

We can use the `contTables` function to obtain a difference in proportions and its 95% CI, by specifying `diffProp=TRUE`:

```
contTables(data=drug,
           rows=group, cols=side_effect,
           pcRow=TRUE, diffProp=TRUE)


 CONTINGENCY TABLES

 Contingency Tables

   group                       Nausea       No nausea    Total

   Active     Observed              15           35          50
              % within row     30.00000     70.00000    100.00000

   Placebo    Observed               4           46          50
              % within row      8.00000     92.00000    100.00000

   Total      Observed              19           81         100
              % within row     19.00000     81.00000    100.00000


 ² Tests

        Value       df     p

   ²    7.862248      1     0.0050478
   N          100


 Comparative Measures

                                   Value          Lower          Upper

   Difference in 2 proportions     0.2200000      0.07238986     0.3676101

     Rows compared
```

6.15 Computing an odds ratio and its 95% confidence interval

We can use the `contTables` function to obtain an odds ratio and its 95% CI, by specifying `odds=TRUE`. Here we will use the summarised HPV data from Module 6.

```
hpv <- data.frame(
  hpv = c("HPV +", "HPV +", "HPV -", "HPV -"),
  cancer = c("Case", "Control", "Case", "Control"),
  n = c(57, 14, 43, 186)
)

hpv$cancer <- factor(hpv$cancer, levels=c("Case", "Control"))
hpv$hpv <- factor(hpv$hpv, levels=c("HPV +", "HPV -"))

contTables(data=hpv,
           rows=hpv, cols=cancer, count=n,
           odds = TRUE)
```

```
 CONTINGENCY TABLES

 Contingency Tables

   hpv      Case         Control      Total

   HPV +     57.00000     14.00000     71.00000
   HPV -     43.00000    186.00000    229.00000
   Total    100.00000    200.00000    300.00000



  ² Tests

         Value       df     p

    ²    92.25660     1     < .0000001
   N          300


 Comparative Measures

                 Value       Lower       Upper

   Odds ratio    17.61130    8.992580    34.49041
```

Activities

Activity 6.1

In a clinical trial involving a dietary intervention, 150 adult volunteers agreed to participate. The investigator wanted to know whether this sample was representative of the general population. One interesting finding was that 90 of the participants drink alcohol regularly compared to 70% of the general population.

a) State the null hypothesis.
b) Calculate the proportion of regular drinkers (and its 95% confidence interval) in the sample using software.
c) Conduct a hypothesis test to decide if the sample of volunteers is representative of the population.
d) Repeat (b) and (c) using the data saved in `Activity_6.1.rds`.

Activity 6.2

A survey was conducted of a random sample of upper primary school children to measure the prevalence of asthma using questionnaires completed by the parents. A total of 514 children were enrolled. Use the dataset `Activity_6.2.rds` for this activity.

a) What type of study was used to collect these data? Based on this, which measure of effect would you use to summarise the association between gender and asthma symptoms?
b) Calculate the relevant measure of effect (with its 95% confidence interval).

Activity 6.3

A study is conducted to test the hypothesis that the observed frequency of a certain health outcome is 30%. If the results yield a CI around the sample proportion that extends from 23.8 to 30.2, what can you say about the evidence against the null hypothesis?

Activity 6.4

In an experiment to test the effect of Vitamin C on IQ scores, the following confidence intervals were estimated around the percentage with improved scores for five different populations (Table 6.7):

Table 6.7

Population	% with improved IQ	95% confidence interval
1	35.0	32.0 to 38.0
2	29.5	25.0 to 34.0
3	43.5	42.0 to 45.0
4	30.5	20.0 to 41.0
5	24.5	21.0 to 28.0

a) Which CI is the most precise?
b) Which CI implies the largest sample size?
c) Which CI is the least precise?
d) Which CI most strongly supports the conclusion that vitamin C increases IQ score and why?
e) Which would most likely to stimulate the investigator to conduct an additional experiment using a larger sample size?

Activity 6.5

In a study to determine the cause of mortality, 89 people were followed up for 5 years. The participants are classified into two groups of those who did or did not have a heart attack. At the end of the follow-up 15 people died among them 10 had a heart attack. Among the 74 survivors 35 had a heart attack.

Present the data in a 2-by-2 table and calculate relative risk of death from heart attack with 95% confidence interval.

Supplementary Activity 6.6

The betel nut, the seed of the areca palm, is grown in the tropical Pacific and Asia and is a commonly use psycho-active substance. Betel nut is often chewed, wrapped inside betel leaves or in combination with tobacco. Chewing betel nut has been linked with a range of health issues.

A case-control study was conducted to assess the association between chewing betel nut and obstructive coronary artery disease. 293 men with obstructive coronary artery disease were recruited, and 88 reported having chewed betel nut. Of the 720 healthy control men recruited, 57 reported having chewed betel nut.

Construct a 2-by-2 table to report the data provided. Calculate the most appropriate measure of effect and its 95% confidence interval.

Supplemntary Activity 6.7

Suppose a clinical trial is conducted to test the effectiveness of a drug, spectinomycin, for treating gonorrhoea in females. Forty-six patients are given a 4-g daily dose of the drug and are seen 1 week later, at which time 6 of the patients still have gonorrhoea.

What is the estimated effectiveness (and 95% confidence interval) of the drug?

Module 7

Hypothesis testing for categorical data

Learning objectives

By the end of this module you will be able to:

- Use and interpret the appropriate test for testing associations between categorical data;
- Conduct and interpret an appropriate test for independent proportions;
- Conduct and interpret a test for paired proportions;

Optional readings

Kirkwood and Sterne (2001); Chapter 17. [UNSW Library Link]

Bland (2015); Chapter 13. [UNSW Library Link]

7.1 Introduction

In Module 6, we estimated the 95% confidence intervals of proportions and measures of association for categorical data and conducted a significance test comparing a sample proportion to a known value.

When both the outcome variable and the exposure variable are categorical, a chi-squared test can be used as a formal statistical test to assess whether the exposure and outcome are related. The P-value obtained from a chi-squared test gives the probability of obtaining the observed association (or more extreme) if there is in fact no association between the exposure and outcome.

In this Module, we also include tests for a difference in proportion for paired data.

Worked Example

We are using the randomised controlled trial as given in Worked Example 6.4 on the nauseating side effect of a drug.

The research question is whether the active drug resulted in a different rate of nausea than the placebo drug. This is equivalent to testing whether there is an association between nausea and type of drug received (active or placebo). Thus, we will test the null hypothesis that the experience of nausea and the treatment are not related to one another. The null hypothesis is:

- H_0: The proportion with nausea in the active drug group is the same as the proportion with nausea in the placebo drug group.

The alternative hypothesis can be stated as:

- H_a: The proportion with nausea in the active drug group is different to the proportion with nausea in the placebo drug group.

7.2 Chi-squared test for independent proportions

A chi-squared test is used to test the null hypothesis that of no association between two categorical variables. First a contingency table is drawn up and then we estimate the counts of each cell (i.e. a, b, c and d) that would be expected if the null hypothesis was true. The row and column totals are used to calculate expected counts in each cell of the contingency table as follows:

Expected count = (Row count × Column count) / Total count

Statistical software will do this for us, as described in the jamovi or R sections in this Module.

A chi-squared value is then calculated to compare the expected counts (E) in each cell with the observed (actual) cell counts (O). The calculation is as follows:

$\chi^2 = \sum \frac{(O-E)}{E}^2$

with [Number of rows $-$ 1] $\times$ [Number of columns $-$ 1] degrees of freedom.

As for many statistics, the deviations between the observed and expected values are squared to prevent the negative and positive values balancing one another out.

If the expected counts are close to the observed counts, the chi-squared statistic will be close to zero, and the P-value will be close to 1. The larger the difference between the observed and expected counts, the larger the chi-squared statistic becomes (and the smaller the P-value). A large chi-squared statistic provides more evidence of an association between the exposure and outcome.

Assumptions for using a Pearson's chi-squared test

The assumptions that must be met when using Pearson's chi-squared test are that:

- each observation must be independent;
- each participant is represented in the table once only;
- at least 80% of the expected cell counts should exceed a value of five;
- all expected cell counts should exceed a value of one.

The first two assumptions are dictated by the study design. The last two assumptions relate to the numbers in the cells and should be explored when running the test. There should not be too many cells with low expected counts.

Worked Example 7.1

We will revisit Worked Example 6.4, investigating the relationship between nausea and drug exposure:

Table 7.1: Nausea status by drug exposure

	Nausea	No nausea	Total
Active	15 (30%)	35 (70%)	50 (100%)
Placebo	4 (8%)	46 (92%)	50 (100%)
Total	19 (19%)	81 (81%)	100 (100%)

We can see from the row percentages that 8% of patients in the placebo group experienced nausea compared to 30% of patients in the active group. If no association existed, we would expect to find approximately the same percent of patients with nausea in each group. Statistical software can calculate the values we would expect if there was no association between nausea and drug exposure (i.e. the expected counts):

Table 7.2: Expected counts of nausea status by drug exposure

	Nausea	No nausea	Total
Active	9.5	40.5	50
Placebo	9.5	40.5	50
Total	19	81	100

For the data being considered from Worked Example 7.1 all cells have an expected count greater than 5 and that the minimum cell count is 9.5. Therefore, it is appropriate to use the Pearson's Chi-Squared test. Note that the 'Expected' counts are higher for the groups with 'No nausea' because 'No nausea' is more prevalent in the sample than 'Nausea'.

The chi-squared statistic is calculated as 7.86 with 1 df, giving a P-value of 0.005. Combining these results with the estimated relative risk (from Module 6), we can state:

The proportion with nausea in those who received the active drug is 30%, compared to 8% in those who received the placebo drug. Nausea was more frequent in those who received the active drug (Relative Risk = 3.75, 95% CI: 1.34 to 10.51). There is strong evidence that the proportion with nausea differs between the two groups (χ^2 = 7.86 with 1 df, P=0.005).

Fisher's exact test

If small expected cell counts are present, Fisher's exact test can be used instead. More information on Fisher's exact test can be found in Chapter 13 of An Introduction to Medical Statistics, Bland (2015), or Section 17.3 of Essential Medical Statistics, Kirkwood and Sterne (2001). The computation of Fisher's exact test is complex, and best conducted by statistical software.

A reasonable question could be posed: why not conduct Fisher's exact test by default? The answer to this is complex.

Fisher's exact test has quite a restrictive assumption: we assume that the totals of the rows and columns are fixed before we conduct the study.

From Worked Example 7.1, this would be saying that we knew we would end up with 50 people in the active treatment arm, and 50 people in the placebo. This seems reasonable, we can design our study to randomise equal groups. However, Fisher's exact test also assumes that we know we will obtain 19 people with nausea and 81 people without nausea. We cannot possibly know this before we do the study.

In the case where we cannot assume that the totals of the rows and columns are fixed before we conduct the study, it can be shown that Fisher's exact test will be conservative (we will be less likely to reject the null hypothesis when it is false, or in other words, the P-value will be larger than it should be).

While there are other tests that perform better than Fisher's exact test, most of the time we live with this conservative test when we have to (i.e. for small expected cell counts) because Fisher's exact test is so widely known.

Pragmatically, we use the standard (Pearson) chi-square when we can, and Fisher's exact test only when we have small expected cell counts.

7.3 Chi-squared tests for tables larger than 2-by-2

Chi-squared tests can also be used for tables larger than a 2-by-2 dimension. When a contingency table larger than 2-by-2 is used, say a 4-by-2 table if there were 4 exposure groups, the Pearson's chi-squared can still be used.

Worked Example 7.2

The file `mod07_allergy.rds` contain information about the severity of allergic reaction, coded as absent, slight, moderate or severe. We can test the hypothesis that the severity of allergy is not

Table 7.3: Association between sex and allergy severity

Table 7.4: Observed counts

Sex	Non-allergenic	Slight allergy	Moderate allergy	Severe allergy	Total
Female	150 (62.0%)	50 (20.7%)	27 (11.2%)	15 (6.2%)	242 (100%)
Male	137 (53.1%)	70 (27.1%)	32 (12.4%)	19 (7.4%)	258 (100%)
Total	287 (57.4%)	120 (24.0%)	59 (11.8%)	34 (6.8%)	500 (100.0%)

Table 7.5: Expected counts

Sex	Non-allergenic	Slight allergy	Moderate allergy	Severe allergy	Total
Female	138.9	58.1	28.6	16.5	242.0
Male	148.1	61.9	30.4	17.5	258.0
Total	287.0	120.0	59.0	34.0	500.0

different between males and females. To do this we can use a two-way tabulation to obtain Table 7.3 which shows the counts, expected counts and the percent of females and males who fall into each severity group for allergy. The table shows that the percentage of males is higher in each of the categories of severity (slight, moderate, severe) than the percentage of females.

The Pearson chi-squared statistic is calculated as 4.31, with 3 degrees of freedom, providing a P-value of 0.23. Therefore, there is little evidence of an association between gender and the severity of allergy.

7.4 McNemar's test for categorical paired data

If a binary categorical outcome is measured in a paired study design, McNemar's statistic is used. This statistic is a form of chi-square applied to a paired situation. A Pearson's chi-squared test cannot be used because the measurements are not independent. However, McNemar's test can be used to assess whether there is a significant change in proportions between two time points or between two conditions, or whether there is a significant difference in proportions between matched cases and controls.

For McNemar's test, the data are displayed as shown in Table 7.6. Cells 'a' and 'd' called concordant cells because the response was the same at both baseline and follow-up or between matched cases and controls. Cells 'b' and 'c' are called discordant cells because the responses between the pairs were different. For a follow-up study, the participants in cell 'c' had a positive response at baseline and a negative response at follow-up. Conversely, the participants in cell 'b' had a negative response at baseline and a positive response at follow-up.

For other types of paired data such as twins or matched cases and controls, the data are similarly displayed with the responses of one of the pairs in the columns and the responses for the other of the pairs in the rows. For paired data, the grand total 'N' is always the number of pairs and not the total number of participants.

Table 7.6: Table layout for testing matched proportions

	Negative at follow-up	Positive at follow-up	Total
Negative at baseline	a	b	a + b
Positive at baseline	c	d	c + d
Total	a + c	b + d	N

Worked Example 7.3

Two drugs labelled A and B have been administered to patients in random order so that each patient acts as their own control. The dataset `mod07_drug_response.rds` is available on Moodle. The null hypothesis is as follows:

- H_0: The proportion of patients who do better on drug A is the same as the proportion of patients who do better on drug B

Counts and overall percentages are presented in . From the "Total" row in the table, we can see that the number of patients who respond to drug A is 41 (68%) and from the "Total" column the number who respond to drug B is less at 35 (58%), that is there is a difference of 10%.

Table 7.7: Worked Example 7.3: Paired data

	Response to Drug B	No response to Drug B	Total
Response to Drug A	21 (35%)	20 (33%)	41 (68%)
No response to Drug A	14 (23%)	5 (8%)	19 (32%)
Total	35 (58%)	25 (42%)	60 (100%)

The difference in the paired proportions is calculated using the simple equation:

$$p_A - p_B = \frac{(b-c)}{N}$$

Here, $p_A - p_B = \frac{(20-14)}{60} = 0.1$

The cell counts show that 20 patients responded to Drug A but not to drug B, and 14 patients responded to Drug B but not to drug A. McNemar's statistic is computed from these two discordant pairs (labelled as 'b' and 'c') as follows:

$$X^2 = \frac{(b-c)^2}{b+c}$$

with 1 degree of freedom. Using our worked example, the McNemar's chi-squared statistic is calculated as 1.06 with 1 degree of freedom, giving a P-value of 0.3.

Note that some packages also calculate an "Exact P-Value". The standard McNemar's chi-squared statistic is generally recommended, unless the sum of the discordant cells is small (Kirkwood and Sterne define small as less than 10; Section 21.3, Kirkwood and Sterne 2001)). Here, $b + c = 34$, so reporting the standard McNemar's chi-squared statistic is appropriate.

As described above, the difference in proportions can be calculated. A 95% confidence interval for this difference can be obtained using statistical software.

In this study of 60 participants, where each participant received both drugs, 41 (68%) responded to Drug A and 35 (58%) responded to Drug B. The difference in the proportions responding is estimated as 10% (95% CI -11% to 31%). There is no evidence that the response differed between the two drugs (McNemar's chi-square=1.06 with 1 degree of freedom, P=0.3).

7.5 Summary

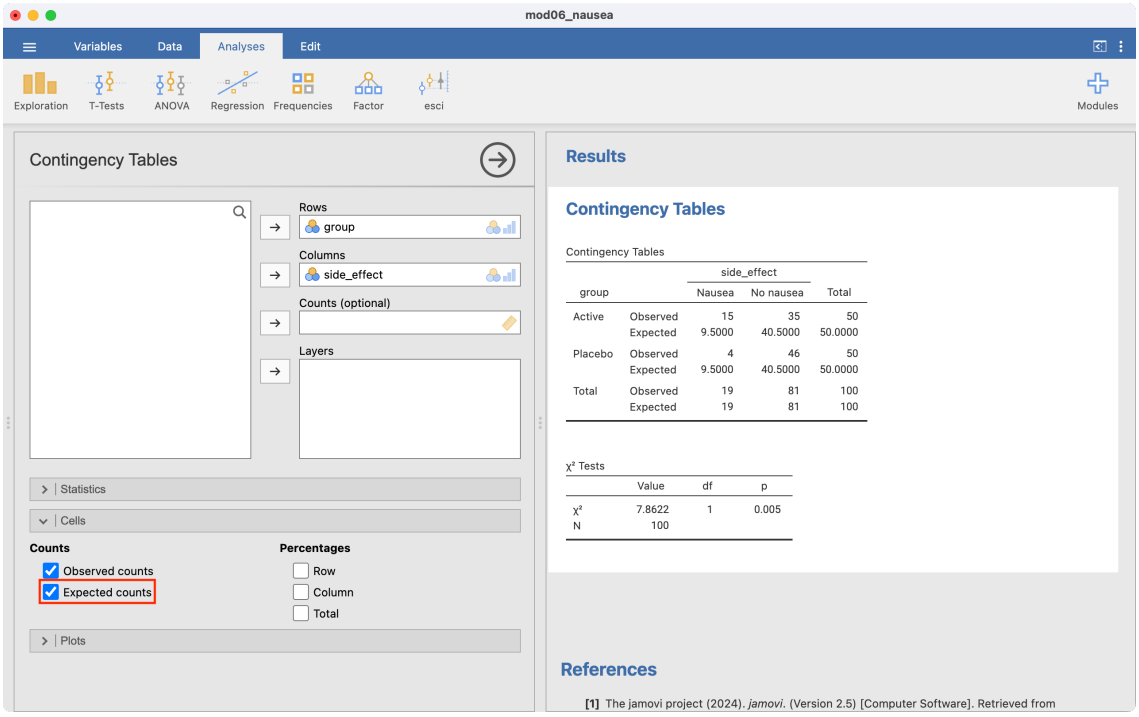
In Module 6, we estimated proportions and measures of association for categorical data and conducted a one-sample test of proportions. In this module, we conduct significance tests for two or more independent proportions using the chi-squared test. The chi-squared test can also be used to conduct a significance test when there are more than two categories in both variables. The McNemar's test is used when we have paired data.

jamovi notes

7.6 Pearson's chi-squared test for individual-level data

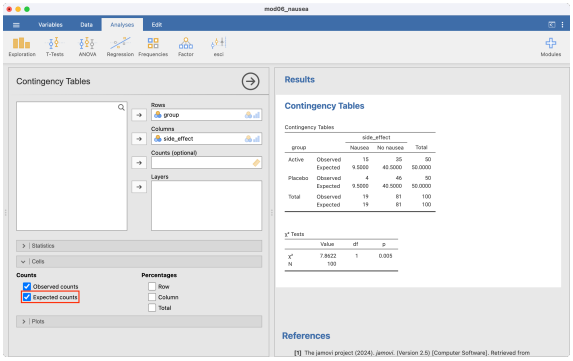
Conducting a Pearson's chi-squared test is done automatically when producing a 2-by-2 table in jamovi. All that we need to do is check that the assumptions of the Pearson's chi-squared test are met, by examining the Expected counts in each cell.

Worked Example 7.1 is completed as in Module 6, but we first examine the Expected counts (remember that we need to change the order of the levels for the exposure and outcome variables):



After confirming that there are no cells with small expected frequencies, we can interpret the chi-square test. The last section reports the chi-squared test statistic which has a value of 7.86 with 1 degree of freedom and a P-value of 0.005.

