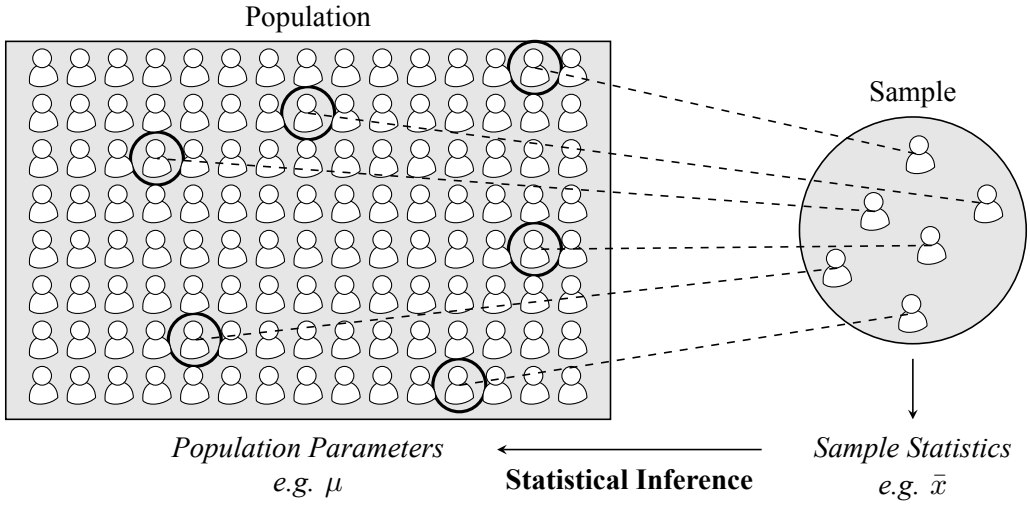
Contents

Introduction

Statistics is the science of collecting, organising, presenting and analysing data. It is vital in helping us develop a better understanding of the world, and to allow decisions to be made in a rational manner based on the information that is available to us. Governments, organisations, businesses and individuals across a diverse range of fields rely on statistics and statisticians.

The *Advanced Higher Statistics* course aims to equip learners with a sound foundation of statistical knowledge. This textbook has been created to serve as a source of key facts, examples, and exercises to support progress through the course.

Population Parameters and Sample Statistics



The diagram above gives a quick overview of the typical role of a statistician wishing to answer questions related to a *population* of interest. This is often in the form of wishing to know more about some *population parameter* such as a population mean, μ. Since it is generally impractical to take a *census* of the entire population, a *sample* is instead obtained from that population, from which *sample statistics* may be calculated, such as the sample mean, $\bar{x}$. *Statistical inference* is the process by which claims about population parameters can be tested using sample statistics.

This course, and textbook, will cover all of the key stages involved in statistical research and investigations, including:

- Gathering data using appropriate *sampling techniques*.
- Initial exploration of data using *descriptive statistics*, including:
  - Visualising data with charts and graphs.
  - Calculating sample statistics.
- Using *statistical inference* to assess the evidence for claims about population parameters, including:
  - An introduction to *probability theory*, which underpins all statistical inference
  - Conducting hypothesis tests

– Constructing confidence intervals

Techniques suitable for a range of contexts will be covered, and it is expected that by the end of the course students should have sufficient knowledge and confidence to collect their own data and conduct their own statistical investigations.

1

Exploratory Data Analysis

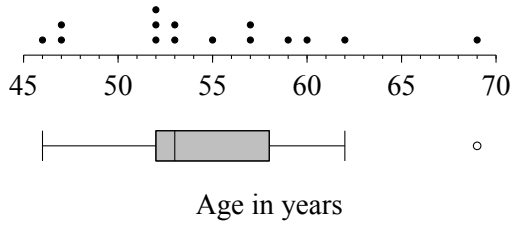
Datasets usually consist rows or columns of values, referred to as *raw data*. Before conducting any *statistical analysis*, a statistician will typically want to get a quick sense of what the data may reveal, as well as to identify any features of note. Common first steps are therefore often to: *organise* the data into suitable lists or tables; *visualise* the data using plots and graphs; and *measure* various properties of the data.

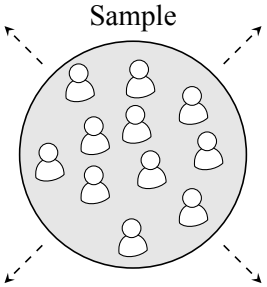
This process is often referred to as *exploratory data analysis*.

Patient data table (first five entries only)

age (yrs)	height (cm)	NHS health board
57	182.2	Fife
43	168.2	Tayside
48	179.4	Tayside
64	162.6	Forth Valley
52	171.5	Fife

Dot plot and box plot (age, NHS Fife patients)

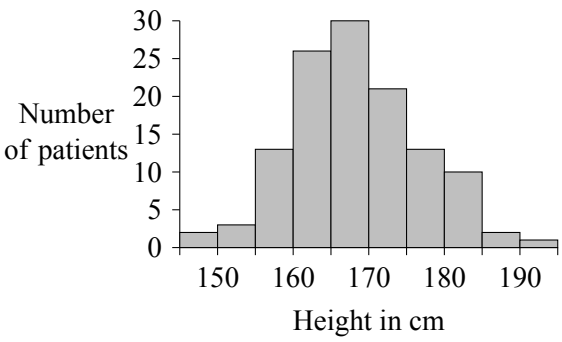




Summary (height in cm, all patients)

sample size	$n = 120$
mean	$\bar{x} = 168.5$
standard deviation	$s = 8.26$
median	$\text{med} = 168.9$
interquartile range	$\text{IQR} = 10.3$
minimum value	$\min = 149.0$
maximum value	$\max = 190.1$

Histogram (height, all patients)

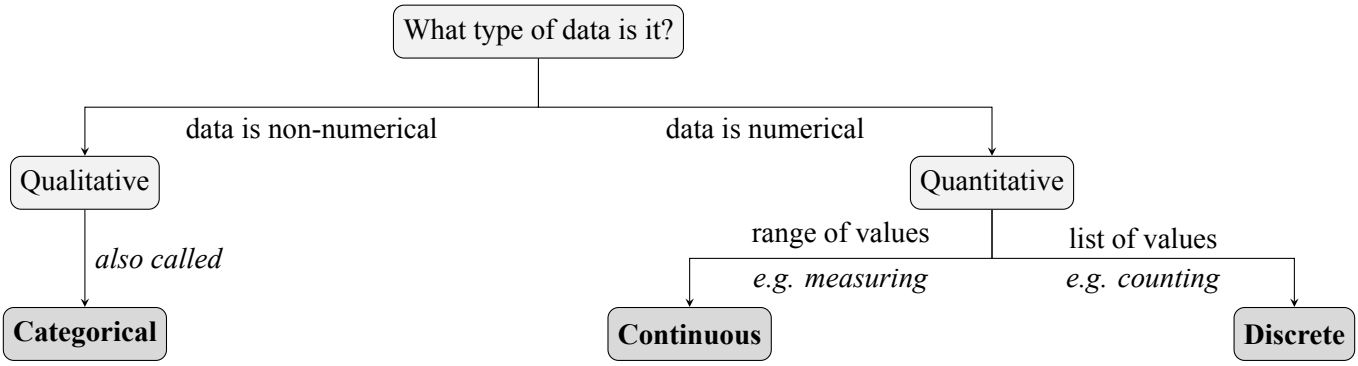


1.1 Types of Data

Knowing what *type* of data has been collected helps inform what kind of graphs and statistics may be appropriate.

Data that is numerical is called *quantitative*, of which there are two types. **Discrete** data consists of values that can be listed, such as someone's shoe size or the number of siblings they have. **Continuous** data can take *any* value within a range, though is typically rounded to reflect the precision of measurements taken, such as the mass of a squirrel.

Qualitative data (also called **categorical**) is *non-numerical*, such as whether a squirrel is red or grey.



For example, the table below shows data relating to some football players:

Position	Team	Height (m)	Caps	Goals
Midfielder	Aston Villa	1.78	66	18
Midfielder	Man United	1.93	49	8
Forward	Southampton	1.75	30	6
Forward	Bournemouth	1.78	49	6
Midfielder	Celtic	1.75	39	5
	Categorical	**Continuous**	**Discrete**	

Exercise 1.1

1. A survey is carried out about the pets owned by a class. Decide whether the following data is qualitative or quantitative.

   (a) How many pets each pupil owns.
   (b) What kind of animal the pets are.
   (c) The pet's age.
   (d) The name of the pet.
   (e) The mass of the pet.
   (f) Whether the pet is a mammal.

2. For each of the following, state whether the data collected would be discrete or continuous.

   (a) The number of people in a room .
   (b) The mass of a book.
   (c) The number of pages in a book.
   (d) The number of songs in an album.
   (e) The volume of water in a bottle.
   (f) The time taken to complete a puzzle.

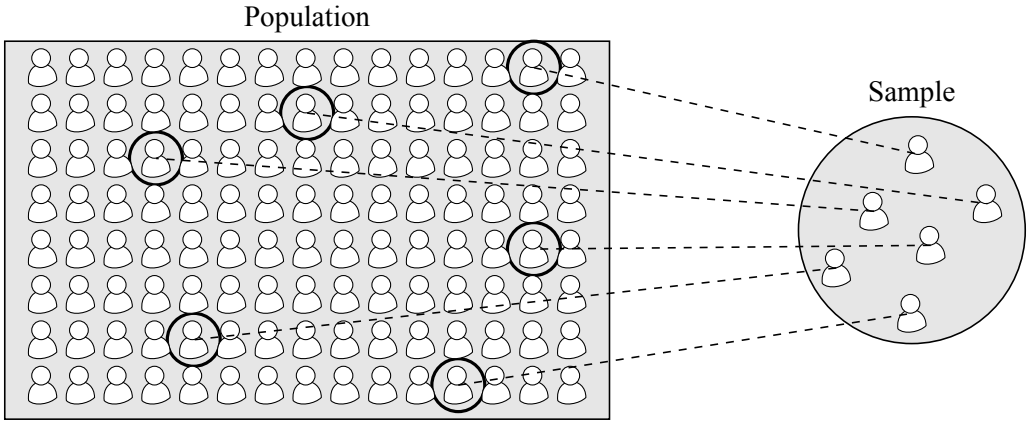
3. A random sample of members of a gym were surveyed and the results are shown in the table below.

Activity level	Smoker	Height (inches)	Mass (pounds)	Pulse (bpm)
Moderate	No	66	140	64
Moderate	No	72	145	58
A lot	Yes	73.5	160	62
Slight	Yes	73	190	66
Moderate	No	69	155	64

   State the type of data that was collected for each variable.

1.2 Population Parameters and Sample Statistics

In a statistical investigation, the term **population** refers to the entire set of objects of interest, whilst a **sample** is a selection of some objects chosen from the population. Understanding the distinction is fundamental to statistical analysis.

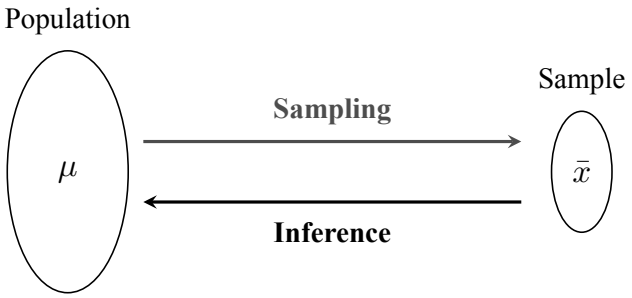


If data were available for an entire *population* then various **population parameters** could be calculated, such as the *population mean*. Since this is never really possible, instead *samples* are taken from populations. Measures calculated from a sample of data are called **sample statistics**, such as the *sample mean*.

Letters from the Greek alphabet are commonly used to represent population parameters, such as:

Population Parameter	**Sample Statistic**
Population mean μ *("mu")*	Sample mean $\bar{x}$ *("x bar")*
Population standard deviation σ *("sigma")*	Sample standard deviation s

It is good practice to refer *clearly* to whether a population parameter or sample statistic is being discussed.



Population parameters are seen as having fixed values, at the time of the study at least, whilst the values of sample statistics will depend on the sample chosen.

Sample statistics may be seen as *estimates* of the true value of their respective population parameters, and **inference** is the process by which claims may be made about population parameters basis on observed sample statistics.

Exercise 1.2

For each, state using the context of the question which population parameter or sample statistic has been obtained.

1. A year group at a school contains 120 pupils. Data is obtained for the number of merits each pupil in the year group obtained, and the mean of those values is calculated to be 3.4.

2. A bus company in Scotland checks the total mileage for some of its buses when they come in for a service. The mean mileage is calculated as 45,300 miles.

3. A medical trial for a new supplement to increase iron levels in the blood involves a number of patients being given the treatment over the course of a month, and then their blood iron levels being checked. The mean increase for the patients is found to be 27.3 micrograms per decilitre (mcg/dL).

1.3 Measures of Location and Spread

Summary statistics are a collection of measures taken of the data from a sample, such as the mean, or standard deviation. These offer a quick way to communicate some general sense of the data. For example, for the *raw data* of $0, 2, 2, 6, 12, 14, 16, 18, 18, 18, 20$, the table below shows a selection of summary statistics.

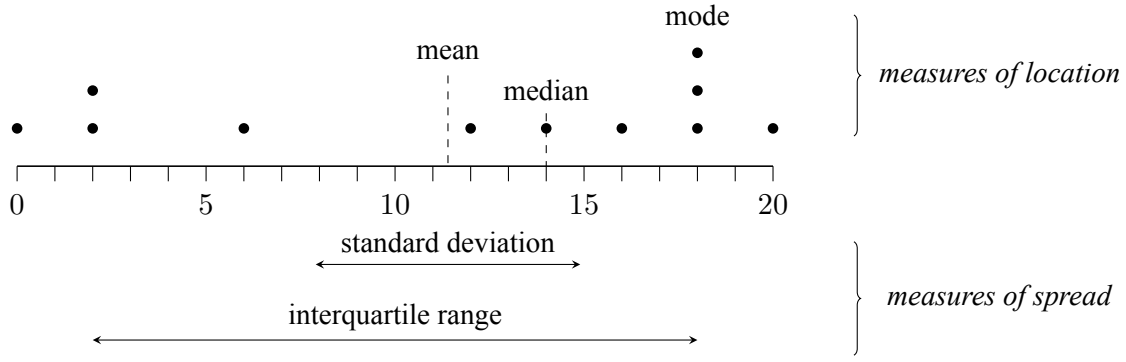
n	mean	sd	sum	sum of squares	min	LQ	median	UQ	max	range	IQR	mode
11	11.45	7.54	126	2012	0	4	14	18	20	20	14	18

There are two main types of summary statistic: those that measure the *central tendency* of data (or its **location**); and those that measure the *variability* of data (or its **spread**). It is common to wish to measure both the location *and* spread.

Measures of Location

The three main measures of location, or *averages*, should already be familiar. To avoid ambiguity, statisticians will typically avoid referring to *"the average"*, and will instead specify exactly *which* average they are referring to:

- **Mode** - the most commonly observed piece of data.
- **Median** - the middle value after the data has been put into numerical order.
- **Mean** - the sum of the values divided by the number of values.



Measures of Spread

The measure of spread which is usually encountered first in a maths classroom is the *range*, which is the difference between the highest and lowest values. In practice this measure is seldom used by statisticians since it only uses the two most extreme values from the data. There are three main measures of spread commonly used, two of which are likely to already be familiar:

- **Interquartile Range** - the difference between the upper quartile and the lower quartile.
- **Variance** - the *mean squared difference* between the values and the mean value.
- **Standard deviation** - the square root of the variance.

The Mean and the Standard Deviation

The *mean* and *standard deviation* are perhaps the most commonly used measures of location and spread, respectively, and both link strongly to the commonly-encountered *bell-shaped curve* (see page 13). The formulae for calculating the sample mean and sample standard deviation from sample data are given below using notation, where the Greek letter Σ (*"Sigma"*) means a *sum* should be calculated, x represents the all of the values in the sample (a shortened abbreviation of $x_1, x_2, \ldots, x_n$ or x_i) and n represents the *number* of values in the sample.

The *sample mean*, $\bar{x}$, is calculated as: $\bar{x} = \dfrac{\Sigma x}{n}$

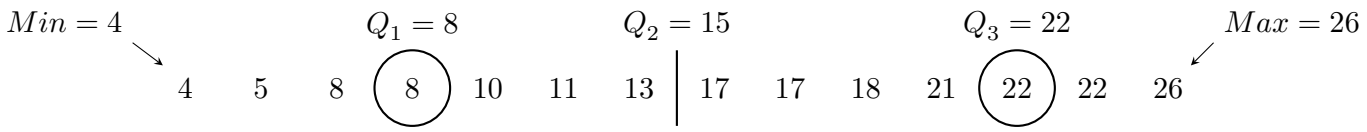
The *sample standard deviation*, s, is calculated as: $s = \sqrt{\dfrac{S_{xx}}{n-1}}$ where: $S_{xx} = \Sigma x^2 - \dfrac{(\Sigma x)^2}{n}$

The *sample variance*, s^2, is simply the *square of the standard deviation*.

The Median and the Interquartile Range

The median, Q_2, and interquartile range, IQR, are often chosen to measure location and spread when the data is particularly *skewed* (see page 13), as they are less influenced by *extreme values* than the mean and standard deviation.

The *lower quartile* (LQ or Q_1) and the *upper quartile* (UQ or Q_3) split a set of values into quarters along with the median. From the lowest value to Q_1 contains the lowest 25% of the values, from Q_1 to Q_2 contains the next 25%, and so on.



The *interquartile range* is calculated as: $\text{IQR} = \text{Q}_3 - \text{Q}_1$

Exercise 1.3

1. The monthly cost of 10 drivers' car insurance (x, in £'s) are recorded. Calculate the mean and standard deviation:

$$\Sigma x = 359 \quad \Sigma x^2 = 15119 \quad n = 10$$

2. Daily fuel spend (f, in £'s) for 15 drivers is summarised below. Calculate the mean and standard deviation:

$$\Sigma f = 87 \quad \Sigma f^2 = 569 \quad n = 15$$

3. The marks, out of 40, of eight pupils sitting a chemistry test are below. Find the mean and standard deviation.

32 25 29 37 16 39 23 32

4. The ages (in years) of 18 customers in a shop are recorded. Calculate the median and interquartile range.

17 63 34 23 40 57 72 70 43
56 29 37 41 50 13 24 30 58

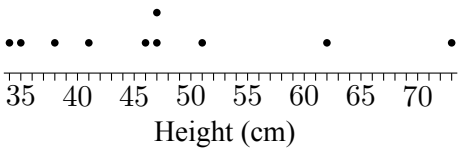
1.4 Displaying Data

Before any analysis of a data set is undertaken, it is common practice to *display* the data in a simple form so that an initial impression of what the data shows can be formed. Some key examples are shown below, using a sample taken by a gardener of the heights of 10 of their sunflowers 8 weeks after germination, measured to the nearest centimetre:

34 35 38 41 46 47 47 51 62 73

Dot Plots

Dot plots offer a quick way of presenting data in a manner similar to a bar chart, stacking repeated values.



Stem-and-Leaf Diagrams

Stem-and-leaf diagrams display data in a manner similar to a grouped bar graph. Leaves should be in order, and a key (or *legend*) and the sample size should be included.

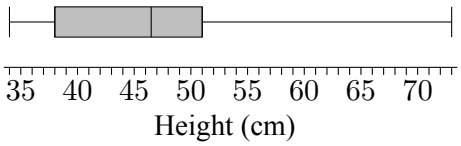
3	4 5 8
4	1 6 7 7
5	1
6	2
7	3

$n = 10$

3 | 5 means 35 cm

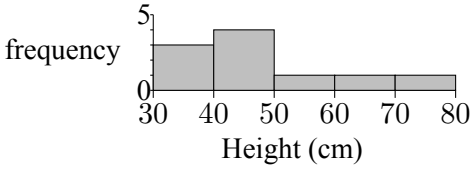
Box Plots

Box plots offer a visual representation of a five-figure summary, consisting of the mininum value, the lower quartile Q_1, the median, the upper quartile Q_3 and the maximum value.



Histograms

Histograms put continuous numerical data into intervals (or *bins*) and display it in a manner similar to a bar graph.



Exercise 1.4

1. The ages (in years) of competitors in a junior golf tournament are below. Draw a dot plot to represent this data:

16 14 19 13 14 17 16 19 16 17 12 16 14 19

2. The price of an ice cream is recorded at 12 stalls, in £'s. Represent the data using a stem-and-leaf diagram.

1.50 2.20 2.50 2.10 1.80 1.40
2.00 1.90 3.60 0.90 1.70 1.50

3. The masses (in kg) of the luggage of 10 passengers on a plane to Barra were recorded. Draw a box plot of these masses.

15 18 14 22 19 14 10 21 17 16

1.5 Tables, Figures and Output

There are many ways in which data can be presented, extending beyond those covered on the previous page. In statistical reports they are often clearly labelled as **tables** or **figures**, and numbered so they can be more easily referred to.

Table 1: Contingency table of mode of transport.

		Transport			
		walk	bus	car	bike
Year Group	3rd	12	15	3	4
	2nd	7	13	8	1
	1st	4	14	11	2

Figure 1: Bar chart of mode of transport.

