PHCM9795: Foundations of Biostatistics

Timothy Dobbins

18 May, 2023

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.

Changelog

2023-07-17

[Changed]

- Section 7.9: Corrected screenshots to test difference in paired proportions in Stata.

2023-07-13

[Changed]

- Section 7.13: tidied up the R function used to calculate the 95% confidence interval for the difference in paired proportions.

2023-07-12

[Changed]

- Worked Example 6.2: removed "This z-statistic does not meet or exceed the critical value of 1.96 for a two tailed test" and re-framed this in terms of interpreting the P-value as calculated from software.
- Worked Example 7.1: corrected column headings for Nausea and No nausea

2023-07-01

[Changed]

- Section 4.2: Fixed typo: "The particular test statistic differs depending on the type of data being ~~analyses~~ analysed"
- Section 4.4: Fixed typo: "This is the situation described in scenario (c) of ~~Figure~~ Figure 4.1."
- Activity 5.2: Added "or R" to the instruction "Use Stata to conduct an appropriate statistical test"
- Activity 5.3: Added "or R" to the instruction "Use Stata to conduct an appropriate statistical test"
- Section 6.3: Fixed formula for testing one sample proportion to: "$z = \frac{(p_{sample} - p_{population})}{\text{SE}(p_{population})}$"
- Activity 7.2: Added "or R" to the instruction "Using Stata, carry out the appropriate significance test"

2023-07-01

[Changed]

- Renamed "Readings" to "Optional readings"
- Module 4, Section 4.8: Corrected the sentence that describes Figure 4.4. "the shaded region for a one-tailed test would be ~~doubled~~ retained on one side of the distribution and eliminated from the other side of the distribution".
- Module 9: Added titles to Tables 9.3 and 9.4.

- Module 9, Section 9.4: Corrected the level of evidence: "providing ~~strong~~ evidence of a difference in the median length of stay between the groups."
- Module 9, Section 9.5.1. Corrected the text under Table 9.4: "The data shows … 10 people who have a ~~negative~~ positive difference."

2023-06-13

[Changed]

- Module 3. Clarified Worked Example 3.1, and moved the example from Section 3.5.1 to Section 3.5.2.
- Section 3.6: Added Stata output for calculating a 95% confidence interval from individual data.

2023-06-01

[Changed]

- Section 1.14.3: RStudio preferences are now located at **Edit > Settings** on MacOS, and **Tools > Global Options** on Windows
- Section 1.14.7.2: Correct layout for commands to install packages (commands must be entered on separate lines)
- Section 1.15.2: Correct the name of the pbc data set to mod_01_pdc.rds
- Activity 2.3(b): Change the request to plot data on a Normal curve, not a standardised Normal curve
- Activity 5.3 and 5.4: added underscores to filenames

Module 1

To do

- Label subheadings consistently between HTML and PDF (e.g. 3 levels deep, like 1.1.1)

Module	Revise main notes	Revise Stata notes	Revise R notes	Revise activities	Revise solutions	Review
1						
2	Y	Y	Y			
3	Y	Y	Y			
4	Y	Y	Y			
5	Y	Y	Y			
6	Y	Y	Y			
7	Y	Y	Y			
8	Y	Y	Y			
9	Y	Y	Y			
10	Y	Y	Y			

Module 2

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]

2.1 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, together with theoretical/pure mathematics and applied mathematics. Within the field of statistics, there are two main divisions: 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.

2.2 Descriptive and inferential statistics

To understand the concepts of statistics, it is important to realise there are two ways of using data: one is via descriptive statistics and the other is via inferential statistics.

Descriptive statistics

Descriptive statistics provide a 'picture' of the characteristics of a population. Examples of descriptive statistics based on the population are given below.

Births

These examples on descriptive statistics consider all the births in Australia in 2019 (Australian Institute of Health and Welfare (2021)). The Australian Institute of Health and Welfare produce comprehensive reports annually on the characteristics of Australia's mothers and babies of the most recent year of data from the National Perinatal Data Collection.

