PHCM9795: Foundations of Biostatistics

Timothy Dobbins

13 June, 2024

# 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

# Module 1

# Summarising and presenting data

### 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 data numerically
- Present and interpret graphical summaries of data using a variety of graphs
- Compute summary statistics to describe the centre and spread of data

### Optional readings

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

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

Acock (2010); Chapter 5.

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)

## 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.

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

**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.

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. ((**australianinstituteofhealthandwelfare24?**)).

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 ((**australianbureauofstatistics24?**)).

> "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.

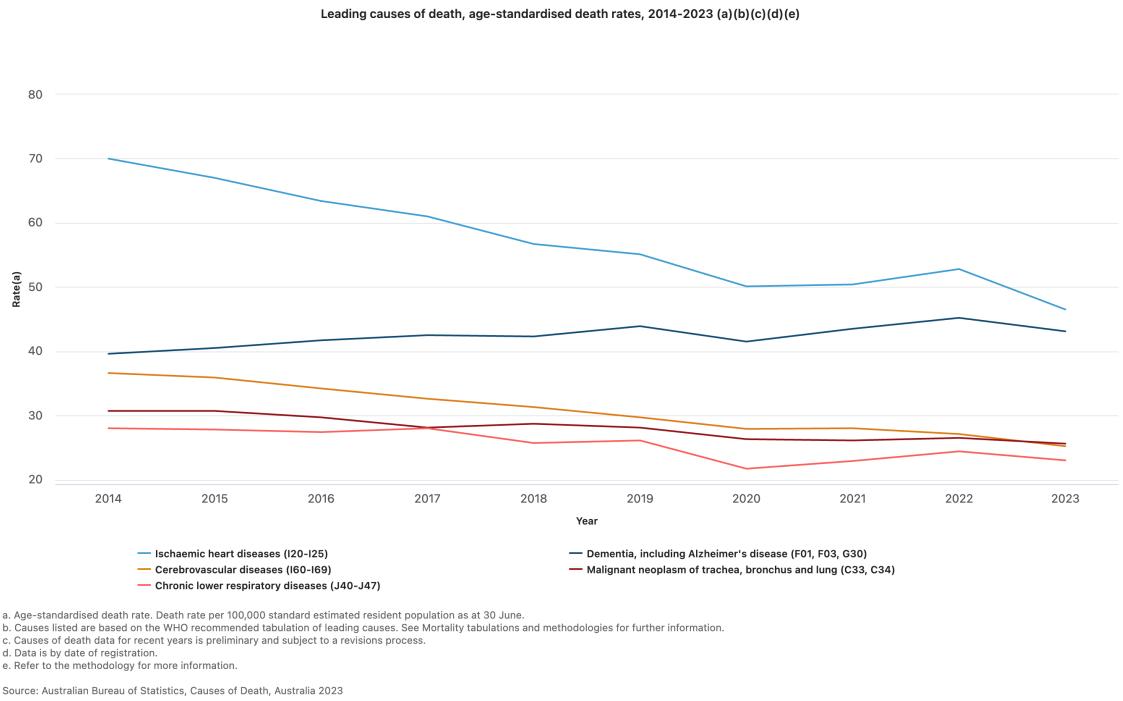



Figure 1.1: Leading causes of death, age-standardised death rates, 2014-2023

### 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 (**australianinstituteofhealthandwelfare25?**)) 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 59 years. Note that R and Stata use slightly different methods to calculate the interquartile range (Stata IQR: 43 to 59 years; R IQR: 43 to 58 years). This difference is not practically important, and either range would be considered correct.

### 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 $\mu$:

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

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 ($\sigma^2$) and standard deviation($\sigma$), 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.

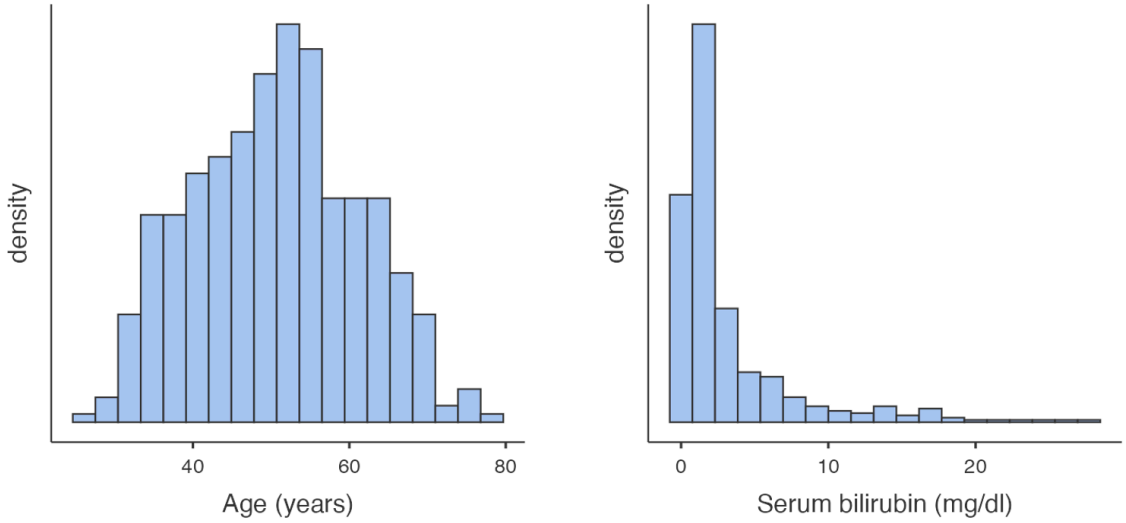



Figure 1.2: Histogram of age (left) and serum bilirubin (right) from a sample of data

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.

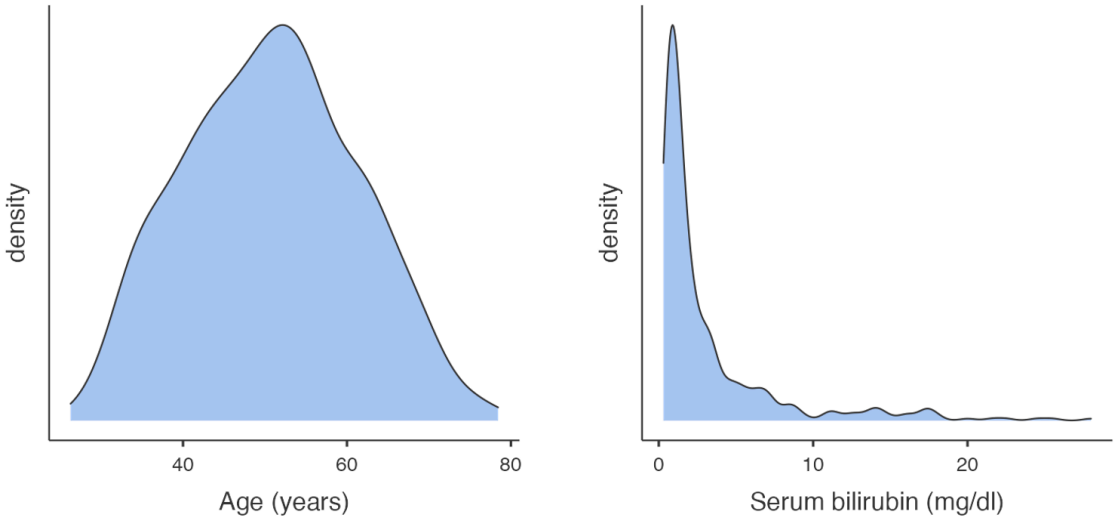



Figure 1.3: Histogram of age (left) and serum bilirubin (right) from a sample of data

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.

### 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.

### 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.

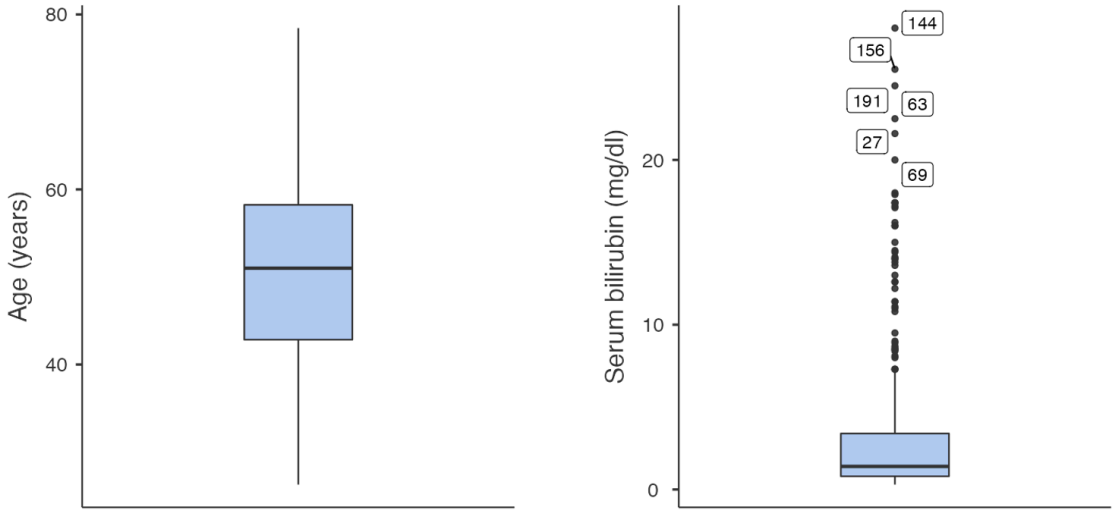



Figure 1.4: Box plot of age (left) and serum bilirubin (right) from PBC study data

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.

# Module 2

# Probability and probability distributions

### Learning objectives

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

- Describe the concept of probability;
- Describe the characteristics of a binomial distribution and a Normal distribution;
- Compute probabilities from a binomial distribution using statistical software;
- Compute probabilities from a Normal distribution using statistical software;
- Decide when to use parametric or non-parametric statistical methods;
- Briefly outline other types of distributions.

### Optional readings

Kirkwood and Sterne (2001); Chapters 5, 14 and 15. [UNSW Library Link]

Bland (2015); Chapters 6 and 7. [UNSW Library Link]

## 2.1 Introduction

In Module 1, we looked at how to summarise data numerically and graphically. In this module, we will introduce the concept of probability which underpins the theoretical basis of statistics, and then introduce the concept of probability distributions. We will look at the binomial distribution, and then look at the most important distribution in statistics: the Normal distribution. Finally, we introduce some other probability distributions commonly used in biostatistics.