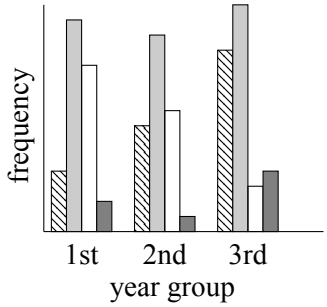
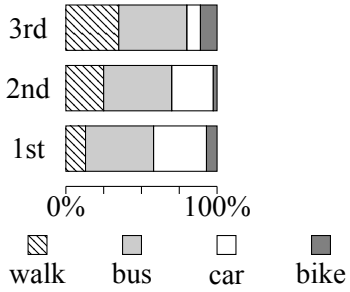


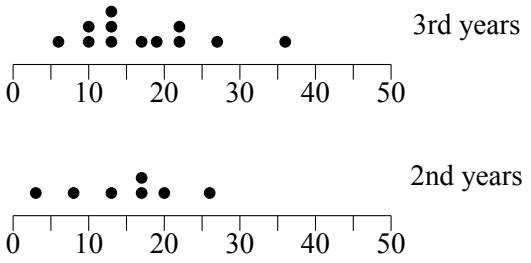
Figure 2: Chart of transport percentages.



Not all graphs encountered in real life will be up to scratch; some graphs may be unclear, accidentally misleading or even intentionally misleading. Being alert to potential flaws in tables and figures and pointing out possible improvements is an important skill.

The widespread availability of computing power means that specialist software is often used by statisticians to avoid the need for time-consuming calculations where possible. Whilst this course requires no use of statistical software, examples of computer **output** will be displayed from which the required information must be extracted.

Figure 3: Dot plots of time (mins) to walk to school for 2nd and 3rd years.



Output 1: Summary data for walk times (mins) for 2nd and 3rd years.

```
data: walk time

3rd yr: n=12    mean=17.3   sd=8.42

2nd yr: n=****  mean=14.9   sd=7.65

alternative hypothesis: means are not equal

t=0.619, df=17, p-value=0.544
```

Whilst this textbook aims to enable familiarity with common figures, tables and computer output, the best way to practise obtaining information from them is to **read a wide range of relevant statistical articles and reports**.

Exercise 1.5

1. Use **Table 1** to calculate the proportion of pupils in the sample that walk to school.
2. Suggest two possible improvements to **Figure 1**.
3. Using **Figure 2**, compare the proportions of pupils in the sample who take the bus between each year group.
4. From **Figure 3**, state the approximate median time taken to walk for 3rd year pupils.
5. One piece of information from **Output 1** has been replace by ****. State the missing value.

1.6 Outliers and Fences

Sometimes a set of data may contain values that seem *extremely high or low* in comparison to the rest of the values. It is important to be wary of such values and the potentially significant impact they could have on any statistical analysis performed. There are a range of objective approaches that can be used to determine whether one or more points are extreme enough to be classed as *outliers*. In this course, *fences* will be used.

Outliers and Fences:

$$\text{Lower Fence} = Q_1 - 1.5 \times \text{IQR}.$$

$$\text{Upper Fence} = Q_3 + 1.5 \times \text{IQR}.$$

Any values *below the lower fence* or *above the upper fence* are **outliers**.

If an outlier is identified in a set of data, an investigation should be conducted to determine the cause. If it is concluded that the outlier is a rogue point, such as an error in measurement, it can be removed from the data set and summary statistics re-calculated. However, if an outlier is a valid point in the data set then it should not be removed.

Returning to the sunflower heights from page 7, the data can now be checked for possible outliers:

Raw data

Heights (cm)

34	35	38	41	46	47	47	51	62	73

Summary data

n	min	Q1	median	Q3	max
10	34	38	46.5	51	73

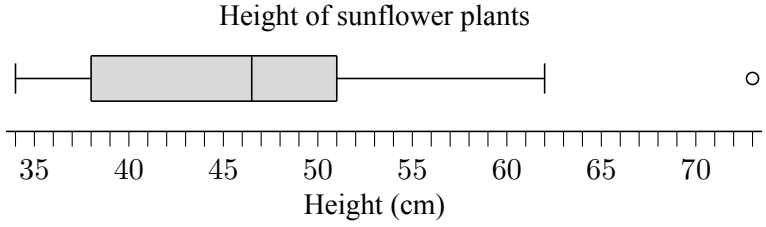
$$\text{IQR} = Q_3 - Q_1 = 51 - 38 = 13$$

$$\text{Lower fence} = Q_1 - 1.5 \times \text{IQR} = 38 - 1.5 \times 13 = 18.5$$

$$\text{Upper fence} = Q_3 + 1.5 \times \text{IQR} = 51 + 1.5 \times 13 = 70.5$$

Since $73 > 70.5$, it is an outlier.

On a box plot, outliers are indicated separately as shown below, with the "whiskers" extending only as far as the highest (and lowest) values which are *not* outliers. However the quartiles remain as those calculated using the *full* set of values, including any outliers. *Plotting outliers separately in this way is to highlight their presence, and is not a judgement on whether those values are rogue or valid.*



The outlier in this data set could be valid if, for instance, the sunflower plant is in a particular sunny spot in the garden. However, if the height was measured incorrectly then it can be removed from the data set.

Example

Problem: A vet conducts a survey, asking 10 randomly chosen cat owners registered with his practice to share the mass of their adult cat, in kilograms. The results are:

4.6 3.6 4.1 4.5 3.7 3.9 4.0 0.3 3.9 4.8

(a) Show that the set of data contains an outlier.

(b) Suggest a possible cause for this outlier which would mean it should be removed.

Solution:

(a) $Q_1 = 3.75$, $Q_3 = 4.4$, IQR $= 4.4 - 3.75 = 0.65$.

Lower fence $= 3.75 - 1.5 \times 0.65 = 2.775$

Upper fence $= 4.4 + 1.5 \times 0.65 = 5.375$

Since $0.3 < 2.775$, the value of 0.3 is an outlier.

(b) The owner could have mistakenly measured the mass of their cat in stones instead of kilograms.

Exercise 1.6

1. Strawberry tarts are sold every day in a bakery over the summer months. The shop manager monitors the sales and takes a random sample of the number of strawberry tarts sold on 10 randomly selected days over one month.

   23 31 45 32 35 19 24 25 30 38

   Calculate the upper and lower fences for this data set and identify any possible outliers.

2. The number of hours spent studying each week by a randomly selected group of 24 students is given below.

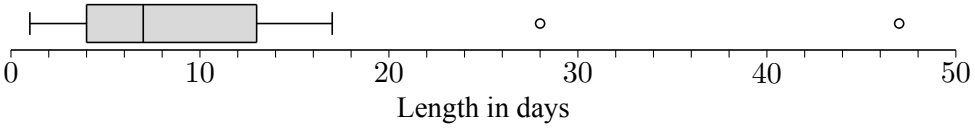
   10 18 15 12 17 27 4 14 14 14 17 15
   15 12 17 17 13 15 15 16 16 9 11 15

   (a) Identify any possible outliers.

   (b) Draw a box plot to display this data.

3. A hospital trust compiled data relating to the length of stay of patients in a particular hospital. A random sample of 18 patients gave the following summary data, measured in days.

n	mean	sd	min	Q1	median	Q3	max
18	10.78	11.17	1	4	7	13	47



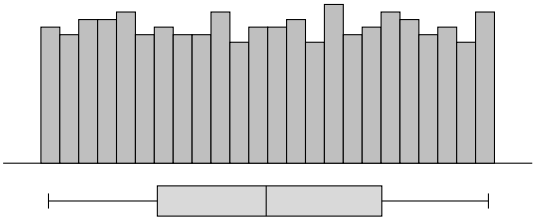
   (a) Show that the two values indicated by circles on the box plot are outliers.

   (b) Suggest a cause for these outliers that would make it valid to remove them from further analysis.

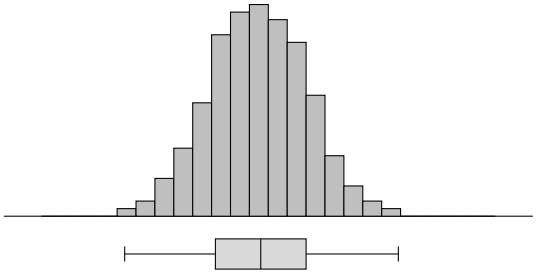
1.7 Describing Data Distributions

Summary statistics alone are not sufficient to describe the full complexity of the way data is distributed. It is also important to consider and be able to describe the *shape* of the distribution, accomplished most easily using a histogram. Whilst real-life rarely *perfectly* follows a precise mathematical model, a number of shapes are commonly observed and a number of descriptions are commonly considered. The following data was generated by *simulating sampling* from populations of data which followed the distributions described.

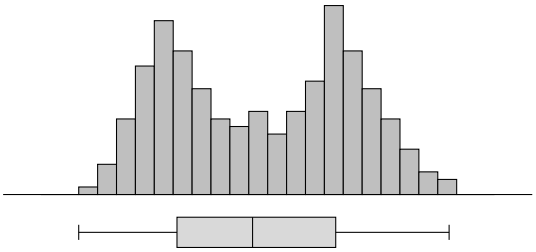
uniformly distributed data ($n = 103$)



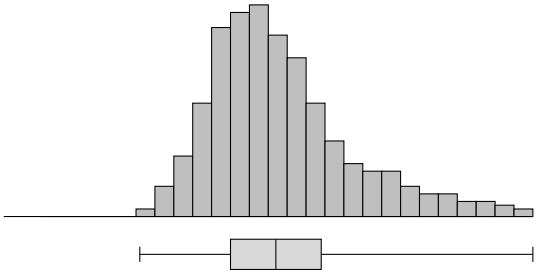
normally distributed data, also referred to as a bell-curve ($n = 215$)



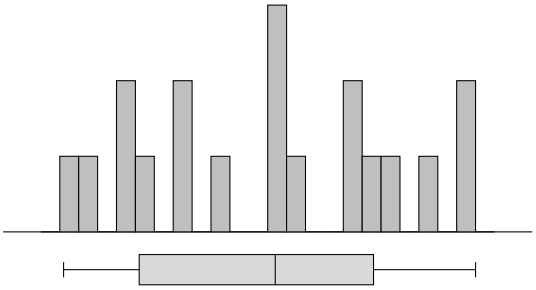
bimodal data ($n = 87$)



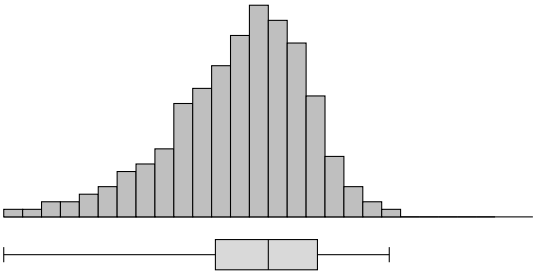
positively skewed data ($n = 134$)



very hard to infer distributions of data from small samples ($n = 19$)



negatively skewed data ($n = 97$)



As well as considering whether the **underlying population distribution** of data appears to be *normal*, *skewed*, *uniform* or *bimodal*, the appropriateness of some statistical inference techniques will depend on whether the distribution is *symmetrical*, or whether two sets of data follow the *same distribution*, with matching shapes and spread.

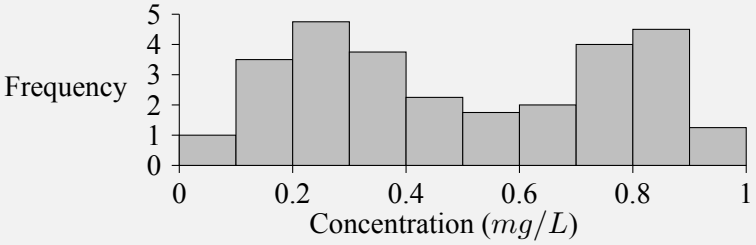
Whilst the following example references techniques that will be introduced later in the course, it is intended to give an idea of the importance of studying distributions in selecting appropriate methods for statistical inference.

Example

Problem: A environmental scientist collected water samples from a number of rivers in a region to explore whether the concentration of a harmful chemical in the water has changed since farming practises used locally have changed. They are considering the following choices of hypothesis test:

- A *t-test*, the most appropriate choice if the data follows a normal distribution.
- A *Wilcoxon Signed Rank test*, appropriate if the data is distributed symmetrically.

Figure 1: Concentration of chemical in river sites



(a) Use **Figure 1** to make a comment on the distribution of chemical concentrations.

(b) Comment on the appropriateness of each choice of hypothesis test suggested for this study.

Solution:

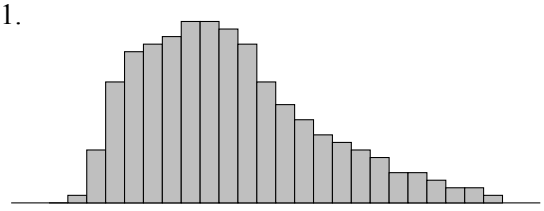
(a) Figure 1 suggests that the population distribution of concentrations of the chemical in river sites may be bimodal.

(b) Since the histogram suggests the population distribution of chemical concentration may be symmetrical but not normal, a Wilcoxon Signed Rank test would be appropriate whilst a *t*-test would not.

Exercise 1.7

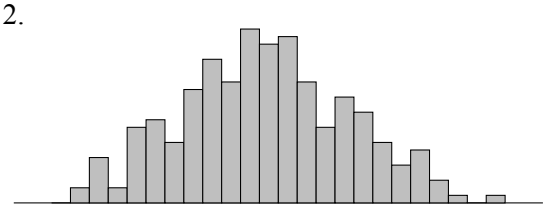
Comment on each population distribution using the simplified histograms, which show data from samples.

1.



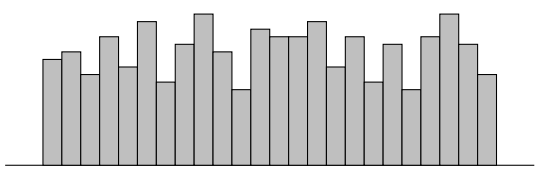
Hazelnut mass (grams)

2.



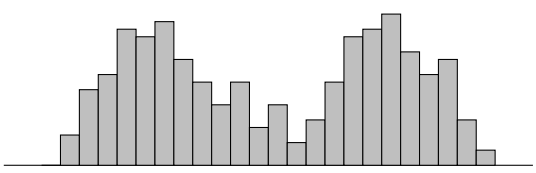
Tree height (metres)

3.



Waiting time (minutes)

4.



Number of successful passes

1.8 Comparing Distributions

When comparing two or more samples of data, a good starting point is to create suitable plots and obtain summary statistics. Differences and similarities should be noted, considering the location, spread and shape of the distributions of each group.

Example

Problem: Within the sport of professional cycling, it is common for riders to specialise in certain types of cycling. *Sprinters* aim to win races that finish on flat road at high speed, whilst *climbers* aim to win races with an uphill finish. A researcher studying the characteristics of elite-level cyclists randomly selects samples of 30 *sprinters* and 30 *climbers*, recording the heights of each. The results are shown summarised in **Table 1** below.

Table 1

variable	n	Min	Q1	Median	Q3	Max	IQR
sprinters	30	165	174	179	184	200	10.5
climbers	30	158	168	174	177	195	9

Box plots for the two groups of cyclists are shown below in **Figure 1**, and dot plots in **Figure 2**.

Figure 1

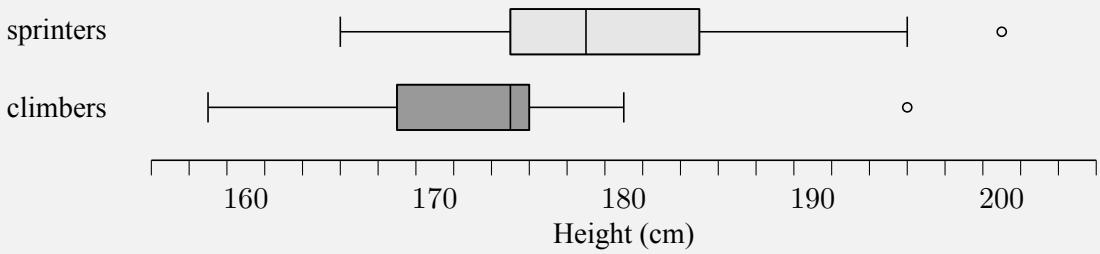
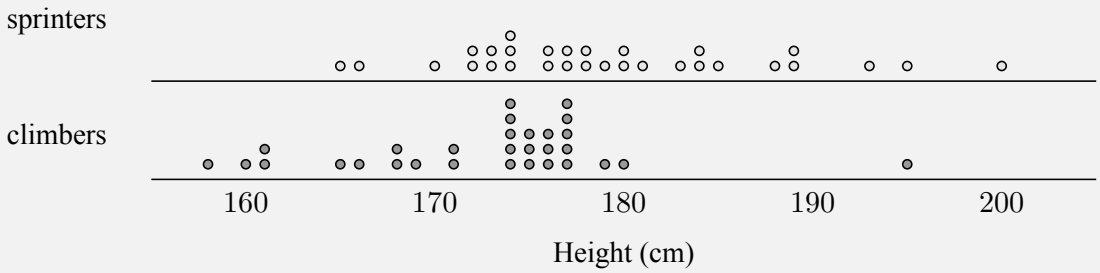


Figure 2



(a) Use **Figure 1** to make two comments comparing the heights of sprinters and climbers.

(b) Use **Figure 2** to comment on the shape of the distribution of the heights of climbers.

Solution:

(a) The sample median height of the sprinters was greater than that of the climbers.

The larger interquartile range suggests that the population heights of sprinters are more varied.

(b) The dot plot suggests the underlying distribution of heights for climbers is negatively skewed.

Note that this data is based on the heights of *samples* sprinters and climbers. Without the use of *statistical inference* techniques covered later in the course, it is only possible to *conjecture* rather than ever being able to make any definitive claims about the heights of the full *populations* of sprinters and climbers.

Exercise 1.8

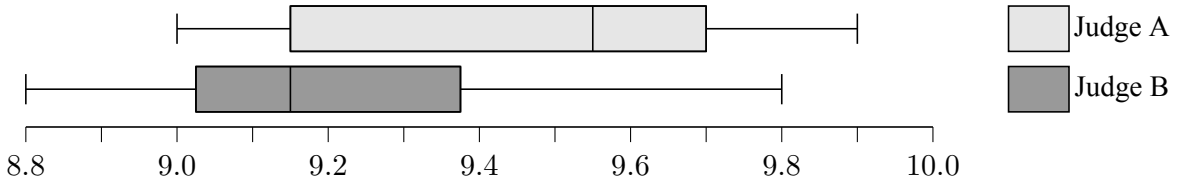
1. Two models of coffee machines produced by the brand *DOLCAFÉ* are supposed produce a white coffee with the same volume of liquid each time, but natural variation means the actual amount dispensed is not quite the same every time. The volumes dispensed, in ml, are recorded for random samples from each model.

Model A	295	298	299	301	303	304	309
Model B	282	294	300	302	306	312	320

   (a) Determine the mean and standard deviation for each model.

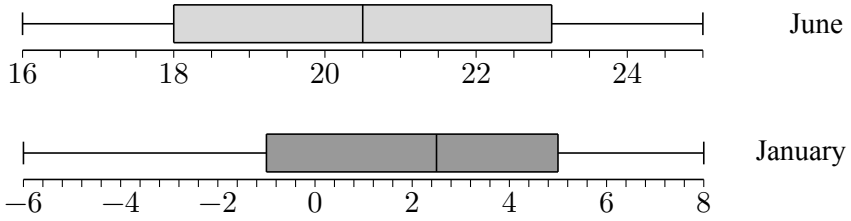
   (b) Make two comments comparing the amount of coffee dispensed by the two models.

2. The scores awarded by two judges in a diving competition are recorded and displayed in box plots, shown below.



   Using the box plots, make two comments comparing the scoring of the two judges.

3. The temperature at 2pm was noted for 14 randomly selected days in January and June at a local tourist spot and box plots drawn to display the results.



   (a) Using the box plots, make two comments comparing temperatures in January and June.

   (b) Explain why the box plots may be misleading.

4. The back-to-back stem-and-leaf diagram below shows the heights of trees in two areas of woodland.

Woodland A					Woodland B			
			5	2	1	3	3	8
		7	4	3	0	2	4	7
9	7	7	2	4	2	5		
	7	6	3	5				

   (a) Suggest one improvement needed for the diagram.

   (b) Make one comment comparing the heights of trees in the two areas of woodland.

5. The manager of a lemonade bottling plant is interested in comparing the performances of two production lines, one of which has only recently been installed. For each line she selects 10 one-hour periods at random and records the number of crates completed in each hour. The table below gives the results:

Production Line	Number of crates completed per hour									
New	78	87	79	82	87	81	85	80	82	83
Old	74	77	78	70	87	83	76	78	81	76

   (a) Draw dot plots to represent the data.

   (b) Calculate the mean and standard deviation for each sample.

   (c) Make two comments comparing the two lemonade production lines.

Review Exercise

1. A music fan wishes to investigate the characteristics of songs that are popular on a well-known music streaming app. She is able to obtain data on the thousand most-streamed songs of all time, a small selection of which are displayed in the table below.

Track	Artist	Released	Length (m:s)	No. of Streams
Creep	Radiohead	1992	3:59	1271293243
Blank Space	Taylor Swift	2014	3:52	1355959075
Something In The Way	Nirvana	1991	20:37	368646862
Nonsense	Sabrina Carpenter	2022	2:43	342897938
Drivers License	Olivia Rodrigo	2021	4:02	1858144199
Iris	The Goo Goo Dolls	1998	4:50	1284942608

   (a) State which of the five variables contain:

      i. Categorical data.

      ii. Discrete data.

      iii. Continuous data.

   (b) If it was suspected that one of the tracks is an outlier in terms of length, state what could be calculated in order to determine whether it is in fact an outlier.

2. A sports clothing store is seeking to better understand its sales of running shoes. For a random sample of twelve sales of pairs of running shoes, the sizes are recorded as follows:

$$4 \quad 7 \quad 9 \quad 11 \quad 6 \quad 9 \quad 8 \quad 4 \quad 10 \quad 9 \quad 6 \quad 5$$

   (a) Contruct a dot plot to display the data.

   (b) Find the mean and standard deviation for the sizes of running shoes sold.

   (c) State the type of data recorded.