One headline from the report is "Nearly two thirds of mothers were aged between 25 and 34 years (185,958 women in 2019)", which is accompanied by a figure illustrating the age distribution of mothers in Australia giving birth in 2019. This example shows descriptive statistics that are presented as the actual number of women giving birth in 2019, together with a comparison of age distributions across Australian states.

Further descriptive statistics provide summary information, about the average (mean) age of women giving birth in 2019 (30.8 years) and the proportion of mothers who were Indigenous (4.8%).

Deaths

In another example, consider characteristics of all the deaths in Australia in 2020 (Australian Bureau of Statistics (2021)).

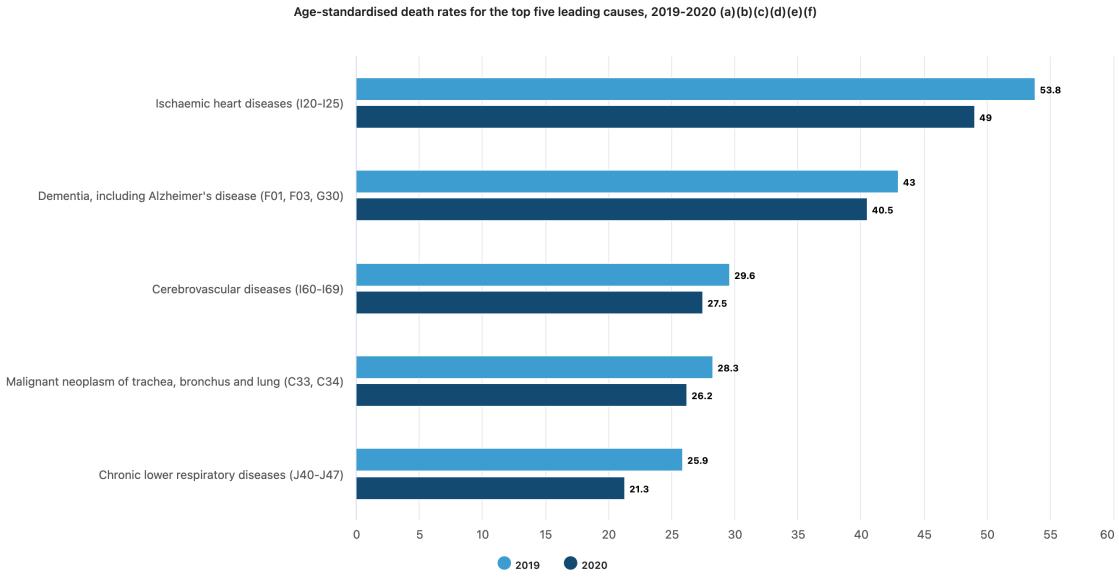
> "During the pandemic many countries saw a change in mortality patterns, including COVID-19 becoming a leading cause of death."

The report presents the leading causes of death in 2020, comparing the age-standardised rates between 2019 and 2020:

- "The top five leading causes of death remained the same as in 2019 (Ischaemic heart disease, Dementia including Alzheimer's disease, Cerebrovascular diseases, Lung cancer and Chronic lower respiratory diseases).

- The age-standardised death rate decreased for all top five leading causes of death from 2019.
- Deaths due to chronic lower respiratory diseases (including emphysema) had the highest proportional rate decrease from 2019 at 17.8%.
- The reduction in acute respiratory conditions such as pneumonia contributed to a decrease in the top five leading causes of death.
- All top five leading causes of death are non-communicable diseases (they are not passed from person to person).”

The information was also presented as a visualisation / infographic, demonstrating a simplistic, yet valuable way of presenting data and enabling rates of death for 2019 and 2020.