### 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 | 5.097087 |
| 2 | 92 | 22.3 | 27.427184 |
| 3 | 155 | 37.6 | 65.048544 |
| 4 | 144 | 35.0 | 100.000000 |

* 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.

### 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.

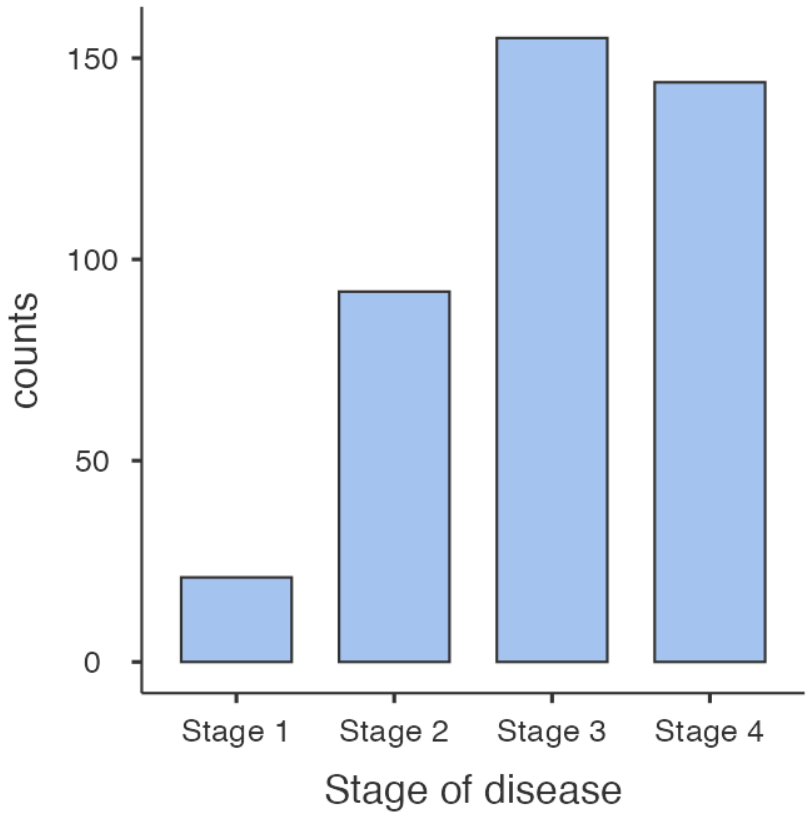



Figure 2.1: Bar graph of stage of disease from PBC study

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.

### 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 of disease * | | | | |
|---|---|---|---|---|---|
| Sex | 1 | 2 | 3 | 4 | Total |
| 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 of disease * | | | | |
|---|---|---|---|---|---|---|
| Sex | | 1 | 2 | 3 | 4 | Total |
| Male | Count | 3 | 8 | 16 | 17 | 44 |
| | Column % | 14.3% | 8.7% | 10.3% | 11.8% | |
| Female | Count | 18 | 84 | 139 | 127 | 368 |
| | Column % | 85.7% | 91.3% | 89.7% | 88.2% | |
| Total | Count | 21 | 92 | 155 | 144 | 412 |

* 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 of disease * | | | | |
|---|---|---|---|---|---|---|
| Sex | | 1 | 2 | 3 | 4 | Total |
| Male | Count | 3 | 8 | 16 | 17 | 44 |
| | Row pct | 6.8% | 18.2% | 36.4% | 38.6% | |

| | | Stage of disease * | | | | |
|---|---|---|---|---|---|---|
| Sex | | 1 | 2 | 3 | 4 | Total |
| Female | Count | 18 | 84 | 139 | 127 | 368 |
| | Row pct | 4.9% | 22.8% | 37.8% | 34.5% | |
| Total | Count | 21 | 92 | 155 | 144 | 412 |

*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 **?@fig-1-2**, 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.

We might also consider a table containing further pieces of information. The table presented in Figure 2.2 (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 2.2: 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.

### 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 2.3). 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.

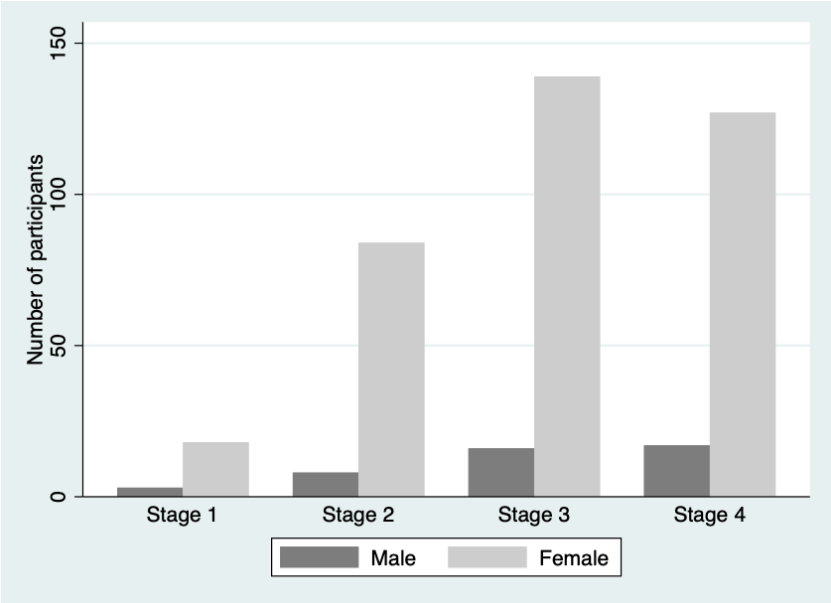



Figure 2.3: Bar graph of stage of disease by sex from PBC study

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 2.4).

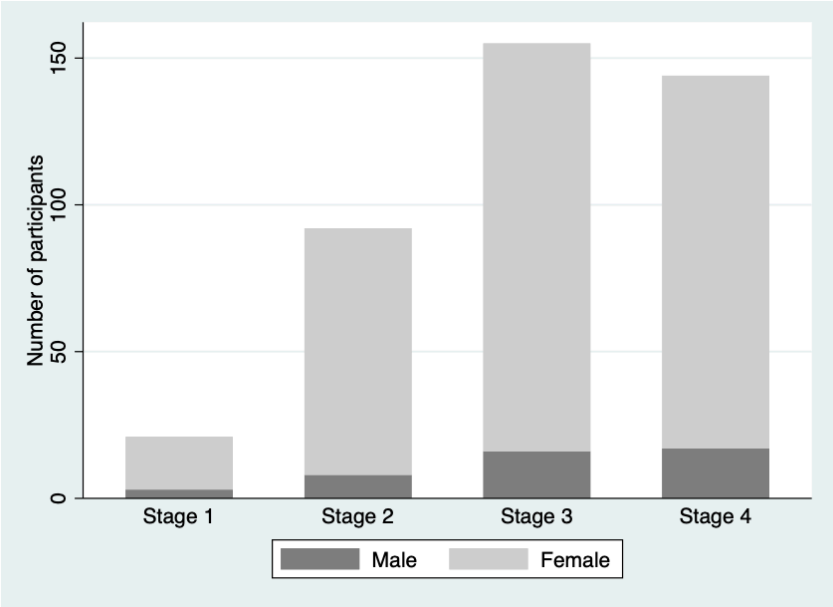



Figure 2.4: Stacked bar graph of stage of disease by sex from PBC study

Finally, a stacked relative bar chart (Figure 2.5) 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 2.5 that the distribution of sex is similar across each stage of disease.

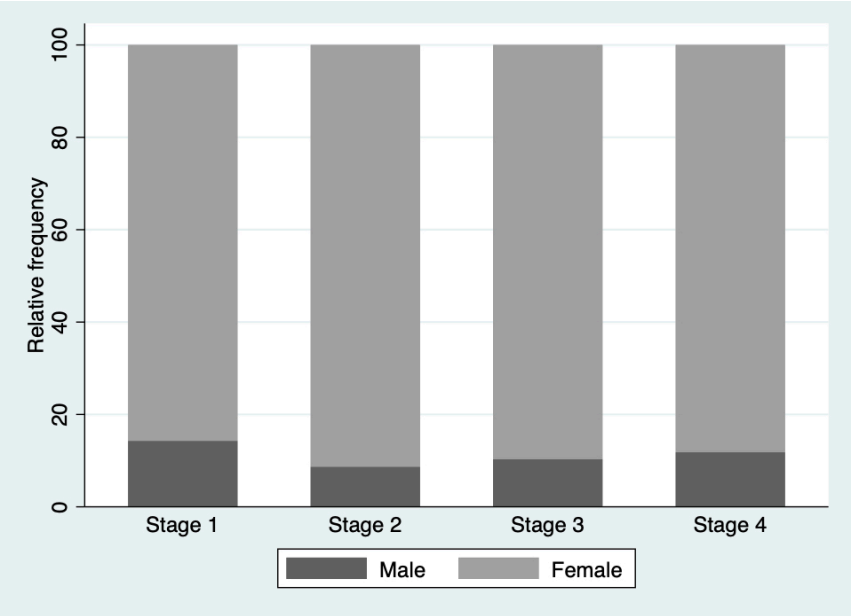



Figure 2.5: Relative frequency of sex within stage of disease from PBC study

## 2.2 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) |

| Summary statistic | Guideline (reference) |
|---|---|
| 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 |

| Summary statistic | Guideline (reference) |
|---|---|
| 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. |

| Summary statistic | Guideline (reference) |
|---|---|
| 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 |

| Summary statistic | Guideline (reference) |
|---|---|
| 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)

## 2.3 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.

## 2.4 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.

## 2.5 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 |

| x (number of heads observed) | P(X=x) |
|---|---|
| 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 × 0.38 × 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 × 0.38 × 0.38 = 0.0549 |
| A | A | Not A | 0.38 × 0.38 × 0.62 = 0.0895 |
| A | Not A | A | 0.38 × 0.62 × 0.38 = 0.0895 |
| Not A | A | A | 0.62 × 0.38 × 0.38 = 0.0895 |
| A | Not A | Not A | 0.38 × 0.62 × 0.62 = 0.1461 |
| Not A | A | Not A | 0.62 × 0.38 × 0.62 = 0.1461 |
| Not A | Not A | A | 0.62 × 0.62 × 0.38 = 0.1461 |

| Person 1 | Person 2 | Person 3 | Probability |
|---|---|---|---|
| Not A | Not A | Not A | 0.62 × 0.62 × 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 |

## 2.6 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 Stata and R notes at the end of this Module.

### Mean and variance of a binomial variable

The properties of the binomial distribution are useful in the statistical modelling of prevalence data. If *X* has a binomial distribution, then the mean of *X* is:

$$E(X) = n \times p$$

and the variance is:

$$var(X) = n \times p \times (1-p)$$

where *n* = the number of trials, and *p* = the probability of the event occurring (or success).