3. The petal lengths (in centimetres) for samples of flowers from two species of iris flowers is recorded: *iris setosa* and *iris versicolor*. The data is summarised in the table below using statistical software:

species	n	min	Q1	Q2	Q3	max
iris setosa	50	1.00	1.40	1.50	1.58	1.90
iris versicolor	50	3.00	4.00	4.35	4.60	5.10

   (a) For the species *iris setosa*, determine whether the dataset contains any outliers.

   (b) Make two comments comparing the petal lengths of the two species of iris.

2

An Introduction to Probability Theory

The purpose of statistics is to make sense of the world by analysing data. Typically however, it is only possible to obtain data from a random sample of the full population of interest, creating an element of uncertainty that must be navigated carefully. For this reason, developing an understanding of statistical analysis begins with learning about probability theory.

Key definitions

In probability theory, an experiment for which all of the possible *outcomes* can be defined, and yet the actual outcome of any given *trial* cannot be known in advance, is called a *random experiment*. The set of all possible outcomes is called the *sample space* of the random experiment. Probability is assigned to individual outcomes or to sets of outcomes, called *events*. Probability itself is defined as the **long-run relative frequency** of an outcome or event occurring, and values range from 0 (*impossible, and will never happen*) to 1 (*certain, and will happen every time*).

For example, in the context of rolling a fair, cubical die:

- the procedure of rolling the die and observing the result is a **random experiment**.
- each roll of the die may be referred to as a **trial**.
- the **sample space** consists of a list of all of the possible **outcomes**: $\{1, 2, 3, 4, 5, 6\}$
- the **probability** of a 5 is $\frac{1}{6}$ since *in the long-run* it can be expected to occur $\frac{1}{6}$ of the time.
- a possible **event** of interest may be that the outcome is *prime*: $\{2, 3, 5\}$

Probability notation

The probability that a die lands on 5 is *one-in-six*, or $\frac{1}{6}$. Using probability notation:

$$\mathrm{P}(5) = \frac{1}{6}$$

2.1 Types of Probability

Experimental Probability

The probability of an event can be estimated through repeated trials - this is referred to as the *experimental probability* of the event. For example, if a motorist passes a set of traffic lights every morning and wishes to know the probability of having to wait at a red light on any given morning, they could record the *proportion* of times they have to wait at a red light over a period of time. If, over the course of 200 mornings, they have to wait at a red light 114 times:

$$\text{P(red light)} \approx \frac{114}{200} = 0.57$$

Theoretical Probability

When each of the outcomes in the sample space of a random experiment can be reasonably assumed to be **equally likely** then the probability of any outcomes or events can be determined. This is called *theoretical probability*.

Equally likely outcomes typically arise from the symmetry of a physical object (such as a coin or die) or a reasonable assumption. The probability of any one *individual outcome* in a sample space containing N equally likely outcomes is $\frac{1}{N}$.

For example, if a playing card is picked at random from a standard pack of 52 cards, containing no jokers, then the probability of picking the Queen of Hearts is $\frac{1}{52}$.

An **event** is a set of one or more outcomes from the sample space. The probability of an event is given by:

$$\text{P(event)} = \frac{\text{number of outcomes from the sample space in the event}}{\text{total number of outcomes in the sample space}}$$

Example 1

Problem: A bag contains 13 red counters, 5 blue counters and 4 green counters. A counter is chosen at random. Find the probability that the counter is green.

Solution:

$$\text{P(green)} = \frac{4}{22} = \frac{2}{11}$$

In this course, probabilities will be given as simplified fractions or **rounded to four decimal places**. In probability theory, *"and"* means that *both* events are true, and *"or"* means that *at least one of* the events are true.

Example 2

Problem: A car dealership has 80 cars in its showroom. They have a mix of electric, hybrid, petrol and diesel cars, both new and second-hand. The table below shows the numbers of each type.

	electric	hybrid	petrol	diesel
new	15	17	28	3
second-hand	1	7	4	5

A car is selected at random. Determine the probability that the car selected is a second-hand hybrid car.

Solution:

$$\text{P(second-hand and hybrid)} = \frac{7}{80} = 0.0875$$

Exercise 2.1

1. A letter of the English alphabet is chosen at random. State the probability that the letter chosen is:

   (a) A vowel. (b) A consonant. (c) In the word *statistics*.

2. A bag contains four blue, five green and six red counters. Find the probability that a randomly chosen counter is:

   (a) Blue. (b) Either green or red. (c) Not red.

3. A box of chocolates contains chocolates of three different flavours. Five are caramel, four are orange and three are strawberry. A chocolate is chosen at random. State the value of:

   (a) P(orange). (b) P(orange and strawberry). (c) P(not caramel)

4. As of 2024, the UK contains 76 cities, of which 55 are in England, 8 in Scotland, 7 in Wales and 6 in Northern Ireland. If a city in the UK is selected at random, find the probability that the chosen city is:

   (a) In Scotland. (b) In Wales or Northern Ireland. (c) Not in England.

5. A board game uses green counters numbered from 1 to 8, blue counters numbered from 1 to 6 and red counters numbered from 1 to 4. A counter is chosen at random. Determine the probability that the counter selected is:

   (a) The blue numbered 5. (b) Numbered 5. (c) Blue or numbered 5.

6. A restaurant grades the fruit contained in a delivery as good or bad. The contents of the previous delivery are shown in the table below.

	Apples	Tangerines	Kiwi
Good	27	19	26
Bad	3	1	4

   Letting G represent that the event the fruit chosen is *good*, and A the event that it is an *apple*, find:

   (a) P(G) (b) P$(G$ and $A)$ (c) P$(\text{Not } A)$

7. A hospital contains a number of doctors, nurses, allied health professionals and support staff, who are each assigned to work on one of Wards 1 to 3. The breakdown of this is shown in the table below.

	Doctor	Nurse	Allied	Support
Ward 1	3	7	5	8
Ward 2	4	10	2	9
Ward 3	2	8	1	5

   If member of staff working at the hospital is chosen at random, find the probability that they are:

   (a) A nurse. (b) Not on Ward 2. (c) On Ward 2 but not a doctor

2.2 Sample Spaces as Lists and Grids

It can help to write out sample spaces. Some sample spaces can be easily described using a simple list, but more organisation may be needed for more complicated sample spaces, such as approaching the list in a structured manner or using a grid. Sample spaces are generally useful only if each of the outcomes is equally likely.

If two coins are tossed, it may be tempting to suggest that the probability of both landing tails is one-in-three, since the experiment can result in either no coins landing tails, only one coin landing tails or two coins landing tails. However a list of the four possible *equally likely* outcomes in the sample space shows that this is not the case:

Not equally likely outcomes: $\{\text{no tails}, \text{one tail}, \textbf{two tails}\}$

Equally likely outcomes: $\{\text{HH}, \text{HT}, \text{TH}, \textbf{TT}\}$

The probability of both coins landing tails is therefore *one-in-four*, or $\text{P}(\text{both tails}) = \frac{1}{4}$.

Example 1

Problem: Three coins are tossed at the same time. Find the probability that exactly two of the coins land tails.

Solution: *A list of equally likely outcomes will help here. Note there are* $2 \times 2 \times 2 = 8$ *outcomes.*

$$HHH \quad HHT \quad HTH \quad \underbrace{HTT} \quad THH \quad \underbrace{THT} \quad \underbrace{TTH} \quad TTT$$

$$\text{P}(\text{exactly two tails}) = \frac{3}{8}$$

Example 2

Problem: Two fair dice are rolled. Determine the probability that the sum obtained is at least nine.

Solution: *A grid of outcomes will help here. Note there are* $6 \times 6 = 36$ *outcomes.*

first die

second die	**+**	**1**	**2**	**3**	**4**	**5**	**6**
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	(9)
	4	5	6	7	8	(9)	(10)
	5	6	7	8	(9)	(10)	(11)
	6	7	8	(9)	(10)	(11)	(12)

$$\text{P}(\text{sum} \geqslant 9) = \frac{10}{36} = \frac{5}{18}$$

Exercise 2.2

For most of the following questions, it is recommended to write out a full list of all the possible equally likely outcomes. There may be times when a full list would be excessive in length, in which case careful multiplication can reveal the *number* of outcomes, and only the *relevant* ones may be needed to be listed.

1. A coin is tossed and a cubical, fair die numbered 1 to 6 is rolled. Write down:
   (a) A list of all outcomes.
   (b) P(even and tails)

2. Each morning, a teacher picks at random whether to start the day with tea, coffee or apple juice, and also picks at random which one of museli, granola, toast or porridge to eat for breakfast. By first writing down all the possible ways the teacher may start their day, determine the probability that on any particular morning the teacher has:
   (a) Tea and toast.
   (b) Museli but not coffee.
   (c) Neither porridge nor tea.

3. A coin is tossed three times. Find the probability that it shows:
   (a) Tails twice in a row.
   (b) Tails at least twice.

4. A random number is chosen from 1 to 140, inclusive. By writing down all relevant results, determine the probability of the number chosen having digits that sum to 3.

5. Two fair, cubical dice are rolled. Find the probability that the product of the numbers obtained is less than 5.

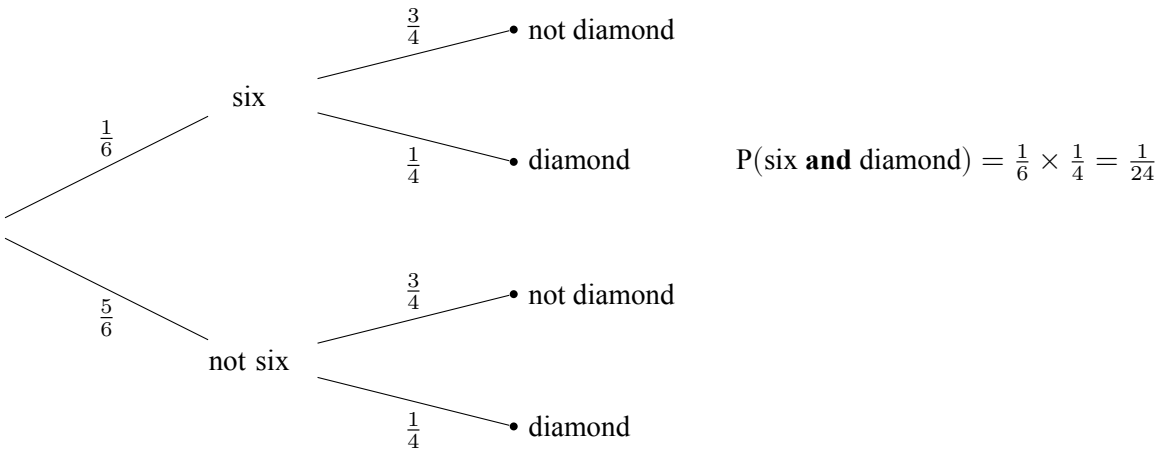
6. A tetrahedral (four-sided) die numbered 1 to 4 and a cubical die numbered 1 to 6 are rolled. Find the probability that at least one of the numbers obtained is 2 or less.

7. Five coins are placed in a bag, with values of 1p, 2p, 5p, 10p and 20p. Two coins are chosen at random, one after another, with the second selected without the first being replaced. Calculate:
   (a) P(the 20p is selected first)
   (b) P(the 20p is selected as one of the two coins)
   (c) P(total of the two coins $\geqslant$ 15p)
   (d) P(total of the two coins $<$ 12p)

8. Two friends, Alex and Charlie, are part of a class which contains eight pupils in total. One pupil from the class is randomly chosen to hand out mini-whiteboards, and different pupil is chosen to hand out pens.
   (a) Determine the number of possible ways the two tasks can be assigned to two pupils.
   (b) Hence, find the probability that:
      i. Alex is chosen to hand out the mini-whiteboards, and Charlie is chosen to hand out the pens.
      ii. Neither Alex nor Charlie are chosen for either task.
      iii. At least one of Alex and Charlie are chosen to perform a task.

2.3 Tree Diagrams

Tree diagrams can be created to map out the possible results of a random experiment, and they are especially useful when considering two or more events. For example, consider a game involving rolling a cubical die and picking a playing card at random from a standard set, with an interest in the chance of the die landing on a *six* and the card picked being a *diamond*:



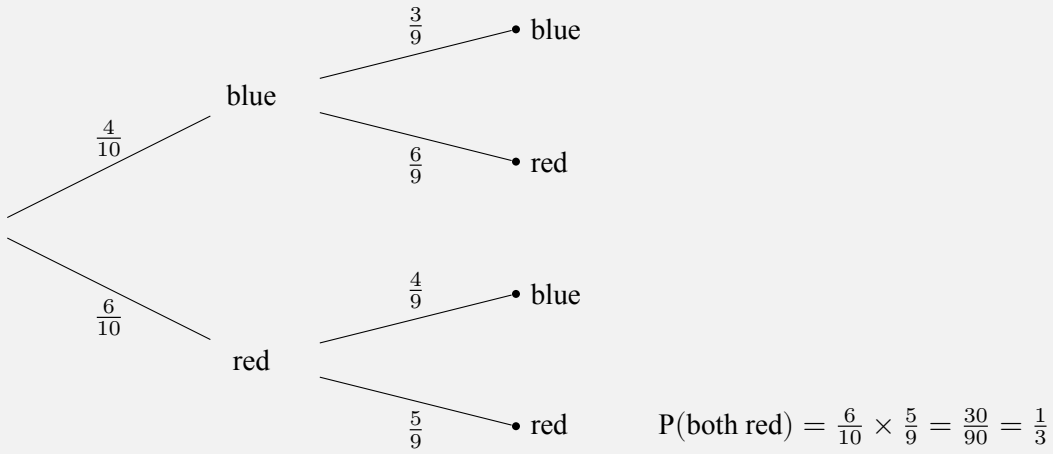
- Every branch must be clearly labelled with a probability.
- The end of each branch must be clearly described, such as *"not diamond"*.
- Sibling branches must add to 1.
- Multiplying the probabilities along branches gives the probability of those events occurring together.

Note that in the tree diagram above the probability of the card being a diamond does not change depending on whether the roll of a die gives a six or not. In the following example the probabilities on the second stage of the tree diagram do change depending on the result of the first stage.

Example

Problem: From a bag containing four blue and six red cubes one is drawn at random, followed by another without replacing the first. Find the probability that two red cubes are drawn.

Solution:



Exercise 2.3

Construct a tree diagram for each question.

1. A counter is selected at random from a bag containing four green counters and five yellow counters, and then a second counter is selected from the bag without replacing the first counter that was taken.
   (a) Construct a tree diagram to represent the possible results.
   (b) Find the probability that both counters selected are green.
   (c) Find the probability that at least one of the counters selected is green.

2. A game involves a player randomly selecting one card then a second card from a set of six blue cards and four red cards, with replacement. Calculate the probability of a player selecting two cards of the same colour.

3. In Scotland, 41% of the population have the blood type *O positive (O+)*. A doctor in Scotland randomly selects two of their patients and asks for them to take part in a blood screening trial, as part of which their blood type will be checked. Calculate the probability that neither of the two patients has blood type *O+*.

4. A type of component manufactured for use in wind turbines is tested at two points in the production process, and marked as either *sufficient quality* or *insufficient quality* on the first testing point and then *pass* or *fail* at the second testing point. A component that is not of sufficient quality at the first testing point will continue to be developed and may still pass at the second testing point. Three-quarters of the components are marked as *sufficient quality* initially, of which nine-tenths go on to pass. Of those initially marked as *insufficient quality*, only three-fifths go on to pass.
   (a) Construct a tree diagram to represent this information.
   (b) Determine the probability that a component selected at random passes at the second testing point.

5. A bus for a particular route that is scheduled to run once per day is cancelled on 3% of days. On the days when it is running, it is running late 12% of the time.
   (a) Construct a tree diagram to represent this information.
   (b) Calculate the probability that the bus is running on time on any given day.

6. A pencil case contains pencils of three different types: four HB pencils, two 2B pencils and one 2H pencil. Two pencils are selected at random, without replacement.
   (a) Construct a tree diagram to represent the types of the two pencils selected.
   (b) Find the probability that both pencils selected are of the same type.
   (c) Determine the probability that the two pencils selected are both of a different type.
   (d) Find the probability that neither pencil selected is an HB pencil.

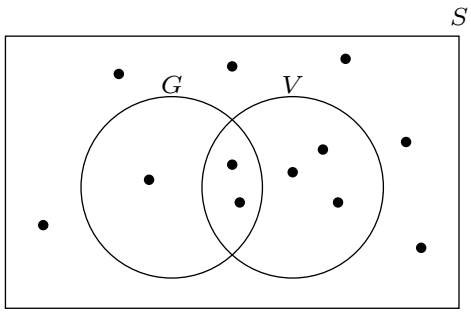
7. A motorist passes through three sets of traffic lights on their way to work each day. If the probability of having to stop at a red light for each set of lights is 0.6, 0.2 and 0.3 respectively then find the probability that:
   (a) The motorist does not have to stop.
   (b) The motorist has to stop at least once.
   (c) The motorist has to stop at exactly two of the sets of lights.

2.4 Venn Diagrams and Set Notation

Venn diagrams are used to show which elements of a set belong or don't belong to one or more subsets. They can offer an alternative way of representing information about a sample space that might otherwise be shown in a table. For instance, the table below shows the twelve dishes from a restaurant's main course menu, broken down by whether or not each is *gluten-free* (G) and whether or not each is *vegan* (V). Beside it is a Venn diagram showing the same information.

	vegan	not vegan
gluten-free	2	1
not gluten-free	3	6

Table



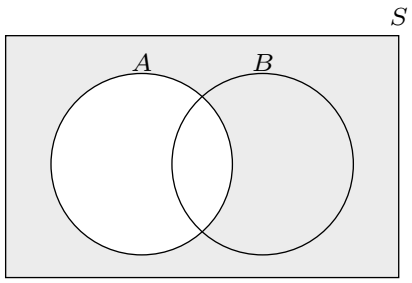
Venn diagram

Complements, Intersections and Unions

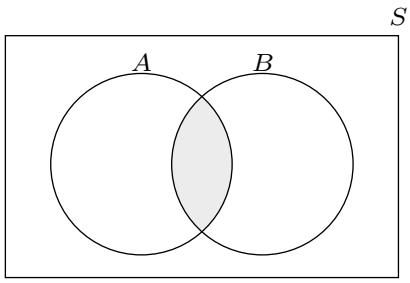
So far, *descriptive* words have been used to refer to different sets of outcomes. Using *set notation*:

Description	Set notation	Name
not A	$\overline{A}$ or A'	Complement of A
A *and* B	$A \cap B$	Intersection of A and B
A *or* B	$A \cup B$	Union of A and B

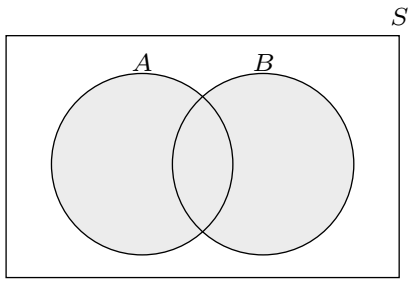
Venn diagrams help illustrate these. Note that *capital letters* are commonly used to represent events.



$\overline{A}$ or A'



$A \cap B$



$A \cup B$

Example

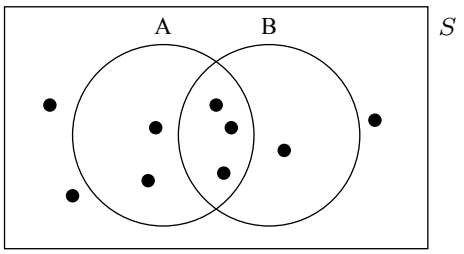
Problem: A dish is picked at random from the twelve on the main course menu shown in the Venn diagram above. State the value of $\mathrm{P}(G \cap V)$.

Solution:

$$\mathrm{P}(G \cap V) = \frac{2}{12} = \frac{1}{6}$$

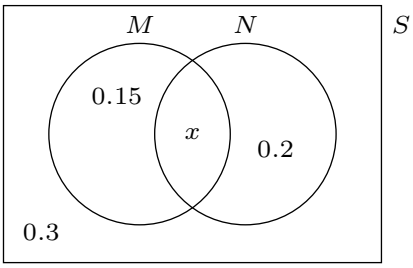
Exercise 2.4

1. One of the nine elements shown on the Venn diagram below is selected at random. State:



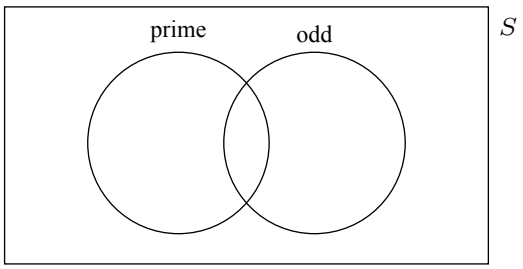
(a) P(A)
(b) P(B)
(c) P($\overline{A}$)
(d) P($\overline{B}$)
(e) P($A \cap B$)
(f) P($A \cup B$)
(g) P($\overline{A} \cap B$)
(h) P($A \cap \overline{B}$)

2. The Venn diagram below, containing events M and N, shows the *probabilities* associated with each subset.



(a) State the value of x.
(b) State:
   i. P(N).
   ii. P($\overline{M}$).
   iii. P($M \cup N$).