If there are small values of expected frequencies, Fisher's exact test can be requested in the **Statistics** section:

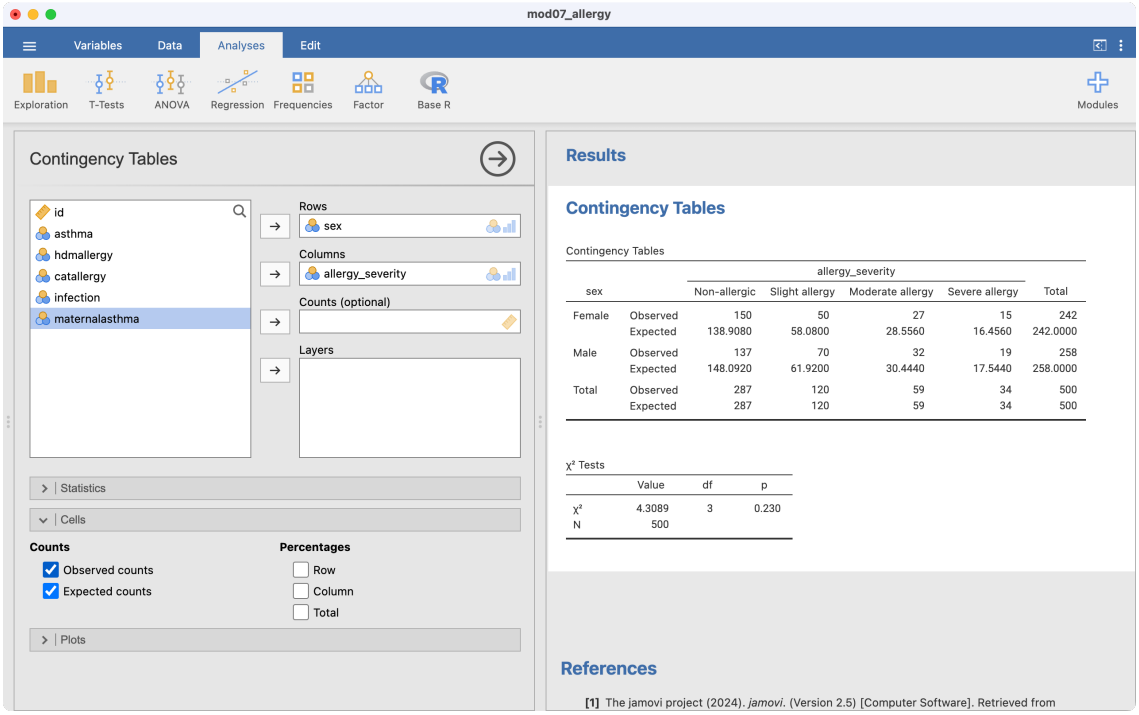


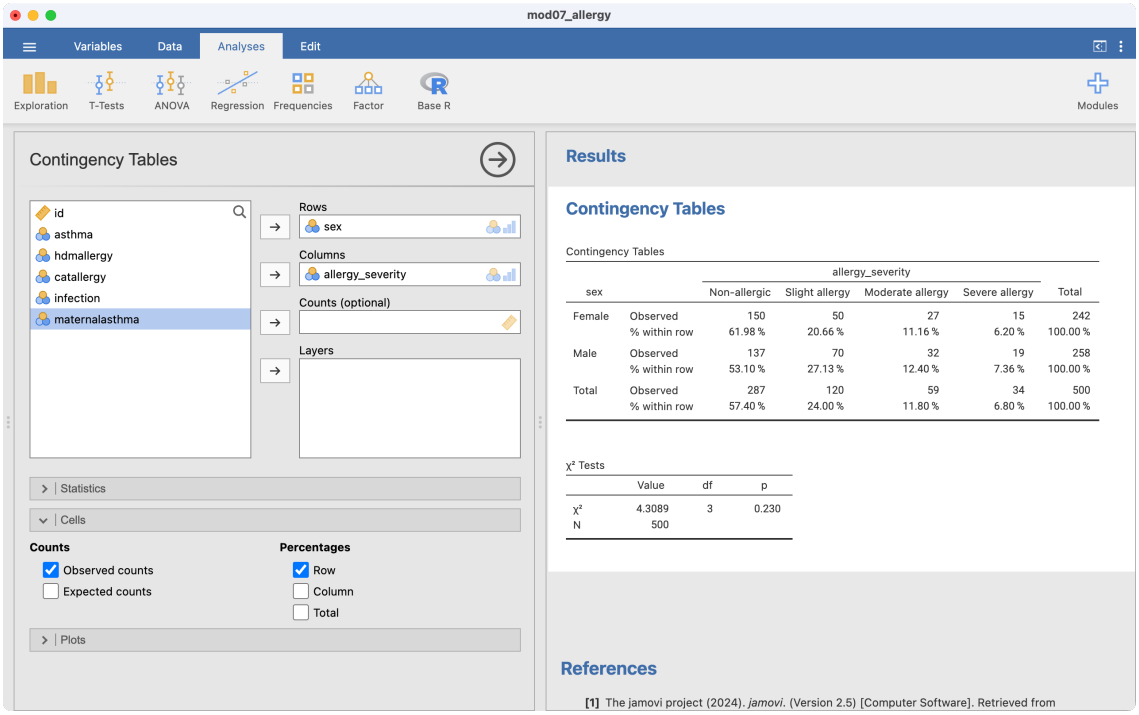
7.7 Pearson's chi-squared test for summarised data

When you only have the summarised date (for example, the cross-tabulated data), you need to enter the summarised data manually as we did in Module 6. After defining the **Count variable**, the Pearson chi-squared test is calculated automatically.

7.8 Chi-squared test for tables larger than 2-by-2

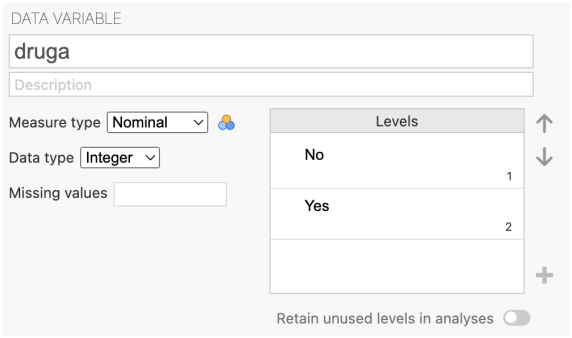
Use the data in `mod07_allergy.rds`. We use similar steps as described above for a 2-by-2 table:



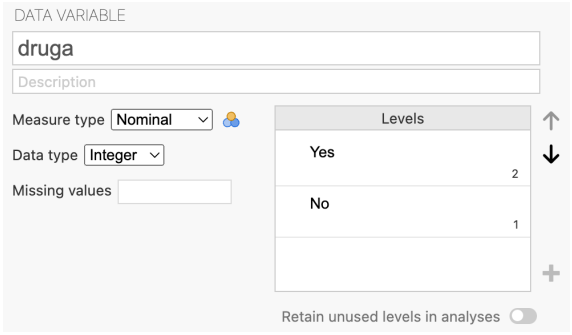


7.9 McNemar's test for paired proportions

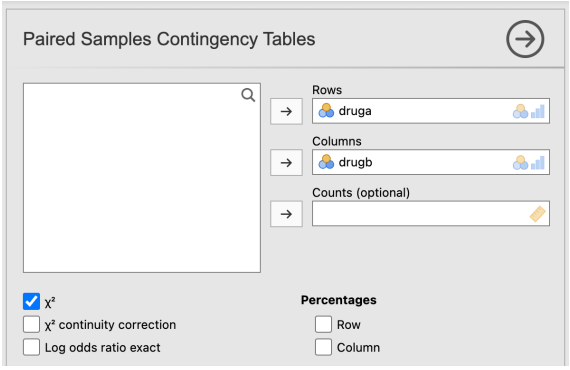
To perform this test in jamovi, we will use the dataset `mod07_drug_response.rds`. As with all 2-by-2 tables, we should check that the variables are set up with the first level of the variable being the outcome of interest using **Data > Setup**:



Both variables, `druga` and `drugb` are set up with "No" being the first level. We need to change the order and define "Yes" as the first level (See Module 6 jamovi notes on how to do this):



The McNemar's test of paired proportions can be conducted at **Analyses > Frequencies > Contingency Tables > Paired Samples > McNemar test**:



Note that jamovi does not currently provide an estimate of the difference in paired proportions, or a confidence interval for the difference.

R notes

7.10 Pearson's chi-squared test for individual-level data

We will demonstrate how to use R to conduct a Pearson chi-squared test using Worked Example 7.1.

```
library(jmv)

nausea <- readRDS("data/examples/mod06_nausea.rds")

head(nausea)
```

```
    group side_effect
1 Placebo      Nausea
2 Placebo      Nausea
3 Placebo      Nausea
4 Placebo      Nausea
5 Placebo   No nausea
6 Placebo   No nausea
```

```
str(nausea$group)
```

```
 Factor w/ 2 levels "Placebo","Active": 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "label")= chr "Group"
```

```
str(nausea$side_effect)
```

```
 Factor w/ 2 levels "No nausea","Nausea": 2 2 2 2 1 1 1 1 1 1 ...
 - attr(*, "label")= chr "Side effect"
```

The columns `group` and `side_effect` have been entered as factors, with "Placebo" and "No nausea" as the first levels. We should use the `relevel()` command to re-order the factor levels.

```
nausea$group <- relevel(nausea$group, ref = "Active")
nausea$side_effect <- relevel(nausea$side_effect, ref = "Nausea")

str(nausea$group)
```

```
 Factor w/ 2 levels "Active","Placebo": 2 2 2 2 2 2 2 2 2 2 ...
```

```
str(nausea$side_effect)
```

```
 Factor w/ 2 levels "Nausea","No nausea": 1 1 1 1 2 2 2 2 2 2 ...
```

After confirming the factors are appropriately defined, we can construct our 2-by-2 table and view the expected frequencies.

```
contTables(
    data = nausea,
    rows = group, cols = side_effect,
    exp = TRUE
)
```

```
 CONTINGENCY TABLES

 Contingency Tables

   group                  Nausea       No nausea    Total

   Active     Observed           15           35          50
              Expected     9.500000     40.50000     50.00000

   Placebo    Observed            4           46          50
              Expected     9.500000     40.50000     50.00000

   Total      Observed           19           81         100
              Expected    19.000000     81.00000    100.00000


  ² Tests

         Value       df    p

    ²    7.862248     1    0.0050478
   N          100
```

After confirming that there are no cells with small expected frequencies, we can interpret the chi-square test. The last section reports the chi-squared test statistic which has a value of 7.86 with 1 degree of freedom and a P-value of 0.005.

If there are small values of expected frequencies, Fisher's exact test can be requested using `fisher = TRUE`:

```
contTables(
    data = nausea,
    rows = group, cols = side_effect,
    fisher = TRUE
)
```

```
 CONTINGENCY TABLES

 Contingency Tables

   group      Nausea    No nausea    Total

   Active         15           35       50
   Placebo         4           46       50
   Total          19           81      100
```

```
  ² Tests

                          Value       df     p

    ²                    7.862248     1     0.0050478
   Fisher's exact test                      0.0094886
   N                          100
```

7.11 Pearson's chi-squared test for summarised data

When you only have the summarised date (for example, the cross-tabulated data), you need to enter the summarised data manually. As we did in Module 6, the 2-by-2 table can be entered as four lines of data:

```
drug_aggregated <- data.frame(
    group = c("Active", "Active", "Placebo", "Placebo"),
    side_effect = c("Nausea", "No nausea", "Nausea", "No nausea"),
    n = c(15, 35, 4, 46)
)
```

The `contTables()` function is used in the usual way, specifying `count=n`.

7.12 Chi-squared test for tables larger than 2-by-2

Use the data in `mod07_allergy.rds`. We use similar steps as described above for a 2-by-2 table.

```
allergy <- readRDS("data/examples/mod07_allergy.rds")

head(allergy)
```

```
  id asthma hdmallergy catallergy infection    sex maternalasthma
1  1     No        Yes         No       Yes Female             No
2  2    Yes         No         No        No Female             No
3  3    Yes         No         No        No Female             No
4  4     No         No         No        No   Male             No
5  4    Yes        Yes        Yes        No Female             No
6  5    Yes        Yes        Yes        No Female             No
  allergy_severity
1 Moderate allergy
2     Non-allergic
3     Non-allergic
4     Non-allergic
5 Moderate allergy
6 Moderate allergy
```

```
contTables(
    data = allergy,
    rows = allergy_severity, cols = sex,
    pcCol = TRUE
)
```

```
 CONTINGENCY TABLES
```

```
 Contingency Tables

   allergy_severity                    Female       Male         Total

   Non-allergic        Observed             150        137           287
                       % within column   61.98347   53.10078   57.40000

   Slight allergy      Observed              50         70           120
                       % within column   20.66116   27.13178   24.00000

   Moderate allergy    Observed              27         32            59
                       % within column   11.15702   12.40310   11.80000

   Severe allergy      Observed              15         19            34
                       % within column    6.19835    7.36434    6.80000

   Total               Observed             242        258           500
                       % within column  100.00000  100.00000  100.00000


  ² Tests

        Value       df    p

    ²   4.308913     3    0.2299813
   N          500
```

7.13 McNemar's test for paired proportions

To perform this test in R, we will use the dataset `mod07_drug_response.rds`.

```
drug <- readRDS("data/examples/mod07_drug_response.rds")

head(drug)
```

```
  druga drugb
1   Yes   Yes
2   Yes   Yes
3   Yes   Yes
4   Yes   Yes
5   Yes   Yes
6   Yes   Yes
```

As usual, we should check that the variables being tabulated are factors, with the first level of the factor being the outcome of interest.

```
str(drug$druga)
```

```
 Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 - attr(*, "label")= chr "Response to Drug A"
```

```
str(drug$drugb)
```

```
 Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 - attr(*, "label")= chr "Response to Drug B"
```

Here we see that the first level of the factor is "No" - we need to use the `relevel()` function to re-order the levels so "Yes" is the first level:

```
drug$druga <- relevel(drug$druga, ref = "Yes")
drug$drugb <- relevel(drug$drugb, ref = "Yes")

str(drug$druga)
```

```
 Factor w/ 2 levels "Yes","No": 1 1 1 1 1 1 1 1 1 1 ...
```

```
str(drug$drugb)
```

```
 Factor w/ 2 levels "Yes","No": 1 1 1 1 1 1 1 1 1 1 ...
```

We can use the `contTablesPaired()` function within the `jmv` library to conduct McNemar's test of paired proportions:

```
contTablesPaired(data = drug, rows = druga, cols = drugb)
```

```
 PAIRED SAMPLES CONTINGENCY TABLES

 Contingency Tables

   druga    Yes    No    Total

   Yes       21    20       41
   No        14     5       19
   Total     35    25       60



 McNemar Test

          Value       df    p

    ²     1.058824     1    0.3034837
   N           60
```

Note that `contTablesPaired()` does not calculate an exact P-value.

To estimate the proportion in each of the paired samples, its difference, and the 95% confidence interval of the difference, we can use the `mcNemarDiff()` function which can be downloaded here.

```
### Copied from gist.githubusercontent.com
mcNemarDiff <- function(data, var1, var2, digits = 3) {
    if (!requireNamespace("epibasix", quietly = TRUE)) {
        stop("This function requires epibasix to be installed")
    }

    tab <- table(data[[var1]], data[[var2]])
```

```
    p1 <- (tab[1, 1] + tab[1, 2]) / sum(tab)
    p2 <- (tab[1, 1] + tab[2, 1]) / sum(tab)
    pd <- epibasix::mcNemar(tab)$rd
    pd.cil <- epibasix::mcNemar(tab)$rd.CIL
    pd.ciu <- epibasix::mcNemar(tab)$rd.CIU
    print(paste0(
        "Proportion 1: ",
        format(round(p1, digits = digits), nsmall = digits),
        "; Proportion 2: ", format(round(p2, digits = digits), nsmall = digits)
    ))
    print(paste0(
        "Difference in paired proportions: ",
        format(round(pd, digits = digits), nsmall = digits),
        "; 95% CI: ", format(round(pd.cil, digits = digits), nsmall = digits),
        " to ", format(round(pd.ciu, digits = digits), nsmall = digits)
    ))
}
### End copy

mcNemarDiff(data = drug, var1 = "druga", var2 = "drugb", digits = 2)


[1] "Proportion 1: 0.68; Proportion 2: 0.58"
[1] "Difference in paired proportions: 0.10; 95% CI: -0.11 to 0.31"
```

In this study of 60 participants, where each participant received both drugs, 41 (68%) responded to Drug A and 35 (58%) responded to Drug B. The difference in the proportions responding is estimated as 10% (95% CI -11% to 31%). There is no evidence that the response differed between the two drugs (McNemar's chi-squared = 1.06 with 1df, P=0.30).

Activities

Activity 7.1

Use the file `Activity_7.1.rds` to further investigate whether there is a gender difference in asthma in a random sample of 514 upper primary school children:

a) Use a contingency table (cross-tabulation) to determine the observed and expected frequencies. Which cell has the lowest expected cell count?
b) Use a chi-squared test to evaluate the hypothesis and interpret the result. Are the assumptions for a chi-squared test met? Calculate the 95% CI of the difference in proportions.

Activity 7.2

The file `Activity_7.2.rds` summarise 5-year mortality (the outcome) for 89 people who did or did not have a heart attack (the exposure).

a) State the null hypothesis.
b) Using jamovi or R, carry out the appropriate significance test to evaluate the hypothesis. Do the data fulfil the assumptions of the statistical test you have used?
c) Estimate the appropriate risk estimate for mortality. Are the confidence intervals of the risk estimates consistent with the P value?
d) Summarise your results and state your conclusion.

Activity 7.3

The effect of two penicillin allergens B and G was tested in a random sample of 500 people. All people were tested with both allergens. For each person, data were recorded for whether or not there was an allergic reaction to the allergen.

Use `Activity_7.3.rds` to test the null hypothesis that the proportion of participants who react to allergen G is the same as the proportion who react to allergen B. Are the 95% CI around the difference consistent with the P value?

Activity 7.4

A study was conducted to assess whether the rate of premature births differed by region. We examined a survey of 200 live births in an urban region in which 2 babies were born prematurely. We also surveyed 80 live births in a rural region and found that 5 babies were born prematurely.

Analyse these data to answer the research question. Write a brief report summarising your results and state your conclusion.

Supplementary Activity 7.5

The betel nut, the seed of the areca palm, is grown in the tropical Pacific and Asia and is a commonly use psycho-active substance. Betel nut is often chewed, wrapped inside betel leaves or in combination with tobacco. Chewing betel nut has been linked with a range of health issues.

A case-control study was conducted to assess the association between chewing betel nut and obstructive coronary artery disease. 293 men with obstructive coronary artery disease were

recruited, and 88 reported having chewed betel nut. Of the 720 healthy control men recruited, 57 reported having chewed betel nut.

Analyse these data to answer the research question. Write a brief report summarising your results and state your conclusion.

Module 8

Correlation and simple linear regression

Learning objectives

By the end of this module you will be able to:

- Explore the association between two continuous variables using a scatter plot;
- Estimate and interpret correlation coefficients;
- Estimate and interpret parameters from a simple linear regression;
- Assess the assumptions of simple linear regression;
- Test a hypothesis using regression coefficients.

Optional readings

Kirkwood and Sterne (2001); Chapter 10. [UNSW Library Link]

Bland (2015); Chapter 11. [UNSW Library Link]

8.1 Introduction

In Module 5, we saw how to test whether the means from two groups are equal - in other words, whether a continuous variable is related to a categorical variable. Sometimes we are interested in how closely two continuous variables are related. For example, we may want to know how closely blood cholesterol levels are related to dietary fat intake in adult men. To measure the strength of association between two continuously distributed variables, a correlation coefficient is used.

We may also want to predict a value of a continuous measurement from another continuous measurement. For example, we may want to know predict values of lung capacity from height in a community of adults. A regression model allows us to use one measurement to predict another measurement.

Although both correlation coefficients and regression models can be used to describe the degree of association between two continuous variables, the two methods provide different information. It is important to note that both methods summarise the strength of an association between variables, and do not imply a causal relationship.

8.2 Notation

In this module, we will be focussing on the association between two variables, denoted x and y.

There may be cases where it does not matter which variable is denoted x and which is denoted y, however this is rare. We are usually interested in whether one variable is associated with another. If we believe that a change in x will lead to a change in y, or that y is influenced by x, we define y as the *outcome variable* and x as the *explanatory variable*.

8.3 Correlation

We use correlation to measure the strength of a linear relationship between two variables. Before calculating a correlation coefficient, a scatter plot should first be obtained to give an understanding of the nature of the relationship between the two variables.

Worked Example

The file `mod08_lung_function.csv` has information about height and lung function collected from a random sample of 120 adults. Information was collected on height (cm) and lung function, which was measured as forced vital capacity (FVC), measured in litres. We can obtain a *scatter-plot* shown in Figure 8.1, where the outcome variable (y) is plotted on the vertical axis, and the explanatory variable (x) is plotted on the horizontal axis.

Figure 8.1 shows that as height increases, lung function also increases, which is as expected. One or two of the data points are separated from the rest of the data but are not so far away as to be considered outliers because they do not seem to stand out of other observations.

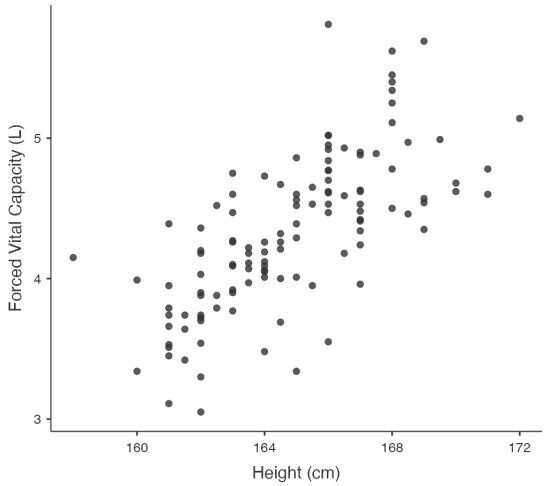


Figure 8.1: Association between height and lung function in 120 adults

Correlation coefficients

A correlation coefficient (r) describes how closely the variables are related, that is the strength of linear association between two continuous variables. The range of the coefficient is from +1 to −1 where +1 is a perfect positive association, 0 is no association and −1 is a perfect inverse association. In general, an absolute (disregarding the sign) r value below 0.3 indicates a weak association, 0.3 to < 0.6 is fair association, 0.6 to < 0.8 is a moderate association, and $\geq$ 0.8 indicates a strong association.

The correlation coefficient is positive when large values of one variable tend to occur with large values of the other, and small values of one variable (y) tend to occur with small values of the other (x) (Figure 8.2 (a and b)). For example, height and weight in healthy children or age and blood pressure.

The correlation coefficient is negative when large values of one variable tend to occur with small values of the other, and small values of one variable tend to occur with large values of the other (Figure 8.2 (c and d)). For example, percentage immunised against infectious diseases and under-five mortality rate.

It is possible to calculate a P-value associated with a correlation coefficient to test whether the correlation coefficient is different from zero. However, a correlation coefficient with a large P-value does not imply that there is no relationship between x and y, because the correlation coefficient only tests for a linear association and there may be a non-linear relationship such as a curved or irregular relationship.