### 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.

In Stata, we used the `binomialp` function with n=6, k=3, and p=0.198. This gives an answer of 0.08 (see **?@sec-binom-stata** for details).

In R, we used the `dbinom` function with x=3, size=6, and prob=0.198. This gives an answer of 0.08 (see **?@sec-binom-r** for details).

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

Alternatively, in Stata we can use the `binomialtail` function (which gives "the probability of observing k or more successes in n trials when the probability of a success on one trial is p"). Again, see **?@sec-binom-stata** for details.

In R we can use the `pbinom` function with the `lower.tail=FALSE` option (**?@sec-binom-r**).

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

This can also be done by using the `binomial` function in Stata (which gives "the probability of observing k or fewer successes in n trials when the probability of a success on one trial is p") or the `pbinom` function in R.

# Module 3

# Precision, standard errors and confidence intervals

### Learning objectives

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

- 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;
- Recognise the importance of the central limit theorem;
- Calculate the standard error of a mean;
- Calculate and interpret confidence intervals for a mean;
- Be familiar with the t-distribution and when to use it.

### Optional readings

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

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

## 3.1 Introduction

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.

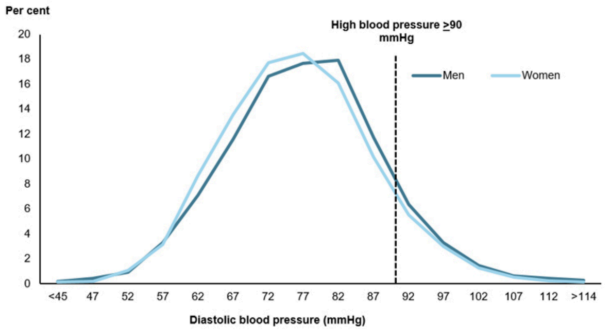
However, for the inferences about the population to be valid, a random sample of the population must be obtained. The goal of using random sampling methods is to obtain a sample that is representative of the target population. In other words, apart from random error, the information derived from the 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 discrete 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

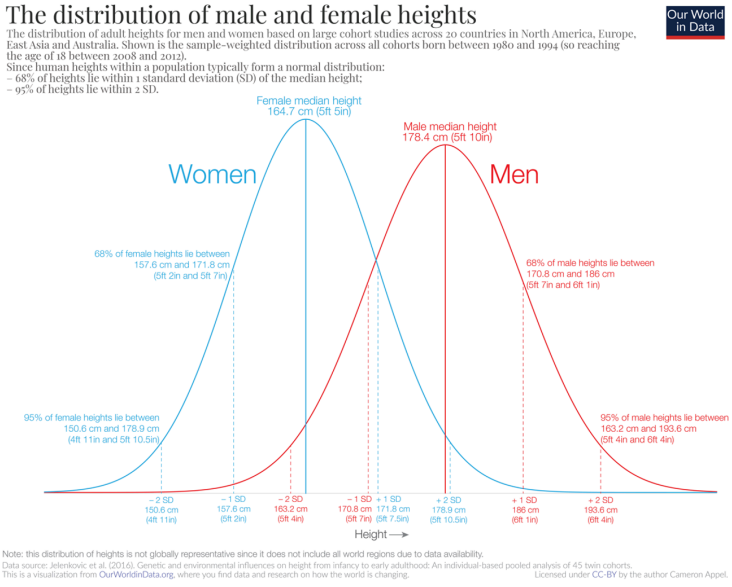



Figure 3.2: Distribution of male and female heights

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.

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 in Module 3.

[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)

The Normal distribution is characterised by two parameters: the mean ($\mu$) and the standard deviation ($\sigma$). 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 $\sigma$ 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.

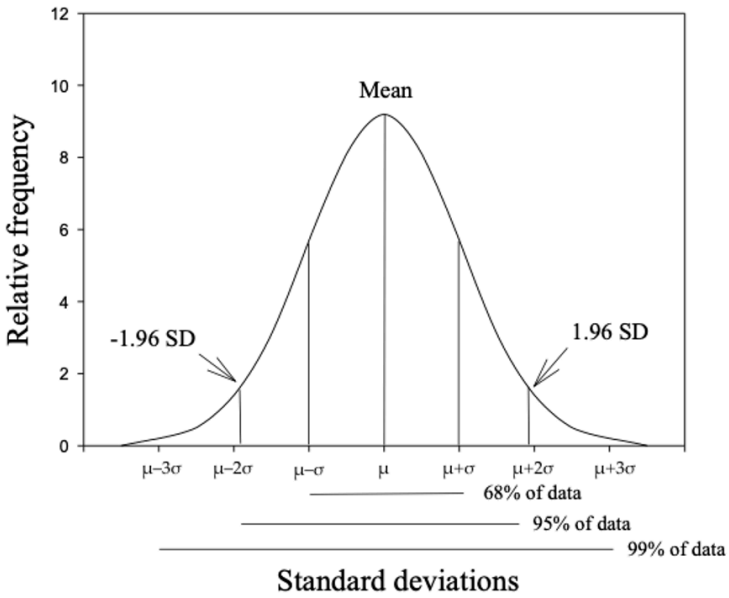



Figure 3.3: Characteristics of a Normal distribution

## 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 $\mu$ and a standard deviation $\sigma$ 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, $\mu$=77.9 and $\sigma$=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.

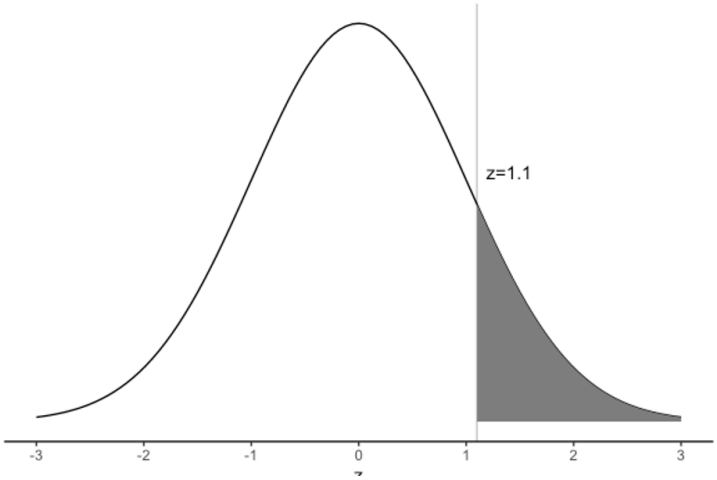



Figure 3.4: Area under the Standard Normal curve (as probability) for Z > 1.1

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 (see **?@sec-normal-stata** and **?@sec-normal-r** for details).

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. With a large sample, simply plotting a histogram is one of the best ways to assess whether a variable is Normally distributed. For smaller samples, examining a histogram can be less clear, particularly for histograms with only a small number of bins. However, if a histogram looks 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 histogram. However a boxplot in isolation is not as useful as a histogram, 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 in Stata and R that test for Normality. These tests are beyond the scope of this course and are not recommended.

The histogram for our 30 weights is approximately bell-shaped and roughly symmetrical. The mean and median (50th percentile) values are similar, as would be expected for a Normal distribution. Thus, it would be reasonable to assume that the data are Normally distributed.

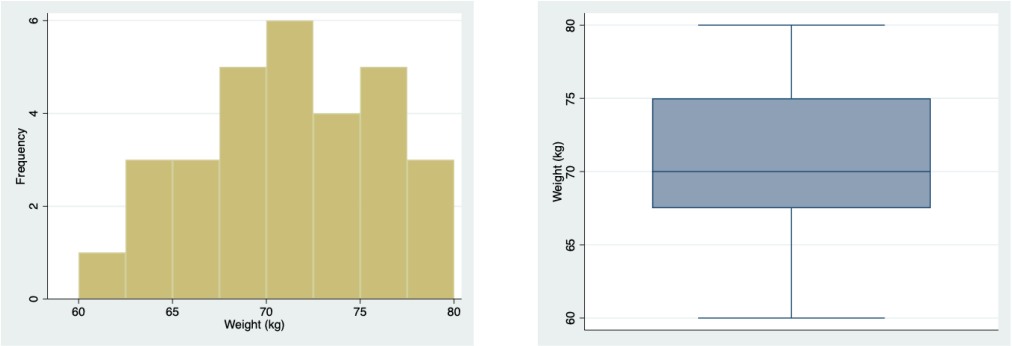



Figure 3.5: Histogram and boxplot of weight of 30 students attending a gym

## 3.6 Non-Normally distributed measurements

In the above example, diastolic blood pressure was Normally distributed with an approximately bell-shaped frequency histogram. However, 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, histograms and boxplots can look very different. For example, data of length of hospital stay in a sample of children are shown as a histogram and as a box plot in Figure 3.6.

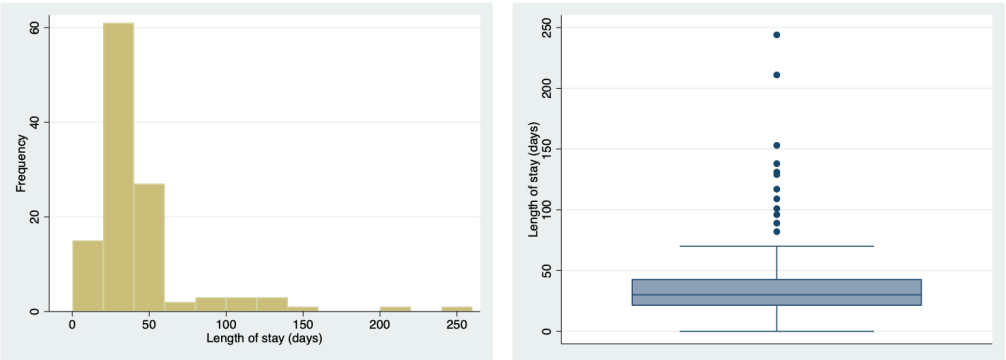



Figure 3.6: Histogram and boxplot of length of stay

In the histogram 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. In the boxplot, the whiskers appear to be fairly symmetric, but there are some unusual values (denoted by dots) above the box and its whiskers. Stata defines these unusual values as being more than 1.5 times the IQR from the edge of the box.

The presence of unusual values may be an indication that the data are not Normally distributed. Both the histogram and the box plot show that the distribution has a marked tail towards high values and that non-parametric statistics should be used to generate summary statistics and analyse the data.