3. Copy the Venn diagram below, writing each of the integers from 1 to 16 inclusive in an appropriate position. Then, given that an integer from 1 to 16 inclusive is selected at random, state the value of:



(a) P(prime)
(b) P(odd)
(c) P($\overline{\text{prime}}$)
(d) P($\overline{\text{odd}}$)
(e) P(prime $\cap$ odd)
(f) P(prime $\cup$ odd)
(g) P($\overline{\text{prime}} \cap$ odd)
(h) P(prime $\cap \overline{\text{odd}}$)

4. 100 pupils are asked whether or not they love cats, and whether or not they love dogs, with the results shown in the table below. One pupil is to be selected at random. Let event C represent "a pupil who loves cats is selected", and event D represent "a pupil who loves dogs is selected". By first representing this information on a Venn diagram, find:

	love cats	do not love cats
love dogs	34	12
do not love dogs	45	9

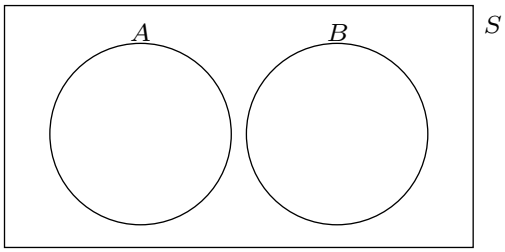
(a) P($C \cap D$)
(b) P($C \cup D$)
(c) P($\overline{C} \cap \overline{D}$)
(d) P($\overline{C} \cup \overline{D}$)
(e) P($C \cap \overline{D}$)
(f) P($\overline{C} \cap D$)
(g) P($C \cup \overline{D}$)
(h) P($\overline{C} \cup D$)

5. Amongst a class of 24 pupils, seven are studying biology *but not chemistry*, five are studying chemistry *but not biology*, and six are studying neither biology nor chemistry. Find the probability that a pupil chosen at random is studying biology.

6. Of the twenty air pollution monitoring stations in an urban area, nine are showing high levels of particulate matter (PM) in the air and eight are showing high levels of nitrogen dioxide (NO_2) in the air, whilst five are not showing high levels of either pollutant. Determine the probability that one of the twenty stations chosen at random is showing high levels of PM *and* NO_2.

2.5 The Addition Rule for the Union

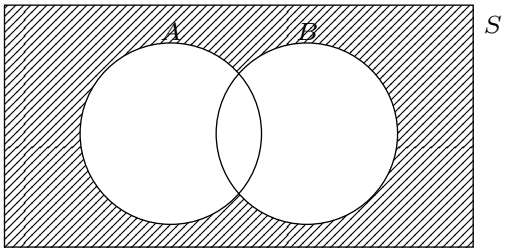
If two sets A and B are such that *no elements lie in both A and B*, then they are said to be **mutually exclusive**. This is sometimes visualised on a digram by two *disjoint* circles, that do not intersect. In terms of probability theory, this means that the probability of the event "*A and B*" is 0 (impossible).



For Mutually Exclusive events A and B:

$$\mathrm{P}(A \cap B) = 0$$

If two sets A and B are such that *every element lies in at least one of A or B*, then they are said to be **exhaustive**. This can be visualised on a Venn digram by crossing out the area of the sample space outside of the two circles. This means that the probability of the event "*A or B*" is 1 (certain).



For Exhaustive events A and B:

$$\mathrm{P}(A \cup B) = 1$$

As well as the two rules above which apply depending on whether A and B are know to be mutually exclusive, exhaustive, or both, the two formulae below apply for *all events A, B*:

The Complement of A:

$$\mathrm{P}(A) + \mathrm{P}(\overline{A}) = 1$$

The Addition Rule for events A and B:

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

It should be noted that there are generally numerous different ways to approach probability problems, including using a Venn diagram, a tree diagram or formulae. Whilst the problems on the next page have been designed to allow practice using formulae, a flexible approach may help.

Example

Problem: Given events F and G such that $\mathrm{P}(\overline{F}) = 0.3$, $\mathrm{P}(G) = 0.4$ and $\mathrm{P}(F \cap G) = 0.1$ show that F and G are exhaustive.

Solution:

$$\begin{aligned} \mathrm{P}(F \cup G) &= \mathrm{P}(F) + \mathrm{P}(G) - \mathrm{P}(F \cap G) \\ &= 0.7 + 0.4 - 0.1 \\ &= 1 \quad \therefore \text{ events } F \text{ and } G \text{ are exhaustive} \end{aligned}$$

Exercise 2.5

1. Given events X and Y such that $\mathrm{P}(X \cap Y) = 0.3$, $\mathrm{P}(X) = 0.4$ and $\mathrm{P}(Y) = 0.5$, find $\mathrm{P}(X \cup Y)$.

2. If $\mathrm{P}(\overline{A}) = \dfrac{1}{5}$, $\mathrm{P}(B) = \dfrac{2}{3}$ and $\mathrm{P}(A \cap B) = \dfrac{1}{2}$, determine the value of $\mathrm{P}(A \cup B)$.

3. Calculate the value of $\mathrm{P}(E \cup F)$ given $\mathrm{P}(E) = 0.32$, $\mathrm{P}(F) = 0.29$ and that events E and F are *mutually exclusive*.

4. Given mutually exclusive events C and D such that $\mathrm{P}(\overline{C}) = \dfrac{5}{7}$ and $\mathrm{P}(D) = \dfrac{1}{2}$, find $\mathrm{P}(C \cup D)$.

5. Calculate the value of $\mathrm{P}(A \cap B)$ if $\mathrm{P}(A) = 0.591$, $\mathrm{P}(B) = 0.472$ and $\mathrm{P}(A \cup B) = 0.9$.

6. If $\mathrm{P}(\overline{X}) = \dfrac{17}{80}$, $\mathrm{P}(\overline{Y}) = \dfrac{29}{40}$ and $\mathrm{P}(X \cup Y) = \dfrac{19}{20}$, determine $\mathrm{P}(X \cap Y)$.

7. Given *exhaustive* events V and W such that $\mathrm{P}(V) = 0.4$ and $\mathrm{P}(\overline{W}) = 0.2$, find $\mathrm{P}(V \cap W)$.

8. Given events H and J such that $\mathrm{P}(H) = \dfrac{3}{8}$, $\mathrm{P}(H \cup J) = \dfrac{7}{9}$ and $\mathrm{P}(H \cap J) = \dfrac{1}{2}$, find $\mathrm{P}(\overline{J})$.

9. If X and Y are mutually exclusive, $\mathrm{P}(X) = 0.3$ and $\mathrm{P}(\overline{Y}) = 0.4$, calculate $\mathrm{P}(X \cup Y)$.

10. Events A and B are exhaustive. If $\mathrm{P}(A) = \dfrac{1}{3}$ and $\mathrm{P}(B \cap A) = \dfrac{1}{6}$, calculate $\mathrm{P}(\overline{B})$.

11. In a fish and chip shop, the probability that a customer's order includes both fish and chips is 0.7, whilst the probability that an order will at least include at least one of fish or chips is 0.95. If the probability that an order contains chips is 0.8, calculate the probability that an order contains fish.

12. If, for exhaustive events M and N, $\mathrm{P}(M \cap N) = \dfrac{1}{4}$ and $\mathrm{P}(M) = \mathrm{P}(N)$, calculate $\mathrm{P}(\overline{N})$.

13. Given events R and Q such that $\mathrm{P}(\overline{Q}) = 0.11$, $\mathrm{P}(R \cap Q) = 0.75$ and $\mathrm{P}(R \cup Q) = 0.95$, find $\mathrm{P}(\overline{R})$.

14. For each pair of events, determine whether they are exhaustive or not:
    (a) X and Y given $\mathrm{P}(X) = 0.4$, $\mathrm{P}(Y) = 0.7$ and $\mathrm{P}(X \cap Y) = 0.2$.
    (b) R and T given $\mathrm{P}(R) = \dfrac{4}{5}$, $\mathrm{P}(\overline{T}) = \dfrac{1}{10}$ and $\mathrm{P}(T \cap R) = \dfrac{7}{10}$.
    (c) E and F given $\mathrm{P}(E) = 0.31$, $\mathrm{P}(F) = 0.82$ and $\mathrm{P}(E \cap F) = 0.28$.

15. For each pair of events, determine whether they are mutually exclusive or not:
    (a) W and V given $\mathrm{P}(W) = 0.35$, $\mathrm{P}(V) = 0.45$ and $\mathrm{P}(W \cup V) = 0.5$.
    (b) X and Y given $\mathrm{P}(X) = \dfrac{1}{3}$, $\mathrm{P}(\overline{Y}) = \dfrac{4}{5}$ and $\mathrm{P}(X \cup Y) = \dfrac{8}{15}$.
    (c) A and B given $\mathrm{P}(\overline{A}) = 0.5$, $\mathrm{P}(\overline{B}) = 0.8$ and $\mathrm{P}(\overline{B} \cap \overline{A}) = 0.3$.

Review Exercise

1. A box contains five red balls numbered from 1 to 5, and four yellow balls numbered from 1 to 4. A ball is selected at random from the box. Let *yellow* be the event "the ball selected is yellow".

   (a) State P($\overline{\text{yellow}} \cup 1$).

   The ball previously selected is returned to the box. Two balls are then randomly taken from the box, without replacement.

   (b) Determine the probability that neither of the two balls taken from the box are yellow.

2. The table below shows the pupil numbers within a school corridor during a particular period, broken down by subject and course level.

	National 5	Higher
English	21	15
History	17	11

   A pupil is selected at random to run an errand. State:

   (a) P(English $\cap$ National 5)

   (b) P(History $\cup$ Higher)

3. For the exhaustive events E and F, such that P$(F \cap E) = 0.7$ and P$(\overline{E}) = 0.2$, determine P(F).

4. A parking ticket machine in a remote location in the Cairngorms National Park takes payment by contactless card reader, which connects to the banking network via a 3G phone signal. One some days, the connection fails and the machine cannot take payment. When the humidity is high, the connection will fail on 18% of days. When the humidity is not high, the connection will fail on 7% of days. Humidity is high in the area on 34% of days. Determine the probability on a randomly selected day of the year that the parking machine can take payment.

5. A board game begins with an *defending* player rolling a tetrahedral die with sides numbered from 1 to 4, and a *attacking* player rolling a cubical die with sides numbered from 1 to 6. The player who obtains the highest number earns a point, with a tie leading to the dice being rolled again.

   (a) Find the probability that the the first point is earned, by either player, without the dice having to be rolled again following a tie.

   (b) Find the probability that the attacking player earns the first point.

   If the attacking player wins three points in a row, they are then awarded a *bonus roll* in which they roll *both* dice and gain a bonus point if the product of the numbers obtained is greater than 16.

   (c) Determine the probability that a bonus roll results in a bonus point.

6. The 100 members of an orchestra are shown in the table below, broken down by the section they are in and whether they can attend the upcoming concert.

	String	Woodwind	Brass	Percussion
attend	56	12	10	4
not attend	8	4	5	1

   A member of the orchestra is selected at random. Let B be the event the member selected is in the *brass* section and A that they can *attend* the upcoming concert. State P$(B \cup \overline{A})$.

7. Exhaustive events A, B and C are such that P$(A) = 0.25$, P$(A \cap B) = 0.1$, P$(\overline{B}) = 0.3$ and P$(A \cup C) = 0.6$. Events A and C are mutually exclusive. Determine the value of P$(B \cap C)$.

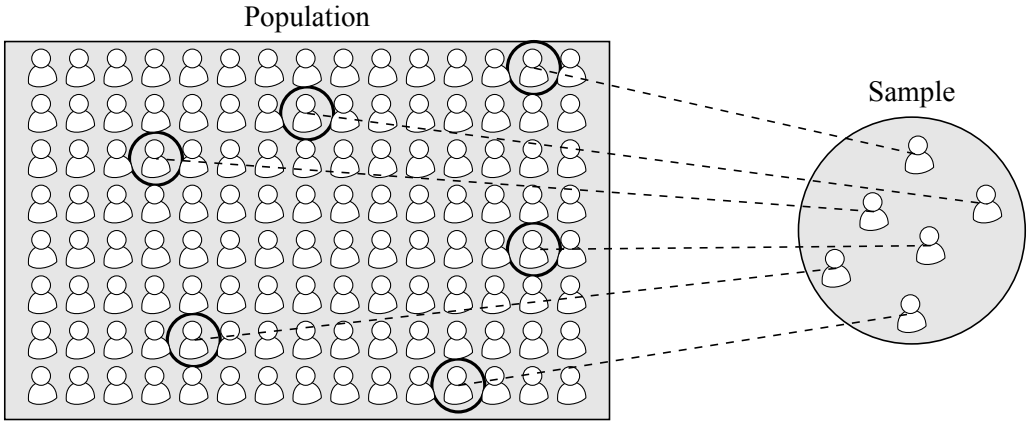
3

Sampling Theory

"Garbage In, Garbage Out" is a phrase used in computer science that may equally be applied to statistics. It should be discouraged to think that the task of a statistician begins with raw data; the manner of the collection of the data itself is of fundamental importance. Any statistical inferences made on the basis of data that has not be collected with an appropriate *sampling method* should not be trusted.

Populations and Samples

To recap from chapter 1, the term **population** refers to the entire set of objects in which we are interested. A **sample** is a selection of some objects chosen from the population.



A **census** is used to collect information from a full population and a **sample survey** is used to collect data on a sample.

It is usually very difficult to carry out a census as it is time consuming, expensive, and requires an accurate and complete list of every member of the population. In some cases, investigating an entire population would destroy it all, such as the burn time of candles. It is therefore usually preferable that a sample survey is carried out and, if representative of the population, it should give an accurate indication of the characteristics of the population.

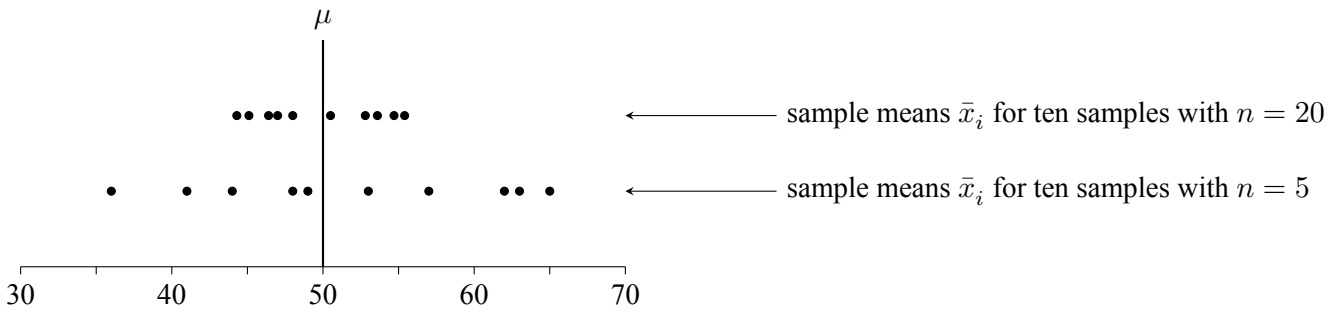
3.1 Sampling Considerations

Samples are used to infer information about the populations from which they are taken, and sample data collected using higher quality sampling techniques should tend to lead to more reliable inferences. There are several key considerations and pitfalls to be aware of whenever the collection of sample data is being planned.

Sample Size

Parameters, such as the *population mean*, are generally considered as fixed, but usually unknown values. If samples are repeatedly taken from a population and their means are are obtained, the values of those sample means will sometimes be higher than that of the true population mean and sometimes lower. This comes from *random sampling error* which, despite its name, is a natural effect due to chance variation.

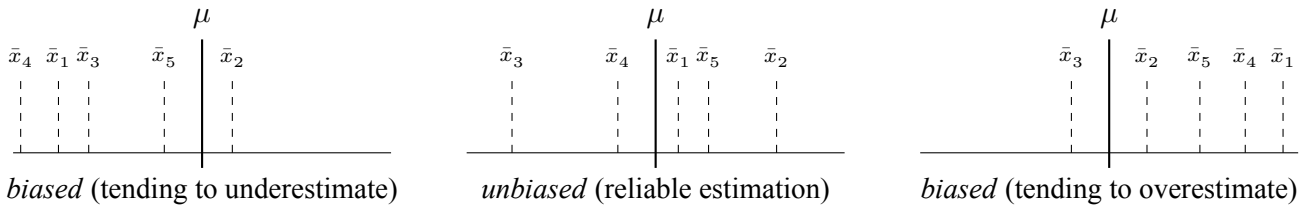
The figure below shows two sets of sample means for random samples taken from the same distribution with a population mean of $\mu = 50$. For one set a sample size of $n = 5$ was used, with $n = 20$ for the other.



It can be seen that larger sample sizes tend to produce sample statistics that better represent the population parameters they are used to estimate. Larger samples are therefore generally preferred, and statistical studies conducting using only small samples of data should typically be treated with more caution than those based on larger samples. When small samples are used, the reasons are usually one or more of cost, time, and availability of data.

Bias

Poor sampling technique may lead to samples that are not *representative* of the population from which they are drawn. The most common effect of this is that *bias* is introduced: a tendency for sample statistics calculated from these samples to overestimate, or underestimate, the population parameter they aim to represent. The diagrams below illustrate the effects various sampling techniques may have on sample means obtained from repeated sampling; the techniques used for the left and right diagrams are causing bias, whilst the sample means obtained for the middle diagram are *unbiased*.



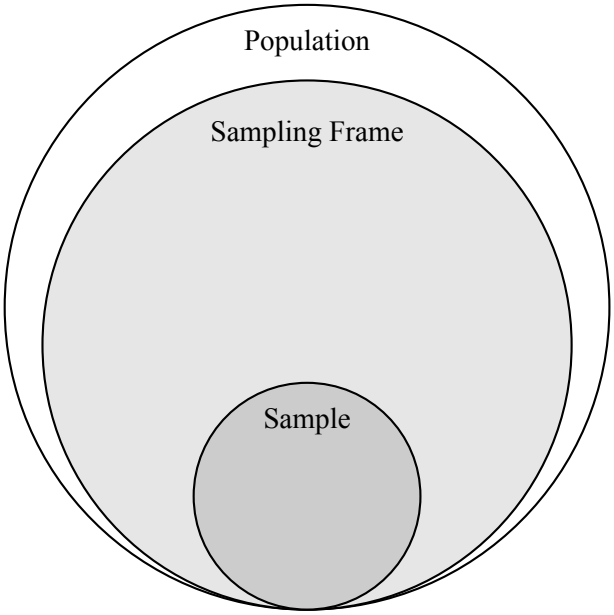
biased (tending to underestimate) *unbiased* (reliable estimation) *biased* (tending to overestimate)

It is very difficult to identify that bias has been introduced, and effectively impossible to correct it. The only reliable way to avoid introducing bias is through application of *appropriate random sampling techniques*.

Sampling Frames

Any real-life experience of sampling for any kind of study is likely to quickly reveal that it is not quite as straight-forward as *"just choosing at random"*. Several key elements must be identified and problems considered.

For example, consider the context of wishing to use registration time one morning to survey a sample of pupils at a large secondary school for their views on the school cafeteria.



The **target population** is the *group being studied*.

e.g. all of the pupils who attend a school

The **sampling frame** is the part of the target population, of size N, from which a sample can be taken, typically in the form of an ordered list.

e.g. a numbered list of the school roll

The **sample** consists of the subset of the sampling frame selected to be studied, with size n.

e.g. those pupils selected to be surveyed

Sampling units, which make up the sampling frame, may be individual people, buildings, measurements taken at different times, square grids on a map, and so on.

The *sampling method* is the way in which sampling units are selected from the sampling frame to create the sample.

Care should be taken to ensure that sampling frames are representative of the target population. For example, if walking habits are being studied, a sampling frame containing only residents of Glasgow city centre is likely to be *not representative* of an intended target population of people living in Scotland.

The impact to a study of an ill-defined target population, an unrepresentative samplng frame or a poor sampling technique is likely to be data which is *biased* and *not representative* of the population, and *unreliable conclusions*.

Exercise 3.1

1. A researcher wants to study the shopping habits of those who use his local supermarket. To collect data, he plans to stand at the exit of the self-checkout area one Sunday morning and ask shoppers if he can take a photo of their receipt. Since he will not have time to ask everyone, he will pick only those he feels are most likely to agree to this.
   (a) Identify:
      i. The target population.
      ii. The sampling frame.
   (b) Suggest two problems with the sampling frame chosen, and explain the impact this may have on their data.
   (c) Comment on the suitability of the sampling method the researcher plans to use.
2. A large field is full of many wildflowers. A botanist wishes to conduct a survey to assess the health and variety of species of the wildflowers. A random number generator will be used to help obtain a sample.
   (a) Identify one problem with a proposal to use individual flowers as sampling units, to make up a numbered sampling frame from which to draw a sample.
   (b) Suggest an alternative sampling unit for the botanist to use instead.
   (c) Explain the purpose of a random number generator in producing a sample.

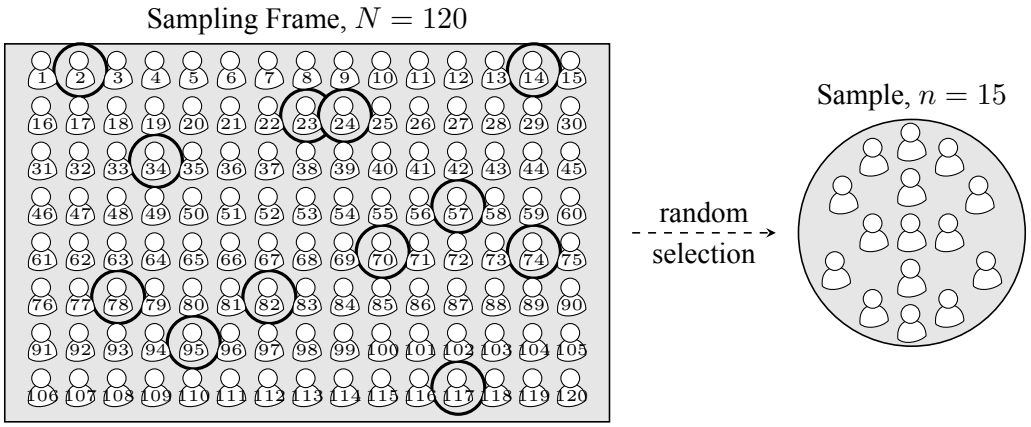
3.2 Six Sampling Methods

In this course, knowledge is required of *four random samplng methods* and *two non-random sampling methods*.