a. Causes listed are based on the WHO recommended tabulation of leading causes. See Mortality tabulations and methodologies for further information.
b. See the Data quality section of the methodology for further information on specific issues related to interpreting time-series and 2020 data.
c. Age-standardised death rate. Death rate per 100,000 estimated resident population as at 30 June (mid year). See the glossary and the Mortality tabulations and methodologies section for further information.
d. All causes of death data from 2006 onward are subject to a revisions process - once data for a reference year are 'final', they are no longer revised. Affected data in this table are: 2011 - 2017 (final), 2018 (revised), 2019 and 2020 (preliminary). See the Data quality section of the methodology and Causes of Death Revisions, 2017 Final Data (Technical Note) and 2018 Revised Data (Technical Note) in Causes of Death, Australia, 2019.
e. Changes in coding processes have been applied to 2020 data. See the Classifications and Mortality coding sections of the methodology for further information.
f. Victorian coroner-referred deaths for Ischaemic heart disease (I20-I25) have been measured on registration year to enable more accurate time-series analysis. Care needs to be taken when interpreting data derived from Victorian coroner-referred deaths. See Technical note: Victorian additional registrations and time series adjustments in Causes of death, Australia, 2019, for detailed information on this issue.

Source: Australian Bureau of Statistics, Causes of Death, Australia 2020

Figure 2.1: Age-standardised death rates for the top five leading causes, 2019-2020

Inferential statistics

Inferential statistics use data collected from a sample of the population, to make conclusions (inferences) about the whole population (that the sample was drawn from).

The following example is about a sample of prisoners, from the National Prisoner Health Data Collection (NPHDC). The NPHDC is the main source of national data about the health of prisoners in Australia. It gathers information over a 2-week period from prison entrants, dischargees, prisoners visiting the prison health clinic, and prisoners taking prescribed medication.

We have information about the population of prisoners, given as the number of prisoners in Australia’s prisons from the ABS website:

> At 30 June 2015: There were 36,134 prisoners in Australian prisons, an increase of 7% (2,345 prisoners) from 30 June 2014.

Characteristics (sex, age group and Indigenous status) of the sample of prisoners from the NPHDC are given in the following table. We can use this information to make inferences about the whole population of prisoners that the sample was drawn from.

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

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

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

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

Notes

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

Figure 2.2: Characteristics of the sample of prisoners from the NPHDC

2.3 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 varaibles

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.

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

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

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

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

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

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 varaibles 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.

2.4 Presenting data

We will now look at ways to summarise and present data. The choice of presentation will depend on the type of variable being summarised. We will use the dataset based on a study into primary biliary cholangitis (PBC) from the Introduction to Stata or Introduction to R exercise to demonstrate the appropriate ways to summarise and present data.

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).

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

Table 2.2: Stage of disease for participants in PBC study

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

* Disease stage was missing for 6 participants

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 2.3). Bar graphs, which are suitable for plotting discrete or categorical variables, are defined by the fact that the bars do not touch.

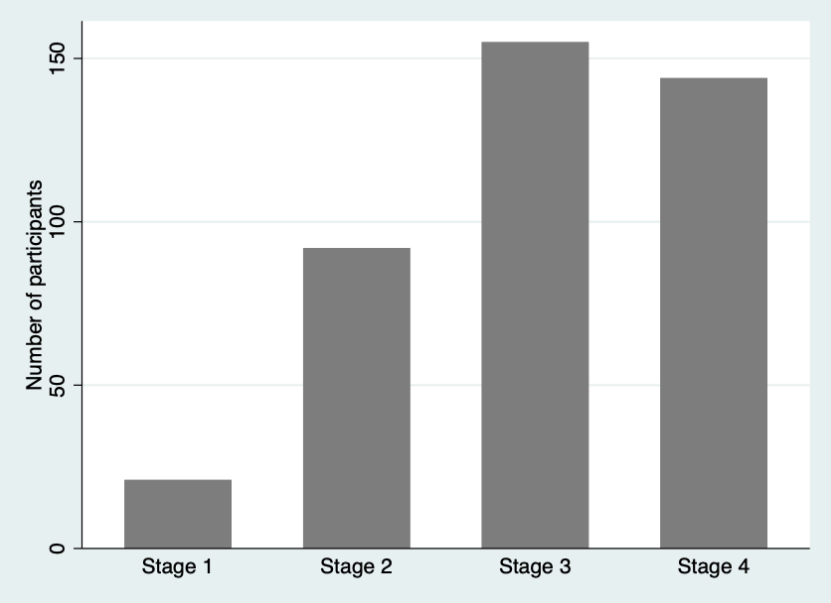


Figure 2.3: Bar graph of stage of disease from PBC study

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 a single continuous variable numerically

2.5 Summary statistics and variation in data

We often collect measurements that are continuous in nature, that is measurements such as height, weight, time, blood pressure etc which can be measured accurately to one or more decimal places. A useful way of describing the continuous data is via summary statistics. These include measures to describe the distribution of data points via the central tendency (e.g. mean, median) and spread (e.g. standard deviation and inter quartile range). We also examine the data visually by graphing it using histograms and box plots.