Note that Stata has defined points as being unusual, or 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, you should not automatically exclude them from the analysis, particularly if the sample was selected randomly from a population. 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 plausible and, if it is, it should be included in the analysis.

## 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 PHCM9796: 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 $\mu$.

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

| v1 | v2 | v3 | v4 | v5 | v6 |
|---|---|---|---|---|---|
| 65 | 70.0 | 70.0 | 67.5 | 65.0 | 80.0 |
| 70 | 72.5 | 67.5 | 62.5 | 67.5 | 72.5 |
| 60 | 65.0 | 72.5 | 77.5 | 75.0 | 75.0 |
| 75 | 70.0 | 67.5 | 77.5 | 67.5 | 62.5 |
| 75 | 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 $\mu$ and standard deviation $\sigma$, the sampling distribution of the sample mean $\bar{x}$ approaches a normal distribution with mean $\mu$ 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

In Module 2, 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 ($\sigma$) is required for calculation of the standard error. Usually, $\sigma$ 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 $\sigma$, 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.

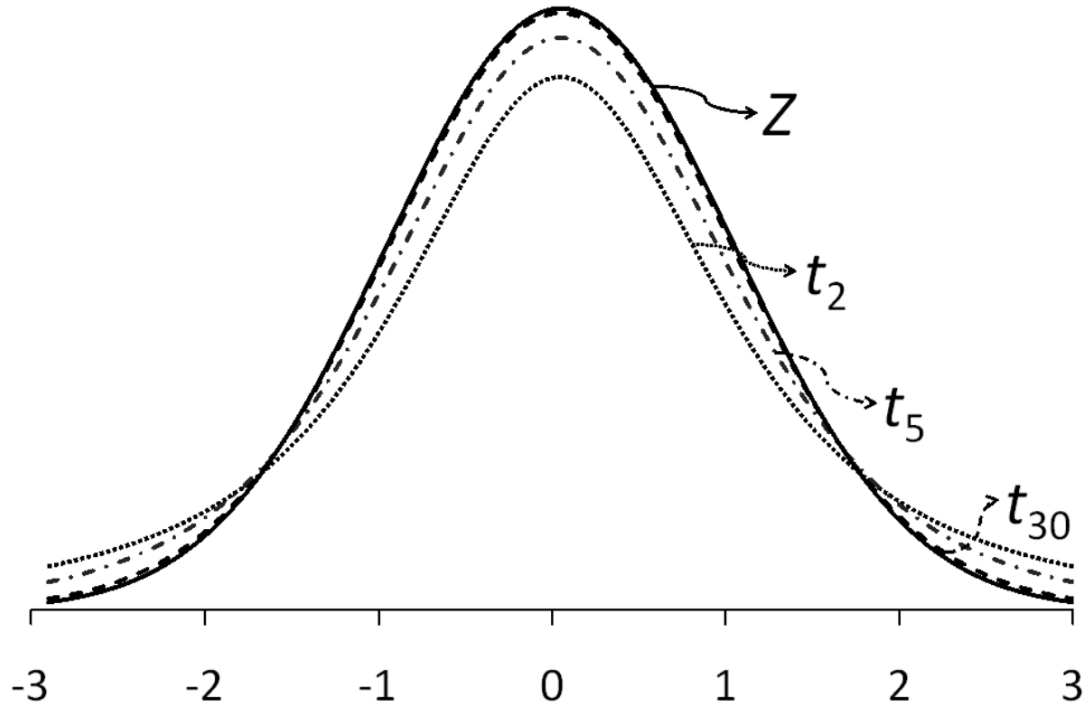



Figure 3.7: The normal (Z) and the student's t-distribution with 2, 5 and 30 degrees of freedom

If a variable $X$ is normally distributed and the population standard deviation $\sigma$ is known, using the normal distribution is appropriate. However, if $\sigma$ 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.

# 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]

Acock (2010); Section 7.4.

## 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

| Size of P-value | Strength of evidence |
|---|---|
| <0.001 | Very strong evidence |

| Size of P-value | Strength of 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.

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

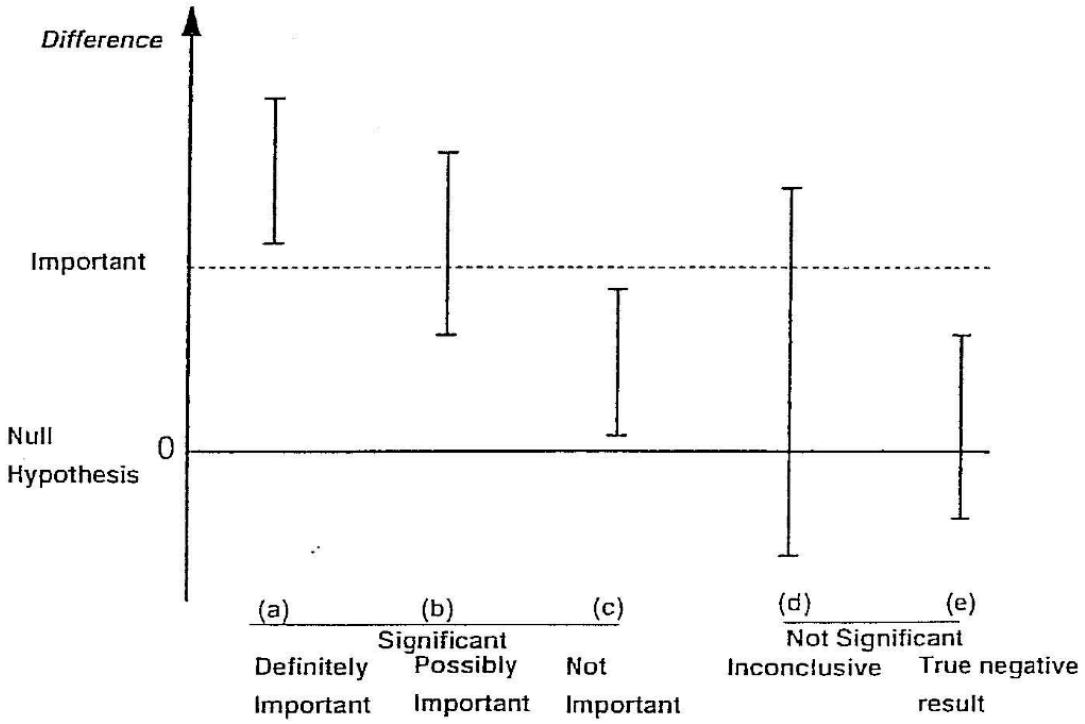



Figure 4.1: Statistical significance vs. clinical importance (Source: Armitage P, Berry G, Matthews JNS. (2001)

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 $\alpha$ (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 $\alpha$.

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 $\beta$ (beta). We will see in Module 10 that $\beta$ 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

**?@tbl-mod4-alpha-beta**. We will return to these concepts in Module 10, when discussing how to determine the appropriate sample size of a study.

## 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:

$$95\% \text{ CI} = \text{Difference in means} \pm 1.96 \times \text{SE(Difference in means)}$$

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.

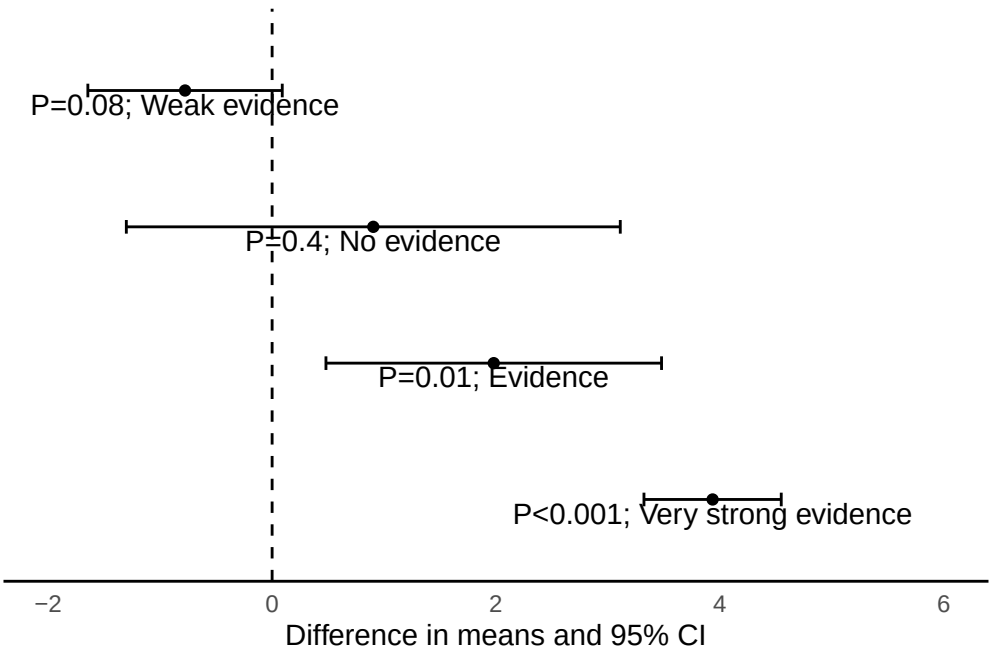



Figure 4.2: Using confidence intervals as informal hypothesis tests

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 4.2: 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 histogram 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 Stata 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 4.3: 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.

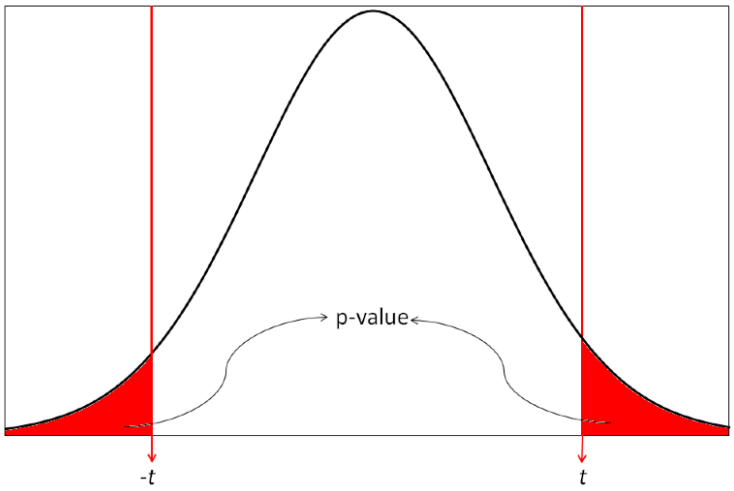



Figure 4.3: P-value for a 2-tailed test

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 occur in one direction only. This makes it easier to reject the null hypothesis because the consequence is that the P value is essentially halved. The P value for a one tailed test would be 0.025 i.e. the shaded region for a one-tailed test would be retained on one side of the distribution and eliminated from the other side of the distribution as shown in Figure 4.4.