The assumptions for using a Pearson's correlation coefficient are that:

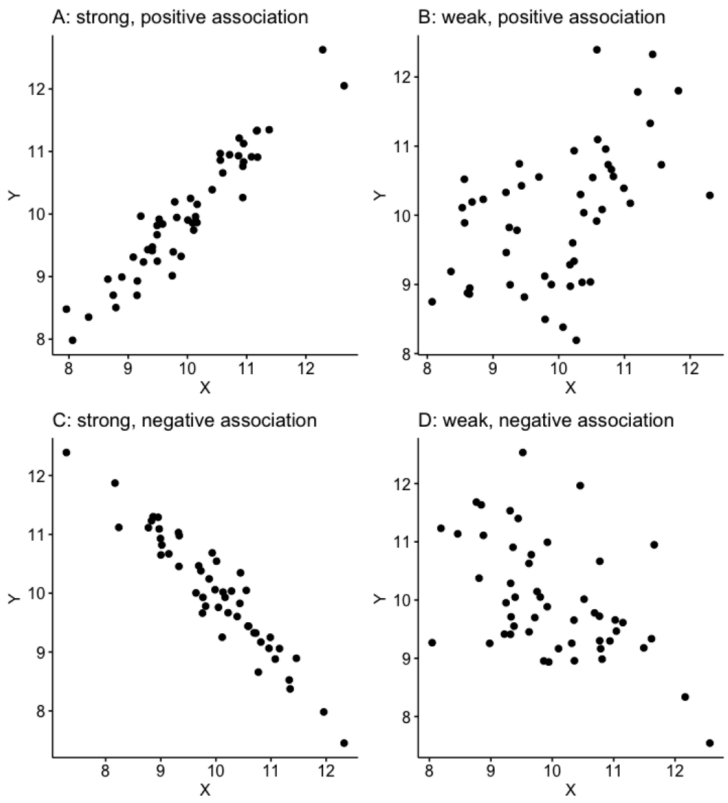


Figure 8.2: Scatter plots demonstrating strong and weak, positive and negative associations

- observations are independent;
- both variables are continuous variables;
- the relationship between the two variables is linear.

There is a further assumption that the data follow a bivariate normal distribution. This assumes: *y* follows a normal distribution for given values of *x*; and *x* follows a normal distribution for given values of *y*. This is quite a technical assumption that we do not discuss further.

There are two types of correlation coefficients– the correct one to use is determined by the nature of the variables as shown in Table 8.1.

Table 8.1: Correlation coefficients and their application

Correlation coefficient	Application
Pearson's correlation coefficient: *r*	Both variables are continuous and a bivariate normal distribution can be assumed
Spearman's rank correlation: ρ	Bivariate normality cannot be assumed. Also useful when at least one of the variables is ordinal

Spearman's ρ is calculated using the ranks of the data, rather than the actual values of the data. We will see further examples of such methods in Module 9, when we consider non-parametric tests, which are often based on ranks.

Correlation coefficients are often presented in the form of a *correlation matrix* which can display the correlation between a number of variables in a single table (Table 8.2).

Table 8.2: Correlation matrix for Height and FVC

	Height	FVC
Height	1	0.70 P < 0.0001
FVC	0.70 P < 0.0001	1

This correlation matrix shows that the Pearson's correlation coefficient between height and lung function is 0.70 with P<0.0001 indicating very strong evidence of a linear association between height and FVC. A correlation matrix sometimes includes correlations between the same variable, indicated as a correlation coefficient of 1. For example, $Height$ is perfectly correlated with itself (i.e. has a correlation coefficient of 1). Similarly, FVC is perfectly correlated with itself.

Correlation coefficients are rarely used as important statistics in their own right because they do not fully explain the relationship between the two variables and the range of the data has an important influence on the size of the coefficient. In addition, the statistical significance of the correlation coefficient is often over interpreted because a small correlation which is of no clinical importance can become statistically significant even with a relatively small sample size. For example, a poor correlation of 0.3 will be statistically significant if the sample size is large enough.

8.4 Linear regression

The nature of a relationship between two variables is more fully described using regression, where the relationship is described by a straight line.

Figure 8.3 shows our lung data with a fitted regression line.

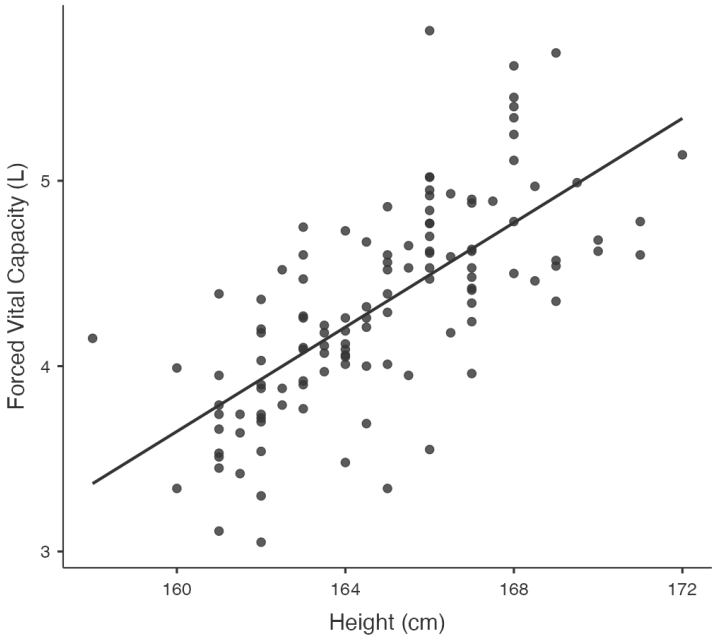


Figure 8.3: Association between height and lung function in 120 adults

The line through the plot is called the line of 'best fit' because the size of the deviations between the data points and the line is minimised in estimating the line.

Regression equations

The mathematical equation for the line explains the relationship between two variables: y, the outcome variable, and x, the explanatory variable. The equation of the regression line is as

follows:

$$y = \beta_0 + \beta_1 x$$

This line is shown in Figure 8.4 using the notation shown in Table 8.3.

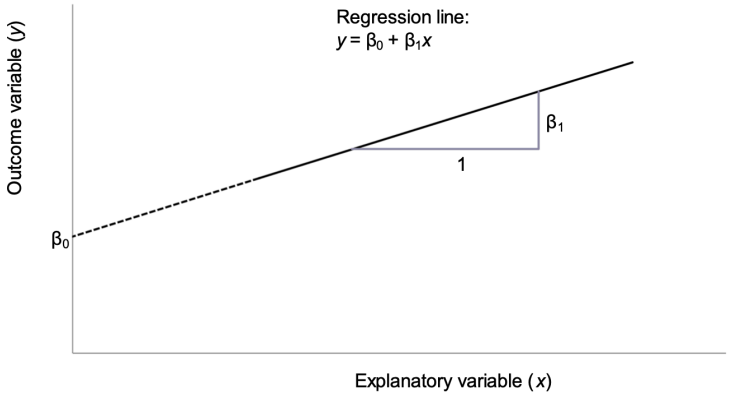


Figure 8.4: Coefficients of a linear regression equation

Table 8.3: Notation for linear regression equation

Symbol	Interpretation
y	The outcome variable
x	The explanatory variable
β_0	Intercept of the regression line
β_1	Slope of the regression line

The intercept is the point at which the regression line intersects with the y-axis when the value of x is zero. In most cases, the intercept does not have a biologically meaningful interpretation as the explanatory variable cannot take a value of zero. In our working example, the intercept is not meaningful as it is not possible for an adult to have a height of 0cm.

The slope of the line is the predicted change in the outcome variable y as the explanatory explanatory variable x increases by 1 unit.

An important concept is that regression predicts an expected value of y given an observed value of x: any error around the explanatory variable is not taken into account.

8.5 Regression coefficients: estimation

The regression parameters β_0 and β_1 are true, unknown quantities (similar to μ and σ), which are estimated using statistical software using the *method of least squares*. This method estimates the intercept and the slope, and also their variability (i.e. standard errors). Software is always used to estimate the regression parameters from a set of data.

Using the method of least squares:

- the intercept is estimated as b_0;
- the slope is estimated as b_1.

8.6 Regression coefficients: inference

We can use the estimated regression coefficients and their variability to calculate 95% confidence intervals. Here, a t-value from a t-distribution with $n - 2$ degrees of freedom is used:

- 95% confidence interval for intercept: $b_0 \pm t_{n-2} \times SE(b_0)$
- 95% confidence interval for slope: $b_1 \pm t_{n-2} \times SE(b_1)$

Note that as the constant (b_0) is not often biologically plausible, the 95% confidence interval for the constant is often not reported.

The significance of the estimated slope (and less commonly, intercept) can be tested using a t-test. The null hypotheses and the alternative hypothesis for testing the slope of a simple linear regression model are:

- H_0: $\beta_1 = 0$
- H_1: $\beta_1 \neq 0$

To test the null hypothesis for the regression coefficient β_1, the following t-test is used:

$$t = b_1 / SE(b_1)$$

This will give a t statistic which can be referred to a t distribution with n − 2 degrees of freedom to calculate the corresponding P-value.

Table 8.4 shows the estimated regression coefficients for our working example.

Table 8.4: Estimated regression coefficients

Term	Estimate	Standard error	t value	P value	95% Confidence interval
Intercept	-18.87	2.194	t=-8.60, 118df	<0.001	-23.22 to -14.53
Height	0.14	0.013	t=10.58, 118df	<0.001	0.11 to 0.17

From this output, we see that the slope is estimated as 0.14 with an estimated intercept of -18.87. Therefore, the regression equation is estimated as:

FVC (L) = − 18.87 + (0.14 $\times$ Height in cm)

There is very strong evidence of a linear association between FVC and height in cm (P < 0.001).

This equation can be used to predict FVC for a person of a given height. For example, the predicted FVC for a person 165 cm tall is estimated as:

FVC = − 18.87347 + (0.1407567 $\times$ 165.0) = 4.40 L.

Note that for the purpose of prediction we have kept all the decimal places in the coefficients to avoid rounding error in the intermediate calculation.

Fit of a linear regression model

After fitting a linear regression model, it is important to know how well the model fits the observed data. One way of assessing the model fit is to compute a statistic called coefficient of determination, denoted by R^2. It is the square of the Pearson correlation coefficient $r : r^2 = R^2$. Since the range of r is from −1 to 1, R^2 must lie between 0 and 1.

R^2 can be interpreted as the proportion of variability in y that can be explained by variability in x. Hence, the following conditions may arise:

If $R^2 = 1$, then all variation in y can be explained by variation of x and all data points fall on the regression line.

If $R^2 = 0$, then none of the variation in y is related to x at all, and the variable x explains none of the variability in y.

If $0 < R^2 < 1$, then the variability of y can be partially explained by the variability in x. The larger the R^2 value, the better is the fit of the regression model.

8.7 Assumptions for linear regression

Regression is robust to moderate degrees of non-normality in the variables, provided that the sample size is large enough and that there are no influential outliers. Also, the regression equation describes the relationship between the variables and this is not influenced as much by the spread of the data as the correlation coefficient is.

The assumptions that must be met when using linear regression are as follows:

- observations are independent;
- the relationship between the explanatory and the outcome variable is linear;
- the residuals are normally distributed.

A residual is defined as the difference between the observed and predicted outcome from the regression model. If the predicted value of the outcome variable is denoted by $\widehat{y}$ then:

$$\text{Residual} = \text{observed} - \text{predicted} = y - \widehat{y}$$

It is important for regression modelling that the data are collected in a period when the relationship remains constant. For example, in building a model to predict normal values for lung function the data must be collected when the participants have been resting and not exercising and people taking bronchodilator medications that influence lung capacity should be excluded. In regression, it is not so important that the variables themselves are normally distributed, but it is important that the residuals are. Scatter plots and specific diagnostic tests can be used to check the regression assumptions. Some of these will not be covered in this introductory course but will be discussed in detail in the **Regression Methods in Biostatistics** course.

The distribution of the residuals should always be checked. Large residuals can indicate unusual points or points that may exert undue influence on the estimated regression slope.

The distribution of the residuals from the model is shown in Figure 8.5. The residuals are approximately normally distributed, with no outlying values.

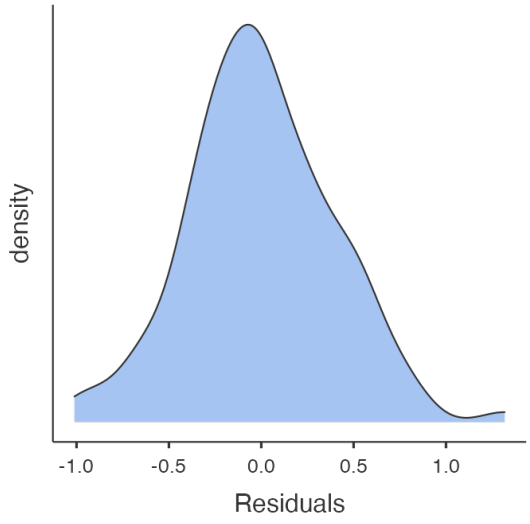


Figure 8.5: Distribution of regression residuals

8.8 Multiple linear regression

In the above example, we have only used a simple linear regression model of two continuous variables. Other more complex models can be built from this e.g. if we wanted to look at the effect of gender (male vs. female) as binary indicator in the model while adjusting for the effect of height. In that case we would include both the variables in the model as explanatory variables.

In the same way we can include any number of explanatory variables (both continuous and categorical) in the model: this is called a multivariable model. Multivariable models are often used for building predictive equations, for example by using age, height, gender and smoking history to predict lung function, or to adjust for confounding and detect effect modification to investigate the association between an exposure and an outcome factor.

Multiple regression has an important role in investigating causality in epidemiology. The exposure variable under investigation must stay in the model and the effects of other variables which can be confounders or effect-modifiers are tested. The biological, psychological or social meaning of the variables in the model and their interactions are of great importance for interpreting theories of causality.

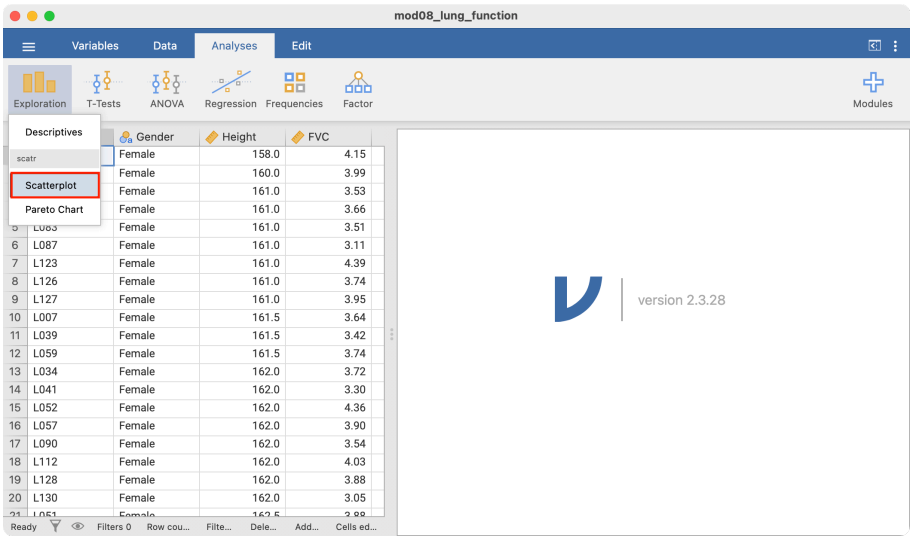
Other multivariable models include binary logistic regression for use with a binary outcome variable, or Cox regression for survival analyses. These models, together with multiple regression, will be taught in **PHCM9517: Regression Methods in Biostatistics**.

Jamovi notes

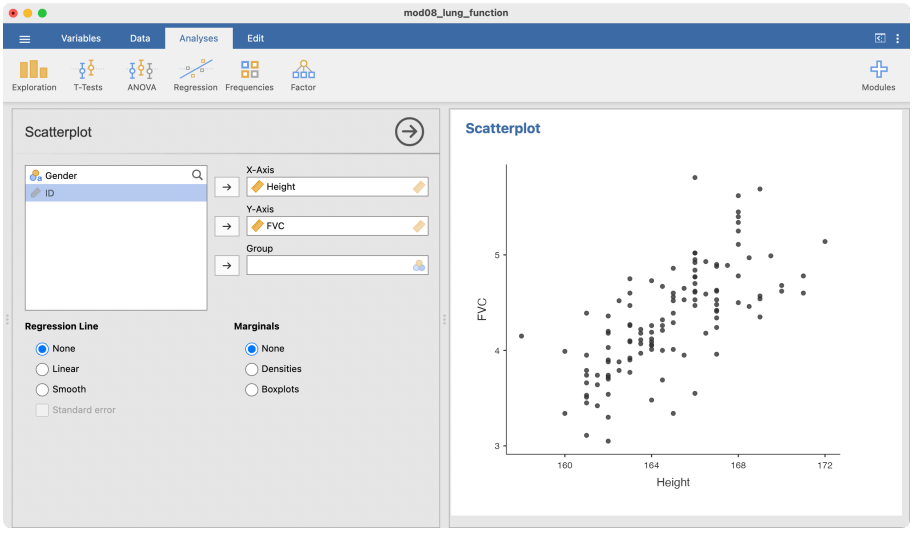
8.9 Creating a scatter plot

We will demonstrate using Jamovi for correlation and simple linear regression using the dataset `mod08_lung_function.csv`.

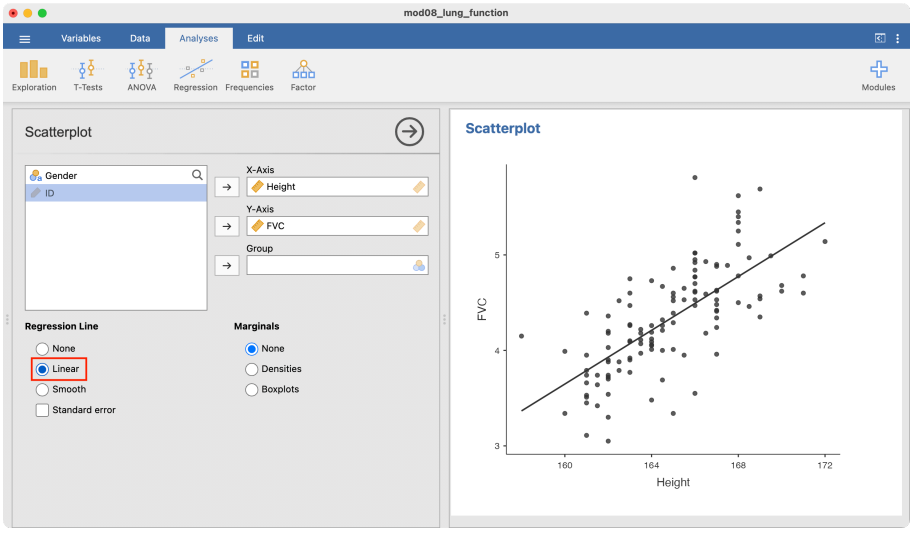
To create a scatter plot to explore the association between height and FVC, we use **Scatterplot** within the **Exploration** menu:



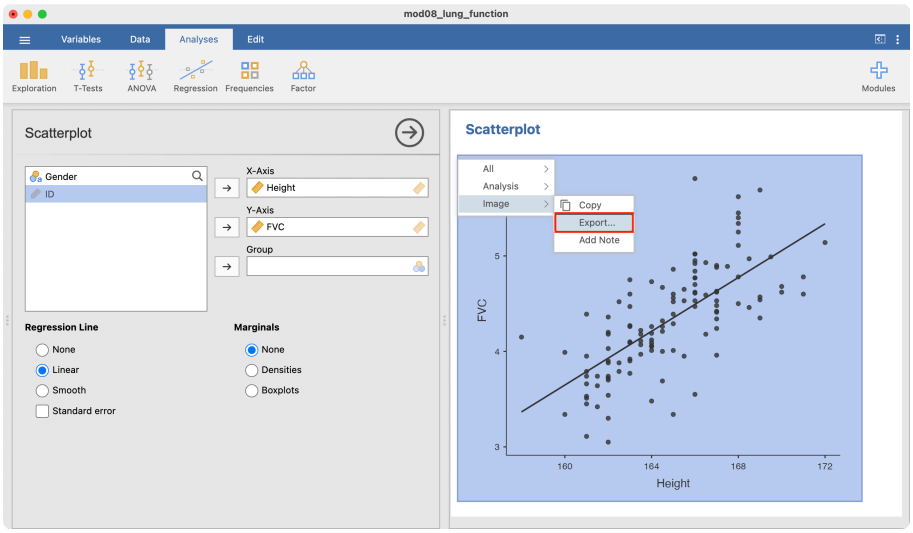
Choose **Height** for the **X-Axis**, and **FVC** for the **Y-Axis**, and the scatter-plot appears in the Output window:



To add a fitted line, click the **Linear** option in the **Regression Line** section:



To save your graph, right-click the graph and choose **Image > Export**, and be sure to save your file as a PNG file:



8.10 Calculating a correlation coefficient

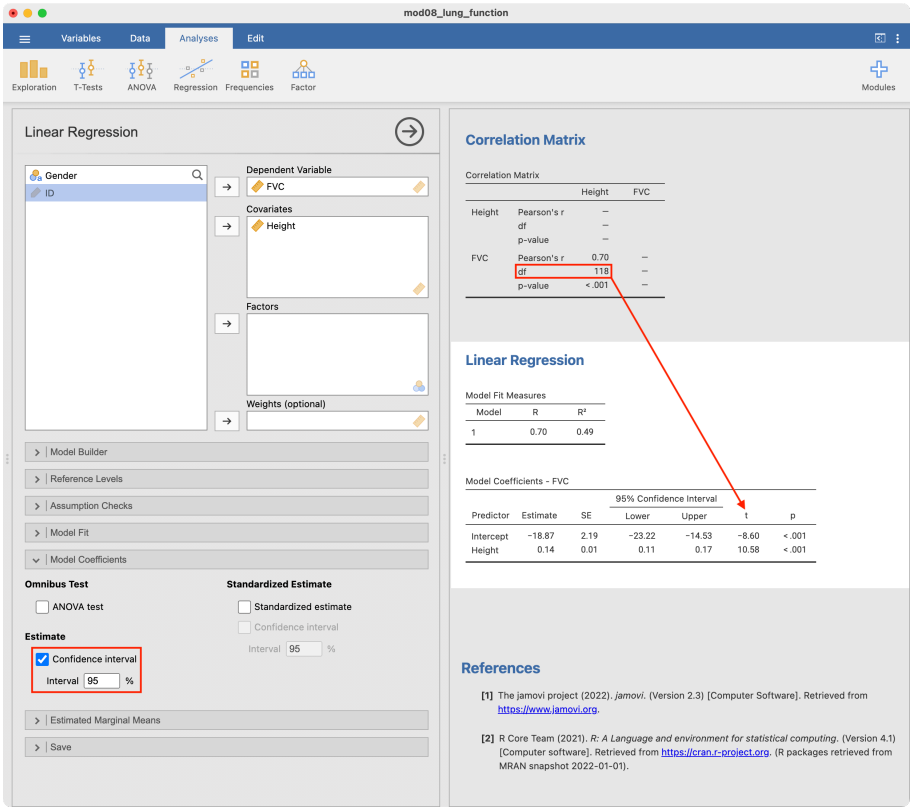
To calculate the Pearson's correlation using the dataset `mod08_lung_function.csv` choose: **Correlation Matrix** within the **Regression** menu.