Mathematical and statistical notation

When computing summary statistics or using more formal statistical methods, mathematical and statistical notation is often used. Below are some of the common statistical terms and interpretation that will be used in the course and which are seen in many text books.

Notation	Interpretation
x	An observation in your sample
$\sum x$	Sum of all the observations
N	Total population size
n	Sample size
μ (mu)	Population mean
σ^2	Population variance
σ	Population standard deviation
$\bar{x}$	Sample mean
s^2	Sample variance
s	Sample standard deviation

Measures of central tendency

Worked example

In our random sample of 30 students attending a university gym on a given day, their weight in kilograms was measured (see below). Weight is a continuous measurement (similar to height, blood pressure etc) that in theory can be measured to infinitely small units, though in practice they can be measured accurately to one or two decimal places.

We will use these data to look at measures of central tendency and spread of the data and other summary statistics.

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

Mean

The most commonly used measure of the central tendency of the data is the mean value. The mean of a set of values is often referred to as the average of all the values. The mean ($\bar{x}$) of a sample dataset is calculated using the following formula:

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

From the weights example: $\bar{x}$ = 2100/30 = 70.0. Thus, the mean weight of this sample is 70.0 kg

Median and mode

Other measures of central tendency include the median and mode. The median is the true centre 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 the weight example, the median would be 70.0 kg.

For a set of eight exam results ranked in order:

48 51 55 59 63 64 69 75

The median is the average of the two middle observations: 59 and 63. So the median is (59+63)/2 = 61

The mode is the most frequent value in the distribution, in the weight example this would be 70.0 kg as this value features most frequently. The mode is not used frequently.

Describing the spread of the data

In addition to measuring the centre of the data, we also need a robust estimate of the 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 weights example, Range = 80.0 - 60.0 = 20.0 kg

Note that while the range is 20.0 kg, the range is often reported as the actual lowest and highest values e.g. Range 60 to 80 kg.

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 (adapted from: Williams, 2015).

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 worked weights example, we would calculate the sample variance:

$$\begin{aligned} s^2 &= \frac{(60.0 - 70.0)^2 + (62.5 - 70.0)^2 + \cdots + (80.0 - 70.0)^2}{30 - 1} \\ &= \frac{737.5}{29} \\ &= 25.43 \text{ kg}^2 \end{aligned}$$

with a sample standard deviation: $s = \sqrt{25.43} = 5.04$ kg.

Thus, in our sample of 30 students, we have an estimated mean weight of 70.0 kg, with a variance of 25.43 kg^2 and a standard deviation of 5.04 kg.

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 around the median i.e. the bottom 25% and top 25% of values are discarded and only the values in the 25%-75% range are quoted. The IQR is the preferred measure of spread when the median has been used to describe central tendency.

In the weights example the IQR would be 67.5 – 75.0 (i.e. the middle 50% of values).

2.6 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 weight measures from 30 people. If we had information on the weight of the total population that the sample was drawn from, we could calculate all the summary statistics described above (for the sample) for the population.

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

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

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

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

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

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

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

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

Worked example

Consider the ages of 30 students visiting a university gym in a particular hour presented in **?@tbl-1-ages**. This information is difficult to interpret in its raw form, but becomes more clear if the ages are grouped in a frequency table as shown in Table 2.4.

Table 2.4: Frequency of ages of students visiting a gym

Age	Frequency
17	1
18	5
19	5
20	7
21	5
22	2
23	4
24	1
Total	30

Tables with more than one variable

So far, we have discussed one-way frequency tables, that is, tables that summarise one variable. We can summarise more than one variable in a table – called a cross tabulation, or a two-way (summarising two variables) table or multi-way (summarising more than two variables) table. However, tables become complex when more than two variables are incorporated (you may need to present the information as two tables or incorporate additional rows and columns).