Simple Random Sampling (Random)

With *simple random sampling* each unit of the sampling frame is *equally likely* to be selected, such as by drawing names out of a hat. In general, a simple random sample of size n can be selected from a population of size N by numbering the units in the sampling frame from 1 to N, then selecting n unique random numbers using a random number generator. Those units in the sampling frame corresponding to those numbers generated are then selected to form the sample.

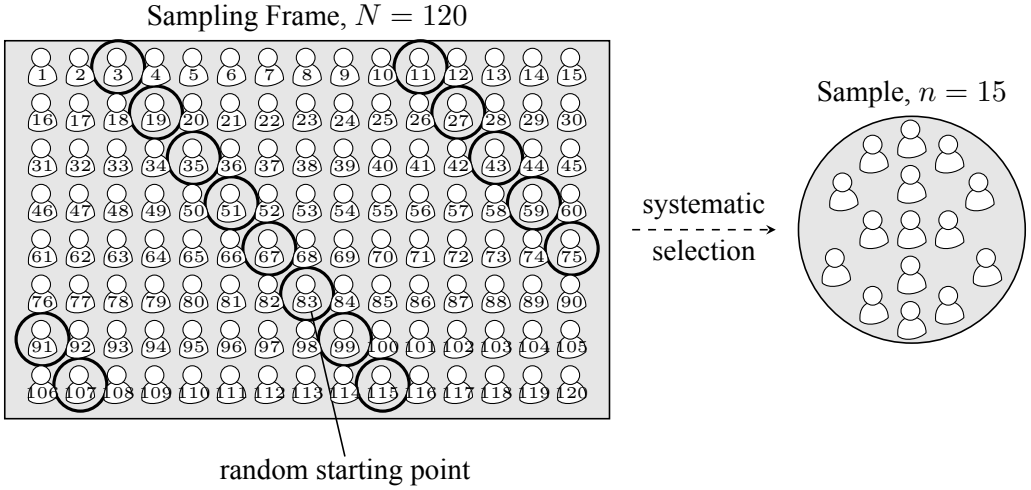
For example, simple random sample of 15 pupils from a year group of 120 can be made by numbering a list of pupils from the year group from 1 to 120, then selecting 15 random numbers using a random number generator.



Systematic Sampling (Random)

Systematic sampling involves using an orderly *pattern* to select units, starting from a randomly selected point. In general, for a numbered sampling frame of size N, a random number generator may select a number from 1 to N, with the corresponding sampling unit selected. After that initial selection, *every subsequent* $\frac{N}{n}th$ *unit is selected*, looping back to the start of the sampling frame as necessary to complete the sample of size n.

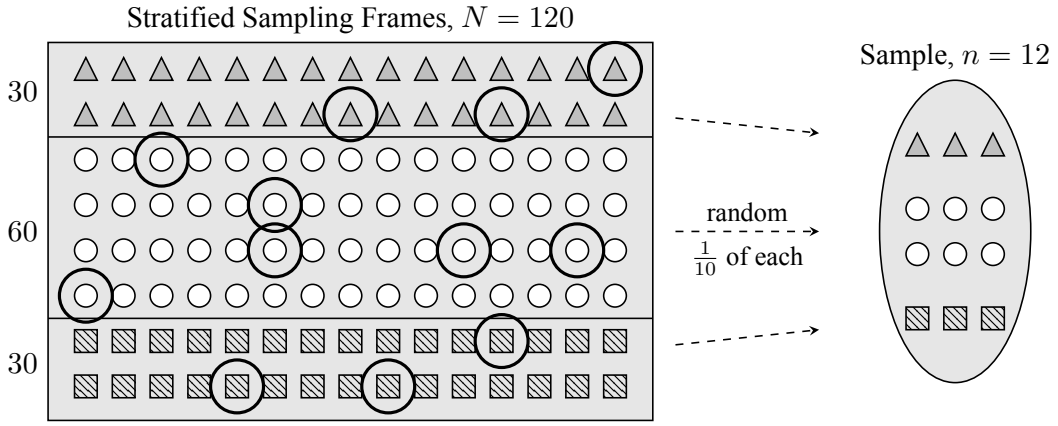
For example, a systematic sample of 15 pupils from a year group of 120 can be made by first selecting a random starting point, such as the $83rd$ pupil, in an alphabetised list of pupils in the year group. From there, every $\frac{120}{15} = 8th$ pupil from the list is chosen, looping back round at the end of the list until 15 pupils have been chosen.



Stratified Sampling (Random)

If a target population contains distinct *strata*, where each stratum may have its own distinct characteristics, *stratified sampling* ensures that each strata will be *proportionally represented*. In general, this involves obtaining an ordered sampling frame for *each stratum* and using the proportions of each stratum in the population to determine how many to sample from each, using simple random sampling. Finally, these samples from the strata are combined.

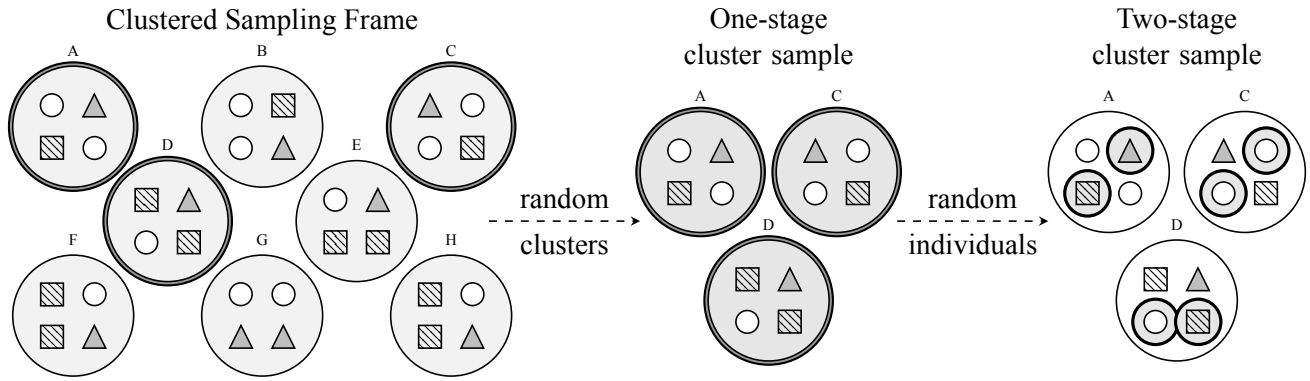
For example, a hospital wishing to survey its staff about the food in the staff canteen may split the $N = 120$ staff into management, doctors/nurses and support staff. For a sample of size $n = 12$, covering $\frac{1}{10}th$ of the sampling frame, simple random sampling is used to select $\frac{1}{10}$ of the staff *from each group*. These are then collected to form the stratified sample.



Cluster Sampling (Random)

If a target population is split into clusters, and each cluster represents a similar mix of characteristics to the full population, *cluster sampling* may be useful. For a *one-stage* cluster sample, one or more clusters are randomly selected, from a sampling frame of clusters, and then *all individual* within those clusters form the sample. For a *two-stage* cluster sample, an additional level of sampling takes place to only sample *some* of the individuals within each cluster.

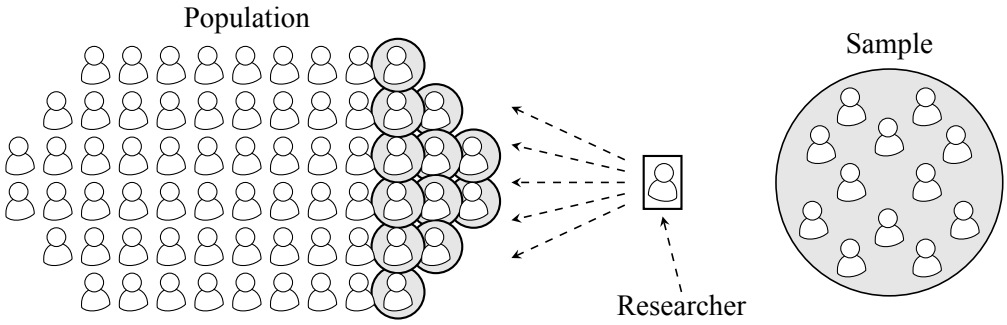
For example, a hospital wishing to survey its staff about the food in the staff canteen may recognise that staff are split into hospital wards A to H. Several wards are randomly selected (A, C and D) and each member of staff working there surveyed. For a two-stage sampling method, simple random sampling could be used to select only *some* of the ward staff from the selected clusters to sample rather than everyone.



The following sampling methods are *non-random*. Whilst they may seem helpful in some instances for quickly gathering data, their lack of random process may lead to *bias*, and any inferences made using such sampling methods may be considered unreliable.

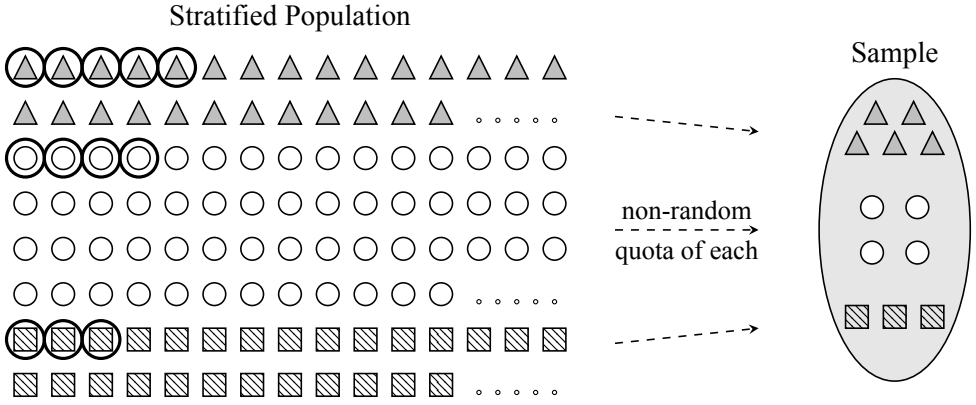
Convenience Sampling (Non-random)

Taking a sample of a target population taken from an easily obtainable group, with no random process taking place, is called *convenience sampling*. For example, a teacher wishes to investigate the heights of pupils in a particular year group and uses pupils within their class in that year group as the sample.



Quota Sampling (Non-random)

Where a population contains distinct strata, *quota sampling* sets a target number of individuals from within each stratum to sample. It differs from stratified sampling in that the number sampled from each stratum is according to a set *quota* rather than proportional to their size, and the selection is often *non-random*. This means that ordered *sampling frames are not required* for each stratum, and the proportion that each makes up within the population may not be known.



For example, a council wishing to know local opinion of proposed changes to a park may see the population as split into strata including pensioners, teenagers, parents of pre-school children, and so on. Interviewer are placed on the town's main shopping street and speak to people who pass, choosing for themselves who to interview from their allocated strata.

Advantages and Disadvantages

The possible advantages and disadvantages of sampling methods depend on the context in which they are used. However, desired *advantages* of a sampling method may include it being: cheap to administer; quick to administer; free from bias due to the use of a random process; representative of the full population; easy to obtain a suitable sampling frame for. Conversely, *disadvantages* may include it being: expensive; time-consuming; non-random, possibly introducing bias; not representative of the full population; difficult to obtain a suitable sampling frame for.

The method of collecting data may also impact the quality of the data obtained. For example, *emailing* selected individuals with a questionnaire to complete may introduce bias by excluding those who don't use email and allowing *self-selection*.

Exercise 3.2

1. For each, state the sampling method that is being described.
   (a) The population is made up of groups, each of which contains broadly similar characteristics to each other, and the population. One or more of these groups are randomly selected, and everyone in those selected groups are chosen to make up the sample.
   (b) It is recognised that the population contains a number of subgroups, each of which may have their own distinct characteristics. Individuals are randomly selected from each subgroup, such that the proportions of each subgroup in the sample match those proportions of the population.
   (c) Once a first random individual from the popualtion has been selected, subsequent individuals are selected at regular intervals until a sample has been obtained.
   (d) Individuals are selected from the population at random, with each having an equal chance of being selected.
   (e) It is recognised that the population contains a number of subgroups, each of which may have their own distinct characteristics. Ten individuals from each subgroup are selected, without the use of a random process.
   (f) The individuals most readily available are selected.

2. The operators of a cruise ship wish to survey the condition of the fittings in the cabins on board by carefully examining a sample of rooms. Of the 360 cabins, 24 are classed as *Deluxe*, 72 are *Premuim* and the remaining are *Standard*.
   (a) Explain why the ship's operators might wish to use stratified sampling for their survey.
   (b) Describe the steps involved in sampling 30 of the ship's cabins.

3. A marine biologist is interested in the health of the coral reefs on the West Coast of Scotland. On his next scuba dive near the coastal town in which he lives, he takes some time to photograph the animals and plants on and around the coral reef surface. Based on the survey, he concludes that the health of coral reefs on the West Coast of Scotland has deteriorated.
   (a) State the type of sampling method used by the marine biologist.
   (b) Suggest one reason why the conclusion he reached may not be reliable.

4. The owners of a nationwide chain of petrol stations wish to measure the accuracy of fuel delivery by its pumps. They own 126 petrol stations, each containing typically 8 pumps, and they wish to sample approximately 40 pumps.
   (a) State one advantage of using cluster sampling for the survey.
   (b) Describe the steps involved in creating a one-stage cluster sample of 40 pumps.

5. A music lover has a collection of vinyl records, arranged along a single shelf across one wall. Concerned that some of the actual records may be missing from their sleeves, she checks the one on the far left of the shelf and then every 10th record until she reaches the end of the shelf.
   (a) State the sampling method which most closely matches the method used.
   (b) Identify one way in which the steps taken do not match the ideal process of the sampling method stated in part (a).

6. A school wishes to survey the views of pupils, parents and staff on their opinions of a proposed new school emblem design. Wanting to ensure each group is included, during the annual Sports Day questionnaires are given to twelve of the pupils and twelve of the parents standing in the queue to register, and to the eight members of staff helping run the event.
   (a) State the sampling method used.
   (b) State one disadvantage of this sampling method and describe the impact this may have on the data obtained.

Review Exercise

1. The owners of a chain of nine restaurants, located in various cities around Scotland, wish to interview a sample of the employees working in those restaurants.

   (a) One of the owners suggests that he asks to interview any employees that he sees when he dines at one of the restaurants next week.

      i. State the type of sampling method that he is suggesting the owners use.

      ii. State one disadvantage of this sampling method.

   (b) Another owner proposes instead that they could type the name of each of the 150 employees into a spreadsheet and use a random cell selection tool to pick those to be interviewed.

      i. State the type of sampling method that she is proposing.

      ii. Describe how the owners could instead use systematic sampling to select 30 employees.

   (c) It is pointed out by the HR Director that the employees can be broken down into three categories, each of which may have their own distinct views on what it is like to work in the restaurant chain:

   - Managers and supervisors.
   - Kitchen staff.
   - Waiting and bar staff.

   Each of the restaurants contains a similar mix of the different types of staff.

      i. Describe how a one-stage cluster sample could be conducted.

      ii. Suggest one advantage and one disadvantage of using this method, compared to using systematic sampling.

2. A company employs the following personnel:

Management	Sales staff	Shop floor staff
100	400	4500

   Explain how a stratified sample of 150 employees could be obtained, ensuring that each type of employee is proportionally represented.

3. A nutritionist is studying the calorific contents of pizzas sold in the UK, having read that the salt content is much higher than it is in Italy. They obtain a list of the takeaways which sell pizza in their local area by searching on social media, and randomly select twelve for their study. Suggest two reasons why the sampling frame may not be sufficiently representative of the target population.

4. A researcher wishes to interview a random sample of pupils taking Higher Mathematics to investigate the proportion of them that are interested in pursuing a career in a mathematical field.

   (a) Suggest a reason why taking a simple random sample of all pupils taking Higher Mathematics in Scotland would not be feasible.

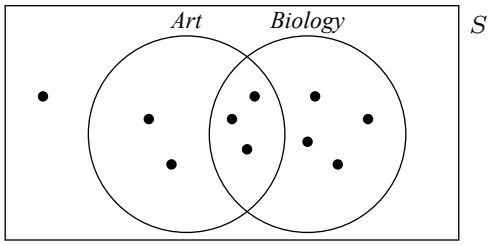
   (b) Explain how a two-stage cluster sample could be taken.

5. A wildflower meadow contains many different species of wildflowers growing together. A botanist wishes to assess the health of the wildflowers in the meadow by studying a sample of the flowers. They have access to a map of the meadow, broken down into boxes with grid references such as A7, and B2. Explain how the botanist could use simple random sampling to obtain a sample.

4

Further Probability Theory

Suppose ten pupils in a class are asked whether they are taking Art, and whether they are taking Biology, and the results are displayed in both a Venn diagram and a table.



	Biology	$\overline{\text{Biology}}$
Art	3	2
$\overline{\text{Art}}$	4	1

From both the table and the Venn diagram, it can be seen that, of the ten pupils, five are taking Art. Therefore, the probability of randomly selecting a pupil from the class and that pupil turning out to be taking Art can be stated:

$$\text{P}(\text{Art}) = \frac{5}{10} = \frac{1}{2}$$

Now suppose a pupil is chosen at random and, before it is known whether or not they are taking Art, they reveal that they are taking Biology. The probability that the pupil is also taking Art is no longer $\frac{1}{2}$, since only three of the seven pupils taking Biology are taking Art. The probability now being sought is the probability that they are taking Art *given that they are taking Biology*. This is called a *conditional probability*, and can be notated as:

$$\text{P}(\text{Art} \mid \text{Biology}) = \frac{3}{7}$$

Conditional probability is fundamental to many of the statistical analysis techniques that will be introduced in this course.

4.1 Conditional Probability from Tables and Diagrams

Example

Problem: A medical trial for drug treating a minor illness has patients given either the drug or a placebo, with a test after one month of treatment determining whether or not each patient has fully recovered.

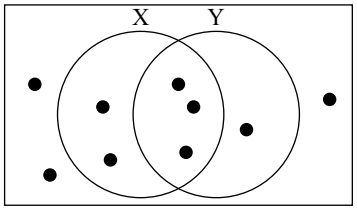
	drug	placebo
fully recovered	42	18
not fully recovered	8	12

A patient is selected at random to be interviewed. State P(fully recovered | placebo).

Solution: $\text{P(fully recovered} \mid \text{placebo)} = \dfrac{18}{30} = \dfrac{3}{5}$

Exercise 4.1

1. The Venn diagram below shows equally likely outcomes in a sample space and events X and Y. State:



(a) $\text{P}(X)$ (e) $\text{P}(Y)$

(b) $\text{P}(X|Y)$ (f) $\text{P}(Y|X)$

(c) $\text{P}(X \cap Y)$ (g) $\text{P}(Y \cap X)$

(d) $\text{P}(X \cup Y)$ (h) $\text{P}(Y \cup X)$

2. A basketball coach records the number of 2-pointer, 3-pointer and free-throw scoring attempts over the course of a season, broken down by whether those attempts scored or missed. An attempt is chosen at random to be analysed. Find:

(a) P(scored) (c) P(3-pointer)

(b) P(scored | 3-pointer) (d) P(3-pointer | scored)

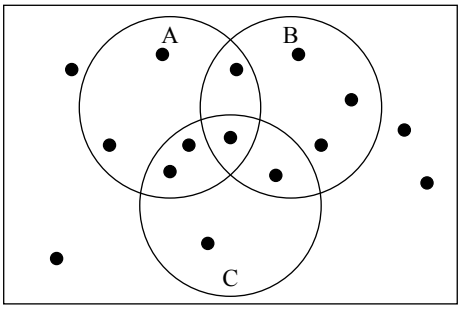
	2-pointer	3-pointer	free throw
scored	72	30	44
missed	60	52	12

3. On mornings when it is raining, the probability that a red squirrel will leave its drey is 0.4. On mornings when it is not raining, the probability that it will leave its drey is 0.9. It is rainy on 30% of mornings. State the value of:

(a) P(rainy) (c) P(leaves | rainy) (e) P(leaves | not rainy)

(b) P(not rainy) (d) P(does not leave | rainy) (f) P(does not leave | not rainy)

4. The Venn diagram below shows equally likely outcomes in a sample space and events A, B and C. Determine:



(a) $\text{P}(A)$ (g) $\text{P}(A|C)$

(b) $\text{P}(B)$ (h) $\text{P}(C|B)$

(c) $\text{P}(C)$ (i) $\text{P}(\overline{B}|C)$

(d) $\text{P}(A \cap B)$ (j) $\text{P}(C|\overline{A})$

(e) $\text{P}(B \cup C)$ (k) $\text{P}((A \cap B)|C)$

(f) $\text{P}(A \cap B \cap C)$

4.2 The Conditional Probability Formula

Where a conditional probability cannot be intuitively stated, the formula for conditional probability may be used:

Conditional Probability:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

It may be helpful to note that $\mathrm{P}(\overline{A}|B) = 1 - \mathrm{P}(A|B)$.

Example

Problem: Given events V and W such that $\mathrm{P}(V) = 0.7236$ and $\mathrm{P}(V \cap W) = 0.1046$, find $\mathrm{P}(W|V)$.