Select the two variables, **FVC** and **Height** in the **Variables** box, and a correlation matrix will be constructed, similar to Table 8.2.

8.11 Fitting a simple linear regression model

To fit a simple linear regression model, choose **Regression > Linear Regression**

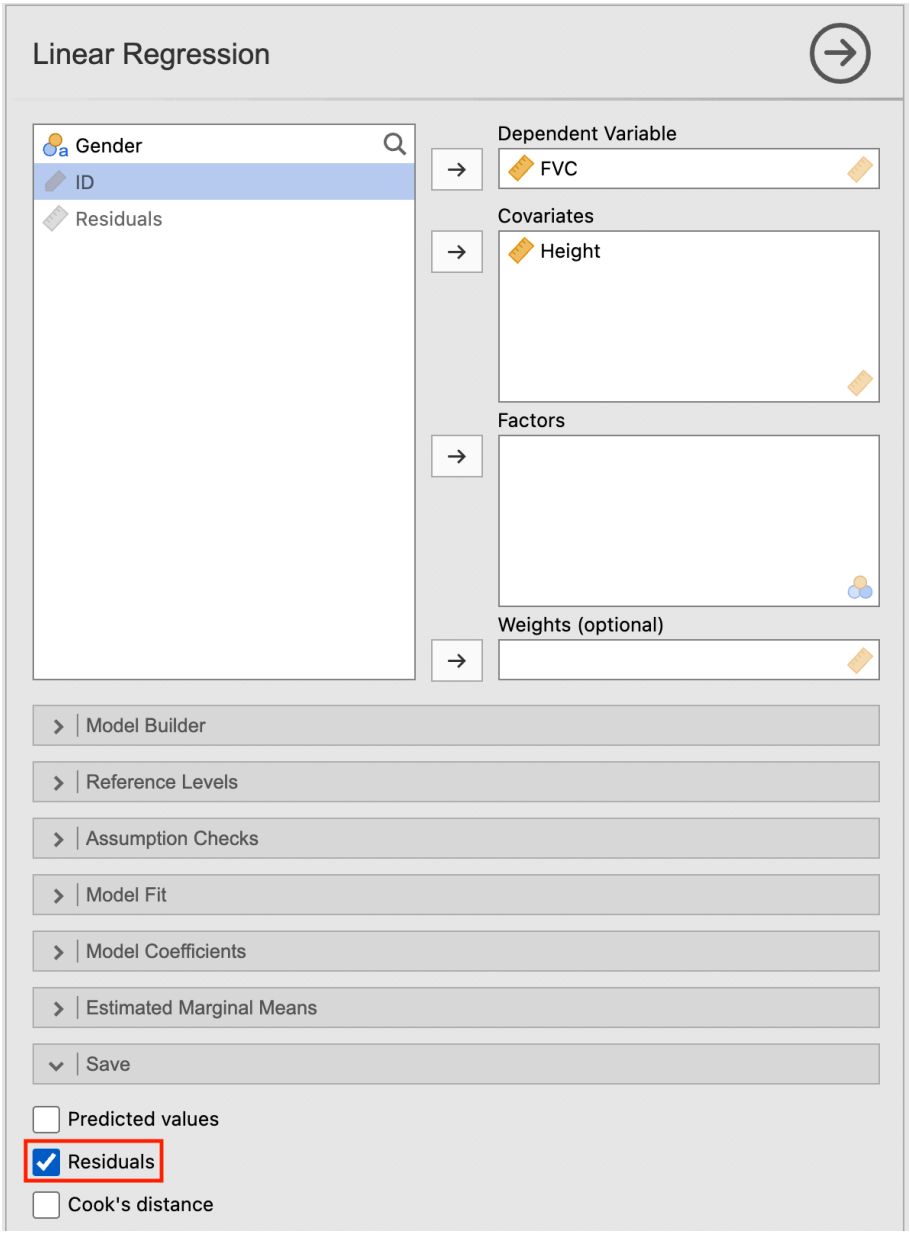
Select `FVC` as the **Dependent variable**, and `Height` as a **Covariate** (Jamovi refers to continuous explanatory variables as *covariates*). To obtain the 95% confidence interval for the regression coefficients, scroll down to the **Model Coefficients** section and click the **Confidence interval** option for Estimate.



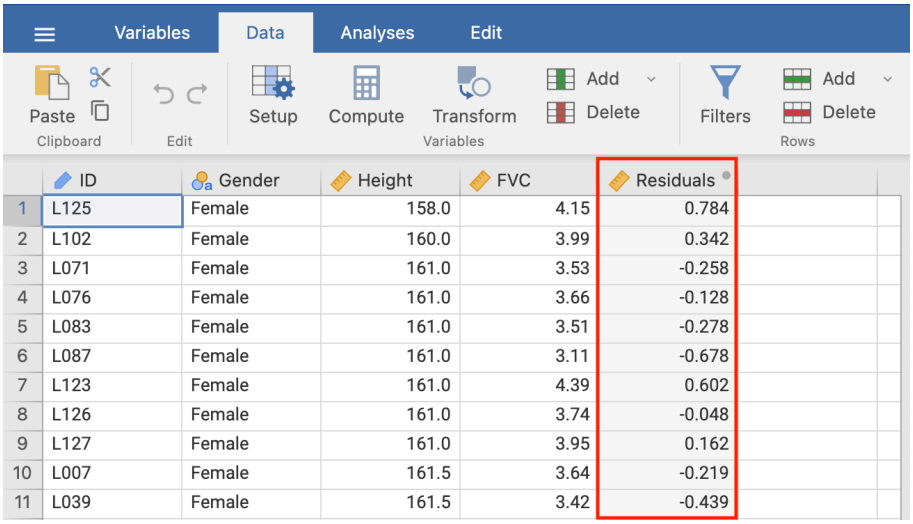
You will notice that the Jamovi output does not provide the degrees of freedom for the regression coefficient t-statistic. This is equivalent to the degrees of freedom in the preceding correlation matrix: in this case, 118 df.

8.12 Plotting residuals from a simple linear regression

To plot the residuals, we first need to save them using the **Save** option within the **Linear Regression** command box:



This creates a new column of *Residuals* within our dataset:



	ID	Gender	Height	FVC	Residuals
1	L125	Female	158.0	4.15	0.784
2	L102	Female	160.0	3.99	0.342
3	L071	Female	161.0	3.53	-0.258
4	L076	Female	161.0	3.66	-0.128
5	L083	Female	161.0	3.51	-0.278
6	L087	Female	161.0	3.11	-0.678
7	L123	Female	161.0	4.39	0.602
8	L126	Female	161.0	3.74	-0.048
9	L127	Female	161.0	3.95	0.162
10	L007	Female	161.5	3.64	-0.219
11	L039	Female	161.5	3.42	-0.439

You can now check the assumption that the residuals are normally distributed by creating a

density plot of the residuals using **Exploration > Descriptives**, as shown in Figure 8.5.

R notes

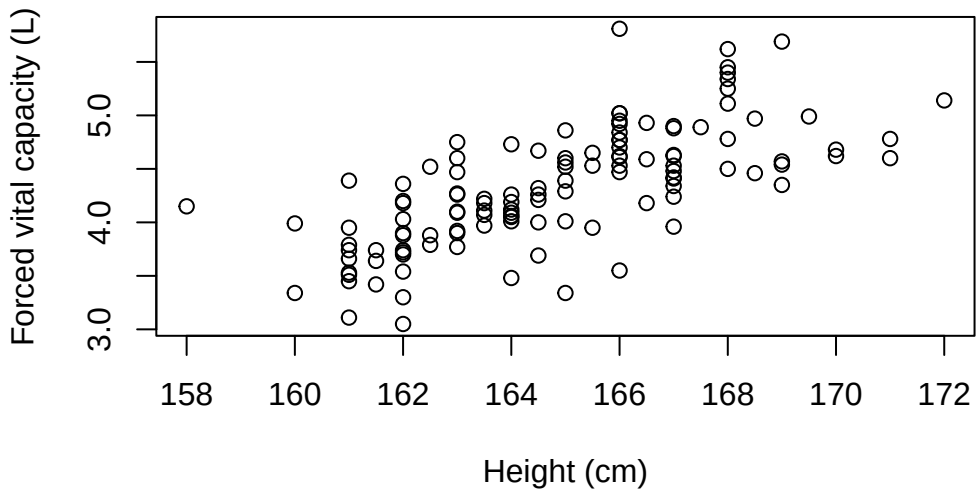
We will demonstrate using R for correlation and simple linear regression using the dataset `mod08_lung_function.csv`.

```
lung <- read.csv("data/examples/mod08_lung_function.csv")
```

8.13 Creating a scatter plot

We can use the `plot` function to create a scatter plot to explore the association between height and FVC, assigning meaningful labels with the `xlab` and `ylab` commands:

```
plot(x=lung$Height, y=lung$FVC,
     xlab="Height (cm)",
     ylab="Forced vital capacity (L)")
```

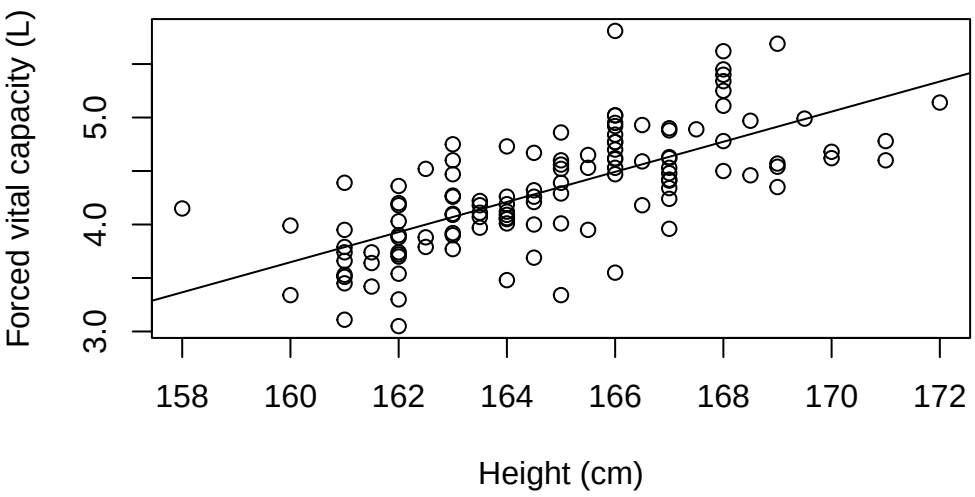


To add a fitted line, we can use the `abline()` function which adds a straight line to the plot. The equation of this straight line will be determined from the estimated regression line, which we specify with the `lm()` function, which fits a *linear model*.

The basic syntax of the `lm()` function is: `lm(y ~ x)` where `y` represents the *outcome* variable, and `x` represents the *explanatory* variable. Putting this all together:

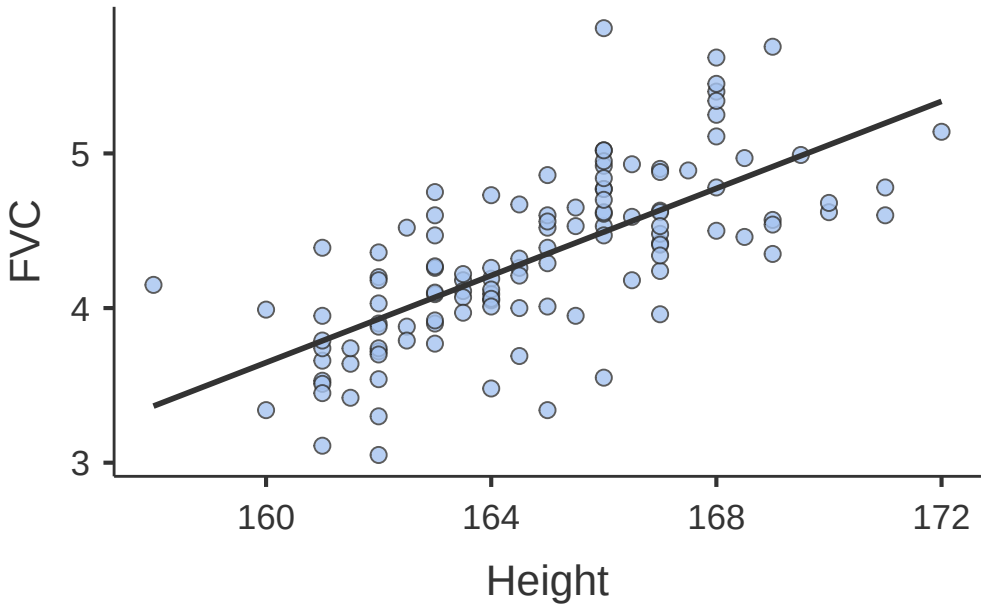
```
plot(x=lung$Height, y=lung$FVC,
     xlab="Height (cm)",
     ylab="Forced vital capacity (L)")
```

```
abline(lm(lung$FVC ~ lung$Height))
```



Note: to obtain output similar to Jamovi output, we can use the `scatr` package:

```
# Install scatr if required:
# install.packages("scatr")
library(scatr)

scat(data=lung, x="Height", y="FVC", line="linear")
```



Calculating a correlation coefficient

We can use the `corrMatrix` function in the Jamovi package to calculate a Pearson's correlation coefficient:

```
corrMatrix(data=lung, vars=c(Height, FVC))
```

```
 CORRELATION MATRIX

 Correlation Matrix

                             Height        FVC

   Height    Pearson's r             -
             df                      -
             p-value                 -

   FVC       Pearson's r     0.6976279           -
             df                    118           -
             p-value         < .0000001          -
```

8.14 Fitting a simple linear regression model

We can use the `lm` function to fit a simple linear regression model, specifying the model as `y ~ x` where `y` represents the *outcome* variable, and `x` represents the *explanatory* variable. Using `mod08_lung_function.csv`, we can quantify the relationship between FVC and height:

```
lm(FVC ~ Height, data=lung)
```

```
Call:
lm(formula = FVC ~ Height, data = lung)

Coefficients:
(Intercept)       Height
   -18.8735       0.1408
```

The default output from the `lm` function is rather sparse. We can obtain much more useful information by defining the linear regression model as an object, then using the `summary()` function:

```
model <- lm(FVC ~ Height, data=lung)
summary(model)
```

```
Call:
lm(formula = FVC ~ Height, data = lung)

Residuals:
     Min       1Q   Median       3Q      Max
-1.01139 -0.23643 -0.02082  0.24918  1.31786

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.87347    2.19365  -8.604 3.89e-14 ***
Height        0.14076    0.01331  10.577  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.3965 on 118 degrees of freedom
Multiple R-squared:  0.4867,    Adjusted R-squared:  0.4823
F-statistic: 111.9 on 1 and 118 DF,  p-value: < 2.2e-16
```

Finally, we can obtain 95% confidence intervals for the regression coefficients using the `confint` function:

```
confint(model)
```

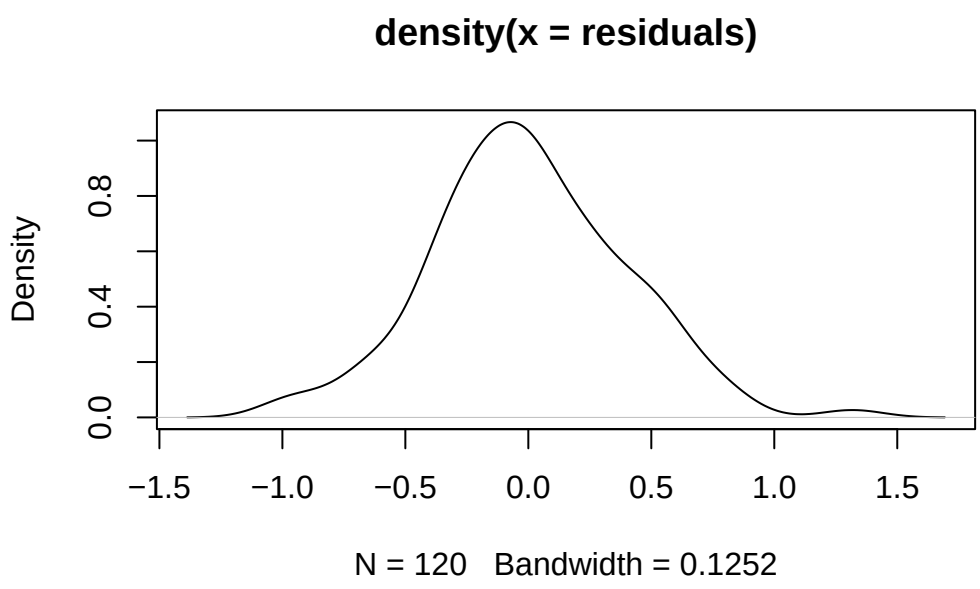
```
                 2.5 %      97.5 %
(Intercept) -23.2174967 -14.5294441
Height        0.1144042   0.1671092
```

Note that the output for R looks slightly different from the Jamovi output, the numerical values are identical.

8.15 Plotting residuals from a simple linear regression

We can use the `resid` function to obtain the residuals from a saved model. These residuals can then be plotted using a density plot in the usual way:

```
residuals <- resid(model)
plot(density(residuals))
```



Activities

Activity 8.1

To investigate how body weight (kg) effects blood plasma volume (mL), data were collected from 30 participants and a simple linear regression analysis was conducted. The slope of the regression was 68 (95% confidence interval 52 to 84) and the intercept was −1570 (95% confidence interval −2655 to −492).

[*You do not need software for this Activity*]

a) What is the outcome variable and explanatory (exposure) variable?
b) Interpret the regression slope and its 95% CI
c) Write the regression equation
d) If we randomly sampled a person from the population and found that their weight is 80kg, what would be the predicted value of plasma volume for this person?

Activity 8.2

To examine whether age predicts IQ, data were collected on 104 people. Use the data in the file `Activity_8.2.rds` to answer the following questions.

a) What is the outcome variable and the explanatory variable?
b) Create a scatter plot with the two variables. What can you infer from the scatter plot?
c) Obtain the correlation coefficient between age and IQ and interpret it.
d) Conduct a simple linear regression and report the relationship between the two variables including the interpretation of the R^2 value. Are the assumptions for linear regression met in this model?
e) What could you infer about the association between age and IQ in the population, based on the results of the regression analysis in this sample?

Activity 8.3

Which of the following correlation coefficients indicates the weakest linear relationship and why?

a) r = 0.72
b) r = 0.41
c) r = 0.13
d) r = −0.33
e) r = −0.84

Activity 8.4

Are the following statements true or false?

a) If a correlation coefficient is closer to 1.00 than to 0.00, this indicates that the outcome is caused by the exposure.
b) If a researcher has data on two variables, there will be a higher correlation if the two means are close together and a lower correlation if the two means are far apart.

Supplementary Activity 8.5

A cross-sectional study was conducted to examine the association between gestational age and systolic blood pressure in low birthweight babies. Data were collected at birth from 60 low birthweight babies and saved in the file Activity_8.5_babies.csv. The dataset contains the following variables:

- ID: participant ID number
- sbp: systolic blood pressure (mmHg)
- gestage: gestational age (weeks)

In this analysis, the researchers are interested in assessing the association between systolic blood pressure (the outcome variable) and gestational age (the explanatory variable).

NOTE: there is no need to apply clinical knowledge to the values of systolic blood pressure.

a) Produce a scatter plot summarising the relationship between systolic blood pressure and gestational age, and describe this relationship.
b) Estimate the correlation coefficient between systolic blood pressure and gestational age, and interpret it.
c) Fit a simple linear regression model predicting systolic blood pressure from gestational age, and present the estimated regression equation.
d) Are the assumptions of simple linear regression satisfied for your regression model? Provide evidence to support your claim.
e) Summarise your regression model in a brief conclusion.

Module 9

Analysing non-normal data

Learning objectives

By the end of this module you will be able to:

- Transform non-normally distributed variables;
- Explain the purpose of non-parametric statistics and key principles for their use;
- Calculate ranks for variables;
- Conduct and interpret a non-parametric independent samples significance test;
- Conduct and interpret a non-parametric paired samples significance test;
- Calculate and interpret the Spearman rank correlation coefficient.

Optional readings

Kirkwood and Sterne (2001); Chapter 13. [UNSW Library Link]

Bland (2015); Chapter 12. [UNSW Library Link]

9.1 Introduction

In general, parametric statistics are preferred for reporting data because the summary statistics (mean, standard deviation, standard error of the mean etc) and the tests used (t-tests, correlation, regression etc) are familiar and the results are easy to communicate. However, non-parametric tests can be used if data are not normally distributed. Non-parametric tests make fewer assumptions about the distribution of the data.