In our example above, if we have two categorical variables (e.g. sex with two categories male and female and BMI status with three categories Normal, Overweight and Obese) measured on each subject (student), we can classify the two variables simultaneously using two-way tables of frequency as shown in Table 2.5.

Table 2.5: Frequency of students visiting a gym by sex and BMI status*

Sex	Not overweight	Overweight	Obese	Total
Male	1	9	2	12
Female	11	6	0	17
Total	12	15	3	29

*BMI was missing for 1 student

Tables containing more than two variables

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

We might also consider a table containing further pieces of information. The table presented in Figure 2.4 (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 2.4: 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.

Table presentation guidelines (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)

2.7 Graphical presentation

Bar graphs

Information from more than one variable can be presented as clustered or multiple bar chart (bars side-by-side) (Figure 2.5). 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.

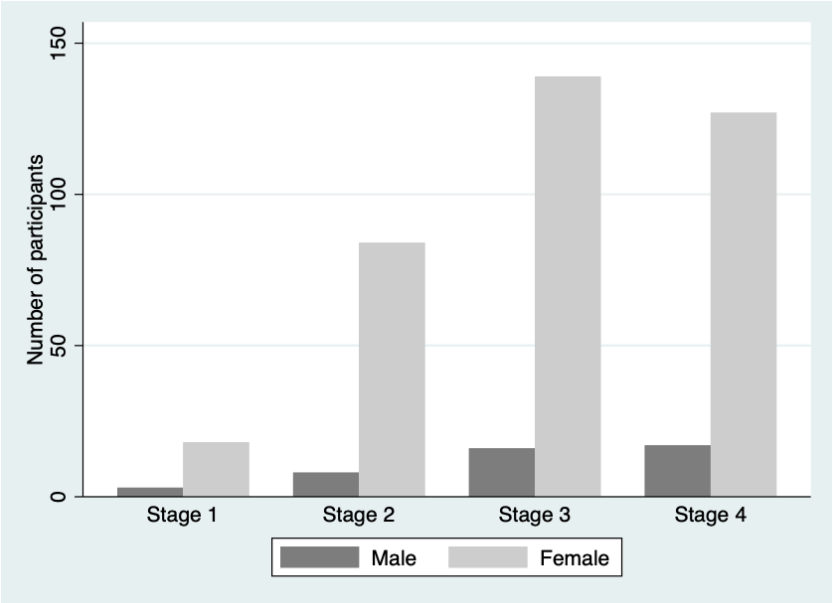


Figure 2.5: Bar graph of stage of disease by sex from PBC study

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 2.6).

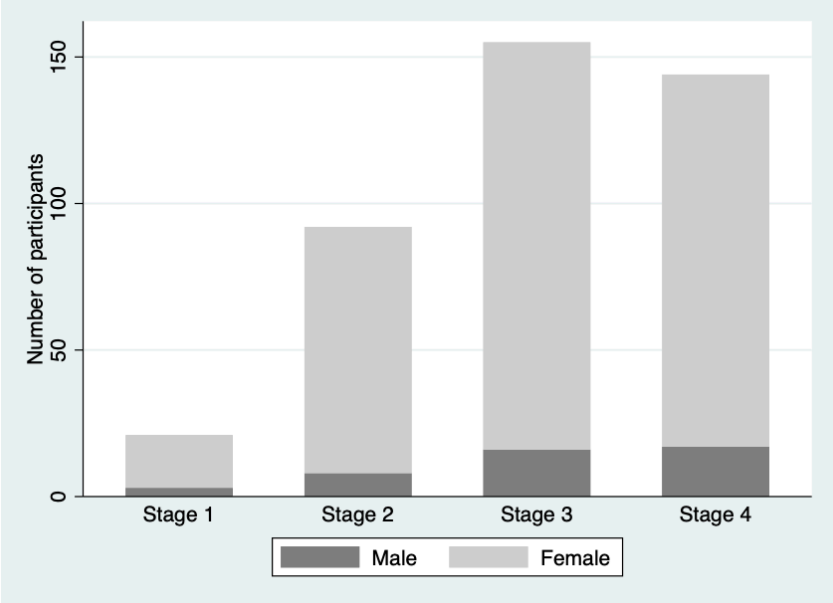


Figure 2.6: Stacked bar graph of stage of disease by sex from PBC study

Finally, a stacked relative bar chart (Figure 2.7) 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 2.7 that the distribution of sex is similar across each stage of disease.