Solution: $\mathrm{P}(W|V) = \dfrac{\mathrm{P}(W \cap V)}{\mathrm{P}(V)} = \dfrac{0.1046}{0.7236} = 0.1446$

Exercise 4.2

1. Find, by first writing out the conditional probability formula for these events:
   (a) $\mathrm{P}(A|B)$ given $\mathrm{P}(B) = 0.8$ and $\mathrm{P}(A \cap B) = 0.16$.
   (b) $\mathrm{P}(\text{red} \mid \text{blue})$ given $\mathrm{P}(\text{red} \cap \text{blue}) = 0.1526$, $\mathrm{P}(\text{blue}) = 0.3147$ and $\mathrm{P}(\text{red}) = 0.2142$.
   (c) $\mathrm{P}(X|Y)$ given $\mathrm{P}(Y) = \frac{1}{2}$ and $\mathrm{P}(Y \cap X) = \frac{1}{3}$.
   (d) $\mathrm{P}(M|N)$ given $\mathrm{P}(\overline{N}) = 0.6137$, $\mathrm{P}(N \cap M) = 0.1249$ and $\mathrm{P}(M \cup N) = 0.4511$.
   (e) Using (a) to (d), state:

   i. $\mathrm{P}(\overline{A}|B)$ ii. $\mathrm{P}(\overline{\text{red}} \mid \text{blue})$ iii. $\mathrm{P}(\overline{X}|Y)$ iv. $\mathrm{P}(\overline{M}|N)$

2. Given events V and W such that $\mathrm{P}(V \cap W) = 0.28$ and $\mathrm{P}(W) = 0.4$, find $\mathrm{P}(\overline{V}|W)$.

3. Numbered cards of different colours are selected at random from a pack during a game such that the probability of a selected card being *yellow* is 20%, and the probability of it being *yellow and a five* is 4%. Find the probability that a selected card is a five *given that it is a yellow*.

4. A football team knows from looking at their past results that they concede at least two goals in $\frac{2}{5}$ of all games. In $\frac{1}{4}$ of all of their games they concede at least two goals and lose the game. Find the probability that, given that they concede at least two goals in a game, they lose.

5. Given events M and N such that $\mathrm{P}(M) = 0.6$, $\mathrm{P}(N) = 0.4$ and $\mathrm{P}(M \cup N) = 0.9$, find $\mathrm{P}(M|N)$.

6. For events V and W, $\mathrm{P}(V) = 0.6$, $\mathrm{P}(V|W) = 0.8$ and $\mathrm{P}(V \cap W) = 0.5$. Calculate $\mathrm{P}(V \cup W)$.

4.3 Independent Events

The probability of tossing a coin and it landing tails is $\frac{1}{2}$. The probability of it landing tails *given that it is a weekday*, notated as P(tails | weekday), is *still* $\frac{1}{2}$. Since it being a weekday or not makes no different to the probability of a coin showing tails, the two events are said to be *independent*.

The probability a randomly selected card from a standard pack of playing cards being a Diamond is $\frac{1}{4}$. The probability the card is a Diamond *given that we discover the card picked is red*, notated as P(Diamond | red), is $\frac{1}{2}$. Since knowing the card picked is red changes the probability of it being a Diamond, the two events are said to be *not independent*.

Independent:

$$\mathrm{P}(A|B) \quad = \quad \mathrm{P}(A)$$

Not Independent:

$$\mathrm{P}(A|B) \quad \neq \quad \mathrm{P}(A)$$

Example

Problem: Given $\mathrm{P}(D) = 0.25$, $\mathrm{P}(E) = 0.6$ and $\mathrm{P}(D \cap E) = 0.15$, show that E and D are independent.

Solution:

$$\mathrm{P}(D|E) = \frac{\mathrm{P}(D \cap E)}{\mathrm{P}(E)} = \frac{0.15}{0.6} = 0.25$$

Since $\mathrm{P}(D|E) = \mathrm{P}(D)$, events D and E are independent.

Exercise 4.3

1. For each pair of events X and Y, state whether you would expect them to be independent or not.

   (a) A person is selected at random in a survey:
   - X: They were born in January.
   - Y: They work in an office.

   (b) A mechanic is inspecting a car's tyre treads:
   - X: The front tyres have low tread.
   - Y: The rear tyres have low tread.

2. For each pair of events, determine whether or not they are independent.

   (a) A and B given:
   - $\mathrm{P}(A) = 0.3$
   - $\mathrm{P}(B) = 0.8$
   - $\mathrm{P}(A \cap B) = 0.16$

   (b) X and Y given:
   - $\mathrm{P}(X) = \frac{3}{4}$
   - $\mathrm{P}(Y) = \frac{7}{12}$
   - $\mathrm{P}(X \cap Y) = \frac{7}{16}$

   (c) M and N given:
   - $\mathrm{P}(M \cup N) = 0.67$
   - $\mathrm{P}(\overline{M}) = 0.55$
   - $\mathrm{P}(M \cap N) = 0.18$

3. A music tutor taught 40 students last year, each learning one of violin, trumpet or saxophone, and each have lessons either midweek or at the weekend. A student is chosen at random to take a survey to give feedback. Let V represent that the student chosen took violin lessons, T that they took trumpet lessons and W that they had lessons at the weekend.

	violin	trumpet	saxophone
midweek	12	11	9
weekend	3	4	1

   (a) Show that V and W are independent.

   (b) Show that T and W are not independent.

4.4 The Multiplication Rule for the Intersection

The conditional probability rule can be rearranged to allow the **intersection** of two events to be calculated using a conditional probability. For independent events, since $\mathrm{P}(B|A) = \mathrm{P}(B)$, this simplifies further.

Multiplication rule:

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B|A)$$

Multiplication rule for independent events:

$$\mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B)$$

Example

Problem: Studies of a species of plant in the UK reveal that 8% of the plants have a particular genetic mutation. Amongst those plants with the genetic mutation, 35% produce yellow flowers. If a plant of this species in the UK is selected at random and observed, determine the probability that it will both have the genetic mutation and produce yellow flowers.

Solution:
$\mathrm{P}(\text{mutation} \cap \text{yellow}) = \mathrm{P}(\text{mutation}) \times \mathrm{P}(\text{yellow} \mid \text{mutation}) = 0.08 \times 0.35 = 0.028$

Exercise 4.4

1. Calculate each probability by first *writing out the formula required*:

   (a) $\mathrm{P}(X \cap Y)$ given:
   - $\mathrm{P}(X) = 0.4$
   - $\mathrm{P}(Y) = 0.26$
   - $\mathrm{P}(Y|X) = 0.18$

   (b) $\mathrm{P}(\text{blue} \cap \text{yellow})$ given:
   - $\mathrm{P}(\overline{\text{blue}}) = \frac{2}{3}$
   - $\mathrm{P}(\text{yellow}) = \frac{1}{2}$
   - $\mathrm{P}(\text{blue} \mid \text{yellow}) = \frac{3}{4}$

   (c) $\mathrm{P}(E \cap F)$ given:
   - E and F are independent
   - $\mathrm{P}(E) = 0.82$
   - $\mathrm{P}(\overline{F}) = 0.7$

2. Events A and B are such that $\mathrm{P}(A) = 0.4$, $\mathrm{P}(B) = 0.7$ and $\mathrm{P}(A|B) = 0.56$. Determine the value of:
   (a) $\mathrm{P}(A \cap B)$
   (b) $\mathrm{P}(A \cup B)$

3. Independent events X and Y are such that $\mathrm{P}(X) = 0.4$ and $\mathrm{P}(Y) = 0.8$. Determine the value of $\mathrm{P}(X \cup Y)$.

4. A large box in a maths classroom contains a number of calculators, with 85% of them made by Arizona Instruments and the rest made by Calculo. 70% of the calculators overall are working, whilst 80% of those made by Calculo are working. Determine the probability that a calculator selected at random from the box is a working Calculo calculator.

5. As part of an analysis of a large company's vulnerability to a cyber attack, it is found that 16% of company-issued employee laptops have out-of-date antivirus protection, whilst 24% are only protected by a weak password. Of those with out-of-date antivirus protection, 60% are only protected by a weak password. If a laptop is selected at random, determine the probability that it both has out-of-data antivirus protection and is only protected by a weak password.

6. Given:
   - $\mathrm{P}(\overline{\text{pink}}) = 0.5$
   - $\mathrm{P}(\overline{\text{pink}} \mid \overline{\text{blue}}) = 0.4$
   - $\mathrm{P}(\overline{\text{pink}} \cap \overline{\text{blue}}) = 0.32$

   Find the value of $\mathrm{P}(\text{blue} \cap \text{pink})$.

4.5 Bayes Theorem and Reversing the Condition

In 1763 the Reverend Thomas Bayes, also a statistician, published *"An Essay towards solving a Problem in the Doctrine of Chances"*. In it, he stated a formula that has come to be known as *Bayes' Theorem*, which can be applied in a particular, specialised way to enable existing knowledge to be updated as new information is collected. In turn, this has resulted in an approach towards statistcal inference that has been growing since the mid-20th century called *Bayesian statistics*, substantially made more feasible by modern computing power.

Whilst knowledge of Bayes' Theorem is required in the Advanced Higher Statistics course, Bayesian *statistics* is *not* part of this course, and so will not be covered here. If you would like to find out more about Bayesian statistics, and whether it may be relevant to you in the future, it is recommended that you discuss this with your teacher at a later point in the course. University modules introducing Bayesian statistics are typically available as part of an undergraduate degree in Mathematics or Statistics.

Bayes' Theorem can be derived directly from the formulae already introduced in this chapter. Note that:

$$\begin{aligned} \mathrm{P}(A \cap B) &= \mathrm{P}(A|B) \times \mathrm{P}(B) \\ \mathrm{P}(B \cap A) &= \mathrm{P}(B|A) \times \mathrm{P}(A) \end{aligned}$$

Since $\mathrm{P}(A \cap B) = \mathrm{P}(B \cap A)$:

$$\mathrm{P}(A|B) \times \mathrm{P}(B) = \mathrm{P}(B|A) \times \mathrm{P}(A)$$

Dividing both sides by $\mathrm{P}(B)$ gives:

Bayes' Theorem:

$$\mathrm{P}(A|B) \quad = \quad \frac{\mathrm{P}(B|A) \times \mathrm{P}(A)}{\mathrm{P}(B)}$$

Bayes' Theorem provides a the link between a conditional probability and its *reverse*. Using it to find a conditional probability given its reverse is sometimes referred to as *reversing the condition*.

Example 1

Problem: Given events V and W such that $\mathrm{P}(V|W) = 0.3$, $\mathrm{P}(W) = 0.8$ and $\mathrm{P}(\overline{V}) = 0.1$, find $\mathrm{P}(W|V)$.

Solution: $\mathrm{P}(W|V) = \dfrac{\mathrm{P}(V|W) \times \mathrm{P}(W)}{\mathrm{P}(V)} = \dfrac{0.3 \times 0.8}{0.9} = \dfrac{4}{15}$

Example 2

Problem: In a neighbourhood, 14% of homes have a garage, whilst 62% of homes have a garden. Of those homes with a garage, 87% have a garden. If a home chosen at random has a garden, determine the probability that it has a garage.

Solution: $\mathrm{P}(\text{garage} \mid \text{garden}) = \dfrac{\mathrm{P}(\text{garden} \mid \text{garage}) \times \mathrm{P}(\text{garage})}{\mathrm{P}(\text{garden})} = \dfrac{0.87 \times 0.14}{0.62} = 0.1965$

Exercise 4.5

1. For each, first write out the formula for Bayes' Theorem explicitly:

   (a) Events X and Y are such that:
   - $\mathrm{P}(X) = 0.7$
   - $\mathrm{P}(Y) = 0.6$
   - $\mathrm{P}(Y|X) = 0.5$

   Calculate $\mathrm{P}(X|Y)$.

   (b) For events V and W:
   - $\mathrm{P}(W) = \frac{4}{5}$
   - $\mathrm{P}(\overline{V}) = \frac{1}{3}$
   - $\mathrm{P}(W|V) = \frac{3}{4}$

   Determine $\mathrm{P}(V|W)$.

2. A bag contains numbered counters of various colours, a quarter of which are red. Three-eighths of the counters are numbered 1, of which two-fifths are red. If a counter is selected at random and it is red, determine the probability that it is numbered 1.

3. Research has shown that 1.7% of people possess a specific gene. A prototype test for the presence of the gene in a person is being developed, and it shows as *positive* for the gene for 7% of everyone people tested. It is later established that the probability a test shows as positive for a person known to possess the gene is 0.9. If a person is selected at random to take the test, and it shows positive, determine the probability that they actually do possess the gene.

4. Given that $\mathrm{P}(M) = \frac{2}{3}$, $\mathrm{P}(N) = \frac{3}{4}$ and $\mathrm{P}(N|M) = \frac{3}{5}$:

   (a) Calculate $\mathrm{P}(M|N)$.

   (b) Calculate $\mathrm{P}(\overline{M}|N)$.

5. An MOT test is an annual examination of cars and other vehicles required to check they are roadworthy. 83% of the vehicles a local garage performs MOT tests on are cars, whilst the rest are vans. Considering only the first MOT test performed on each vehicle, before any repairs are carried out:

   - 15% of all tests result in a fail
   - Cars account for 86% of all failed tests.

   Determine the proportion of cars that *pass* their first MOT.

6. A region has access to a local radio stations, *Pacific 525*, as well as a number of national stations. When a resident of the region directs their car radio to automatically find a station, the probability that it will connect to *Pacific 525* is 0.64, otherwise it will instead connect to a national station. 72% of the time, when their car automatically connects there will be music playing. When their car radio connects to *Pacific 525* the probability music will be playing is 0.84. If the car radio connects to a station and music is playing, determine the probability that the station is *Pacific 525*.

7. Events E and F are such that:

   - $\mathrm{P}(\overline{F}|E) = 0.23$.
   - $\mathrm{P}(F) = 0.87$.
   - $\mathrm{P}(\overline{E}) = 0.09$.

   Calculate $\mathrm{P}(\overline{E}|F)$.

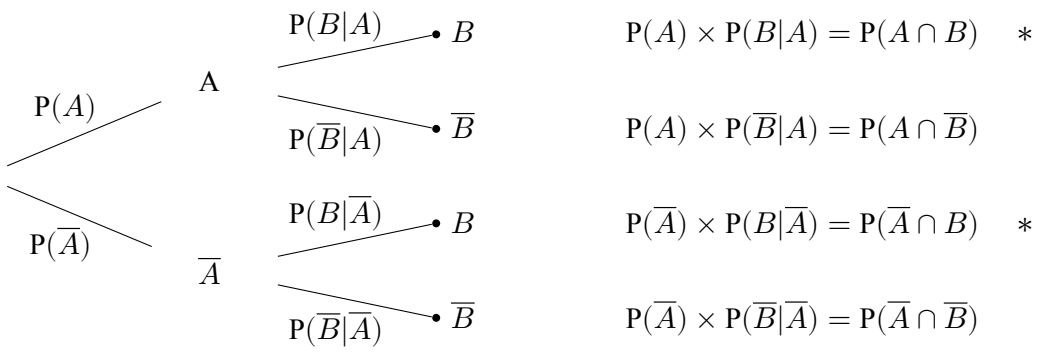
4.6 Total Probability and Tree Diagrams

The *law of total probability* for X conditional on n mutually exclusive, exhaustive events Y_i is:

$$\mathrm{P}(X) = \sum_{i=1}^{n} \mathrm{P}(Y_i) \times \mathrm{P}(X|Y_i)$$

It may help to recognise the right hand side of the equation as containing the *multiplication rule for the intersection*: the total probability of X is obtained by adding together the probabilities of all (disjoint) intersections in which X occurs.

The tree diagram below shows the appropriate notation for each stage, in which first event A and then event B are considered:



The *total probability* of B, $\mathrm{P}(B)$, can be calculated as:

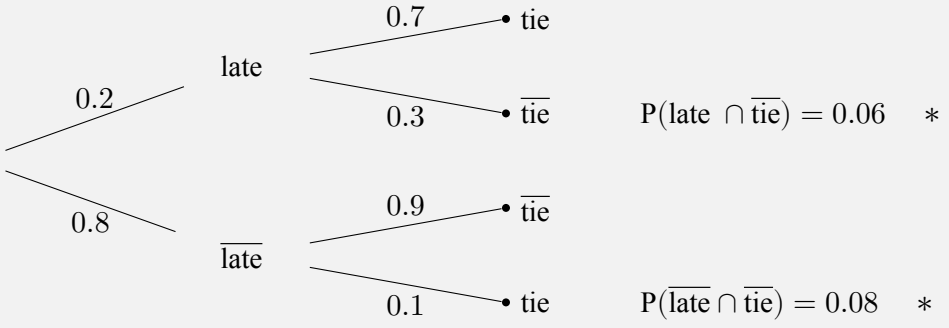
$$\begin{aligned} \mathrm{P}(B) &= \mathrm{P}(A \cap B) &+ \mathrm{P}(\overline{A} \cap B) \\ &= \mathrm{P}(A) \times \mathrm{P}(B|A) &+ \mathrm{P}(\overline{A}) \times \mathrm{P}(B|\overline{A}) \end{aligned}$$

The advantage of the use of tree diagrams, for questions where they are suitable, is that they can offer a more intuitive and visual approach to tackling problems than solely using formulae. If they are used, they must be annotated carefully with events and probabilities, and accurate probability notation must still be used when giving answers.

Example

Problem: A teacher observes that one of their pupils is late to school 20% of the time, and when this happens they have their tie with them 70% of the time. When they are on time they have their tie 90% of the time. Construct a tree diagram and calculate the probability that, on a randomly chosen day, they have no tie.

Solution:



$\mathrm{P}(\overline{\text{tie}}) = 0.06 + 0.08 = 0.14$

Exercise 4.6

Construct a tree diagram for each question.

1. A UK-based car company makes a special edition model with a colour choice of blue or yellow. It sells cars both within the UK as well as exporting to Europe.

   - 72% of the special edition models manufactured by the company are blue.
   - The rest are yellow.
   - 84% of the blue cars are sold within the UK and 95% of the yellow cars are sold within the UK.

   Calculate the probability that a randomly selected special edition model is sold within the UK.

2. A large number of trees are planted as part of a tree-planting project.

   - Three-quarters of trees are oak.
   - The rest are alder trees.
   - The probability an oak tree failing is 0.356.
   - The probability of failure for an alder tree is 0.234.

   Calculate the probability that a tree randomly selected for observation survives.

3. The star striker for a football team has a probability of 0.7 of scoring in any given game. When they do, the probability that the team goes on to win is 0.8, whilst when they don't the probability of a win drops to 0.45. Calculate the probability that the team win a randomly chosen game.

4. When Betty's hears a song on the radio:

   - The probability that it is a song she knows is 0.8.
   - When she knows it, the probability that it is a song she likes is 0.6.
   - When she doesn't know it, the probability she will like it is 0.3.

   Determine the proportion of songs Betty hears on the radio that she likes.

5. An obstacle course consists of three parts, with each part needed to be completed successfully for a competitor to be allowed to move on to the next. 30% of competitors fail to complete the first part, and of those who make it onto the second part only 40% manage to make it onto the third part. Of those that do make it on to the third part, however, 65% successfully complete it. Calculate the probability that a competitor makes it onto the third part.

6. As part of a study examining the survival rates of hedgehogs over winter hibernation periods, a number of hedgehogs are captured, weighed, tagged and then released. Depending on their weight they are classed as having a "good weight", an "average weight" or being "underweight", with 18%, 76% and 6% observed respectively in each category. From the study it is found that the probability of survival for each category is 91%, 72% and 35% respectively. Assuming that the distribution of hedgehog weights and probabilities of survival can be taken as representative of hedgehog populations in general, find the proportion of hedgehogs that do not survive winter hibernation.

7. The probability of being dealt a pair (such as two fives) in an opening hand in a game of poker is 0.06. If this happens, a particular player knows that based on their game history they have a probability of 0.83 of winning the hand, whilst otherwise they have a probability of 0.36. Calculate the probability they win a hand.

4.7 Bayes Theorem and Tree Diagrams

In this chapter, Bayes' Theorem has already been used to calculate the *reverse* of a conditional probability: calculating $\mathrm{P}(B|A)$ when $\mathrm{P}(A|B)$ is known. This process can be made more intuitive and more readily extended to more complex problems by instead using tree diagrams combined with the formula for conditional probability, introduced on Page 44:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

A key misunderstanding often made is that a high value for $\mathrm{P}(B|A)$ must lead to a similarly high value for $\mathrm{P}(A|B)$. The following example demonstrates that this is not always the case, and gives an insight into a common problem that must be considered whenever medical testing is being undertaken, especially for rare conditions or diseases:

Example

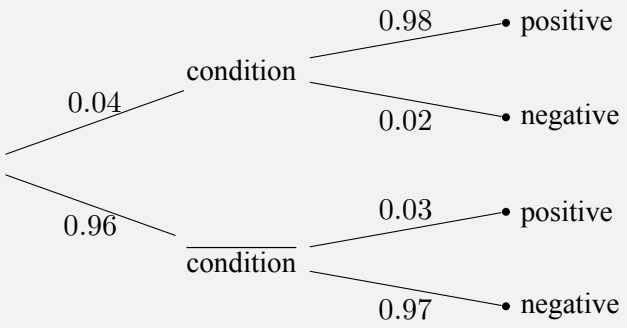
Problem: A medical test for a condition has a *sensitivity* of 98%, meaning that when the patient *has the condition*, the test will correctly return a *positive* result with a probability of 0.98. This means there is a *false negative* rate of 2%. The test has *specificity* of 97%, meaning that when the patient *does not* have the condition, the test will correctly return a *negative* result with a probability of 0.97. This means there is a *false positive* rate of 3%. The *prevalence* in the general population is 4%, meaning only 4% of people have the condition.

If a person is selected at random from the general population to be tested for the condition:

(a) Calculate the probability that test comes back *positive*.

(b) Given that the test comes back positive, determine the probability that they do actually *have* the condition.