9.2 Transforming non-normally distributed variables

When a variable has a skewed distribution, one possibility is to transform the data to a new variable to try and obtain a normal or near normal distribution. Methods to transform non-normally distributed data include logarithmic transformation of each data point, or using the square root or the square or the inverse (i.e. 1/x) etc.

Worked Example

We have data from 132 patients who had a hospital stay following admission to ICU available on Moodle (`mod09_infection.rds`). The distribution of the length of stay for these patients is shown in the density plot in Figure 9.1. As is common with variables that record time, the data are skewed with many patients having relatively short stays and a few patients having very long hospital stays. Clearly, it would be inappropriate to use parametric statistical methods for these data.

When data are positively skewed, as shown in Figure 9.1, a logarithmic transformation can often make the data closer to being normally distributed. This is the most common transformation

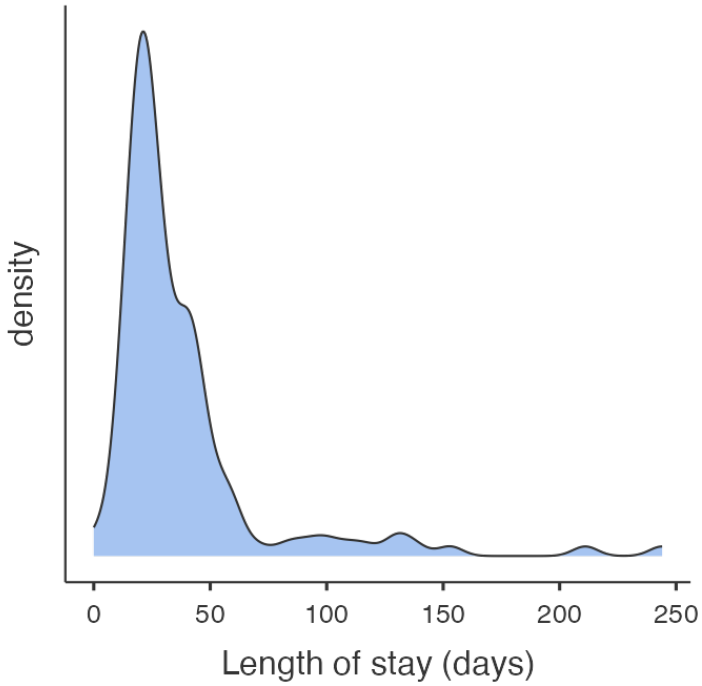


Figure 9.1: Length of hospital stay for 132 patients

used. You should note, however, that the logarithmic function cannot handle 0 or negative values. One way to deal with zeros in a set of data is to add 1 to each value before taking the logarithm.

We would generate a new variable, as shown in the jamovi or R notes. As the minimum length of stay in these sample data was 0, we have added 1 to each length of stay before taking the logarithm. The distribution of the logarithm of (length of stay + 1) is shown in Figure 9.2.

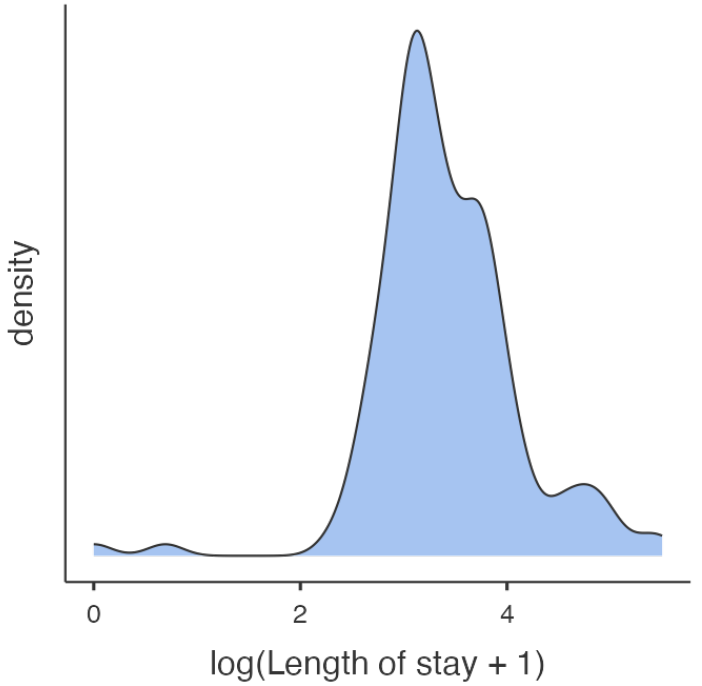


Figure 9.2: Distribution of log transformed (length of stay + 1)

The distribution now appears much more bell shaped. Table 9.1 shows the descriptive statistics

for length of stay before and after logarithmic transformation. Before transformation, the SD is almost as large as the mean value which indicates that the data are skewed and that these statistics are not an accurate description of the centre and spread of the data.

Table 9.1: Summary statistics for untransformed and transformed length of stay

	Length of stay	log(Length of stay + 1)
Mean (Standard deviation)	38.1 (35.78)	3.41 (0.715)
Mean: 95% confidence interval	31.9 to 44.2	3.29 to 3.53
Median [Interquartile range]	27 [21 to 42]	3.3 [3.1 to 3.8]
Range	0 to 244	0 to 5.5

The mean and standard deviation of the transformed length of stay are in log base e (i.e. ln) units. If we raise the mean of the log of length of stay to the power of e, it returns a value of 30.2 days ($e^{3.41} = 30.2$).

Technically, this is called the geometric mean of the data, and it has a different interpretation to the usual mean, the arithmetic mean. This is a much better estimate in this case of the "average" length of stay than the mean of 38.1 days (95% CI 31.9, 44.2 days) obtained from the non-transformed positively skewed data. Note that, if you have added 1 to your data to deal with 0 values, the back-transformed estimate is *approximately* equal to the geometric mean.

This set of data also includes a variable summarising whether a patient acquired a nosocomial infection (also known as healthcare-associated infections), which are infections that develop while undergoing medical treatment but were absent at the time of admission.

If we were testing the hypothesis that there was a difference in length of stay between groups (status of nosocomial infection), t-tests should not be used with length of stay, but could be used for the log transformed variable, which is approximately normally distributed. The output from the t-test of the log-transformed length of stay is shown in Table 9.2. This is done using the t-test shown in Module 5.

Table 9.2: Summary statistics for transformed length of stay

Nosocomial infection	n	Mean (SE)	95% Confidence interval
No	106	3.33 (0.068)	3.19 to 3.46
Yes	26	3.73 (0.136)	3.45 to 4.01
Difference (Yes - No)		0.39 (0.153)	0.09 to 0.70

Here, a two-sample t-test gives a test statistic of 2.59 with 130 degrees of freedom, and a P-value of 0.01.

As explained above, the estimated statistics would need to be converted back to the units in which the variable was measured. From Table 9.2, we can take the exponential of the corresponding log-transformed values:

- the geometric mean of the infected group is approximately 41.5 days with a 95% confidence interval from 31.4 to 55.0 days.
- the geometric mean of the uninfected group is approximately 27.9 days with a 95% confidence interval from 24.4 to 31.9 days.

9.3 Non-parametric significance tests

It is often not possible or sensible to transform a non-normal distribution, for example if there are too many zero values or when we simply want to compare groups using the unit in which the measurement was taken (e.g. length of stay). For this, non-parametric significance tests can be used but the general idea behind these tests is that the data values are replaced by ranks. This also protects against outliers having too much influence.

Ranking variables

Table 9.3 shows how ranks are calculated for the first 21 patients in the length-of-stay data. First the data are sorted in order of their magnitude (from the lowest value to the highest) ignoring the group variable. Each data point is then assigned a rank. Data points that are equal are assigned the mean of their ranks. Thus, the two lengths of stay of 11 days share the ranks 4 and 5, and have a mean rank of 4.5. Similarly, there are 5 people with a length of stay of 14 days and these share the ranks 9 to 13, the mean of which is 11. Once ranks are computed they are assigned to each of the two groups and summed within each group.

Table 9.3: Transforming data to ranks (first 21 participants)

ID	Infection	Length of stay	Rank	Infection=nd	Infection=Yes
32	No	0	1.0	1.0	
33	No	1	2.0	2.0	
12	No	9	3.0	3.0	
22	No	11	4.5	4.5	
16	No	11	4.5	4.5	
28	Yes	12	6.0		6.0
27	No	13	7.5	7.5	
20	No	13	7.5	7.5	
24	No	14	11.0	11.0	
11	No	14	11.0	11.0	
130	No	14	11.0	11.0	
10	No	14	11.0	11.0	
25	No	14	11.0	11.0	
19	No	15	15.5	15.5	
30	No	15	15.5	15.5	
23	No	15	15.5	15.5	
14	No	15	15.5	15.5	
15	No	17	20.5	20.5	
13	No	17	20.5	20.5	
21	Yes	17	20.5		20.5
17	No	17	20.5	20.5	

By assigning ranks to individuals, we lose information about their actual values and this makes it more difficult to detect a difference. However, outliers and extreme values in the data are brought back closer to the data so that they are less influential. For this reason, non-parametric tests have less power than parametric tests and they require much larger differences in the data to show statistical significance between groups.

9.4 Non-parametric test for two independent samples (Wilcoxon ranked sum test)

The non-parametric equivalent to an independent samples t-test (Module 5) is the Wilcoxon ranked sum test, also known as the Mann-Whitney U test. This can be obtained using the `Mann-Whitney U` option in jamovi, and the `wilcox.test` in R.

The assumption for this test is that the distributions of the two populations have the same general shape. If this assumption is met, then this test evaluates the null hypothesis that the medians of the two populations are equal. This test does not assume that the populations are normally distributed, nor that their variances are equal.

Conducting the Wilcoxon ranked sum test for our length of stay data yields a P-value of 0.014, providing evidence of a difference in the median length of stay between the groups.

This P-value should be provided alongside non-parametric summary statistics such as medians and inter-quartile ranges. In our example, we can obtain the median length of stay values of 24 (Interquartile Range: 19 to 40 days) in the group with no infection and 37 (Interquartile Range: 24 to 50 days) in the group with infection.

9.5 Non-parametric test for paired data (Wilcoxon signed-rank test)

There are two types of non-parametric tests for paired data, called the Sign test and the Wilcoxon signed rank test. In practice, the Sign test is rarely used and will not be discussed in this course.

If the differences between two paired measurements are not normally distributed, a non-parametric equivalent of a paired t-test (Module 5) should be used. The equivalent test is the Wilcoxon matched-pairs signed rank test, also simply called the Wilcoxon matched-pairs test. This test is resistant to outliers in the data, however the proportion of outliers in the sample should be small. This test evaluates the null hypothesis that the median of the paired differences is equal to zero.

In this test, the absolute differences between the paired scores are ranked and the difference scores that are equal to zero (i.e. scores where there is no difference between the pairs) are excluded. Note that the power of the test (the ability to detect an effect if there truly is an effect) reduces in the presence of zero differences, as the effective sample size (the number of non-zero differences) is reduced.

Worked Example

A crossover trial is done to compare symptom scores for two drugs in 11 people with arthritis (higher scores indicate more severe symptoms). The data are contained in datafile file `mod09_arthritis.csv`. The data are shown in Table 9.4.

Table 9.4: Arthritis symptom scores for 11 patients after administering two drugs

Patient ID	Score: Drug 1	Score: Drug 2	Difference (Drug 2 - Drug 1)
1	3	4	1
2	2	7	5
3	3	4	1
4	8	10	2
5	6	8	2
6	6	1	-5
7	2	6	4
8	3	7	4
9	5	8	3
10	9	10	1
11	7	8	1

The data shows that there is 1 person who has a negative difference, where the symptom score on drug 2 that is smaller than that for drug 1 (i.e., drug 2 is better than drug 1); and 10 people who have a positive difference. No one has the same score for both drugs.

Before doing the analysis let us examine the distribution of the difference of symptom scores between the two drugs. As in Module 5, we first need to compute the difference between the symptom scores. To examine the distribution, we construct a distribution plot as shown in Figure 9.3.

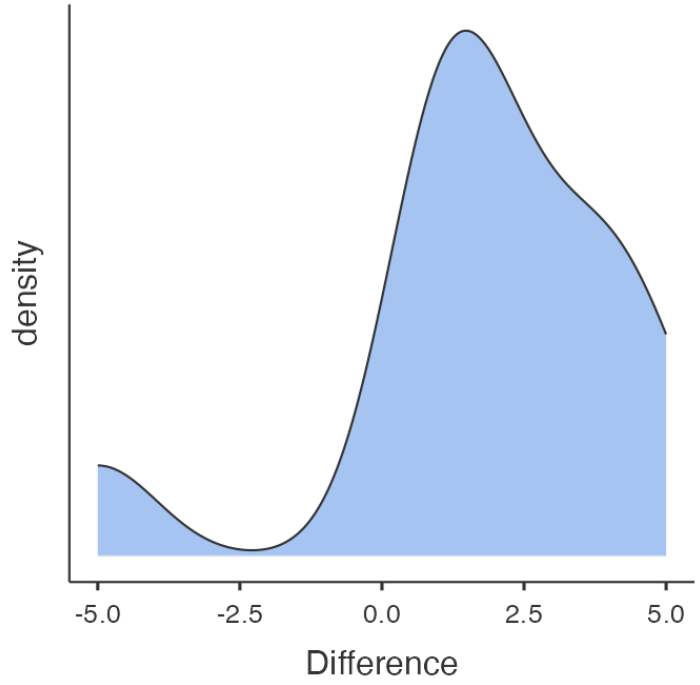


Figure 9.3: Distribution of difference in symptom scores between Drug 1 and Drug 2

The plot shows that the differences are not normally distributed. The data looks weirdly negatively skewed with a gap in values around -2.5. Therefore, it would not be appropriate to conduct a paired t-test. Hence, we conduct a non-parametric paired test (Wilcoxon matched-pairs signed-rank test).

The P-value obtained from this test is 0.049. Thus, there is evidence of a difference in symptom scores between the two drugs.

9.6 Non-parametric estimates of correlation

Estimating correlation using Pearson's correlation coefficient can be problematic when bivariate Normality cannot be assumed, or in the presence of outliers or skewness. There are two commonly used non-parametric alternatives to Pearson's correlation coefficient: Spearman's rank correlation (ρ or rho), and Kendall's rank correlation (τ or tau).

When estimating the correlation between x and y, Spearman's rank correlation essentially replaces the observations x and y by their ranks, and calculates the correlation between the ranks. Kendall's rank correlation compares the ranks between every possible combination of pairs of data to measure concordance: whether high values for x tend to be associated with high values for y (positively correlated) or low values of y (negatively correlated).

In terms of which is the more appropriate measure to use, the following passage from An Introduction to Medical Statistics (Bland (2015)) provides some guidance:

> "Why have two different rank correlation coefficients? Spearman's ρ is older than Kendall's τ, and can be thought of as a simple analogue of the product moment

> correlation coefficient, Pearson's r. Kendall's τ is a part of a more general and consistent system of ranking methods, and has a direct interpretation, as the difference between the proportions of concordant and discordant pairs. In general, the numerical value of ρ is greater than that of τ. It is not possible to calculate τ from ρ or ρ from τ, they measure different sorts of correlation. ρ gives more weight to reversals of order when data are far apart in rank than when there is a reversal close together in rank, τ does not. However, in terms of tests of significance, both have the same power to reject a false null hypothesis, so for this purpose it does not matter which is used."

We will illustrate estimating rank correlation using the data `mod08_lung_function.csv`, which has information about height and lung function collected from a sample of 120 adults.

The Spearman rank correlation coefficient is estimated as 0.75, demonstrating a positive association between height and FVC. The Kendall rank correlation coefficient is estimated as 0.56, again demonstrating a positive association between height and FVC.

9.7 Summary

In this module, we have presented methods to conduct a hypothesis test with data that are not normally distributed. Non-parametric methods do not assume any distribution for the data and use significance tests based on ranks or sign (or both). A non-parametric test is always less powerful than its equivalent parametric test if the data are normally distributed and so whenever possible parametric significance tests should be used. In some cases when data are not normally distributed with a reasonably large sample size, the data can be transformed (most commonly by log transformation) to make the distribution normal. A parametric significance test should then be used with the transformed data to test the hypothesis.

jamovi notes

9.8 Transforming non-normally distributed variables

One option for dealing with a non-normally distributed varaible is to transform it into its square, square root or logarithmic value. The new transformed variable may be normally distributed and therefore a parametric test can be used. First we check the distribution of the variable for normality, e.g. by plotting a density plot.

You can calculate a new, transformed, variable using **Data > Compute**. For example, to create a new column of data based on the log of length of stay:

- click into an empty column at the end of the spreadsheet
- click **Data > Compute**
- provide a name for the new variable, here `log(length of stay + 1)`
- enter the formula for the new variable in the f_x box, here `LN(los + 1)`
- hit enter to create the new column:

You can now check whether this logged variable is normally distributed as described in Module 2, for example by plotting a density plot as shown in Figure 9.2.

To obtain the back-transformed mean, we can use a calculator to calculate the exponential mean:

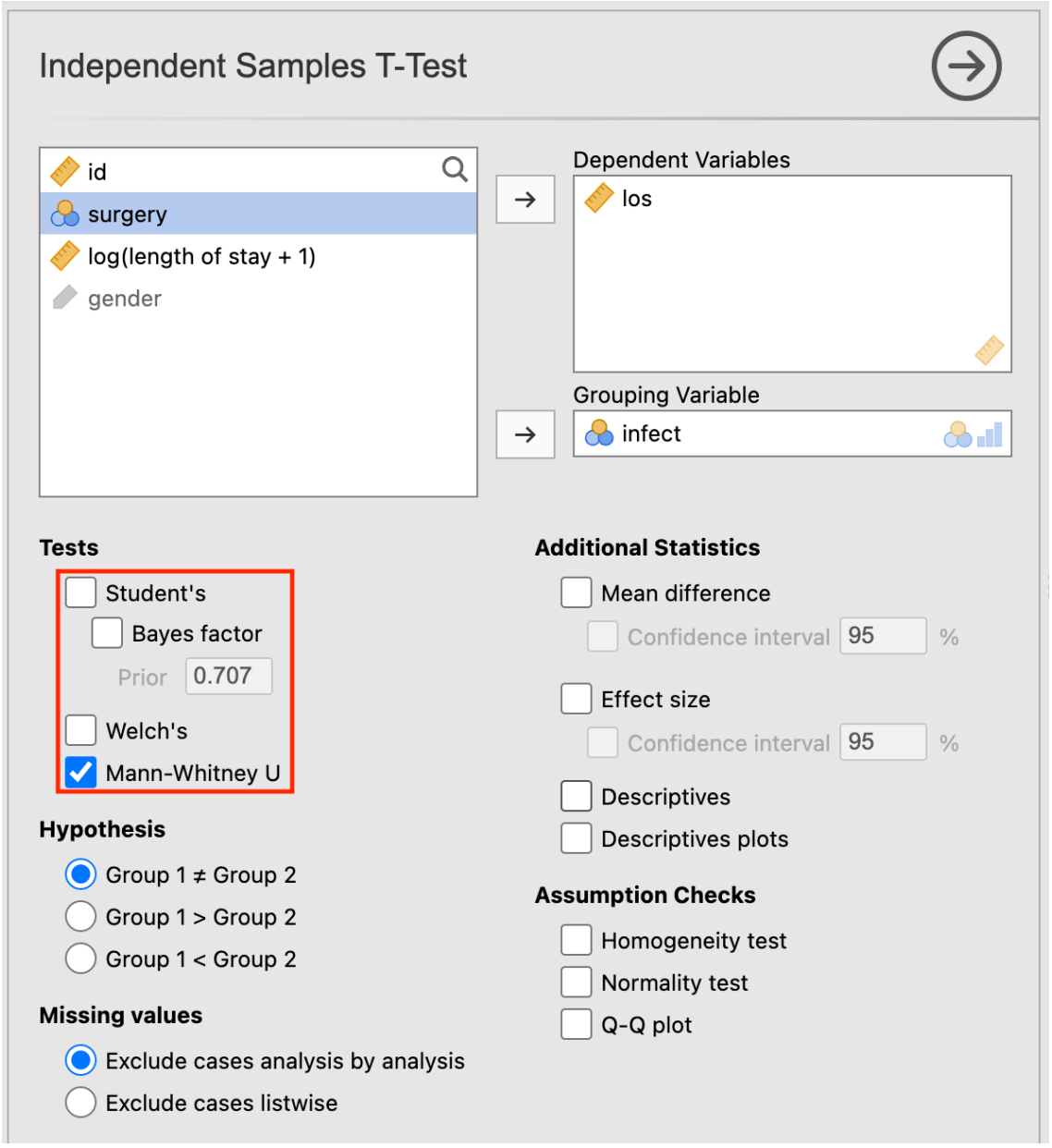
$$e^{3.407232} = 30.18$$

If your transformed variable is approximately normally distributed, you can apply parametric tests such as the t-test. In the Worked Example 9.1 dataset, the variable `infect` (presence of nosocomial infection) is a binary categorical variable. To test the hypothesis that patients with nosocomial infection have a different length of stay to patients without infection, you can conduct a t-test on the `log(length of stay + 1)` variable. You will need to back transform your mean values, as shown in Worked Example 9.1 in the course notes when reporting your results.

9.9 Wilcoxon ranked-sum test

The Wilcoxon ranked-sum test will be demonstrated using the length of stay data in `mod09_infection.rds`. Here, our continuous variable is `los` and the grouping variable is `infect`. Note that jamovi uses calls the Wilcoxon ranked-sum test the **Mann-Whitney U** test. The two tests are equivalent.

