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?**)).

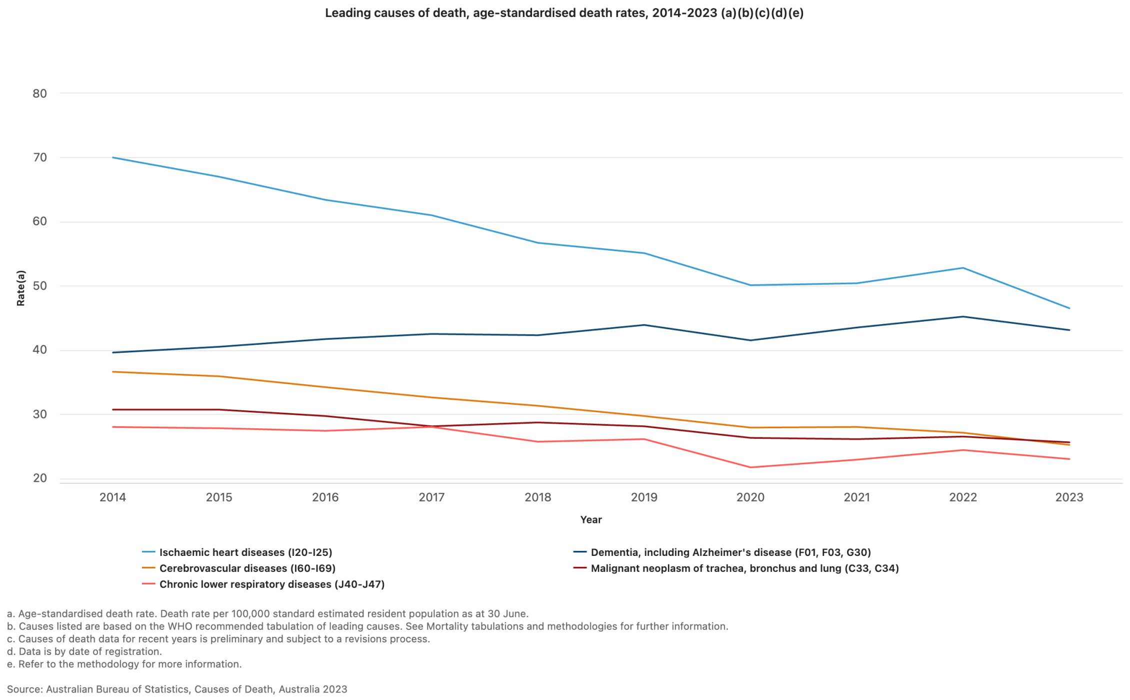
> “COVID-19 was the ninth leading cause of death in 2023, after ranking third in 2022.”

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.



Leading causes of death, age-standardised death rates, 2014-2023 (a)(b)(c)(d)(e)

a. Age-standardised death rate. Death rate per 100,000 standard estimated resident population as at 30 June.
b. Causes listed are based on the WHO recommended tabulation of leading causes. See Mortality tabulations and methodologies for further information.
c. Causes of death data for recent years is preliminary and subject to a revisions process.
d. Data is by date of registration.
e. Refer to the methodology for more information.

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

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

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

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

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

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

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

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

Summarising a single continuous variable graphically

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

Frequency histograms

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

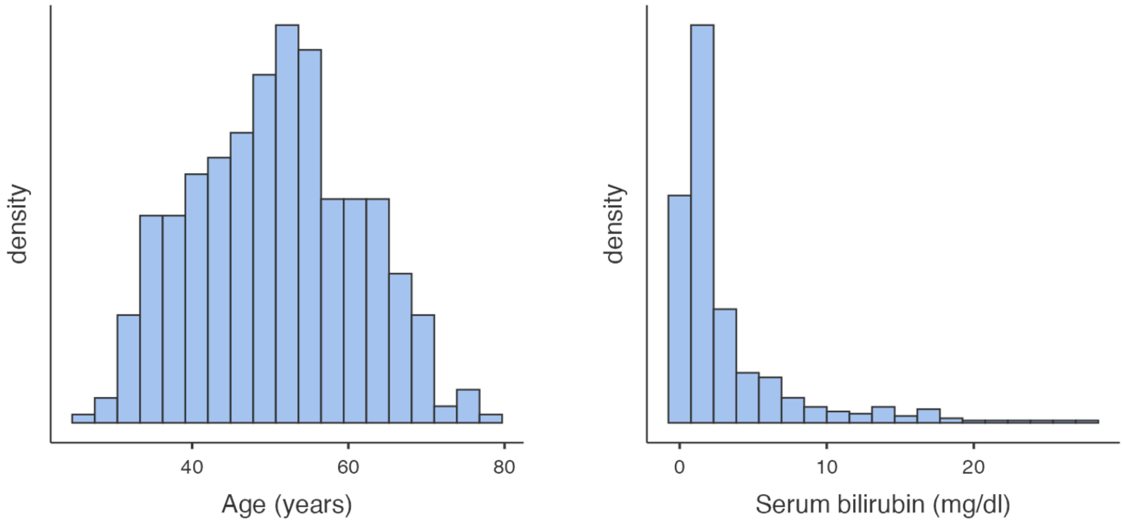


Figure 1.2: Histogram of age (left) and serum bilirubin (right) from a sample of data

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.