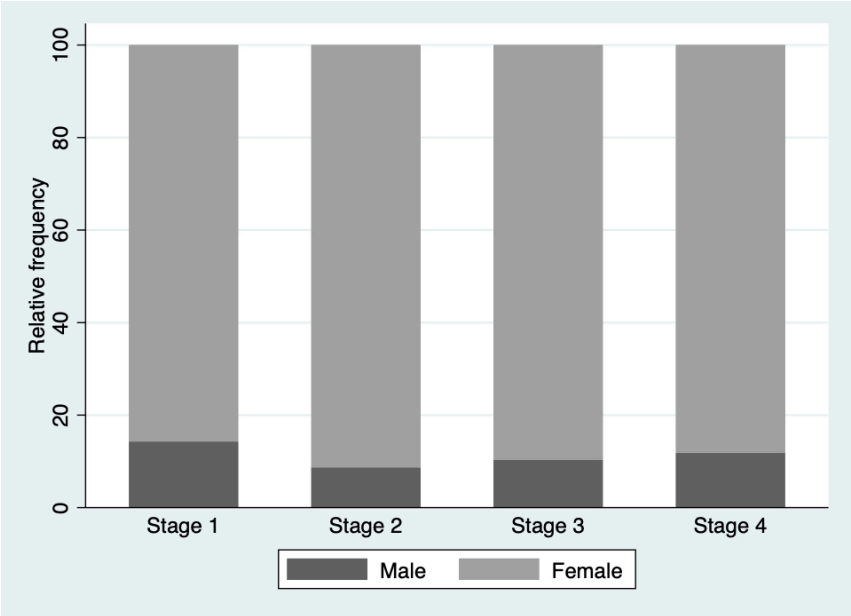


Figure 2.7: Relative frequency of sex within stage of disease from PBC study

Line graphs

A line graph is effective to illustrate trends over time (e.g. change over several years). Let's look at an example from cancer epidemiology.

Cancer incidence is the number of new cases of cancer diagnosed in a population in a given time period. A useful comparison with the incidence rate is the mortality rate, revealing information about the deaths from cancer in the same period. Figure 2.8 shows the prostate cancer trend in the NSW male population in the period 1972-2014, specifically the age-standardised incidence and mortality rate per 100,000.

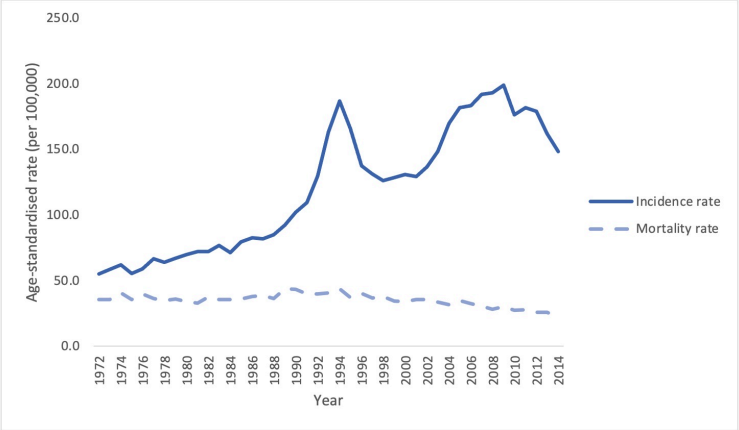


Figure 2.8: Prostate cancer age-standardised incidence and mortality rates (per 100,000), NSW, 1972-2014

Source: The Cancer Institute NSW (2018) Cancer statistics NSW. https://www.cancer.nsw.gov.au/cancer-statistics-nsw (Accessed: 24 Jan 2019).