The Wilcoxon ranked-sum test is conducted in **Analyses > T-Tests > Independent Samples T-Test** in jamovi. The screen is set up in the same way as for a two-sample t-test. To conduct the Wilcoxon ranked-sum test, untick the **Student's** box and click the **Mann-Whitney U** box in the **Tests** section:



Note that the result appears without any descriptive analyses. You should obtain the appropriate descriptive statistics in the usual way.

Independent Samples T-Test

Independent Samples T-Test

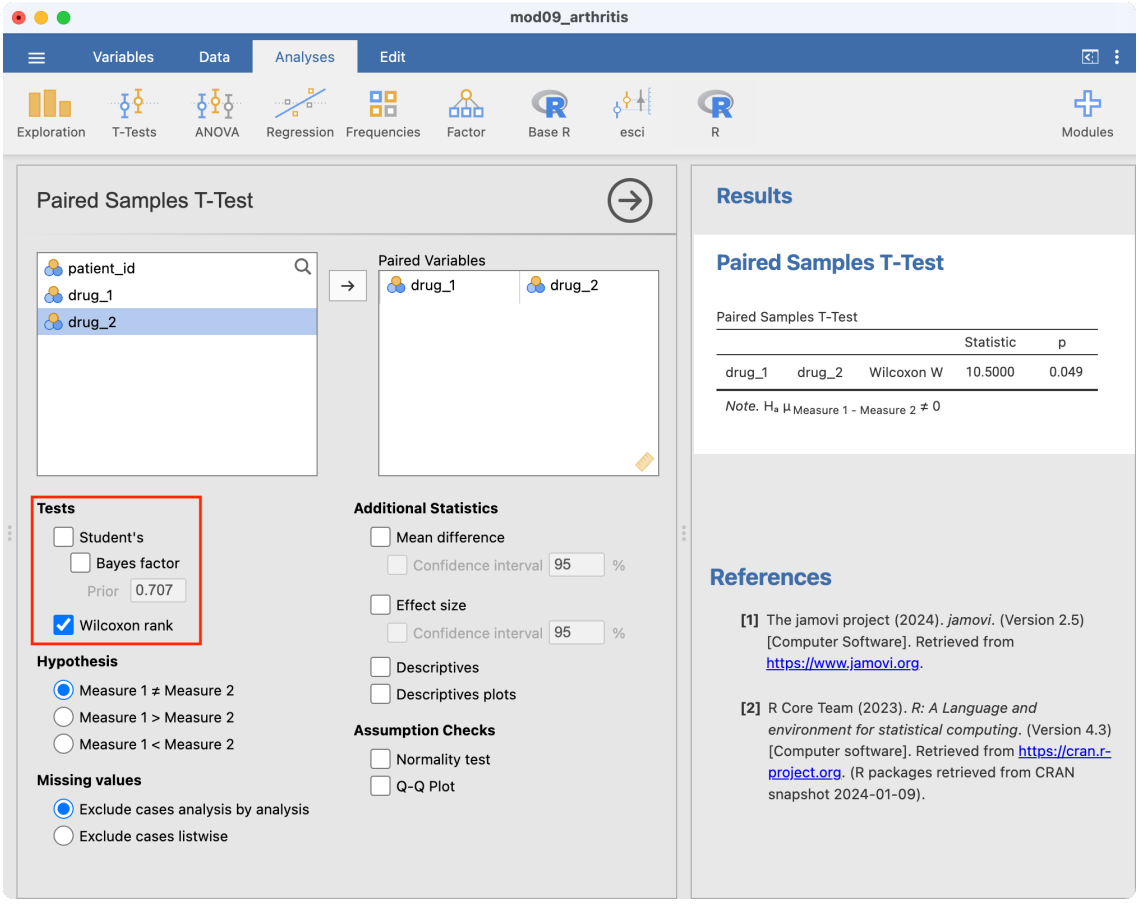
		Statistic	p
los	Mann-Whitney U	949.0000	0.014

Note. $H_a\ \mu_{No} \neq \mu_{Yes}$

9.10 Wilcoxon matched-pairs signed-rank test

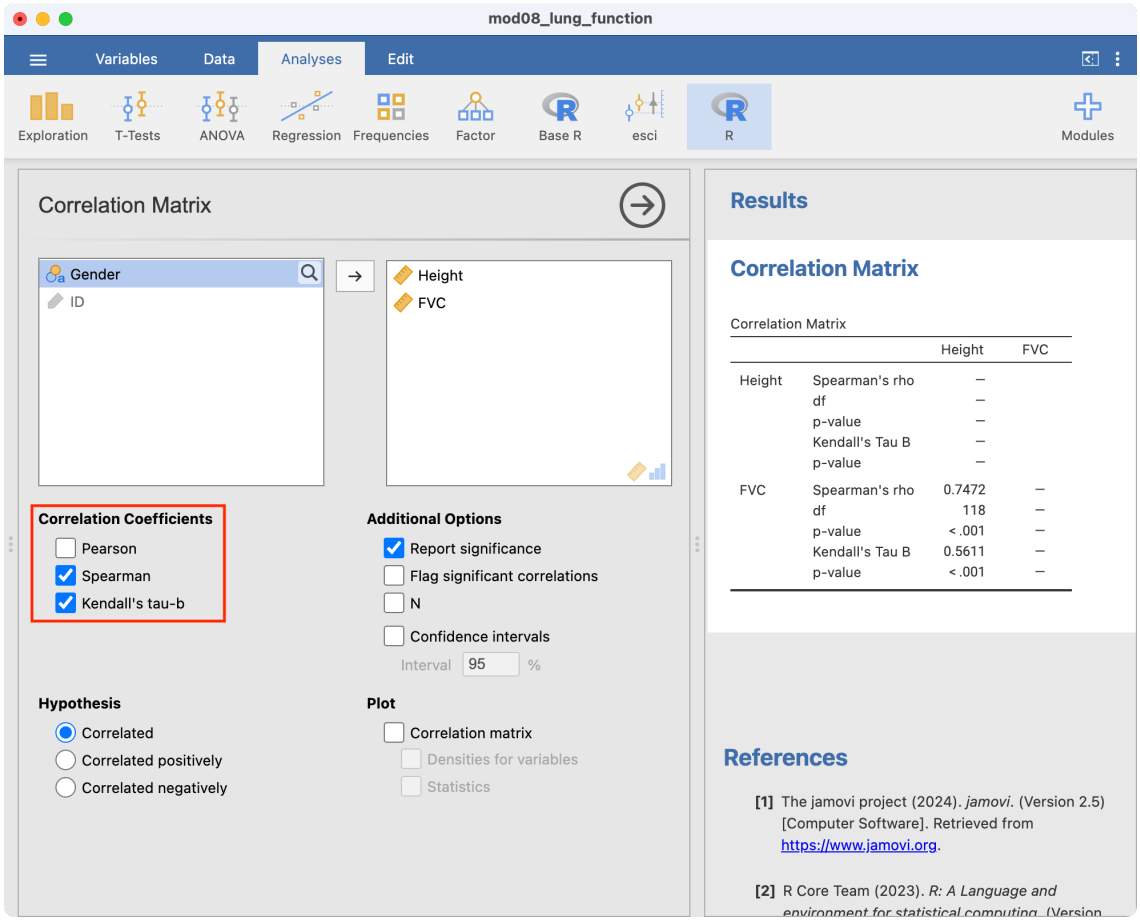
The Wilcoxon ranked-sum test is conducted in jamovi using **Analyses > T-Tests > Paired Samples T-Test**.

We will demonstrate using the dataset on the arthritis drug cross-over trial (`mod09_arthritis.csv`). Like the paired t-test the paired data need to be in separate columns.



9.11 Estimating rank correlation coefficients

The analyses for Spearman's and Kendall's rank correlation are conducted in similar ways using **Regression > Correlation Matrix**:


mod08_lung_function
Variables
Data
Analyses
Edit
Exploration
T-Tests
ANOVA
Regression
Frequencies
Factor
Base R
esci
R
Modules
Correlation Matrix
Gender
ID
Height
FVC
Correlation Coefficients
Pearson
Spearman
Kendall's tau-b
Additional Options
Report significance
Flag significant correlations
N
Confidence intervals
Interval
95
%
Hypothesis
Correlated
Correlated positively
Correlated negatively
Plot
Correlation matrix
Densities for variables
Statistics
Results
Correlation Matrix
Correlation Matrix
Height
FVC
Height
Spearman's rho
df
p-value
Kendall's Tau B
p-value
FVC
Spearman's rho
0.7472
df
118
p-value
< .001
Kendall's Tau B
0.5611
p-value
< .001
References
[1] The jamovi project (2024). jamovi. (Version 2.5) [Computer Software]. Retrieved from https://www.jamovi.org.
[2] R Core Team (2023). R: A Language and

R notes

9.12 Transforming non-normally distributed variables

One option for dealing with a non-normally distributed varaible is to transform it into its square, square root or logarithmic value. The new transformed variable may be normally distributed and therefore a parametric test can be used. First we check the distribution of the variable for normality, e.g. by plotting a histogram.

You can calculate a new, transformed, variable using standard commands. For example, to create a new column of data based on the log of length of stay:

```
library(jmv)

hospital <- readRDS("data/examples/mod09_infection.rds")

hospital$ln_los <- log(hospital$los+1)
descriptives(data=hospital, vars=c(los, ln_los))
```

```
 DESCRIPTIVES

 Descriptives

                        los         ln_los

   N                        132          132
   Missing                    0            0
   Mean                38.05303     3.407232
   Median              27.00000     3.332205
   Standard deviation  35.78057    0.7149892
   Minimum             0.000000     0.000000
   Maximum             244.0000     5.501258
```

You can now check whether this logged variable is normally distributed as described in Module 2, for example by plotting a histogram as shown in Figure 9.2.

To obtain the back-transformed mean, we can use the `exp` command to `anti-log` the mean:

```
exp(3.407232)
```

```
[1] 30.18159
```

If your transformed variable is approximately normally distributed, you can apply parametric tests such as the t-test. In the Worked Example 9.1 dataset, the variable `infect` (presence of nosocomial infection) is a binary categorical variable. To test the hypothesis that patients with nosocomial infection have a different length of stay to patients without infection, you can conduct a t-test on the `ln_los` variable. You will need to back transform your mean values, as shown in Worked Example 9.1 in the course notes when reporting your results.

9.13 Wilcoxon ranked-sum test

We use the `wilcox.test` function to perform the Wilcoxon ranked-sum test:

```
wilcox.test(continuous_variable ~ group_variable, data=df)
```

The Wilcoxon ranked-sum test will be demonstrated using the length of stay data in `mod09_infection.rds`. Here, out continuous variable is `los` and the grouping variable is `infect`.

```
wilcox.test(los ~ infect, data=hospital)
```

```
    Wilcoxon rank sum test with continuity correction

data:  los by infect
W = 949, p-value = 0.01413
alternative hypothesis: true location shift is not equal to 0
```

9.14 Wilcoxon matched-pairs signed-rank test

The `wilcox.test` function can also be used to conduct the Wilcoxon matched-pairs signed-rank test. The specification of the variables is a little different, in that each variable is specified as `dataframe$variable`:

```
wilcox.test(df$continuous_variable_1, df$continuous_variable_1, paired=TRUE)
```

We will demonstrate using the dataset on the arthritis drug cross-over trial (`mod09_arthritis.csv`). Like the paired t-test the paired data need to be in separate columns.

```
arthritis <- read.csv("data/examples/mod09_arthritis.csv")

wilcox.test(arthritis$drug_1, arthritis$drug_2,
            paired=TRUE)
```

```
Warning in wilcox.test.default(arthritis$drug_1, arthritis$drug_2, paired =
TRUE): cannot compute exact p-value with ties
```

```
    Wilcoxon signed rank test with continuity correction

data:  arthritis$drug_1 and arthritis$drug_2
V = 10.5, p-value = 0.04898
alternative hypothesis: true location shift is not equal to 0
```

9.15 Estimating rank correlation coefficients

The analyses for Spearman's and Kendall's rank correlation are conducted in similar ways:

```
lung <- read.csv("data/examples/mod08_lung_function.csv")

cor.test(lung$Height, lung$FVC, method="spearman")
```

```
Warning in cor.test.default(lung$Height, lung$FVC, method = "spearman"): Cannot
compute exact p-value with ties
```

```
    Spearman's rank correlation rho

data:  lung$Height and lung$FVC
S = 72796, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7472186

cor.test(lung$Height, lung$FVC, method="kendall")



    Kendall's rank correlation tau

data:  lung$Height and lung$FVC
z = 8.818, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.5611396
```

Activities

Activity 9.1

There is a hypothesis that university students who live and dine in the university hall consume less vitamin C than the students who live and dine at home. To test the hypothesis, 30 students were randomly selected and their urinary ascorbic acid level was measured in mg over 3 hours. Urinary excretion of ascorbic acid is a measure of vitamin C nutrition in humans. The data is given in the following table and a copy of the data set, `Activity_9.1.rds` is also available on Moodle.

Living and dining in Hall (n.=.17)	Living and dining at Home (n = 13)
34	163
62	205
37	83
27	372
38	50
20	22
7	47
53	255
22	30
37	89
14	96
28	48
28	25
70	163
16	
9	
121	

a) Examine the distribution of the data using a box-plot and histogram, and obtain descriptive statistics. How would you describe the distribution of ascorbic acid?
b) Which statistical test would be appropriate to test the hypothesis mentioned in the question and why?
c) State the hypotheses appropriate to the analytical method you mentioned in (b).
d) Carry out the statistical test you have mentioned in (b) and write your conclusion.

Activity 9.2

A drug was tested for its effect in lowering blood pressure. Fifteen women with hypertension were enrolled and had their systolic blood pressure measured before and after taking the drug. The data are available in the file `Activity_9.2.rds` on Moodle.

a) State the research question and the null hypothesis.
b) Obtain suitable summary statistics and test the null hypothesis. Describe the reason for choosing the test.
c) Write a brief conclusion.
d) What are the main limitations of this study? Consider both epidemiological and statistical aspects.

Activity 9.3

It is often necessary to take blood from newborn babies, which causes them pain. A group of researchers wanted to test whether a pacifier (dummy) would provide some pain relief. Newborn babies undergoing venepuncture were randomised to one of two treatment groups. In the placebo group, babies had blood taken as per the usual protocol. In the pacifier group, babies were allowed to suck on a pacifier while having their blood taken. Each venepuncture was observed by an assessor, who rated the baby's pain on the DAN scale which ranges from 0 (no pain) to 10. The DAN is based on observation of the baby's facial expression, limb movement, and vocal expression. The data are provided in the file `Activity_9.3.csv`.

Analyse these data to answer the research question. Write a brief report summarising your results and state your conclusion.

Supplementary Activity 9.4

A study was conducted among 201 people who had recently quit smoking. The dataset `Activity_9.4_smoke.csv` contains the following variables:

- ID: participant ID number
- age: participant age (years)
- gender: participant gender (1=Male, 2=Female)
- day_abs: number of days abstinent (i.e., number of days since last cigarette)

Analyse these data to assess whether there is a difference in the number of days abstinent between males and females. Write a brief report summarising your results and state your conclusion.

Module 10

An introduction to sample size estimation

Learning objectives

By the end of this module you will be able to:

- Explain the issues involved in sample size estimation for epidemiological studies;
- Estimate sample sizes for descriptive and analytic studies;
- Compute the sample size needed for planned statistical tests;
- Adjust sample size calculations for factors that influence study power.

Optional readings

Kirkwood and Sterne (2001); Chapter 35. [UNSW Library Link]

Bland (2015); Chapter 18. [UNSW Library Link]

For interest: Woodward (2013); Chapter 8. [UNSW Library Link]

10.1 Introduction

Determining the appropriate sample size (the number of participants in a study) is one of the most critical issues when designing a research study. A common question when planning a project is "How many participants do I need?" The sample size needs to be large enough to ensure that the results can be generalised to the population and will be accurate, but small enough for the study question to be answered with the resources available. In general, the larger the sample size, the more precise the study results will be.

Unfortunately, estimating the sample size required for a study is not straightforward and the method used varies with the study design and the type of statistical test that will be conducted on the data collected. In the past, researchers calculated the sample size by hand using complicated mathematical formula. More recently, look-up tables have been created which has removed the need for hand calculations. Now, most researchers use computer programs where parameters relevant to the particular study design are entered and the sample size is automatically calculated. In this module, we will use an abbreviated look-up table to demonstrate the parameters that need to be considered when estimating sample sizes for a confidence interval and use software for all other sample size calculations. The look-up table allows you to see at a glance, the impact of different factors on the sample size estimation.

Under and over-sized studies

In health research, there are different implications for interpreting the results if the sample size is too small or too large.

An under-sized study is one which lacks the power to find an effect or association when, in truth, one exists. If the sample size is too small, an important difference between groups may not be statistically significant and so will not be detected by the study. In fact, it is considered unethical to conduct a health study which is poorly designed so that it is not possible to detect an effect or association if it exists. Often, Ethics Committees request evidence of sample size calculations before a study is approved.

A classic paper by Freiman et al examined 71 randomised controlled trials which reported an absence of clinical effect between two treatments.(Freiman et al. 1978) Many of the trials were too small to show that a clinically important difference was statistically significant. If the sample size of an analytic study is too small, then only very limited conclusions can be drawn about the results.

In general, the larger the sample size the more precise the estimates will be. However, large sample sizes have their own effect on the interpretation of the results. An over-sized study is one in which a small difference between groups, which is not important in clinical or public health terms, is statistically significant. When the study sample is large, the null hypothesis could be rejected in error and research resources may be wasted. This type of study may be unethical due to the unnecessary enrolment of a large number of people.

10.2 Sample size estimation for descriptive studies

To estimate the sample size required for a descriptive study, we usually focus on specifying the width of the confidence interval around our primary estimate. For example, to estimate the sample size for a study designed to measure a prevalence we need to:

- nominate the expected prevalence based on other available evidence;
- nominate the required level of precision around the estimate. For this, the width of the 95% confidence interval (i.e. the distance equal to 1.96 $\times$ SE) is used.

Table 10.1 is an abbreviated look-up table that we can use to estimate the sample size for this type of study. Note that the sample size required to detect an expected population prevalence of 5% is the same as to detect a prevalence of 95%. Similarly 10% is equivalent to 90% etc. It is symmetric about 50%. From Table 10.1, you can see that the sample size required increases as the expected prevalence approaches 50% and as the precision increases (i.e. the required 95% CI becomes narrower).

Table 10.1: Sample size required to calculate a 95% confidence interval with a given precision

	Required confidence interval width									
Prevalence	1%	1.5%	2%	2.5%	3%	3.5%	4%	5%	10%	15%
5% or 95%	1,825	812	457	292	203	149	115			
10% or 90%	3,458	1,537	865	554	385	283	217	139		
15% or 85%	4,899	2,177	1,225	784	545	400	307	196	49	
20% or 80%	6,147	2,732	1,537	984	683	502	385	246	62	28
25% or 75%	7,203	3,202	1,801	1,153	801	588	451	289	73	33

Worked Example 10.1

A descriptive cross-sectional study is designed to measure the prevalence of bronchitis in children age 0-2 years with a 95% CI of $\pm$ 4%. The prevalence is expected to be 20%. From the table, a sample size of at least 385 will be required for the width of the 95% CI to be $\pm$ 4% (i.e. the reported precision of the summary statistic will be 20% (95% CI 16% to 24%)).

If the prevalence turns out to be higher than the researchers expected or if they decided that they wanted a narrower 95% CI (i.e. increase precision), a larger sample size would be required.

- What sample size would be required if the prevalence was 15% and the desired 95% CI was ± 3%?
- Answer: 545

10.3 Sample size estimation for analytic studies

Analytic study designs are used to compare characteristics between different groups in the population. The main study designs are analytic cross-sectional studies, case-control studies, cohort studies and randomised controlled trials. For analytic study designs, the outcome measure of interest can be a mean value, a proportion or a relative risk if a random sample has been enrolled. For case-control studies the most appropriate measure of association is an odds ratio.

Factors to be considered

The first important decision in estimating a required sample size for an analytic study is to select the type of statistical test that will be used to report or analyse the data. Each type of test is associated with a different method of sample size estimation.

Once the statistical method has been determined, the following issues need to be decided:

- Statistical power: the chance of finding a difference if one exists, e.g. 80%;
- Level of significance: the P value that will be considered significant, e.g. $P<0.05$;
- Minimum effect size of interest: the size of the difference between groups e.g. the difference in the proportion of parents who oppose immunisation in two different regions or the difference in mean values of blood pressure in two groups of people with different types of cardiac disease;
- Variability: the spread of the measurements, e.g. the expected standard deviation of the main outcome variable (if continuous), or the expected proportions;
- Resources: an estimate of the number of participants available and amount of funding to run the study.

In addition to deciding the level of power and probability that will be used, the difference between groups that is regarded as being important has to be estimated. The smallest difference between study groups that we want to detect is described as the minimum expected effect size. This is determined on the basis of clinical judgement, public health importance and expertise in the condition being researched, or may it be need to be determined from a pilot study or a literature review. The smaller the expected effect or association, the larger the sample size will need to obtain statistical significance. We also need some knowledge of how variable the measurement is expected to be. For this we often use the standard deviation for a continuous measure. As measurement variability increases, the sample size will need to increase in order to detect the expected difference between the groups. Above all, a study has to be practical in terms of the availability of a population from which to draw sufficient numbers for the study and in terms of the funds that are available to conduct the study.

Power and significance level

The power of a study, which was discussed in Module 4, is the chance of finding a statistically significant difference when one exists, i.e. the probability of correctly rejecting the null hypothesis. The relationship between the power of a study and statistical significance is shown in Table 10.2.

Table 10.2: Comparison of study result with the truth

	Effect	No effect
Evidence	Correct	α
No evidence	β	Correct

The power of a study is expressed as 1 - β where β is the probability of a false negative (that is, the probability of a Type II error - incorrectly not rejecting the null hypothesis. In most research, power is generally set to 80% (a Type II error rate of 20%). However, in some studies, especially in some clinical trials where rigorous results are required, power is set to 90% (a Type II error rate of 10%).