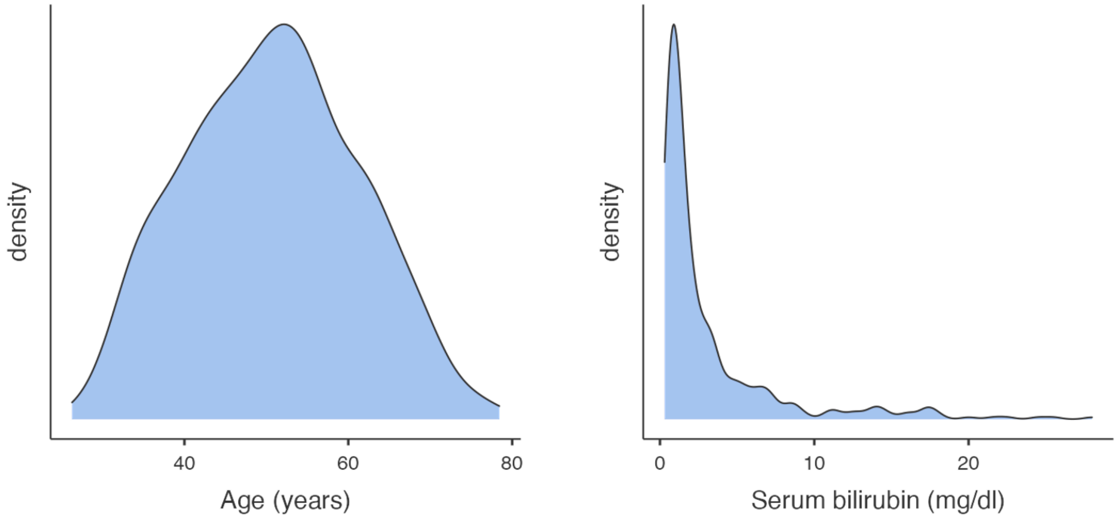


Figure 1.3: Histogram of age (left) and serum bilirubin (right) from a sample of data

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.

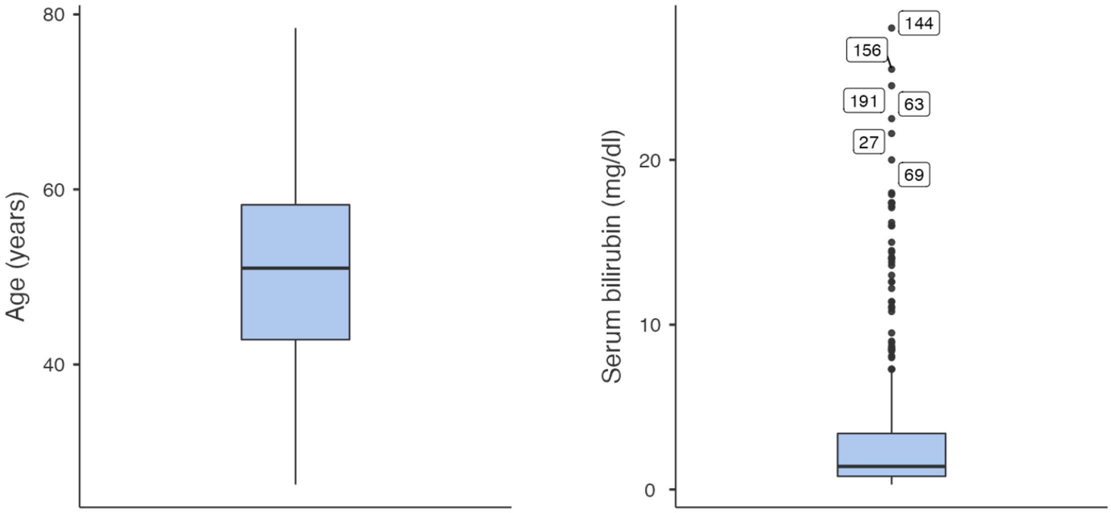


Figure 1.4: Box plot of age (left) and serum bilirubin (right) from PBC study data

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	5.1
2	92	22.3	27.4
3	155	37.6	65.0
4	144	35.0	100.0

* Disease stage was missing for 6 participants

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

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.