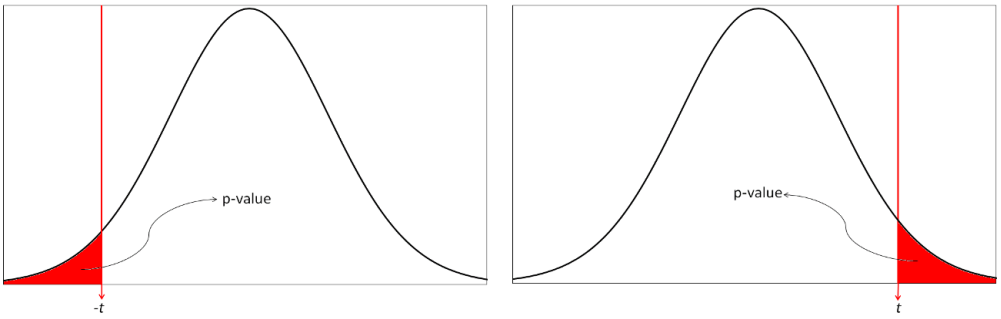



Figure 4.4: P-value for 1-tailed tests

If a one tailed P value is reported, and is considered to be an invalid decision, it is usually easily converted to a two tailed value by doubling its numeric value. For example, for the same test statistic and sample size:

- One tailed P value = 0.042 i.e. statistically significant
- Two tailed P value = 0.084 i.e. non-significant

Obviously, the choice of whether to use a one or two tailed test is not as important when the P value is highly significant or clearly non-significant but can make a difference to the conclusions when the P value is on the margins of significance.

In most health research, the use of a one tailed test is rarely justified because it is unusual to be certain of the direction of effect prior to the research study being undertaken. It has been

suggested that if the researchers were sure enough to consider using a one-tailed test, the research study would not be needed.

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 (including Stata) 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:

$$\begin{aligned} 6.478546e-05 &= 6.478546 \times 10^{-5} \\ &= 6.478546 \times 0.00001 \\ &= 0.00006478546 \end{aligned}$$

As for the Stata output, 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.

# 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]

Acock (2010); Section 7.7, 7.8.

## 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

distribution of the analysis variable should be examined for each group, as shown in Figure 5.1. The datasets `mod05_birthweight.dta` and `mod05_birthweight.rds` are available on Moodle.

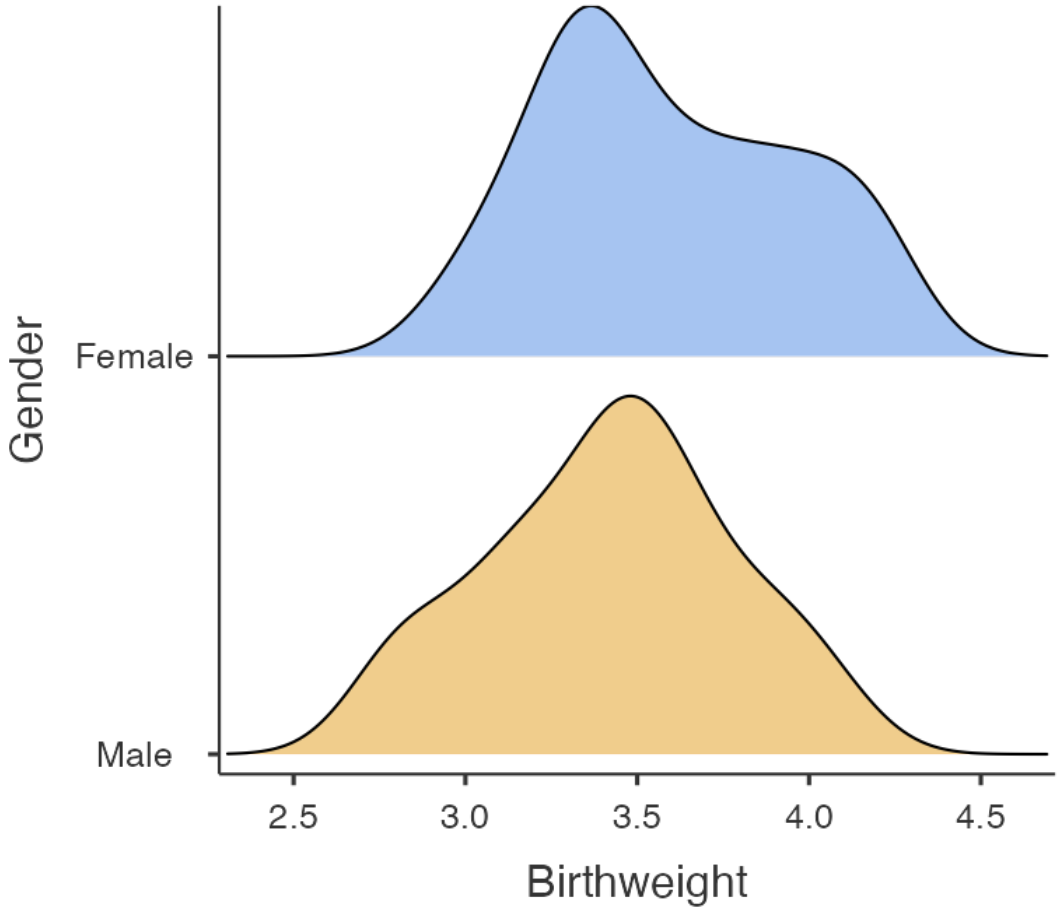



Figure 5.1: Distribution of birth weight by gender

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 | | |
| N Non-missing | 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) |
| Min, Max | 2.95, 4.25 | 2.75, 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.dta` or `mod05_birthweight.rds`. The process of conducting the t-test is summarised for Stata 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.xls 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.

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.

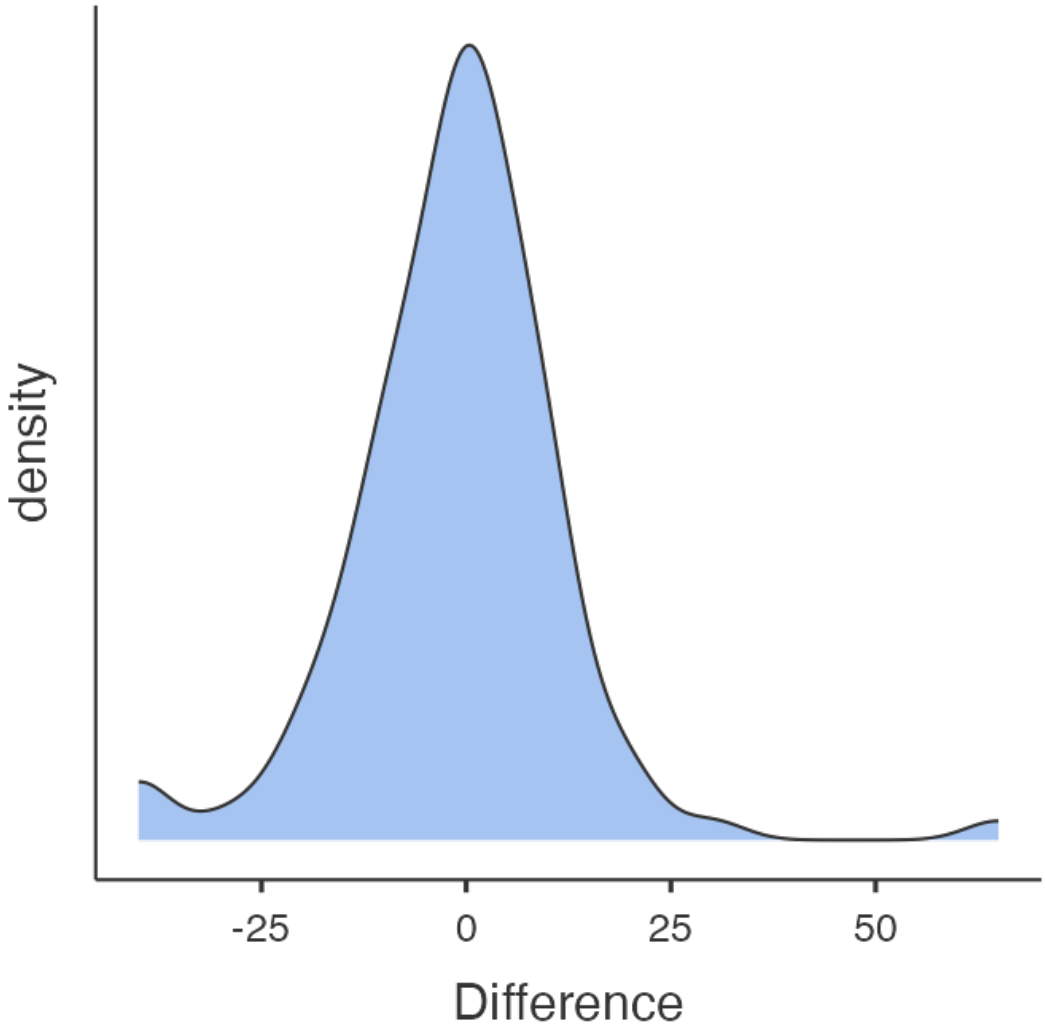



Figure 5.2: Distribution of differences in ankle SBP between two sites of 107 participants

# 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]

Acock (2010); Section 7.5.

## 6.1 Introduction

In Modules 4 and 5, we discussed methods used to test hypotheses when the data are continuous. In Modules 6 and 7, we will focus on hypothesis testing for binary 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 Stata 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.**

## 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%.

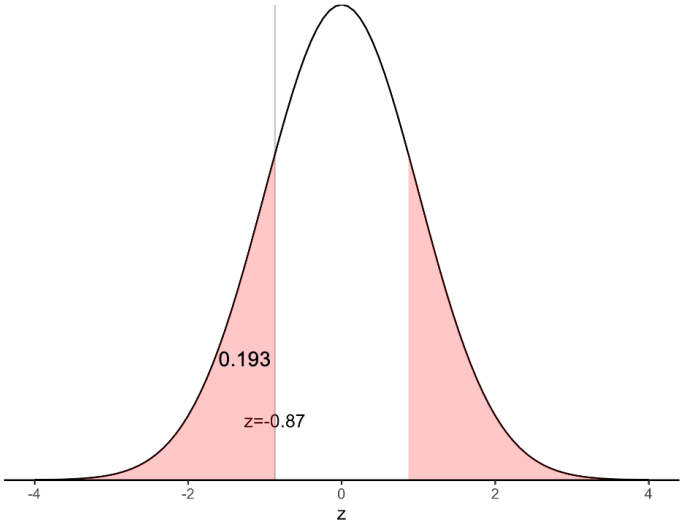
The researchers are interested in whether the proportion of smoking in this community is the same as the population prevalence of smoking of 20%. The null hypothesis can be written as: $H_0$: the proportion of smokers in the community is 20% (the same as in the national census).