The significance level, or α level, is the level at which the P value of a test is considered to be statistically significant. The α level is usually set at 5% indicating a probability of <0.05 will be regarded as statistically significant. Occasionally, especially if several outcome measures are being compared, the α level is set at 1% indicating a probability of <0.01 will be regarded as statistically significant.

The calculation of sample sizes for analytic studies are based on calculations that are somewhat tedious to compute by hand. Software packages or online sample size calculators are the standard method of calculating sample sizes for analytic studies. In this module, we will demonstrate the use of an online calculator called PS, available at https://cqsclinical.app.vumc.org/ps/

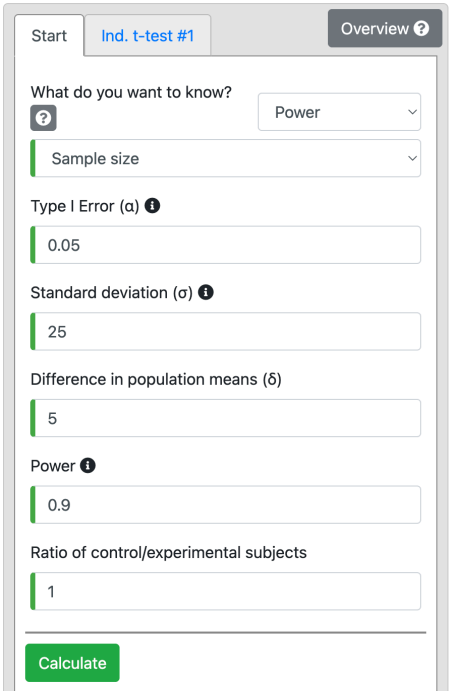
10.4 Detecting the difference between two means

The test that is used to show that two mean values are significantly different from one another is the independent samples t-test (Module 5). The sample size needed for this test to have sufficient power can be calculated using PS as shown in the Worked Example below.

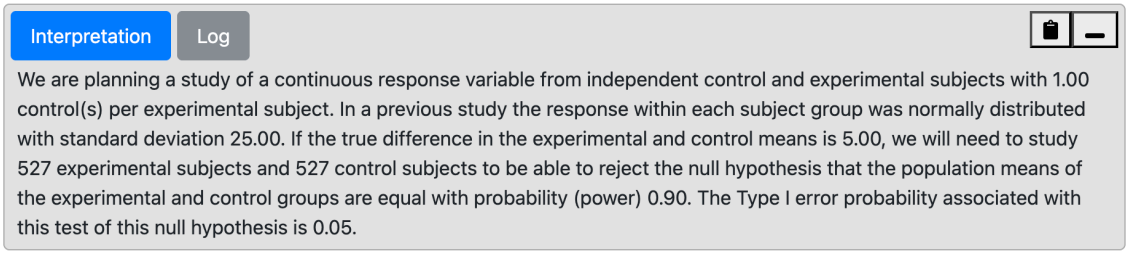
Worked Example 10.2

There is a hypothesis that the use of the oral contraceptive (OC) pill in premenopausal women can increase systolic blood pressure. A study was planned to test this hypothesis using a two sided t-test. The investigators are interested in detecting an increase of at least 5 mm Hg systolic blood pressure in the women using OC compared to the non-OC users with 90% power at a 5% significance level. A pilot study shows that the SD of systolic blood pressure in the target group is 25 mm Hg and the mean systolic blood pressure of non-OC user women is 110 mm Hg. What is the minimum number of women in each group that need to be recruited for the study to detect this difference?

Solution The effect size of interest is 5 mm Hg and the associated standard deviation is 25 mm Hg. For power of 90% and alpha of 5%, the sample size calculation using the **t-test > Independent** tab of PS:



Output 10.2



From the output, we can see that with 90% power we will need 527 participants in each group, i.e., 1054 participants in total.

If the above were carried out by taking baseline measures of systolic blood pressure, and then again when the women were taking the OC pills, it would be a matched-pair study. Computing sample sizes for paired studies requires an estimate of the correlation between the paired observations, or an estimate of the standard deviation of the differences. Calculating sample sizes for paired studies is a little more complex than for independent studies, and is outside the scope of this course.

10.5 Detecting the difference between two proportions

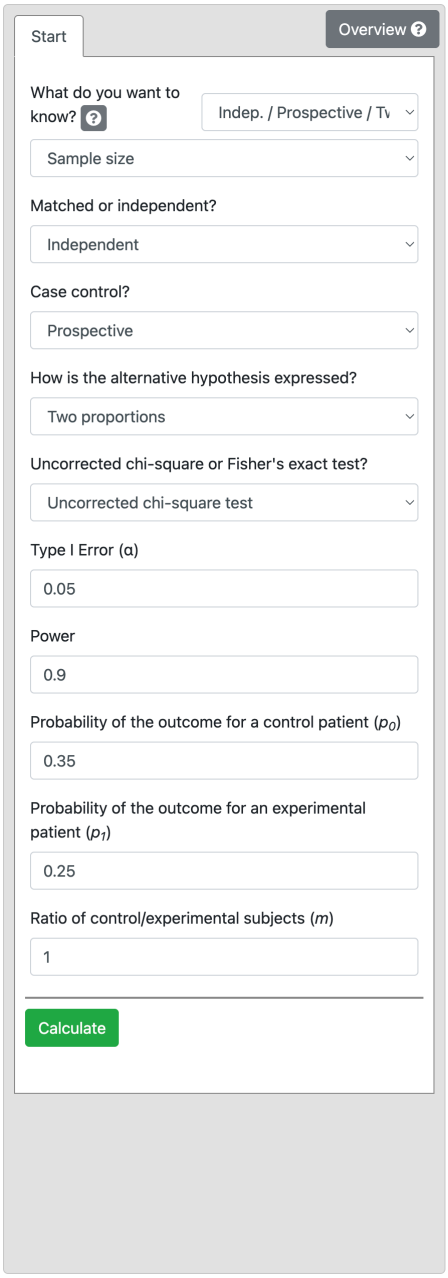
The statistical test for deciding if there is a significant difference between two independent proportions is a Pearson's chi-squared test (Module 7).

Other than the power and alpha required for the test, the expected prevalence or incidence rate of the outcome factor needs to be estimated for each of the two groups being compared, based on what is known from other studies or what is expected. Occasionally, we may not know the expected proportion in one of the groups, e.g. in a randomised control trial of a novel intervention. In the sample size calculation for such a study, we should instead justify the minimum expected difference between the proportions based on what is important from a clinical or public health perspective. Based on the minimum difference, we can then derive the expected proportion for both groups. Note that the smaller the difference, the larger the sample size required.

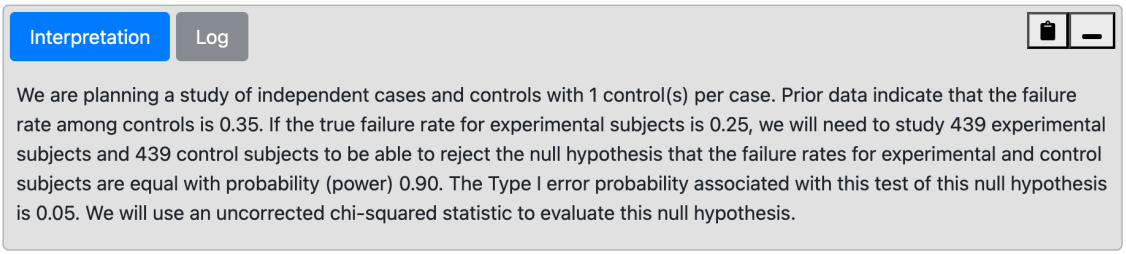
The sample size required in each group to observe a difference in two independent proportions can be calculated using the **Dichotomous** tab of PS.

Worked Example 10.3

A health promotion campaign is being developed to reduce smoking in a community, and will be tested in a randomised controlled trial. The researchers would like to detect a reduction in smoking from 35% to 25%. How many participants should be recruited to detect a difference at a 5% significance level, with a power of 90%?



Output 10.3: Sample size calculation for two independent proportions



From Output 10.3, we see that we would need 439 intervention and 439 control participants (i.e. a total sample size of 878 participants).

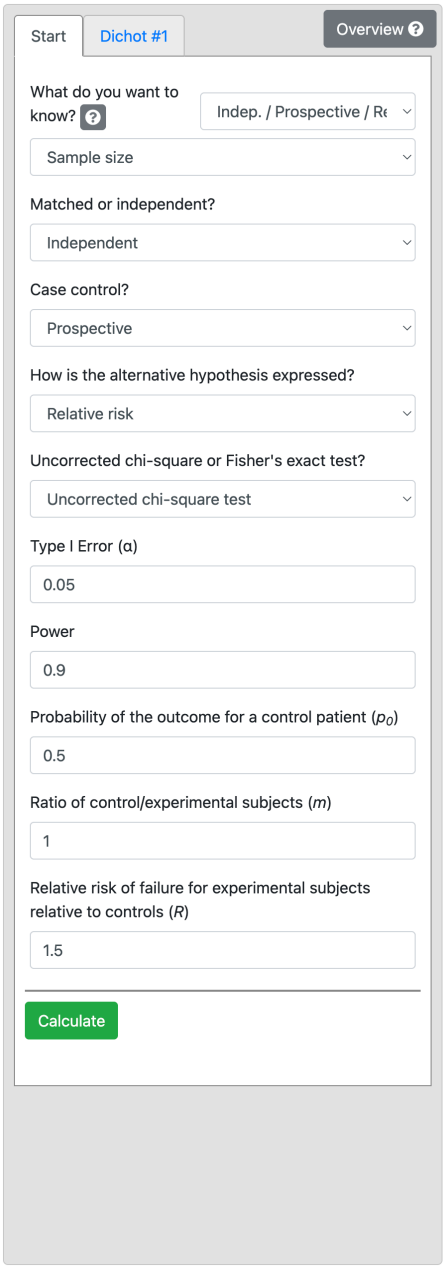
10.6 Detecting an association using a relative risk

The relative risk is used to describe the association between an exposure and an outcome variable if the sample has been randomly selected from the population. This statistic is often used to describe the effect or association of an exposure in a cross-sectional or cohort study or the effect/association of a treatment in an randomised controlled trial. To estimate the sample size required for the RR to have a statistically significant P value, i.e. to show a significant association, we need to define: - the size of the RR that is considered to be of clinical or public health importance; - the event rate (rate of outcome) among the group who are not exposed to the factor of interest (reference group); - the desired level of significance (usually 0.05); - the desired power of the study (usually 80% or 90%).

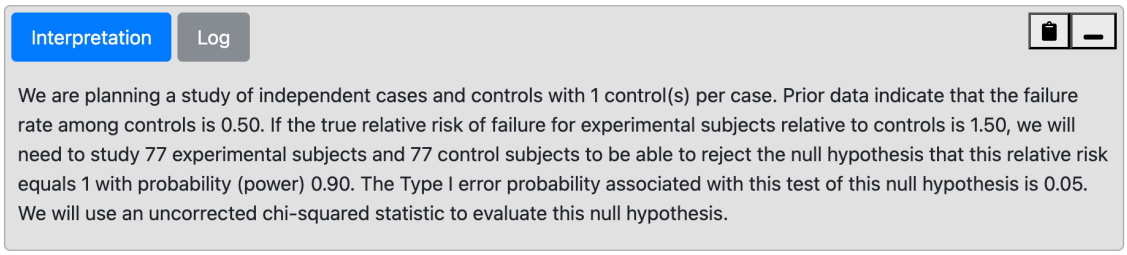
In general, a RR of 2.0 or greater is considered to be of public health importance, however, a smaller RR can be important when exposure is high. For example, there may be a relatively small risk of respiratory infection among young children with a parent who smokes (RR ~ 1.2). If 25% of children are exposed to smoking in their home, then the high exposure rate leads to a very large number of children who have preventable respiratory infections across the community.

Worked Example 10.4

A study is planned to investigate the effect of an environmental exposure on the incidence of a certain common disease. In the general (unexposed) population the incidence rate of the disease is 50% and it is assumed that the incidence rate would be 75% in the exposed population. Thus the relative risk of interest would be 1.5 (i.e. 0.75 / 0.50). We want to detect this effect with 90% power at a 5% level of significance.



Output 10.4: Sample size calculation for relative risk



From Output 10.4, we can see that for a control proportion of 0.5 and RR of 1.5, we need a total sample size of 154, that is 77 people would be needed in each of the exposure groups.

10.7 Detecting an association using an odds ratio

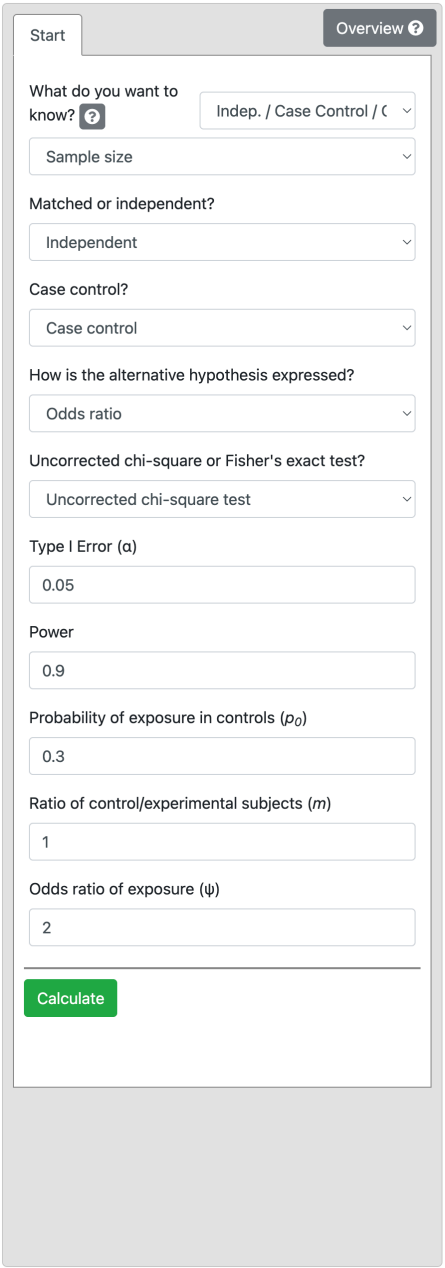
If we are designing a case-control study, the appropriate measure of effect is an odds ratio. The method for estimating the sample size required to detect an odds ratio of interest is slightly

different to that for the relative risk. However, the same parameters are required for the estimation:

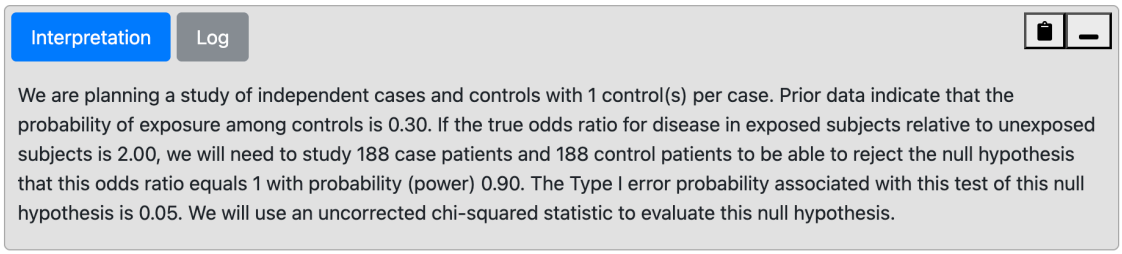
- the minimum OR to be considered clinically important;
- the proportion of exposed among the control group;
- the desired level of significance (usually 0.05);
- the desired power of the study (usually 80% or 90%).

Worked Example 10.5

A case-control study is designed to examine an association between an exposure and outcome factor. Existing literature shows that 30% of the controls are expected to be exposed. We want to detect a minimum OR of 2.0 with 90% power and 5% level of significance.



Output 10.5



We find that 188 controls and 188 cases are required i.e. a total of 376 participants.

This sample size would be smaller if we increased the effect size (OR) or reduced the study power to 80%. You could try this yourself (answer: 141 per group if power is reduced to 80%).

10.8 Factors that influence power

Dropouts

It is common to increase estimated sample sizes to allow for drop-outs or non-response. To account for drop-outs, the estimated sample size can be divided by (1 minus the dropout rate). Consider the following case:

- n-completed: the number who will complete the study (i.e. n after drop-out)
- n-recruited: the number who should be recruited (i.e. n before drop-out)
- d: drop-out rate (as a proportion - i.e. a number between 0 and 1)

Then n-completed = n-recruited × (1 - d)

Re-arranging this formula gives: n-recruited = n-completed ÷ (1 - d).

Unequal groups

Many factors that come into play in a study can reduce the estimated power of a study. In clinical trials, it is not unusual for recruitment goals to be much harder to achieve than expected and therefore for the target sample size to be impossible to realise within the timeframe planned for recruitment.

In case-control studies, the number of potential case participants available may be limited but study power can be maintained by enrolling a greater number of controls than cases. Or in an experimental study, more participants may be randomised to the new treatment group to test its effects accurately when much is known about the effect of standard care and a more precise estimate of the new treatment effect is required.

However, there is a trade-off between increasing the ratio of group size and the total number that needs to be enrolled. Consider Worked Example 10.5: selecting an equal number of controls and cases would require 188 cases and 188 controls, a total of 376 participants.

We may want to reduce the number of cases required, by selecting 2 controls for every case. Selecting 2 controls (N1) per case (N2) would require 140 cases and 280 controls, a total of 420 participants. We can extend this example and investigate the impact of changing the ratio of controls per case.

Table 10.3: The effect of unequal groups on Worked Example 10.5

Controls per case	Number of cases required	Number of controls required	Total participants required
1	188	188	376
2	140	280	420
3	124	372	496
4	115	460	575
5	111	555	666
6	107	642	749
7	105	735	840
8	103	824	927
9	102	918	1,020
10	101	1,010	1,111

This can be visualised graphically, as in Figure 10.1.

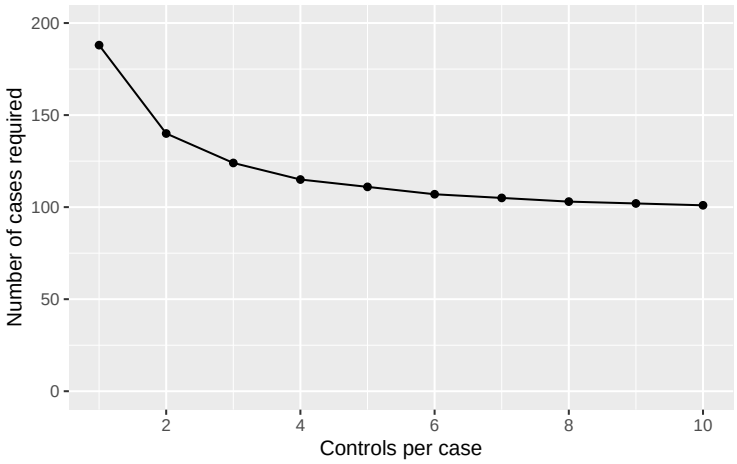


Figure 10.1: Increasing the number of controls per case

We can see that the number of cases required drops off if we go from 1 to 2 controls per case, and again from 2 to 3 controls per case. Once we go from 3 to 4 controls per case, we only reduce the number of cases by 9 (124 vs 115 cases), but at an increase of 88 (372 vs 460) controls. Clearly, this reduction in cases is not offset by the extra controls required.

10.9 Limitations in sample size estimations

In this module we have seen how to use a sample size calculator to estimate the sample size requirement of a study given the statistical test that will be used and the expected characteristics of the sample. However, once a study is running, it is not unusual for sample size to be compromised by the lack of research resources, difficulties in recruiting participants or, in a clinical trial, participants wanting to change groups when information about the new experimental treatment rapidly becomes available in the press or on the internet.

One approach that is increasingly being used is to conduct a blinded interim analysis say when 50% of the total data that are planned have been collected. In this, a statistician external to the research team who is blinded to the interpretation of the group code is asked to measure the effect size in the data with the sole aim of validating the sample size requirement. It is rarely a good idea to use an interim analysis to reduce the planned sample size and terminate a trial

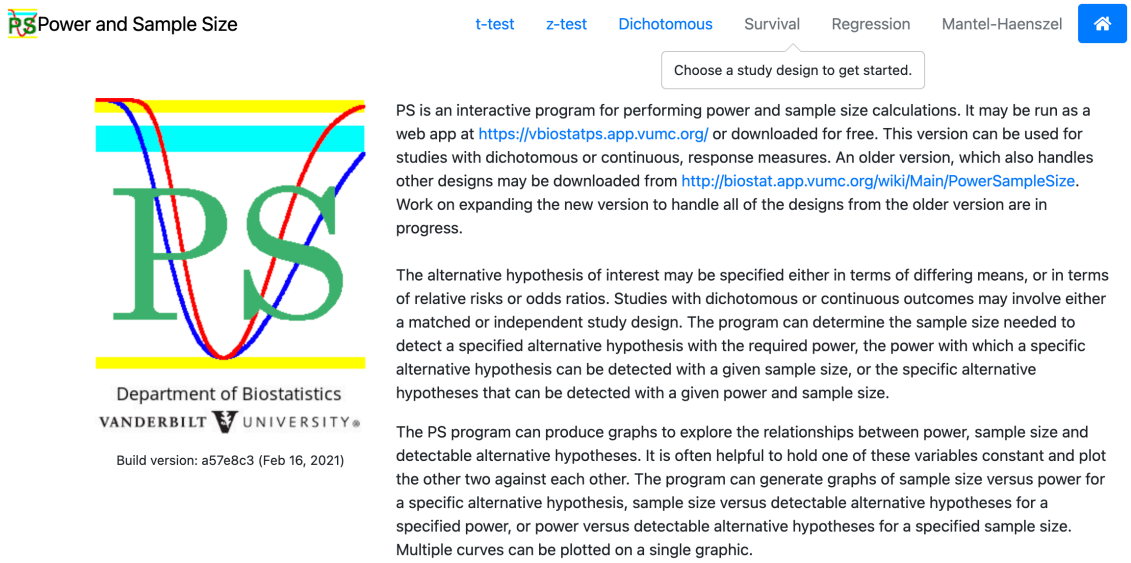
early because the larger the sample size, the greater the precision with which the treatment effect is estimated. However, interim analyses are useful for deciding whether the sample size needs to be increased in order to answer the study question and avoid a Type II error.