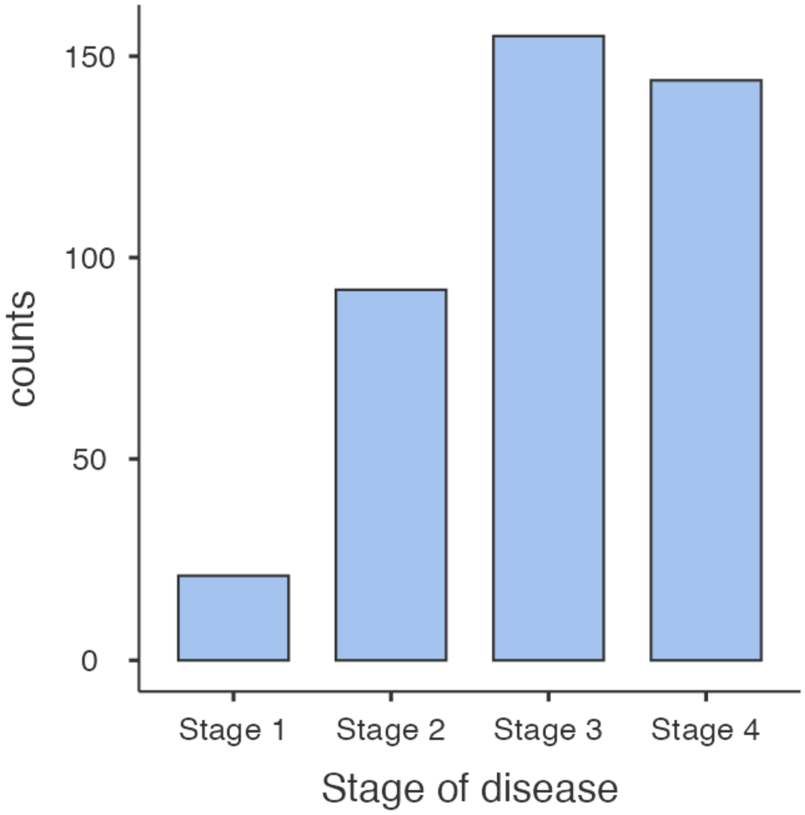


Figure 2.1: Bar graph of stage of disease from PBC study

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).

TKTK - fix missing "100%" cells

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).

TKTK - fix missing "100%" cells

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%	
Female	Count	18	84	139	127	368
	Row pct	4.9%	22.8%	37.8%	34.5%	
Total	Count	21	92	155	144	412

* Disease stage was missing for 6 participants

Tables containing more than two variables

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

In **?@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?

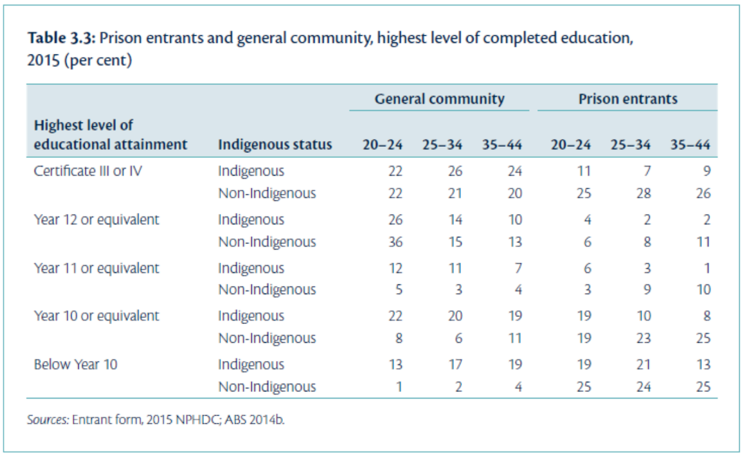


Table 3.3: Prison entrants and general community, highest level of completed education, 2015 (per cent)

		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.