The age standardised incidence rate for prostate cancer increased steadily in the period 1972 – 1991, from 55.2 cases per 100,00 to 109.3 cases per 100,000. There were two notable peaks in incidence in the period 1972-2014. In particular, there was an increase between 1992-1994, and also between 2002-2009. Since 2009 (to 2014) the rates decreased from 198.9 per 100,000 to 148.2 per 100,000. Whilst the incidence rate for prostate cancer has fluctuated over the period, the age standardised mortality rate remained relatively stable (around 35 deaths per 100,000). Since 2009 the mortality rate appears to be decreasing and was at its lowest in 2014 at 22.1 per 100,000.

[The increase in prostate cancer incidence in the early 1990's occurred at a time when blood testing of men for Prostate Specific Antigen (PSA) became more widespread. The more recent peak in incidence in the early 2000's maybe explained by PSA being increasingly used as a screening test for men who did not have symptoms of prostate cancer.]

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

2.8 Using graphs to display the centre and spread of the data

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

Frequency histograms

A histogram that plots the frequency of the grouped observations is called a frequency histogram. Some features of a frequency histogram:

- The area under each rectangle is proportional to the frequency
- The rectangles are drawn without gaps between them (unlike a bar graph)
- The data are 'binned' into discrete intervals (of (usually of equal width)
- The mid-point of the histogram represents the centre (mean, median) of the data

If the rectangles are symmetrically distributed about the middle of the histogram, we say that the data are symmetric, and the mean and median will be approximately equal.

If the histogram has a longer tail to the right, then the data are said to be positively skewed (or skewed to the right), and the mean will be greater than the median.

If the histogram has an extended tail to the left, then the data are negatively skewed (or skewed to the left) and the mean will be smaller than the median.

> The skewness of a distribution is defined by the location of the longer tail, not the location of the peak of the data.

Figure 2.9 presents two histograms from the PBC data from the Introduction to Stata exercise: for age and serum bilirubin. 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).

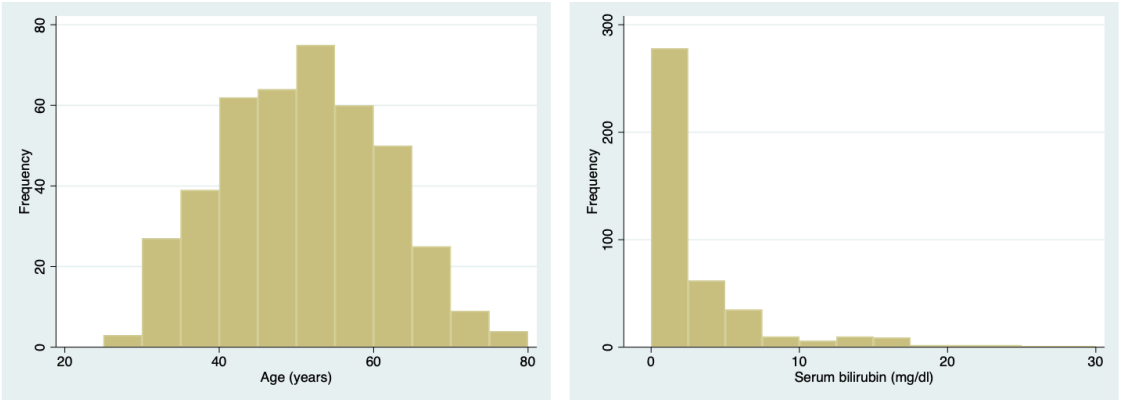


Figure 2.9: Histogram of age (left) and serum bilirubin (right) from PBC study data

Boxplots

Another useful 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.

If the data are symmetric, the line across the box (the median value) will be in the centre of the box, and the tails will be roughly equal.