Solution:



$\mathrm{P}(\text{condition} \cap \text{positive}) = 0.0392 \quad * \checkmark$

$\mathrm{P}(\overline{\text{condition}} \cap \text{positive}) = 0.0288 \quad *$

(a) $\mathrm{P}(\text{positive}) = 0.0392 + 0.0288 = 0.068$

(b) $\mathrm{P}(\text{condition} \mid \text{positive}) = \dfrac{\mathrm{P}(\text{condition} \cap \text{positive})}{\mathrm{P}(\text{positive})} = \dfrac{0.0392}{0.068} = 0.5765$ $\qquad \dfrac{\checkmark}{*+*}$

It may be surprising that a test which appears so effective can lead to such a high proportion of false positives, and calculations such as this should be considered when healthcare professionals establish testing procedures. In practice, the false positive rate can be drastically reduced by deciding when a test may be appropriate. GPs, for example, will ask questions about a patient's symptoms and medical history to determine whether they have reason to suspect that the chance that they have the condition is significantly greater than a random member of the population, and therefore whether a test is appropriate.

Exercise 4.7

Construct a tree diagram for each question.

1. A company sends customers' orders by either regular or express delivery.

   - 35% of orders are sent by express delivery.
   - Of those sent by express delivery, 98% arrive undamaged,
   - 91% of the orders sent by regular deliveries arrive undamaged.

   Customers are requested to submit a form online to confirm whether their order arrived damaged or undamaged.

   (a) Determine the probability that a form received by the company states that an order arrived damaged.

   (b) Given that an order arrived damaged, calculate the probability that the order was sent by express delivery.

2. Anti-virus software is being developed to detect whether attachments sent by email are malicious or safe, and flag those that are malicious. In its current state it can correctly flag that an attachment is malicious 92% of the time. It incorrectly flags safe attachments 2% of the time. A company is interested in using the software, and 1.2% of all attachments emailed to their employees are actually malicious.

   (a) Determine the proportion of attachments that are flagged as malicious.

   (b) Given that an attachment is flagged, calculate the probability that is malicious.

3. A pupil sitting a multiple choice test is able to answer $\frac{3}{4}$ of the questions and guesses the rest. They know from trying practice tests that the probability that are correct in the questions they are able to answer is 0.9, whilst they have a one-in-five chance of getting a correct answer by guessing. Determine the probability that a question which they get wrong was one that they guessed.

4. A marketing campaign for a new product is estimated to have been seen by three times as many people through its social media advert as by its promotional email. The probability that seeing the social media advert leads to a person visiting their website is 0.037, whilst for the promotional email that figure drops to 0.021.

   (a) Determine the probability that someone who sees the marketing campaign in either form visits the website.

   (b) Determine the probability that a visit to the website due to the marketing campaign came from that person seeing the social media advert.

5. A sports team is given an extra rest day half of the time after winning a game, a quarter of the time after drawing a game and only a tenth of the time when they lose a game. They win 55% of their games, draw 25% of their games and lose the rest. Given that they are given an extra rest day, find the probability that they lost the last game.

6. All precision-engineered parts at a small factory undergo two inspections as part of quality control. During Inspection A, 8% of parts are flagged for a more thorough inspection during Inspection B. Of those that are more thoroughly inspected during Inspection B, 15% are scrapped due to flaws in the manufacturing process. 3% of those that had not been flagged initially are also then scrapped during Inspection B as flaws are spotted that were initially missed. Finally, of those parts that have not been scrapped, 1% are set aside for stress-testing of the final products and never leave the factory. Given that a part doesn't leave the factory, find the probability that it was due to a flaw.

7. For medical tests for conditions that are known to have the given specificity and sensitivity, as well as the prevalence of the condition, calculate the probability that a randomly tested person has the condition, given that they test positive:

   (a) Sensitivity = 94%, Specificity = 89%, Prevalence = 9%

   (b) Sensitivity = 98%, Specificity = 99%, Prevalence = 1%

Review Exercise

1. Events E and F are such that:
   - $\mathrm{P}(F \cap E) = 0.7$
   - $\mathrm{P}(\overline{E}) = 0.2$
   - E and F are exhaustive.

   Calculate $\mathrm{P}(E|F)$.

2. A second-hand online computing retailer has 40 laptops in stock, broken down by whether they use Windows or Mac, and the grades A, B and C depending on the condition they are in:

	A	B	C
Windows	7	18	5
Mac	4	5	1

   A laptop is selected at random.
   W is the event that the laptop selected uses Windows.
   A is the event that the laptop selected is in grade A condition.
   B is the event that the laptop selected is in grade B condition.

   (a) Find:
      i. $\mathrm{P}(W \cup A)$.
      ii. $\mathrm{P}(\overline{W} \cap B)$.

   80% of the Windows laptops and 90% of the Mac laptops come with the original power supply.

   (b) Calculate the probability that a randomly selected laptop comes with its original power supply.

   (c) Determine the probability that a randomly selected laptop is a Mac laptop, given that it does not come with its original power supply.

3. At the start of each round of a board game played by four players, a random process determines who gets to go 1st, who goes 2nd, who goes 3rd and who goes 4th. Sam knows that when she plays against her two friends, she wins the round two-thirds of the time when she goes 1st, half of the time when she goes 2nd and only a fifth of the time when she goes either 3rd or 4th.

   (a) Assuming that each round is independent of the result of previous rounds, calculate the proportion of rounds that Sam wins.

   (b) Given that Sam lost the last round, calculate the probability that she went 1st.

4. For the events M and B: $\mathrm{P}(M \cap B) = \frac{3}{25}$, $\mathrm{P}(M) = \frac{1}{5}$ and $\mathrm{P}(\overline{B}) = \frac{2}{5}$. Show that events M and B are independent.

5. As part of a machine learning project, a program is created that tries to correctly identify whether it is being shown a picture of a dog, a cat or a human. It has been tested and found to correctly identify a human with the probability 0.8, correctly identify a dog with the probability 0.6 and correctly identify a cat with the probability 0.55. When subsequently put to use by the public, the program is shown a picture of a human 60% of the time, a dog 25% of the time and a cat 15% of the time. Assuming the probabilities of successful detection are the same when used by the public as when under testing, calculate:

   (a) the proportion of times that the program is incorrect

   (b) the probability that the program was shown a cat, given that is incorrect.

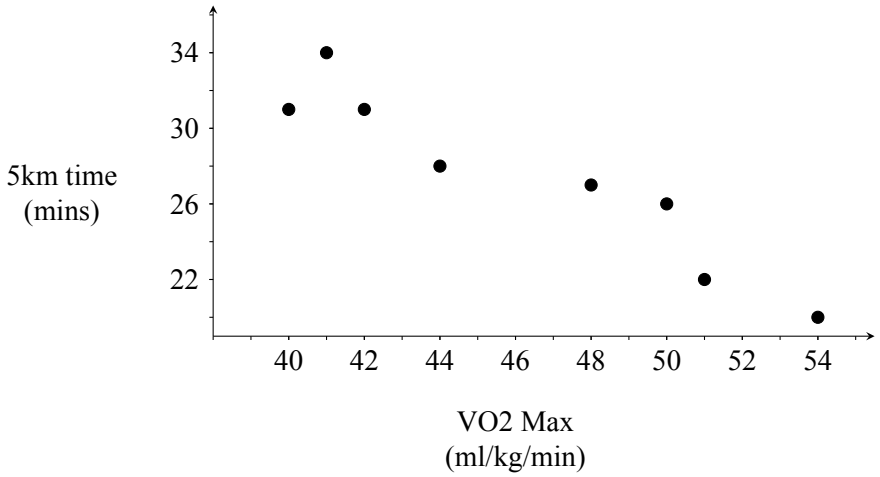
5

An Introduction to Linear Regression

A common interest for anyone working with data is a possible relationship between two numerical variables. For example, suppose a sports scientist wishes to explore the relationship between runners' *VO2 Max*, which is measure of aerobic fitness, and their *5km running time*. From a sample of eight runners they may record, for each, the values of the two *variables*, recording the results in a table as follows:

Runner	A	B	C	D	E	F	G	H
VO2 Max (ml/kg/min)	50	41	54	42	44	48	40	51
5km time (mins)	26	34	20	31	28	27	31	22

Before applying any statistical analysis techniques to this *bivariate data* the sports scientist constructs a *scatterplot* to visualise it, looking for any apparent patterns or problems:



Whilst the sample size is very small, the scatterplot seems to suggest that there is a *linear relationship* between a runner's VO2 Max and their 5km running time. This chapter will introduce the use of a *correlation coefficient* to evaluate the strength of any linear relationship between numerical variables and the construction of a *linear regression model* to describe this relationship and make predictions.

5.1 Scatterplots and Linear Correlation

When deciding whether a relationship exists between two numerical variables, a scatterplot should always be drawn before any analysis is undertaken using the bivariate data collected. This is the first step in determining the level of linear relationship, or *linear correlation*, between the variables. Whilst other kinds of relationship may exist between two numerical variables, in this course the focus is on the existence and strength of *linear* relationships. Consider the following:

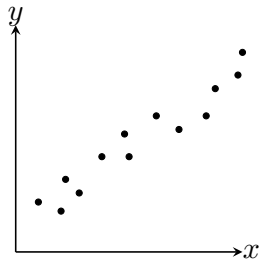


Figure 1

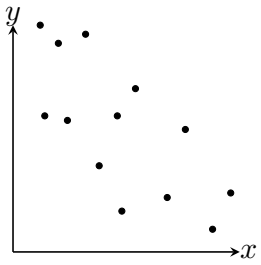


Figure 2

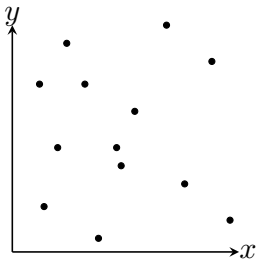


Figure 3

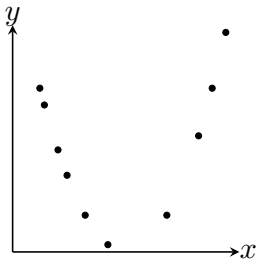


Figure 4

Figure 1 illustrates a *strong, positive linear correlation* between x and y, where both variables increase together.

Figure 2 shows a *weak, negative linear correlation* between x and y, where one decreases as the other increases.

In **Figure 3**, as one variable increases, there appears to be no clear pattern as to how the other variable behaves. There appears to be *no linear relationship* between x and y.

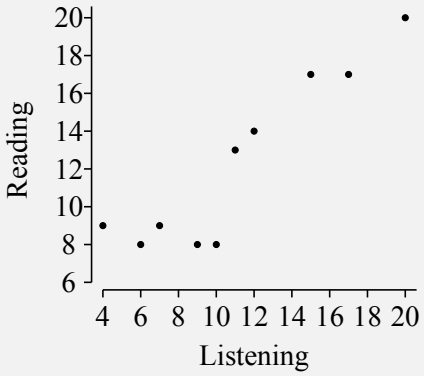
Figure 4 is also an example of *no linear relationship* between x and y, although there appears to be some non-linear (possibly quadratic) association between the variables.

Note that **correlation doesn't imply causation**. A linear relationship can exist between variables without meaning that one causes the other. Seemingly linear relationships can also appear by chance - this is referred to as *spurious correlation*.

Example

Problem: A French class takes two tests, one *listening* and one *reading*. The results of the 10 pupils are shown in the table and scatterplot below, along with some summary statistics. Comment on the scatterplot.

Listening	Reading
9	8
7	9
11	13
20	20
4	9
6	8
17	17
12	14
10	8
15	17



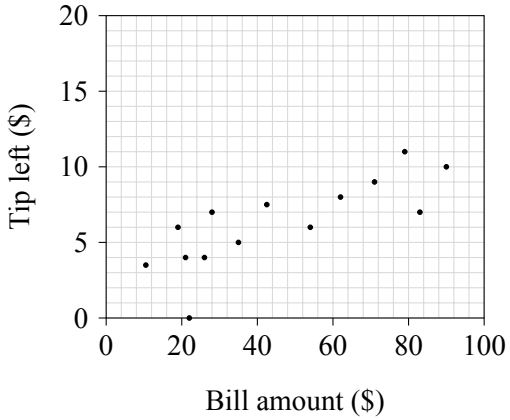
Solution:
The scatterplot appears to show a strong, positive linear relationship between pupils' listening and reading scores.

Scatterplots should always be constructed and closely scrutinised before any analysis of bivariate data is undertaken. Through this visualisation, it may become apparent for instance that the dataset may actually be clearly separated into two or more *distinct groups*, or that possible *outliers* are present which should be carefully considered.

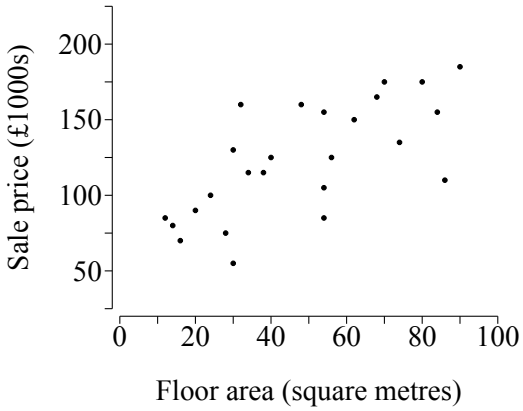
Exercise 5.1

1. For each, make a comment on the scatterplot.

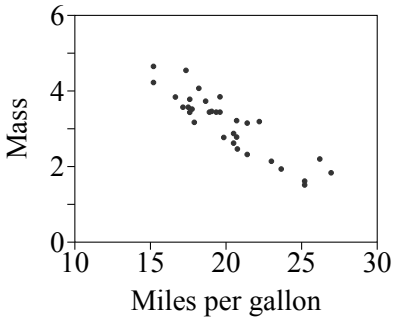
   (a) Data is collected by a diner in the USA on the *bill amount* and *tip left* by its guests.

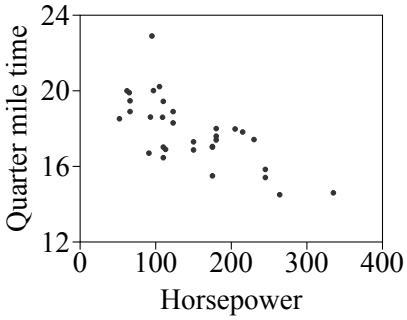


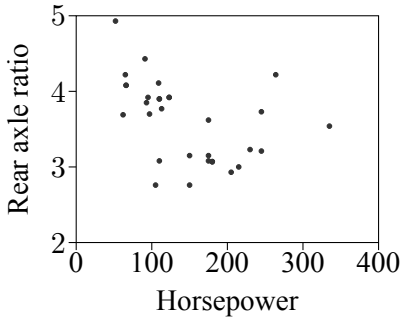
   (b) The *sale price* and *floor area* are recorded for a number of 1-bedroom city centre apartments.



2. Information on the designs of 32 different models of cars from the 1970s was collected by a motoring magazine. Data collected included *miles per gallon*, *mass* (in thousands of pounds), *quarter mile time* (in seconds), *rear axle ratio* and *horsepower*. Three scatterplots are shown below to explore possible relationships between the variables.

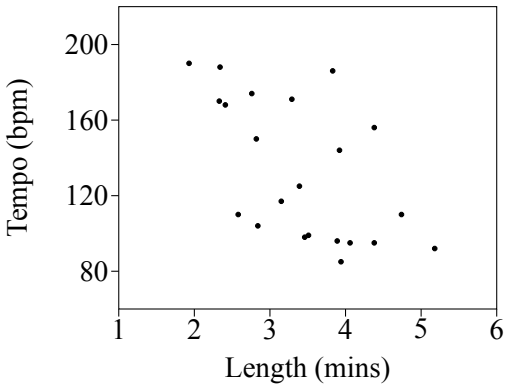






   Comment on each scatterplot.

3. A statistician is keen to learn what can be learned by studying data on songs provided by a music streaming platform. For a random sample of pop songs on the platform, their *length* (x, in minutes) and *tempo* (y, in bpm) are recorded, and a scatterplot is constructed. Some summary statistics were also calculated.



$$\Sigma x = 75.13$$
$$\Sigma y = 2919$$
$$\Sigma x^2 = 272.1$$
$$\Sigma y^2 = 416279$$
$$\Sigma xy = 9583.94$$
$$n = 22$$

   (a) Comment on the scatterplot.

   (b) Use the summary statistics to calculate the mean song length and the mean tempo for the sample .

5.2 Pearson's Product Moment Correlation Coefficient

The strength of the *linear relationship* between two variables is measured using a *correlation coefficient*. Whilst there are a number of different correlation coefficients used in statistics, the only one used in this course is *Pearson's Product Moment Correlation Coefficient (PMCC)*, which can now simply be referred to as the **correlation coefficient**. Given a sample of bivariate data, the *sample correlation coefficient*, r, can be calculated as:

Pearson's Correlation Coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

The formulae for the *"Three Ss"* and the *"Five Sigmas"* are given in the data booklet:

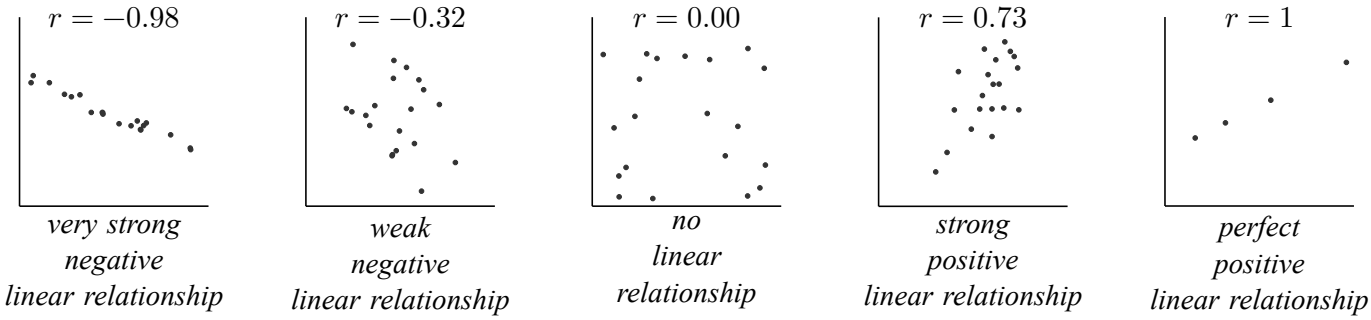
Three Ss:

$$S_{xx} = \Sigma(x-\bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$
$$S_{yy} = \Sigma(y-\bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$
$$S_{xy} = \Sigma(x-\bar{x})(y-\bar{y}) = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$$

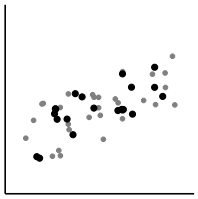
Five Sigmas:

$$\Sigma x$$
$$\Sigma y$$
$$\Sigma x^2$$
$$\Sigma y^2$$
$$\Sigma xy$$

For all sets of data, $-1 \leqslant r \leqslant 1$. A correlation coefficient of zero means there is no correlation, or *no linear relationship*, between the two variables, whilst any non-zero correlation coefficient means *a linear relationship exists*. Each of figures below show a scatterplot of sample data along with its sample correlation coefficient, r, and an interpretation.

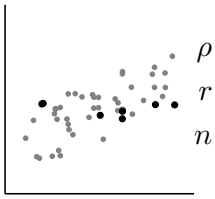


As the correlation coefficient, r, is calculated for a *sample*, it is a **sample statistic**. It gives an estimate for the *population correlation coefficient*, ρ (the Greek letter, *"rho"*). For this reason, a non-zero value of r calculated from a sample only *suggests* a linear relationship exists between the variables. Establishing *evidence* will be covered in a later chapter.



$\rho = 0.66$
$r = 0.71$
$n = 20$

Here, the **sample** has resulted in a value for r which approximates well the **population** correlation coefficient, ρ.



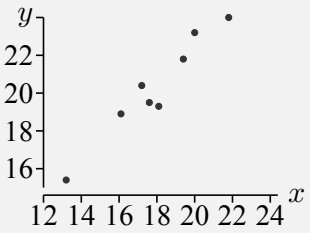
$\rho = 0.66$
$r = -0.18$
$n = 7$

Here, the **sample** has resulted in a value for r which is far from the true **population** correlation coefficient, ρ.

Example

Problem: A biologist records the tail length (x) and total height (y), both in centimetres, for a random sample of 8 red squirrels. A scatterplot and summary statistics are for the data produced and shown below.

$$\Sigma x = 143.4 \qquad \Sigma y = 163.3 \qquad \Sigma xy = 2978.5$$
$$\Sigma x^2 = 2618.3 \qquad \Sigma y^2 = 3391.8 \qquad n = 8$$



Calculate the correlation coefficient and interpret its value.

Solution:

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 2618.3 - \frac{143.4^2}{8} = 47.82$$

$$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 3391.8 - \frac{163.3^2}{8} = 58.43$$

$$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 2978.5 - \frac{143.4 \times 163.3}{8} = 51.39$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{51.39}{\sqrt{47.82 \times 58.43}} = 0.9354$$

This suggests a strong, positive linear correlation between the length of squirrels' tails and their heights.

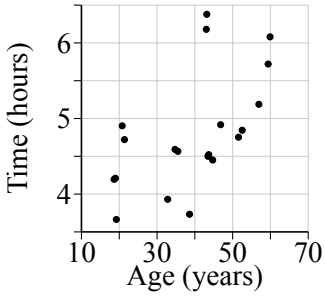
Exercise 5.2

1. Students' average homework mark and their test score are recorded. Calculate the correlation coefficient.

Homework (average, out of 20)	17.1	11.2	19.6	15.4	19.8	8.1
Test (%)	81	46	94	83	94	64

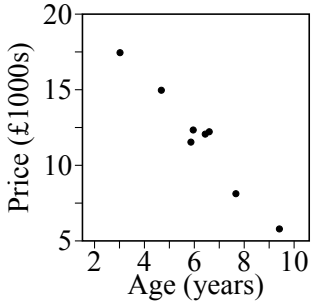
2. For each, calculate the correlation coefficient and interpret its value.

(a) For 20 first-time marathon runners, their age (x, in years) and finishing time (y, in hours) is recorded.