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 Table A2.1 in the Appendix), or using the hand-calculator in

Stata. 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.csv` contains the data for this example. In the data file, smokers are coded as 1 and non-smokers are coded as 0.

In Stata, we can use the `prtest` command to perform a z-test, or the `bitest` command to perform the exact binomial test. In R, we can use the `prop.test` function to perform a z-test, or the `binom.test` function to perform the exact 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 is

needed for some of the commands to work (e.g. the epidemiology table commands). This coding ensures that measures of association, such as the odds ratio or relative risk, are computed correctly by Stata. While R does not 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.

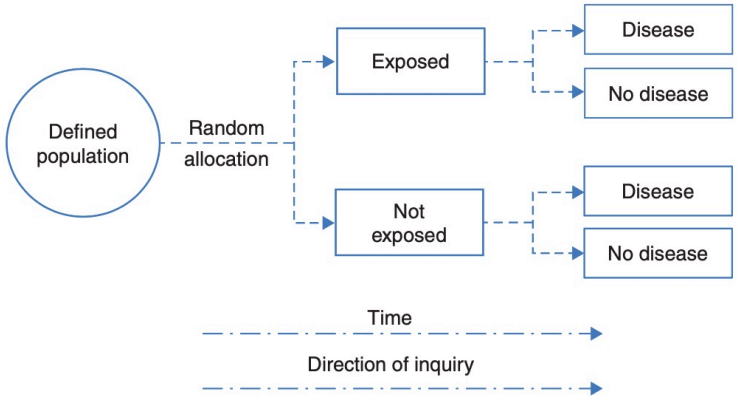



Figure 6.1: The design of a randomised controlled trial [Figure 4.1, Essential Epidemiology]

### 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.

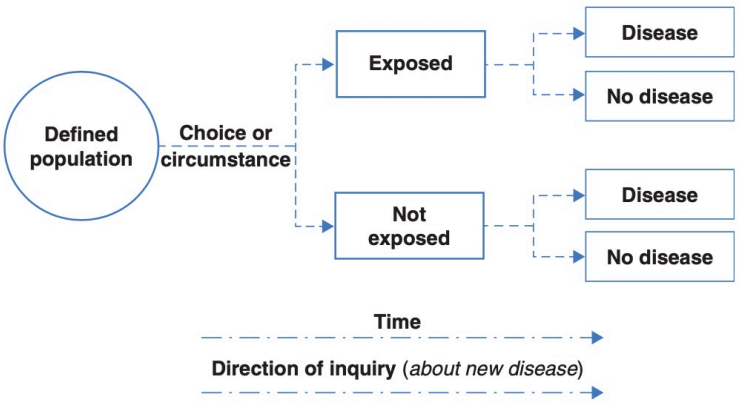



Figure 6.2: The design of a cohort study [Figure 4.2, Essential Epidemiology]

### 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.

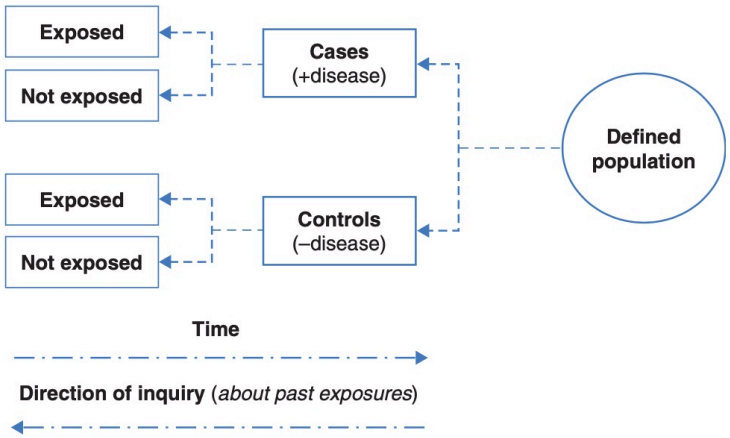



Figure 6.3: The design of a case-control trial [Figure 4.3, Essential Epidemiology]

### 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.

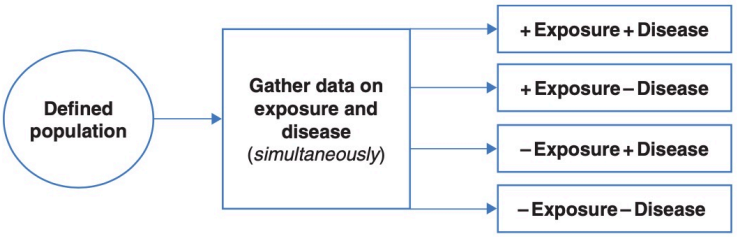



Figure 6.4: The design of a cross-sectional study [Figure 4.4, Essential Epidemiology]

## 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 files `mod06_nausea.dta` and `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 Stata 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 Stata or R to estimate the odds ratio and its 95% confidence interval. We should use the Cornfield option in Stata to provide a better estimate of the 95% confidence interval. It appears that the `jmv` package in R does not use the Cornfield approximation to estimate the 95% confidence interval, but uses the Woolf method.

The odds ratio is estimated as 17.6, and its 95% confidence interval is estimated 9.0 to 34.3 (Cornfield, using Stata) or 9.0 to 34.5 (Woolf, using R).

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.

# 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]

Acock (2010); Section 7.6.

## 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 Stata 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 ($\chi^2$ = 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.

Table 7.3: Allergy data

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 |

### Worked Example 7.2

The files `mod07_allergy.dta` and `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 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 datasets `mod07_drug_response.dta` and `mod07_drug_response.rds` are 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: 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.

# 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]

Acock (2010); Chapter 8.

## 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.

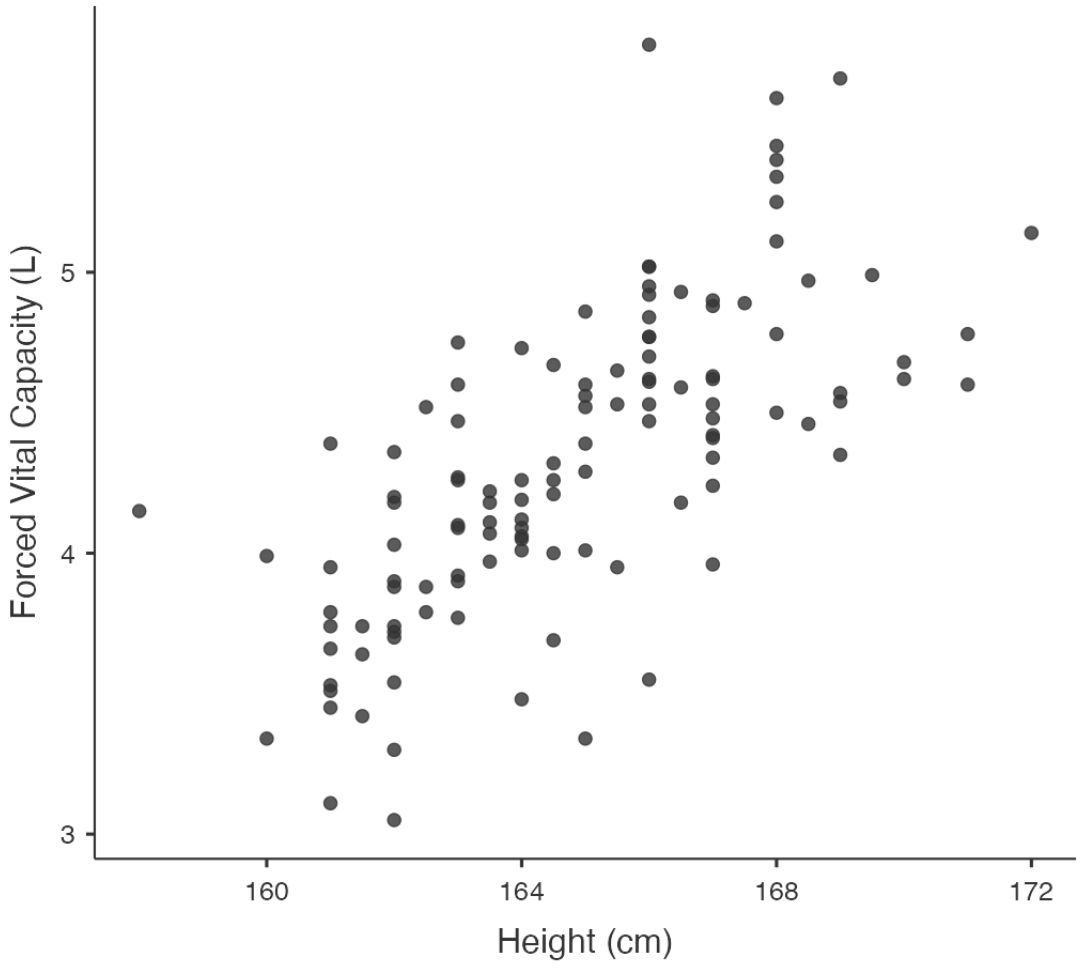



Figure 8.1: Association between height and lung function in 120 adults

### 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.

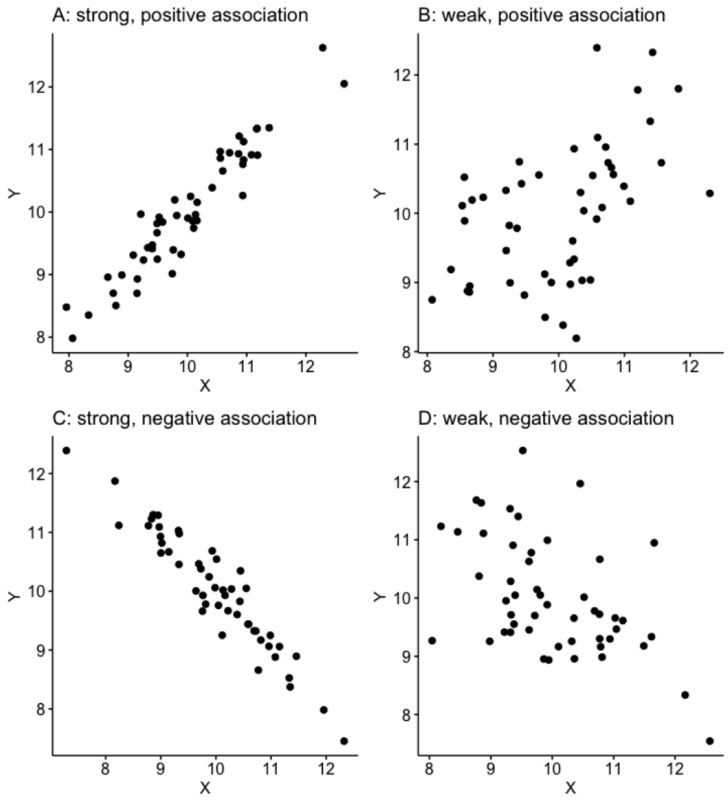



Figure 8.2: Scatter plots demonstrating strong and weak, positive and negative associations

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:

- 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: *rho* | Bivariate normality cannot be assumed. Also useful when at least one of the variables is ordinal |

Spearman's $\rho$ 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<br>P < 0.0001 |
| FVC | 0.70<br>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.