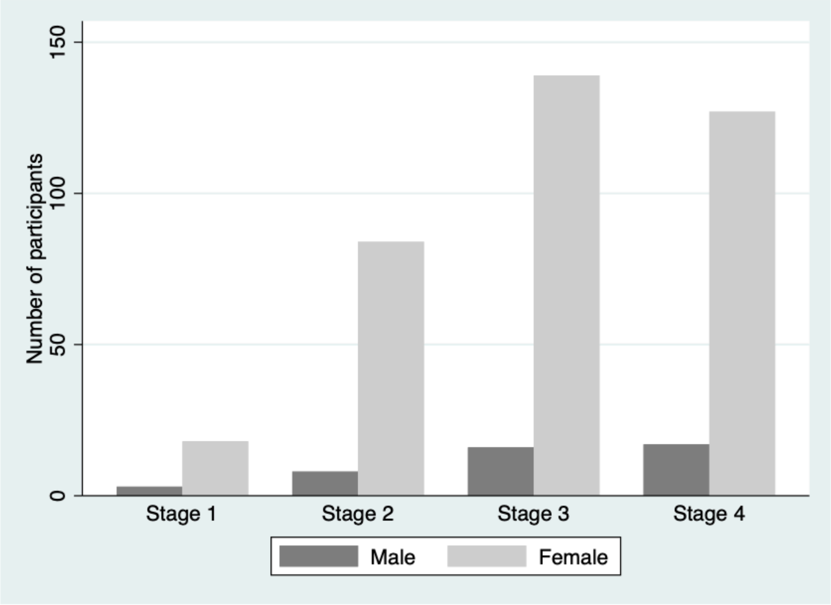


Figure 2.3: Bar graph of stage of disease by sex from PBC study

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).

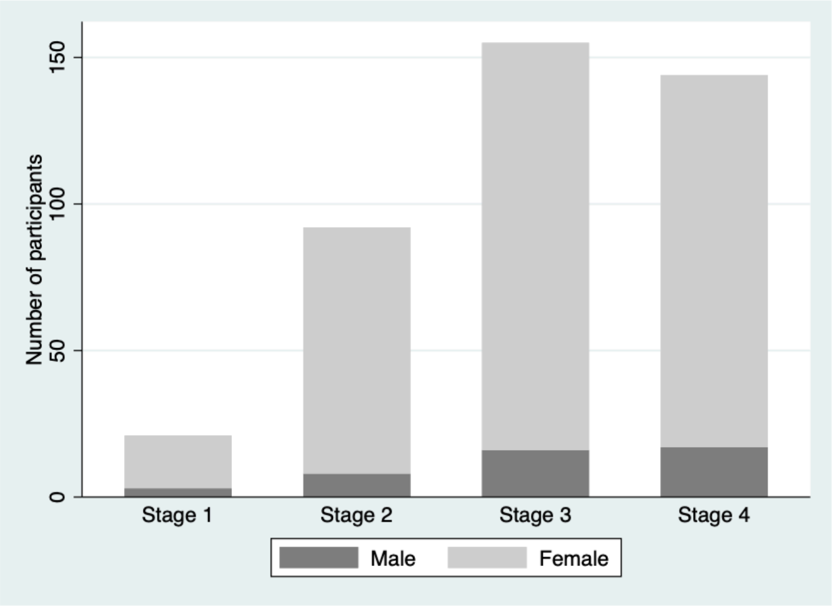


Figure 2.4: Stacked bar graph of stage of disease by sex from PBC study

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.

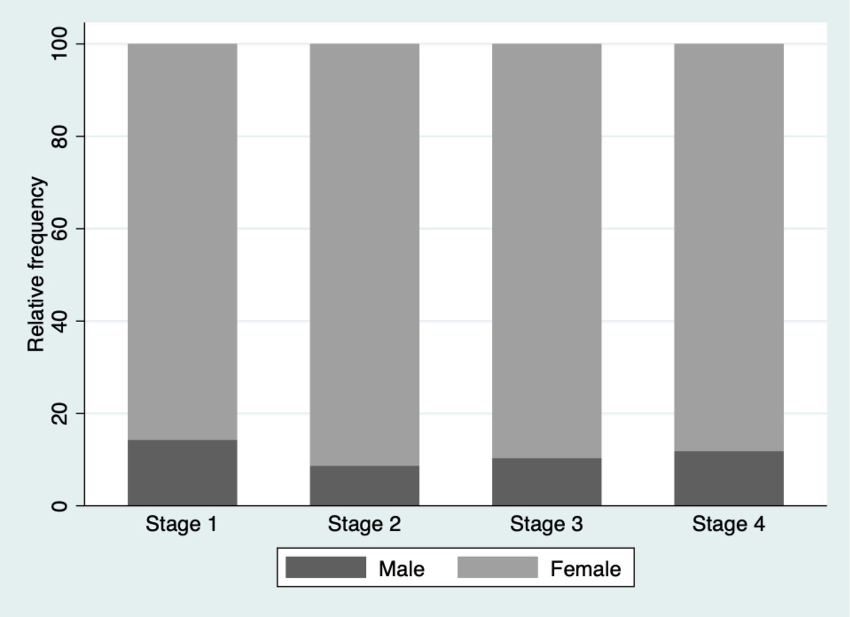


Figure 2.5: Relative frequency of sex within stage of disease from PBC study

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) × 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.

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.

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.