10.10 Summary

In this module we have discussed the importance of conducting a clinical or epidemiological study with enough participants so that an effect or association can be identified if it exists (i.e. study power), and how this has to be balanced by the need to not enrol more participants than necessary because of resource issues. We have looked at the parameters that need to be considered when estimating the sample size for different studies and have used a look-up table to estimate required sample size for a prevalence study and a sample size calculator to estimate appropriate sample sizes in epidemiological research under the most straightforward situations. The common requirement in all the situations is that the researchers need to specify the minimum effect measure (e.g. difference in means, OR, RR etc) they want to detect with a given probability (usually 80% to 90%) at a certain level of significance (usually $P<0.05$). The ultimate decision on the sample size depends on a compromise among different objectives such as power, minimum effect size, and available resources. To make the final decision, it is helpful to do some trial calculations using revised power and the minimum detectable effect measure.

Software notes

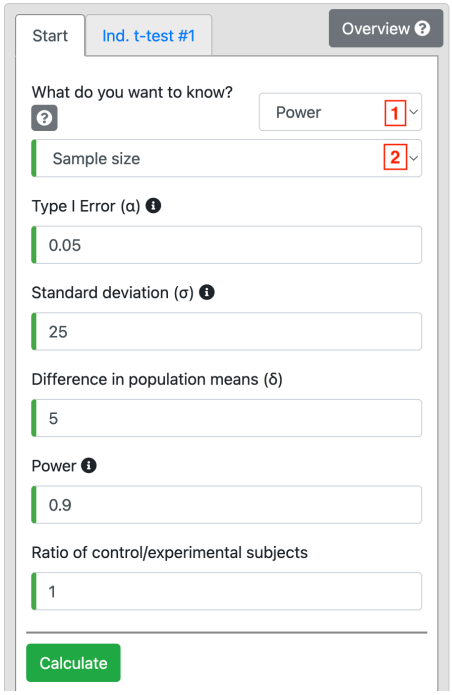
In this module, we will demonstrate the use of an online calculator called PS, available at https://cqsclinical.app.vumc.org/ps/ While sample size calculations can be conducted in R (using the `EpiR` and `pwr` packages in particular), there is currently an inconsistency in `EpiR` which means that I do not recommend this package.



The web-app allows for a variety of variable types; we will focus on **t-test** and **Dichotomous** analyses.

10.11 Sample size calculation for two independent samples t-test

To do the problem discussed in Worked Example 10.2, choose **t-test > Independent**. Choose **Power** in the first drop-down, and then choose **Sample size**:



Based on the information in Worked Example 10.2, change **Power** to `0.9` and enter `25` as the **Standard deviation (σ)**. We are interested in detecting a difference of 5 mmHg, so this is entered as **Diference in population means (δ)**. We use equal numbers in each group, so **Ratio** is entered as 1.

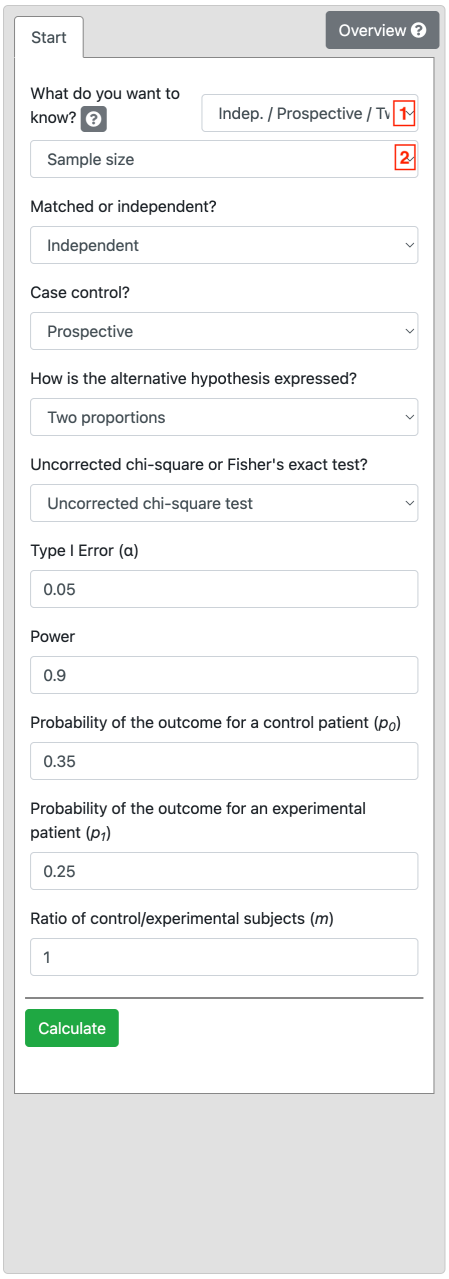
Click **Calculate** to produce Output 10.2.

10.12 Sample size calculation for difference between two independent proportions

All sample size calculations for analyses based on 2-by-2 tables are conducted under the **Dichotomous** tab, choosing:

- **Indep / Prospective / Two Prop.** for analyses based on estimating the difference in proportions;
- **Indep / Prospective / Rel. Risk** for analyses based on estimating a relative risk;
- **Indep / Case Control / Odds Ratio** for case-control studies.

To do the problem discussed in Worked Example 10.3, choose **Indep / Prospective / Two Prop.** and **Sample size** in the first two drop-down options.



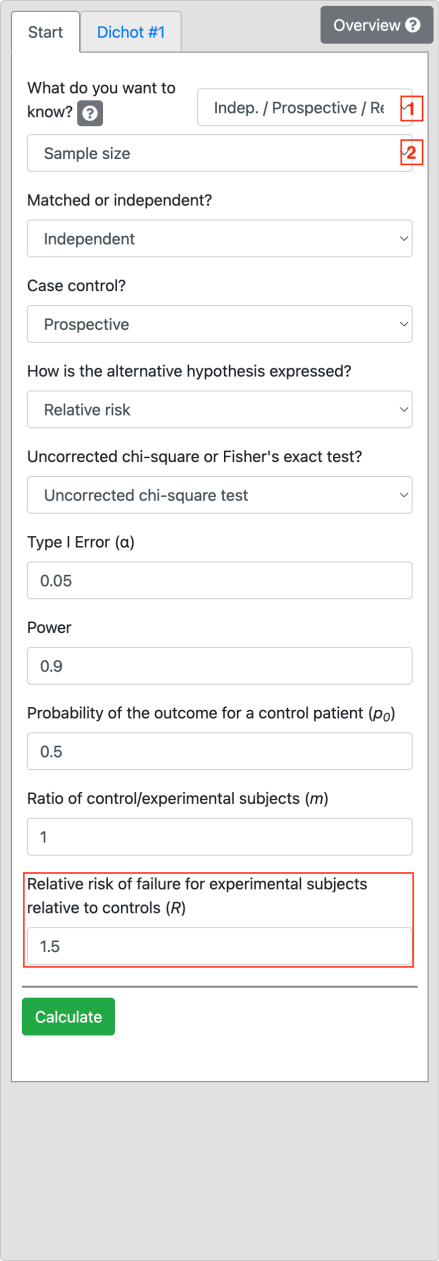
Based on the information in Worked Example 10.5, the **Power** is `0.9` and **Significance level** is `0.05`, and the two proportions are entered as `0.35` and `0.2`. Define equal numbers in each group (**Ratio** of `1`).

Click **Calculate** obtain Output 10.3.

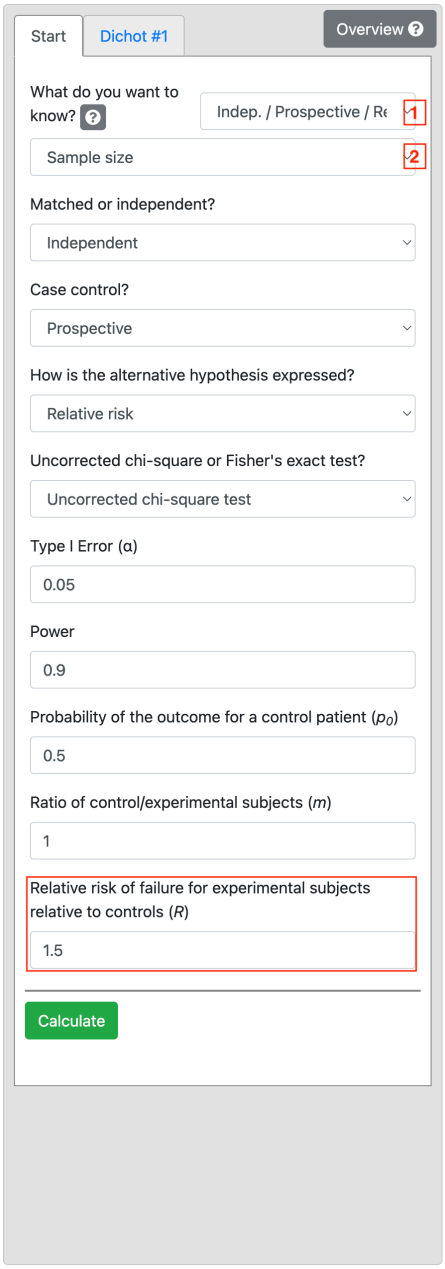
Note: It doesn't matter if you swap the proportions for the **Control** and **Experimental** group, you will get the same result.

10.13 Sample size calculation with a relative risk or odds ratio

Using information from Worked Example 10.4, we select **Indep / Prospective / Rel. Risk**, and enter **Power** as `0.9`, enter `0.5` as the **Probability of the outcome for a control patient (p0)**. Finally, enter `1.5` as the **Relative risk of failure for experimental subjects relative to controls (R)** shown below:



Now we calculate the sample size for Worked Example 10.5. Select **Indep / Case Control / Odds Ratio** and **Sample size**, and enter **Power** as `0.9`, enter `0.3` as the **Probability of exposure in controls (p0)**. Finally, enter `2` as the **Odds ratio of exposure (ψ)** shown below:



10.14 Estimating power or effect size

Note that in all PS examples provided here, we have chosen to estimate the sample size (given the study power and effect size of interest). PS also allows us to estimate the power (given the study sample size and effect size) or the effect size (given the study sample size and power).

Activities

Activity 10.1

We are planning a study to measure the prevalence of a relatively rare condition (say approximately 5%) in children age 0-5 years in a remote community.

a) What type of study would need to be conducted?
b) Use the correct sample size table included in your notes to determine how many children would need to be enrolled for the confidence interval to be
    i. 2%
    ii. 4% around the prevalence?

What would the resulting prevalence estimates and 95% CIs be?

Activity 10.2

We are planning an experimental study to test the use of a new drug to alleviate the symptoms of the common cold compared to the use of Vitamin C. Participants will be randomised to receive the new experimental drug or to receive Vitamin C. How many participants will be required in each group (power = 80%, level of significance = 5%).

a) If the resolution of symptoms is 10% in the control group and 40% in the new treatment group?
b) How large will the sample size need to be if we decide to recruit two control participants to every intervention group participant?
c) If we decide to retain a 1:1 ratio of participants in the intervention and controls groups but the resolution of symptoms is 20% in the control group and 40% in the new treatment group?
d) How many participants would we need to recruit (calculated in c) if a pilot study shows that 15% of people find the new treatment unpalatable and therefore do not take it?

Activity 10.3

In a case-control study, we plan to recruit adult males who have been exposed to fumes from an industrial stack near their home and a sample of population controls in whom we expect that 20% may also have been exposed to similar fumes through their place of residence or their work. We want to show that an odds ratio of 2.5 for having respiratory symptoms associated with exposure to fumes is statistically significant.

a) What statistical test will be needed to measure the association between exposure and outcome?
b) How large will the sample size need to be to show that the OR of 2.5 is statistically significant at $P < 0.05$ with 90% power if we want to recruit equal number of cases and controls?
c) What would be the required sample size (calculated in b) if the minimum detectable OR were 1.5?
d) If there are problems recruiting cases to detect an OR of 1.5 (as calculated in c), what would the sample size need to be if the ratio of cases to controls was increased to 1:3?

Activity 10.4

In the above study to measure the effects of exposure to fumes from an industrial stack, we also want to know if the stack has an effect on lung function which can be measured as forced vital capacity in 1 minute (FEV1). This measurement is normally distributed in the population.

a) If the research question is changed to wanting to show that the mean FEV1 in the exposed group is lower than the mean FEV1 in the control group what statistical test will now be required?
b) Population statistics show that the mean FEV1 and its SD in the general population for males are 4.40 L (SD=1.25) which can be expected in the control group.

We expect that the mean FEV1 in the cases may be 4.0 L. How many participants will be needed to show that this mean value is significantly different from the control group with $P < 0.05$ with an 80% power if we want to recruit equal number in each group?

c) How much larger will the sample size need to be if the mean FEV1 in the cases is 4.20 L?

Supplementary Activity 10.5

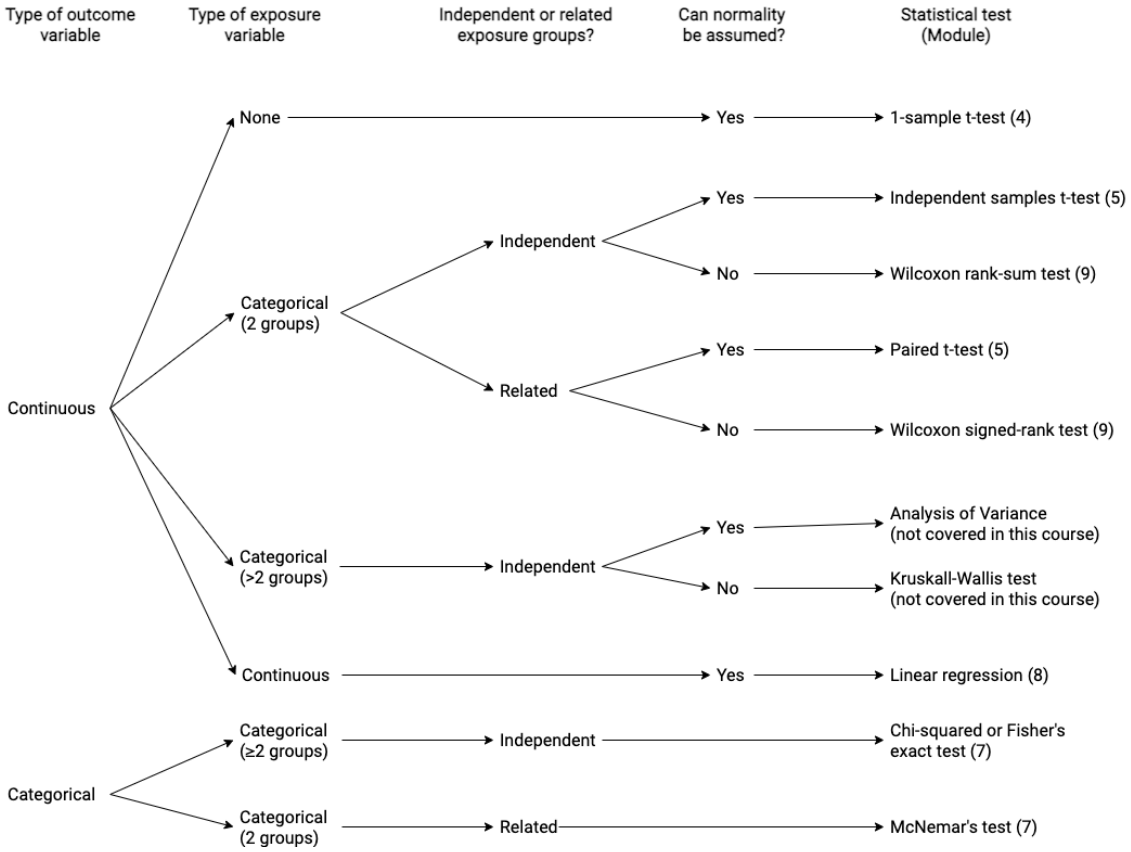
Your colleague, Clancy, wants to test an intervention consisting of behavioural techniques to reduce the amount of time that primary school children spend on small-screen recreation (i.e., use of television, computers, and tablets). Clancy would like to use a randomised controlled trial to assess the impact of the intervention compared to providing an information pamphlet, and will use the relative risk to summarise the intervention effect.

Clancy's study will use the categorisation of 2 or more hours per day of screen time as the primary outcome. Evidence suggests that 53% of children in primary school spend 2 or more hours per day on small-screen recreation.

a) How many participants should be recruited to detect a relative risk of 0.5, with 90% power and a 2-sided significance level of 0.05?
b) Clancy's colleague suggests Clancy should instead aim to detect a relative risk of 0.75 (keeping the power and 2-sided significance level as in part (a)). How many children should be recruited for this new effect size? What has happened to the sample size and why?
c) In talking with parents, you find out that they would be more likely to enrol their child in the study if their child had a greater chance of receiving the behavioural intervention than the pamphlet. Assuming Clancy wants to detect a relative risk of 0.75 (i.e., the scenario in part (b)), how many participants should Clancy recruit to have twice as many participants in the behavioural intervention arm compared to the pamphlet arm?
d) Clancy's colleague suggests that approximately 25% of children drop out of trials in school-based studies, regardless of whether they receive the control or the intervention. How many participants should Clancy recruit to allow for this, based on scenario (c)?

Appendix

Analysis flowchart



References

Altman, Douglas G. 1990. *Practical Statistics for Medical Research*. 1st ed. Boca Raton, Fla: Chapman and Hall/CRC.

Armitage, Peter, Geoffrey Berry, and J. N. S. Matthews. 2013. *Statistical Methods in Medical Research*. 4th ed. Wiley-Blackwell.

Assel, Melissa, Daniel Sjoberg, Andrew Elders, Xuemei Wang, Dezheng Huo, Albert Botchway, Kristin Delfino, et al. 2019. "Guidelines for Reporting of Statistics for Clinical Research in Urology." *BJU International* 123 (3): 401–10. https://doi.org/10.1111/bju.14640.

Australian Bureau of Statistics. Thu, 10/10/2024 - 11:30. "Causes of Death, Australia, 2023." https://www.abs.gov.au/statistics/health/causes-death/causes-death-australia/latest-release.

Australian Institute of Health and Welfare. 2024. "Australia's Mothers and Babies." *Australian Institute of Health and Welfare*. https://www.aihw.gov.au/reports/mothers-babies/australias-mothers-babies/contents/about.

———. 2025. "Australia's Health." https://www.aihw.gov.au/reports-data/australias-health.

Bland, Martin. 2015. *An Introduction to Medical Statistics*. 4th Edition. Oxford, New York: Oxford University Press.

Boers, Maarten. 2018. "Graphics and Statistics for Cardiology: Designing Effective Tables for Presentation and Publication." *Heart* 104 (3): 192–200. https://doi.org/10.1136/heartjnl-2017-311581.

Brown, Lawrence D., T. Tony Cai, and Anirban DasGupta. 2001. "Interval Estimation for a Binomial Proportion." *Statistical Science* 16 (2): 101–17. https://www.jstor.org/stable/2676784.

Cole, T. J. 2015. "Too Many Digits: The Presentation of Numerical Data." *Archives of Disease in Childhood* 100 (7): 608–9. https://doi.org/10.1136/archdischild-2014-307149.

Deeks, Jon. 1998. "When Can Odds Ratios Mislead?" *BMJ* 317 (7166): 1155. https://doi.org/10.1136/bmj.317.7166.1155a.

Delacre, Marie, Daniël Lakens, and Christophe Leys. 2017. "Why Psychologists Should by Default Use Welch's t-Test Instead of Student's t-Test" 30 (1): 92. https://doi.org/10.5334/irsp.82.

Freiman, Jennie A., Thomas C. Chalmers, Harry Smith, and Roy R. Kuebler. 1978. "The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial." *New England Journal of Medicine* 299 (13): 690–94. https://doi.org/10.1056/NEJM197809282991304.

Kirkwood, Betty, and Jonathan Sterne. 2001. *Essentials of Medical Statistics*. 2nd edition. Malden, Mass: Wiley-Blackwell.

Ruxton, Graeme D. 2006. "The Unequal Variance t-Test Is an Underused Alternative to Student's t-Test and the Mann–Whitney U Test." *Behavioral Ecology* 17 (4): 688–90. https://doi.org/10.1093/beheco/ark016.

Schmidt, Carsten Oliver, and Thomas Kohlmann. 2008. "When to Use the Odds Ratio or the Relative Risk?" *International Journal of Public Health* 53 (3): 165–67. https://doi.org/10.1007/s00038-008-7068-3.

Therneau, Terry M., and Patricia M. Grambsch. 2010. *Modeling Survival Data: Extending the Cox Model*. New York Berlin Heidelberg: Springer.

Vickers, Andrew J., Melissa J. Assel, Daniel D. Sjoberg, Rui Qin, Zhiguo Zhao, Tatsuki Koyama, Albert Botchway, et al. 2020. "Guidelines for Reporting of Figures and Tables for Clinical Research in Urology." *European Urology*, May. https://doi.org/10.1016/j.eururo.2020.04.048.

Webb, Penny, Chris Bain, and Andrew Page. 2016. *Essential Epidemiology: An Introduction for Students and Health Professionals*. 3rd edition. Cambridge: Cambridge University Press.

West, Robert M. 2021. "Best Practice in Statistics: Use the Welch t-Test When Testing the

Difference Between Two Groups." *Annals of Clinical Biochemistry* 58 (4): 267–69. https://doi.org/10.1177/0004563221992088.
Woodward, Mark. 2013. *Epidemiology: Study Design and Data Analysis*. 3rd edition. Chapman and Hall/CRC.