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.

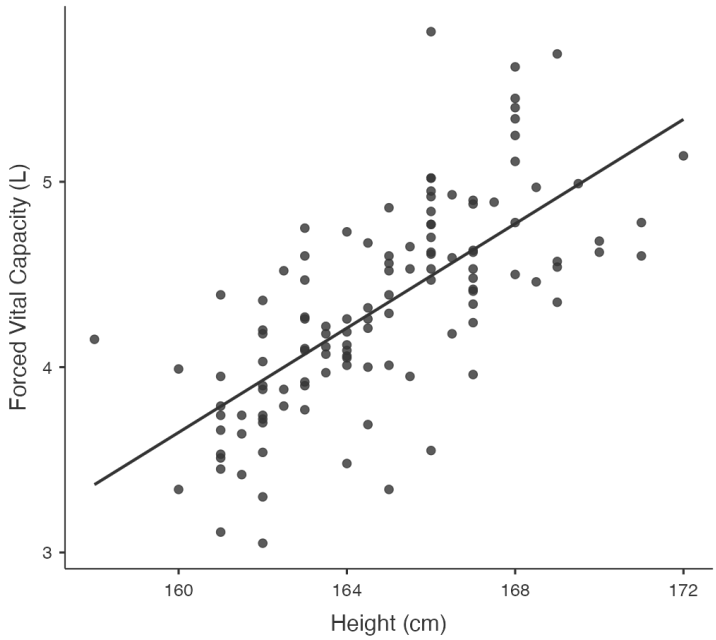



Figure 8.3: Association between height and lung function in 120 adults

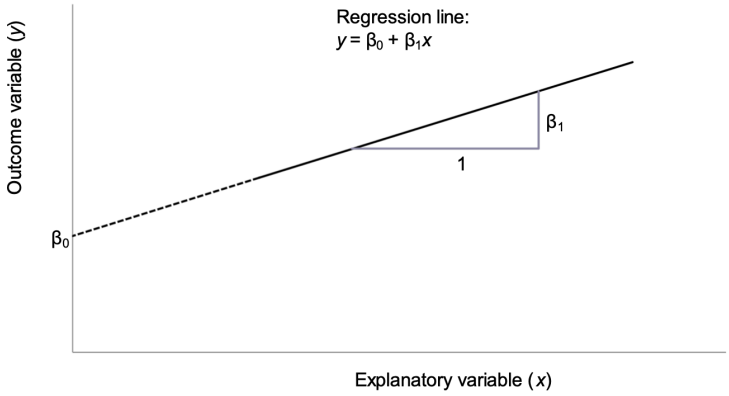



Figure 8.4: Coefficients of a linear regression equation

Table 8.3: Notation for linear regression equation

| Symbol | Interpretation |
|---|---|
| $y$ | The outcome variable |
| $x$ | The explanatory variable |
| $\beta_0$ | Intercept of the regression line |
| $\beta_1$ | 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 $\beta_0$ and $\beta_1$ are true, unknown quantities (similar to $\mu$ and $\sigma$), 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 $\beta_1$, 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.

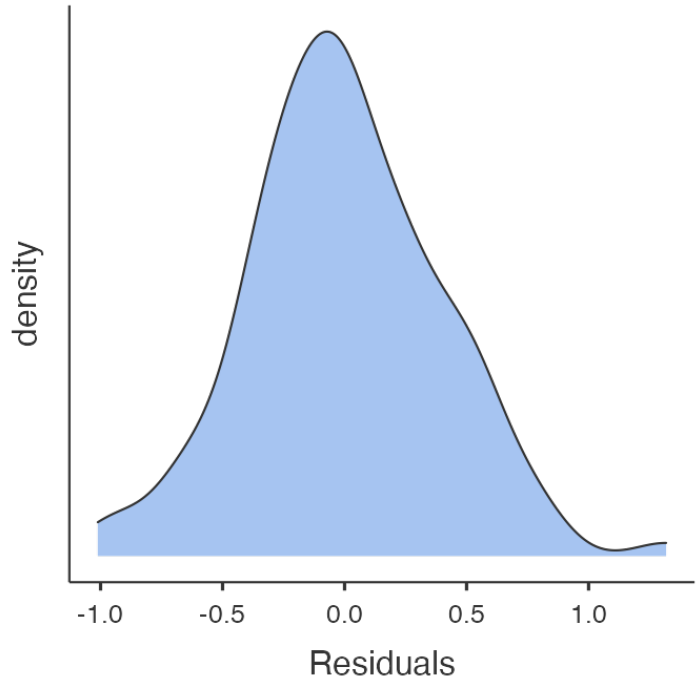



Figure 8.5: Distribution of regression residuals

## 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**.

–>

# 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]

Acock (2010); Section 7.11.

## 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.dta` and `mod09_infection.rds`). The distribution of the length of stay for these patients is shown in the histogram 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

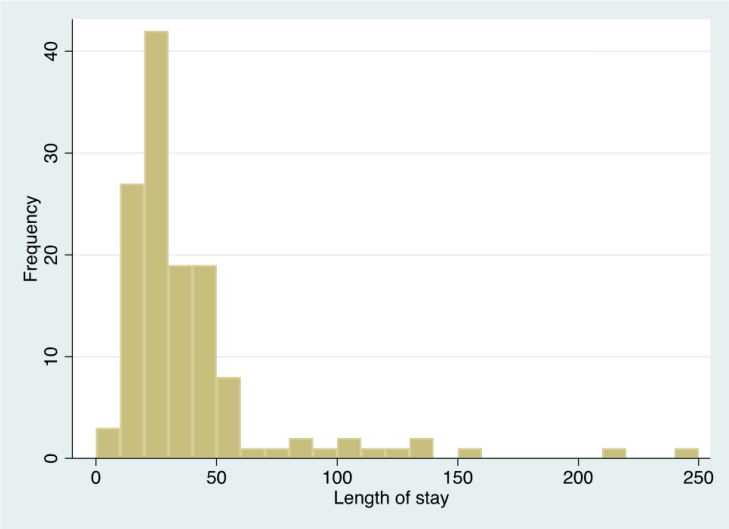



Figure 9.1: Length of hospital stay for 132 patients

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 Stata 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.

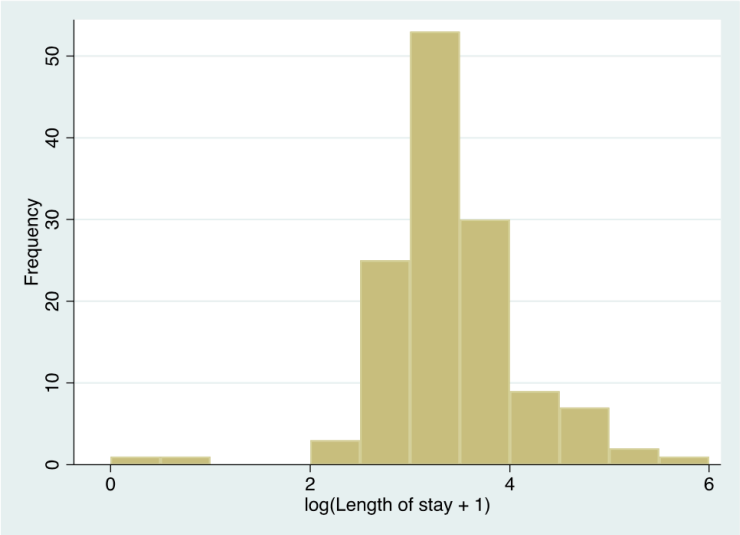



Figure 9.2: Distribution of log transformed (length of stay + 1)

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 |

| | Length of stay | log(Length of stay + 1) |
|---|---|---|

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=No | 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 `ranksum` command in Stata, 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 plot a histogram as shown in Figure 9.3.

The histogram shows that the differences are not normally distributed. The data looks negatively skewed with a gap in the histogram between the values of -5 and 0. Therefore, it

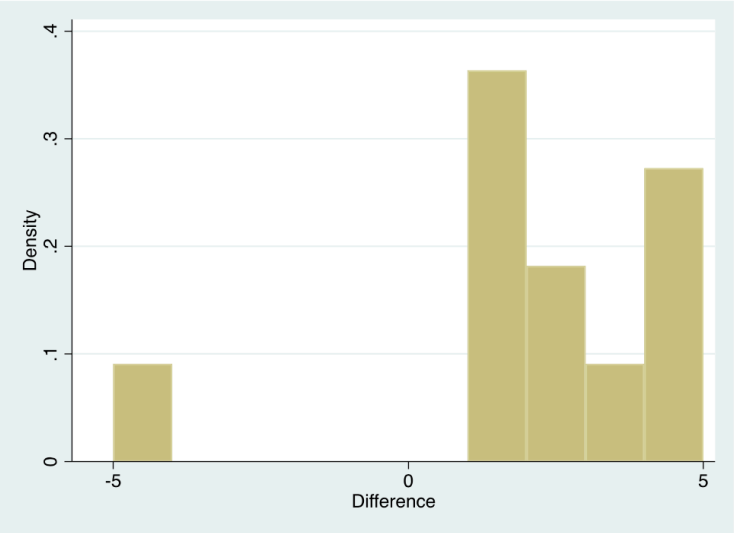



Figure 9.3: Distribution of difference in symptom scores between Drug 1 and Drug 2

would not be appropriate to conduct a paired t-test. Hence, we conduct a non-parametric paired test (Wilcoxon matched-pairs signed-rank test).