Figure 2.10 presents two boxplots from the PBC data: for age and serum bilirubin. We can see that the boxplot for age has roughly equal tails, and the median (the horizontal line) lies roughly in the middle of the interquartile range (the shaded box). It would be reasonable to assume that age follows a symmetric distribution from this plot. The boxplot for serum bilirubin shows a much longer upper tail, and a median much closer to the bottom of the shaded box than the middle. The boxplot also shows a number of points above the 75th percentile plus 1.5 × IQR. As the upper tail is longer than the lower tail, this distribution is positively skewed.

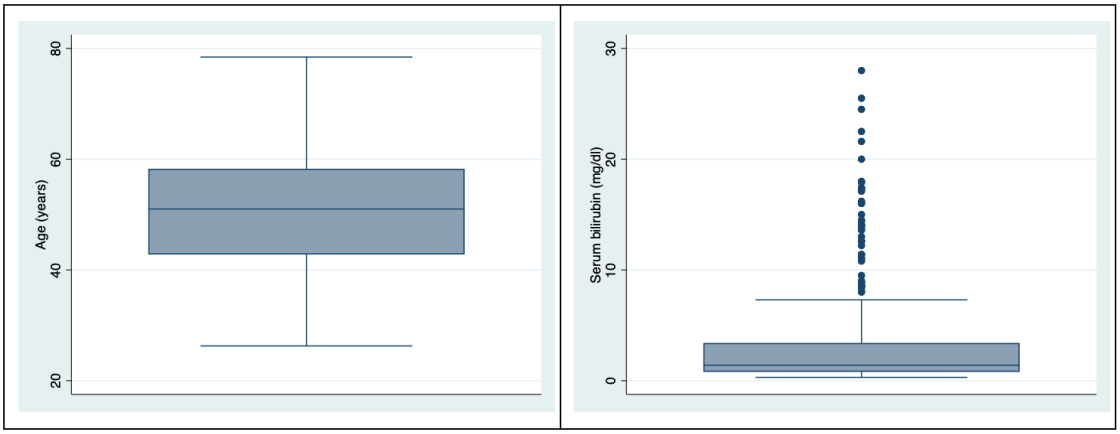


Figure 2.10: Box plot of age (left) and serum bilirubin (right) from PBC study data

2.9 How to report 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 below.

Sources:

Altman (1990)

Cole (2015)

Assel et al. (2019)

Table 2.6: Guidelines for presentation of statistical results

Summary statistic	**Guideline (reference)**
Mean	It is usually appropriate to quote the mean to one extra decimal place compared with the raw data. (Altman)
Median, Interquartile range, Range	As medians, interquartile ranges and ranges are based on individual data points, these values should be presented with the same precision as the original data.
Percentage	Percentages do not need to be given with more than one decimal place at most. When the sample size is less than 100, no decimal places should be given. (Altman)
Standard deviation	The standard deviation should usually be given to the same accuracy as the mean, or with one extra decimal place. (Altman)
Standard error	As per standard deviation
Confidence interval	Use the same rule as for the corresponding effect size (be it mean, percentage, mean difference, regression coefficient, correlation coefficient or risk ratio) (Cole)
Test statistic	Test statistics should not be presented with more than two decimal places.
P-value	Report p values to a single significant figure unless the p value is close to 0.05 (say, 0.01 – 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 similar. A p value can indeed be 1, although some investigators prefer to report this as >0.9. (Assel)
Difference in means	As for the estimated means
Difference in proportions	As for the estimated proportions
Odds ratio / Relative risk	Hazard and odds ratios are normally reported to two decimal places, although this can be avoided for high odds ratios (Assel)
Correlation coefficient	One or two decimal places, or more when very close to ±1 (Cole)
Regression coefficient	Use one more significant figure than the underlying data (adapted from Cole)

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.

Australian Bureau of Statistics. 2021. "Causes of Death, Australia, 2020." https://www.abs.gov.au/statistics/health/causes-death/causes-death-australia/latest-release.

Australian Institute of Health and Welfare. 2021. "Australia's Mothers and Babies." https://www.aihw.gov.au/reports/mothers-babies/australias-mothers-babies.

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.

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.

Kirkwood, Betty, and Jonathan Sterne. 2001. *Essentials of Medical Statistics*. 2nd edition. Malden, Mass: Wiley-Blackwell.

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.