A non-parametric paired test can be obtained in Stata using the `signrank` command, or in R using the `wilcox.test` (specifying `paired=TRUE`).

The P-value obtained from Stata is 0.044, and the P-value obtained from R is 0.049. The reason these two P-values differ is that R uses a so-called "continuity correction" which makes the test more conservative (i.e. makes the test less likely to reject the null-hypothesis when the null-hypothesis is true). In both cases, 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 ($\rho$ or rho), and Kendall's rank correlation ($\tau$ 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 $\rho$ is older than Kendall's $\tau$, and can be thought of as a simple analogue of the product moment correlation coefficient, Pearson's r. Kendall's $\tau$ 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 $\rho$ is greater than that of $\tau$. It is not possible to calculate $\tau$ from $\rho$ or $\rho$ from $\tau$, they measure different sorts of correlation. $\rho$ gives more weight to reversals of order when data are far apart in rank than when there is a reversal close together in rank, $\tau$ 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.dta` or `mod08_lung_function.rds`, 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.

# 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

| Prevalence | 1% | 1.5% | 2% | 2.5% | 3% | 3.5% | 4% | 5% |
|---|---|---|---|---|---|---|---|---|
| 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 |
| 20% or 80% | 6,147 | 2,732 | 1,537 | 984 | 683 | 502 | 385 | 246 |
| 25% or 75% | 7,203 | 3,202 | 1,801 | 1,153 | 801 | 588 | 451 | 289 |

### Worked Example

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 $\pm$ 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 are the standard method of calculating sample sizes for these types of study, and examples from both R and Stata will be provided.

## 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 R and Stata as shown in the Worked Example below.

### Worked Example

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 Stata can be calculated using the `power twomeans` command:

### Output 10.1: Two independent samples t-test sample size calculation

```
. power twomeans 110 115, sd(25) power(0.9)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1 versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.9000
```

```
        delta =    5.0000
           m1 =  110.0000
           m2 =  115.0000
           sd =   25.0000

Estimated sample sizes:

            N =     1,054
  N per group =       527
```

We can use the `epi.sscompc` function within the `epiR` package in R to calculate the sample size:

```
library(epiR)
library(pwr)

epi.sscompc(treat=110, control=115, n=NA, sigma=25, power=0.9)


$n.total
[1] 1052

$n.treat
[1] 526

$n.control
[1] 526

$power
[1] 0.9

$delta
[1] 5
```

Note that Stata and R provide slightly different estimated sample sizes. This difference is immaterial from a practical point of view, and highlights the importance of referencing which software package has been used when writing up results.

From the output, we can see that with 90% power we will need 526 or 527 participants in each group, i.e., 1052 or 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 obervations. If we do not have any estimates for this correlation, we can assume a value of 0. If the correlation is positive, a zero for correlation would give a more conservative estimate of sample size required (i.e. estimate a sample size larger than necessary). While a negative correlation would require a bigger sample size than a zero correlation, it is relatively uncommon to encounter negative correlations between pairs. Any discussions on the effect of correlation on sample size is beyond the scope of this course. Thus, we will always assume a correlation of zero between paired measurements in this course.

We can compute the required sample size in Stata using the `power pairedmeans` command:

**Output 10.2: Paired samples t-test sample size using Worked Example 10.2**

```
. power pairedmeans 110 115, corr(0) power(0.9) sd(25)

Performing iteration ...
```

```
Estimated sample size for a two-sample paired-means test
Paired t test assuming sd1 = sd2 = sd
Ho: d = d0  versus  Ha: d != d0

Study parameters:

        alpha =    0.0500          ma1 =  110.0000
        power =    0.9000          ma2 =  115.0000
        delta =    0.1414           sd =   25.0000
           d0 =    0.0000         corr =    0.0000
           da =    5.0000
         sd_d =   35.3553

Estimated sample size:

            N =       528
```

As discussed in the R notes, calculating the sample size required for a paired t-test is a little more cumbersome in R. Here, only the output of the process is provided - refer to the R notes for detail on the code.

**Output 10.2: Paired samples t-test sample size using Worked Example 10.2**

```
     Paired t test power calculation

              n = 527.2954
              d = 0.1414214
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number of *pairs*
```

Assuming a correlation of 0 between the two sets of measurements, we can see that we will need 528 pairs of measurements to achieve a power of 90% (virtually the same as for an independent samples study).

## 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 power twoproportions command in Stata.

### Worked Example

If we expect that the prevalence of smoking in two comparison groups (e.g. males and females) will be 35% and 20%. The sample size required in each group to show that the prevalences are significantly different at P<0.05 with 80% power is shown in Output 10.3.

### Output 10.3: Sample size calculation for two independent proportions

```
Estimated sample sizes for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =   -0.1500  (difference)
           p1 =    0.3500
           p2 =    0.2000

Estimated sample sizes:

            N =       276
  N per group =       138
```

From Output 10.3, we see that we would need 138 males and 138 females (i.e. a total sample size of 276 participants).

What sample size would be required if the prevalence of smoking among men was 30%?

Answer = 294 men and 294 women would be needed.

[Command: `power twoproportions .3 .2, test(chi2)`]

To do the same problem using R, we use the `epi.sscohortc` function. We need to specify the risk of the the outcome each group (labelled group 0 and 1) and the desired power:

```
epi.sscohortc(irexp1=0.35, irexp0=0.2, n=NA, power=0.8)


$n.total
[1] 276

$n.exp1
[1] 138

$n.exp0
[1] 138

$power
[1] 0.8

$irr
[1] 1.75

$or
[1] 2.153846
```

## 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

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.

We can use the Stata command `power twoproportions` as below:

### Output 10.4: Sample size calculation for relative risk

```
Estimated sample sizes for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
        power =    0.9000
        delta =    1.5000  (relative risk)
           p1 =    0.5000
           p2 =    0.7500
        rrisk =    1.5000

Estimated sample sizes:

            N =       154
  N per group =        77
```

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.

The `epiR` package does not have a function to estimate sample size and power directly for a relative risk, but we can use the `epi.sscohortc` function. To do this, we recognise that the assumed rate in the exposed group will equal the rate in the unexposed group multiplied by the relative risk.

Here we define the risk in the unexposed group as 0.5 and the desired relative risk to detect is 1.5. So we specify `irexp0 = 0.5` and `irexp1 = 1.5 * 0.5`:

```
epi.sscohortc(irexp1=1.5*0.5, irexp0=0.5, n=NA, power=0.9)


$n.total
[1] 154

$n.exp1
[1] 77

$n.exp0
[1] 77

$power
[1] 0.9

$irr
[1] 1.5

$or
[1] 3
```

## 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

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.

```
. power twoproportions .3, test(chi2) oratio(2) power(0.9)
Estimated sample sizes for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
        power =    0.9000
        delta =    2.0000  (odds ratio)
           p1 =    0.3000
           p2 =    0.4615
   odds ratio =    2.0000

Estimated sample sizes:

            N =       376
  N per group =       188
```

We can use the `epi.sscc` function in R to calculate a sample size based on an odds ratio in a case-control study:

```
epi.sscc(OR=2, p0=0.3, n=NA, power=0.9)


$n.total
[1] 376

$n.case
[1] 188

$n.control
[1] 188

$power
[1] 0.9

$OR
[1] 2
```

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 in either Stata or R (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 @ref(wex10-5): 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) (corresponding to a ratio of N2/N1 of 0.5) 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.

| Controls per case | Allocation ratio (N2/N1) | Number of cases required | Number of controls required | Total participants required |
|---|---|---|---|---|
| 1 | 1.0000 | 188 | 188 | 376 |
| 2 | 0.5000 | 140 | 280 | 420 |
| 3 | 0.3333 | 124 | 371 | 495 |
| 4 | 0.2500 | 116 | 462 | 578 |
| 5 | 0.2000 | 111 | 553 | 664 |
| 6 | 0.1666 | 108 | 644 | 752 |
| 7 | 0.1429 | 105 | 734 | 839 |
| 8 | 0.1250 | 104 | 825 | 929 |
| 9 | 0.1111 | 102 | 916 | 1,018 |
| 10 | 0.1000 | 101 | 1,006 | 1,107 |

This can be visualised graphically, as in Figure 10.1.

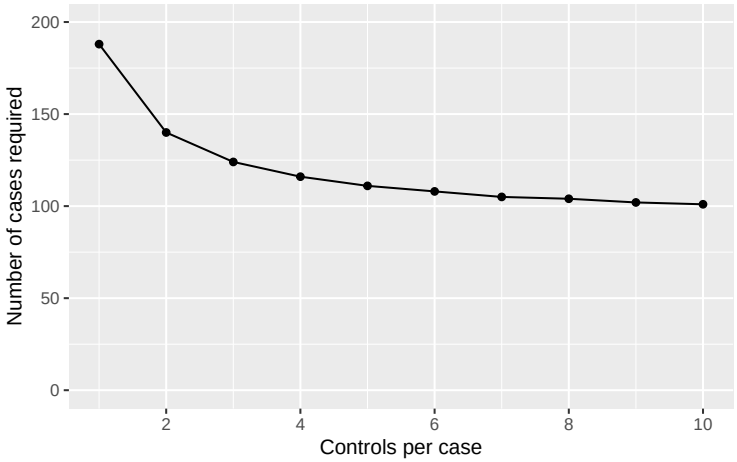



Figure 10.1: Increasing the number of controls per case

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 8 (124 vs 116 cases), but at an increase of 91 (371 vs 462) controls. Clearly, this reduction in cases is not offset by the extra controls required.

In Stata, the allocation ratio is the number in the experimental group / the number in the control group. There is some inconsistency in R, which as of May 2023, leads to inconsistent results between Stata and R for case-control studies with unequal group sizes.

| | Stata | R |
|---|---|---|
| Difference in means | ratio = n-exposed / n-controls | ratio = n-exposed / n-unexposed |
| Difference in proportions | ratio = n-exposed / n-controls | ratio = n-exposed / n-unexposed |
| Relative risk | ratio = n-exposed / n-controls | ratio = n-exposed / n-unexposed |
| Odds ratio | ratio = n-cases / n-controls | ratio = n-controls / n-cases |

## 10.9 Limitations in sample size estimations

In this module we have seen how to use Stata for estimating 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 underway, 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 Stata 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.

# References


Acock, Alan C. 2010. *A Gentle Introduction to Stata*. 3rd ed. College Station, Tex: Stata Press.

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.

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.

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.