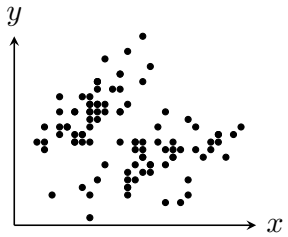
$$\Sigma x = 785.9$$
$$\Sigma y = 96$$
$$\Sigma x^2 = 34481$$
$$\Sigma y^2 = 472.6$$
$$\Sigma xy = 3888.7$$
$$n = 20$$

(b) The age (x, in years) and selling price (y, £1000s) of 8 used cars is recorded.



$$\Sigma x = 49.63$$
$$\Sigma y = 94.48$$
$$\Sigma x^2 = 333.1$$
$$\Sigma y^2 = 1208.0$$
$$\Sigma xy = 538.87$$
$$n = 8$$

3. A *sepal* is part of a flower. The lengths (x, in cm) and widths (y, in cm) are measured for a random sample of 100 *iris* flowers, of two different varieties. A scatterplot is plotted, and a correlation coefficient of -0.206 is obtained. The researcher concludes that this proves that there is a weak negative linear relationship.



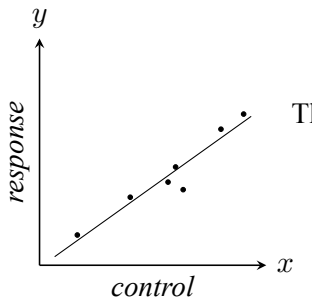
(a) Without considering the scatterplot, explain why the researcher should not have used the word "proves" in their conclusion.

(b) Comment on what might be observed by inspecting the scatterplot, and explain what the researcher should do with the data as a consequence.

(c) Suggest two possible improvements to the design of the scatterplot.

5.3 Least Squares Linear Regression

One goal of linear regression is to allow *predictions* to be made. This typically involves an independent (or *control*) variable on the x-axis, and a dependent (or *response*) variable on the y-axis, with the aim of constructing a *trend line* such that the model can predict unknown y values based on known x values. To determine the equation of this line, the method of *least squares regression* can be used.

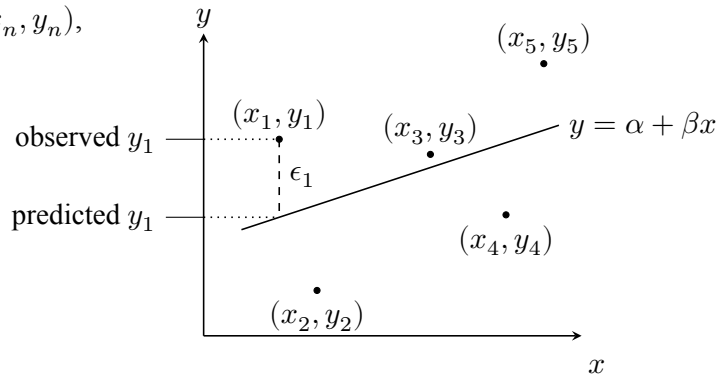


This line is referred to as a y *on* x *regression line*.

Given a full population of bivariate data $(x_1, y_1), (x_2, y_2)...(x_n, y_n)$, the equation of the best fitting line is of the form:

$$y = \alpha + \beta x$$

- α (*"alpha"*) represents the *intercept parameter*
- β (*"beta"*) represents the *slope parameter*



Differences between the y values predicted by this line and the actual y_i values are called *errors*, or ϵ_i (*"epsilon"*). The parameters α and β are calculated such that the *sum of squared errors*, $\Sigma\epsilon_i^2$, is minimised.

Since regression lines are generally calculated from a random *sample* of bivariate data points, the values for the population parameters α and β have to be estimated, with the sample statistics a and b respectively calculated as *estimators*:

Estimation of Slope and Intercept Parameters

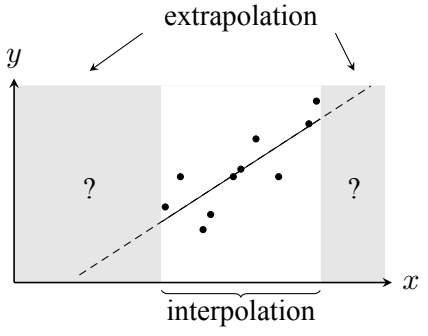
$$\hat{\beta} = b = \frac{S_{xy}}{S_{xx}} \qquad \text{and} \qquad \hat{\alpha} = a = \bar{y} - b\bar{x}$$

Once values for a and b have been calculated, and so the least squares regression line of y on x obtained as $y = a + bx$, unknown values of y may be predicted for given values of x. This equation (and the line it defines) is *not* suitable for predicting x values based on y, the process for which will be covered in a later chapter.

Whether a linear relationship will continue to hold for x values lower or higher than any observed within the range of the data can be unpredictable, and should not be assumed.

Interpolation, prediction within the range of the control data, is generally reliable if the correlation is high.

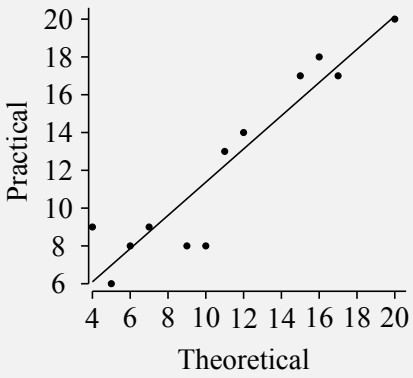
However, **extrapolation**, prediction out with the range of the data, should be avoided as unreliable and unsupported.



Example

Problem: A biology class takes two tests, one theoretical and one practical. The results of the 12 pupils are shown in the table and scatterplot below, along with some summary statistics.

Theoretical test result (x)	5	9	7	11	20	4	6	17	12	10	15	16
Practical test result (y)	6	8	9	13	20	9	8	17	14	8	17	18



$$\Sigma x = 132, \quad \Sigma y = 147,$$
$$\Sigma x^2 = 1742, \quad \Sigma y^2 = 2057, \quad \Sigma xy = 1872$$

(a) Calculate the equation of the least squares regression line of y on x.
(b) Calculate a predicted mark for the practical test for a pupil who missed it, having scored 8 in the theoretical test.
(c) Comment on the reliability of this prediction.

Solution:

(a) $$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$
$$= 1742 - \frac{132^2}{12}$$
$$= 290$$

$$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$$
$$= 1872 - \frac{132 \times 147}{12}$$
$$= 2555$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{255}{290} = 0.879$$

$$a = \bar{y} - b\bar{x} = \frac{147}{12} - 0.879 \times \frac{132}{12} = 2.578$$

$$y = 2.578 + 0.879x$$

(b) $y = 2.578 + 0.879 \times 8 = 9.61$

(c) This is an example of interpolation as the value of x used is within the range of values of the other students, and consequently it can be considered a robust prediction, especially given the high value for r.

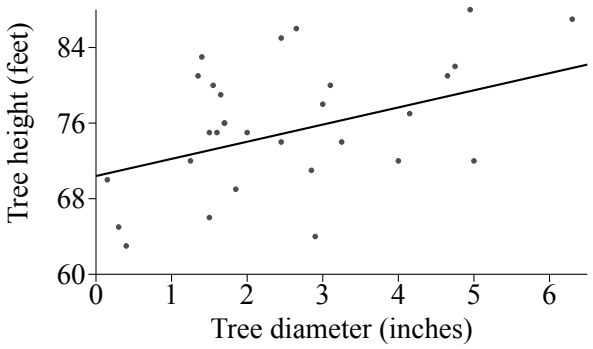
Exercise 5.3

1. A bivariate sample has summary statistics:

$$\sum x = 117 \qquad \sum x^2 = 1869 \qquad \sum y = 660 \qquad \sum y^2 = 54638 \qquad \sum xy = 9519 \qquad n = 8$$

It is determined from inspecting a scatterplot that there appears to be a linear relationship between the variables. Obtain the equation of the least squares regression line of y on x.

2. 31 randomly selected black cherry trees have their diameter at a set distance from the ground measured, in inches, with the height of each tree also measured, in feet. The equation of the least squares regression line for y on x is obtained, and shown on the scatterplot below.



Summary statistics:

$\sum x = 410.7$

$\sum x^2 = 5736$

$\sum y = 2356$

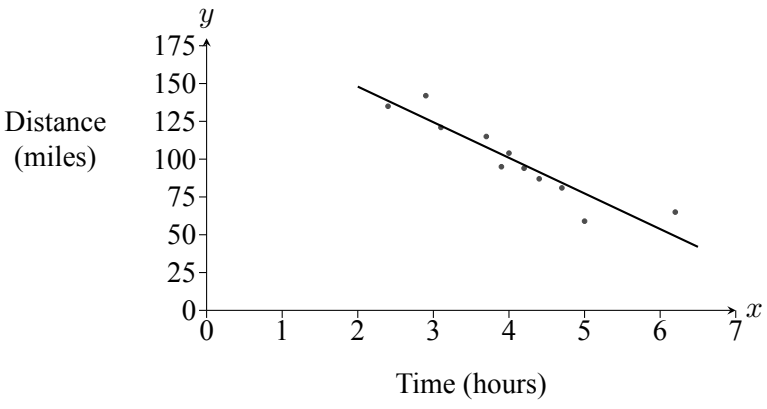
$\sum y^2 = 180274$

$\sum xy = 31524$

$n = 31$

Calculate the equation of the least squares regression line of y on x.

3. A running website randomly selected eleven amateur runners who recently completed a marathon and asked them to complete a survey which included their finishing time in the marathon and information about their training routine. The scatterplot below shows, for each runner, their finishing time (x, in hours) and the total distance they ran in the final four weeks of their training (y, in miles).



Summary statistics:

$\sum x = 44.5$

$\sum x^2 = 191.21$

$\sum y = 1098$

$\sum y^2 = 116768$

$\sum xy = 4179.2$

$n = 11$

A least squares regression line for y on x is shown on the graph.

(a) Obtain the equation of the least squares regression line.

(b) Calculate the correlation coefficient, and comment on its value.

The website wishes to allow users the ability to input the number of miles in their training plan for the four weeks leading up to the rate, and receive in response an estimate for their finishing time.

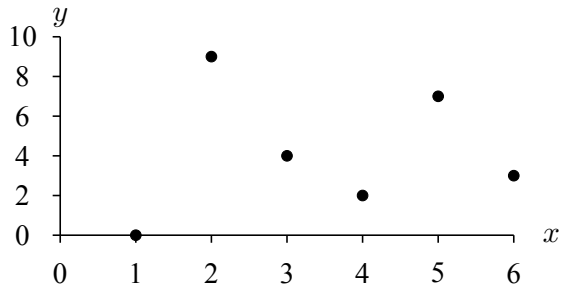
(c) Explain why the equation provided above is unsuitable for this purpose.

4. A researcher laid a transect line from the base of a tress and placed a quadrat at one metre intervals along the transect line. The number of seeds dropped by the tree contained in each quadrat was recorded.

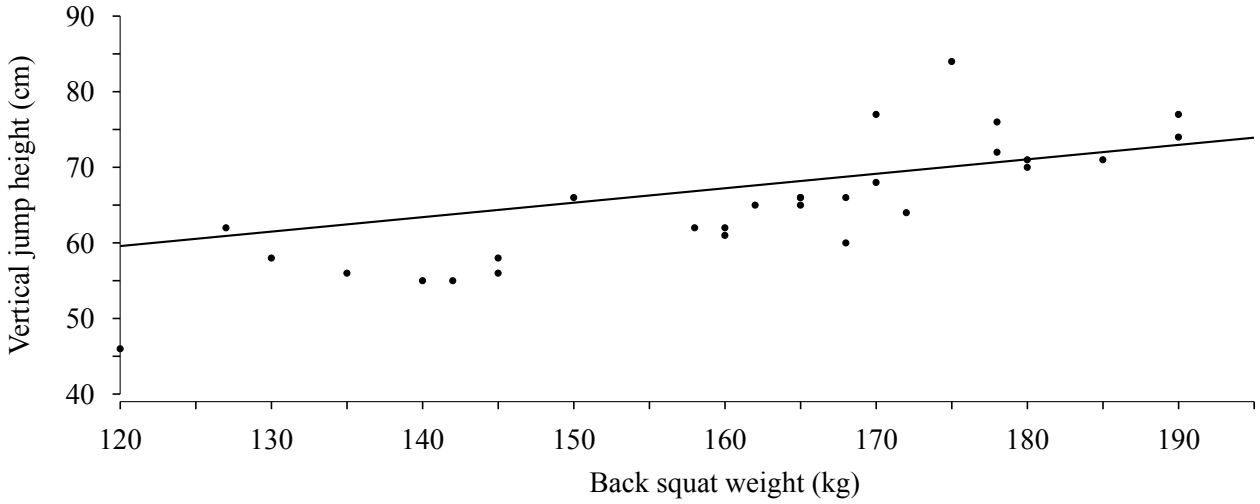
Distance from base of tree (x)	1	2	3	4	5	6
No. of seeds found (y)	0	9	4	2	7	3

(a) Calculate the correlation coefficient.

(b) Explain why the researcher may have decided not to obtain a regression line.



5. A strength and conditioning coach wants to increase the vertical jump height performance in their trainees and considers whether back squat weight has an impact on vertical jump height performance. The summary statistics below are taken from the back squat weight (x in kg) and vertical jump height (y in cm) achieved by a random sample of 29 of the coach's trainees.



$$\Sigma x = 4673 \quad \Sigma x^2 = 763101 \quad \Sigma xy = 308020$$
$$\Sigma y = 1889 \quad \Sigma y^2 = 124945 \qquad n = 29$$

(a) Use the scatterplot to comment on the relationship between vertical jump height and back squat weight achieved by the trainees.

(b) Calculate the product moment correlation coefficient for this data set.

(c) The coach wishes to use the data to predict vertical jump height.

i. Find the equation of the least squares regression line of vertical jump height on back squat weight.

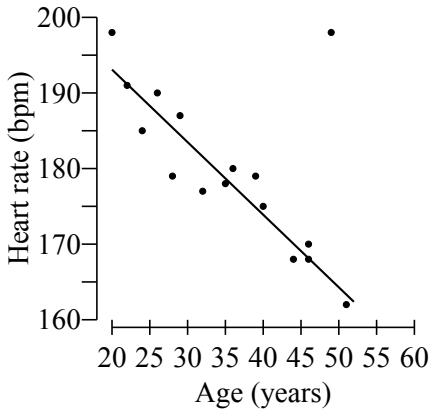
ii. Estimate the vertical jump height of a trainee who back squats 165kg and comment on the reliability of this estimate.

Based on the correlation coefficient the coach advises that increasing their squat back weight will increase their vertical jump height.

(d) Give a reason why this advice is not necessarily correct.

Review Exercise

1. As part of a study on intensive exercise, a sports scientist recorded the peak heart rates (y) of a random selection of sixteen volunteers of different ages (x) who took regular exercise. The linear regression equation was calculated for the data shown in the scatter diagram and found to be $y = 203.9 - 0.67x$.



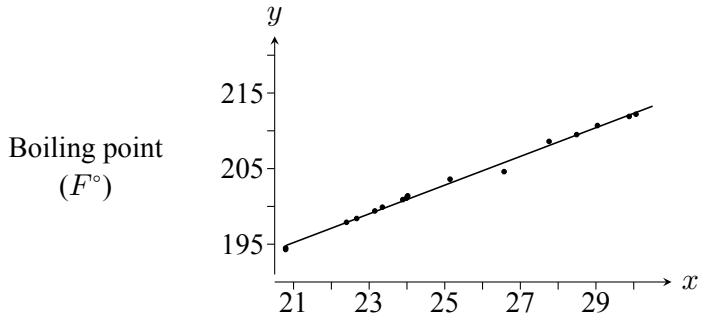
However, after considering the scatter diagram for the data, it was realised that one piece of data has been misrecorded, and this volunteer's data was removed from the data set.

(a) State the approximate age of the volunteer whose data was removed.

(b) Calculate the new equation of the least squares regression line using the revised summary data:
$\Sigma x = 518, \Sigma y = 2687, \Sigma x^2 = 19196, \Sigma y^2 = 482691, \Sigma xy = 91534$

(c) Comment on the difference this makes to the prediction for the average peak heart rate of a 45 year old volunteer.

2. James Forbes was a Scottish physicist who, amongst other things, was interested in finding a way to estimate altitude from measurements of the boiling temperature of water. Some of his observations on boiling point (y, in F°) and barometric pressure (x, in inches of mercury) are summarised below and shown in the scatterplot.



Summary data for sample:

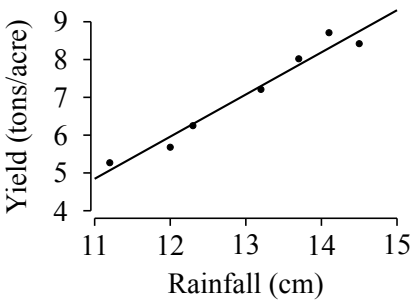
$S_{xx} = 145.94$
$S_{yy} = 530.78$
$S_{xy} = 277.54$

$n = 17$

The least squares regression line of $y = 155.29 + 1.9018x$ was fitted. Show the calculations required to obtain the values of the sample intercept and sample slope parameter.

3. The yield of a particular crop on a farm is thought to depend principally on the amount of rainfall in the growing season. The values of the yield, y, in tons per acre, and the rainfall, x, in centimetres, for seven successive years are given in the table and displayed in the scatterplot.

Rainfall (x)	12.3	13.7	14.5	11.2	13.2	14.1	12.0
Yield (y)	6.25	8.02	8.42	5.27	7.21	8.71	5.68



(a) Comment on what the scatterplot shows.

(b) Calculate the product moment correlation coefficient for this data set.

(c) Find the equation of the least squares regression line of y on x.

(d) Comment on the appropriateness of using the regression line found to estimate the rainfall in a year if the yield for the year is known.

6

Random Variables

A random variable is a mathematical object that describes the possible outcomes, and their probabilities, of a random experiment. They are used to create a *model* of the behaviour of random processes, stated formally and algebraically. Random variables take a defined list or range of *numerical* values, either *discrete* or *continuous*.

Random variables are notated using *capital letters*, such as X and Y, whilst *specific values* that they can take are notated with small letters such as x and y. Time spent paying careful attention to which is required when first encountering random variables will help greatly as the course progresses.

Discrete Random Variables

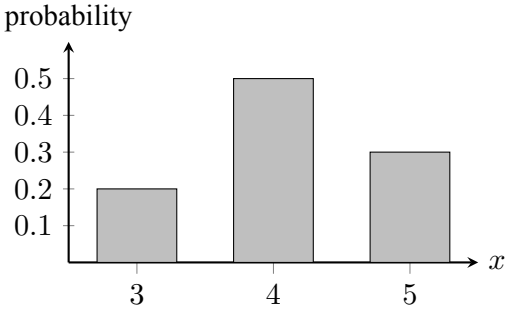
If a random variable takes *discrete* numerical values, and a probability is assigned to each, then it is a **discrete random variable**. The outcomes are often integers, but this not necessary, and whilst discrete numbers are often described as ones that can be *"listed"*, the list does not have to be finite. Some examples of discrete random variables are:

Random variable (e.g. X)	**Discrete values** (e.g. x)
The outcome of a roll of a cubical die	$1, 2, 3, 4, 5, 6$
The number of goals scored in a game of football	$0, 1, 2, 3, 4, 5, 6, 7...$
The value of a randomly picked UK coin, in pounds	$0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2$

6.1 Probability Distributions for Discrete Random Variables

A *probability function*, or *pf*, assigns probabilities to each possible value of a discrete random variable, and there are a number of ways of presenting the *probability distribution* created. For example, consider a discrete random variable X that takes the value 3 with probability 0.2, the value 4 with probability 0.5 and the value 5 with probability 0.3. Below, the probability distribution for X has been *tabulated* and *graphed*.

x	3	4	5
$P(X = x)$	0.2	0.5	0.3



Note that a discrete probability distribution must define *mutually exclusive and exhaustive* outcomes, and hence *the sum of all probabilities must equal 1*.

$0 \leqslant P(X = x_i) \leqslant 1$ for all i and $\sum P(X = x_i) = 1$

Example

Problem: The probability distribution table for the discrete random variable D is below.

d	6	7	8	9	10
$P(D = d)$	0.35	0.3	c	0.2	0.1

(a) Calculate the value of c.

(b) Calculate $P(D \leqslant 7)$.

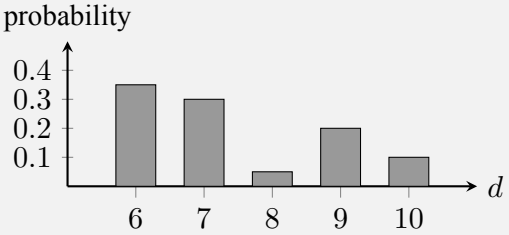
(c) Calculate $P(D > 7)$.

Solution:

(a) $c = 1 - (0.35 + 0.3 + 0.2 + 0.1) = 0.05$

d	6	7	8	9	10
$P(D = d)$	0.35	0.3	0.05	0.2	0.1

(b) $P(D \leqslant 7) = P(D = 6) + P(D = 7)$
$= 0.35 + 0.3$
$= 0.65$



(c) $P(D > 7) = 1 - P(D \leqslant 7)$
$= 1 - 0.65$
$= 0.35$

Exercise 6.1

1. Explain why each of the following tables do not represent a tabulated probability distribution.

(a)

a	7	8	9
$\mathrm{P}(A = a)$	0.6	0.3	0.2

(b)

b	2	5	10
$\mathrm{P}(B = b)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{2}{5}$

(c)

c	*red*	*blue*	*green*
$\mathrm{P}(C = c)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$

(d)

d	2	4	6
$\mathrm{P}(D = d)$	0.9	-0.3	0.4

2. Each table below represents the probability distribution of a discrete random variable. Determine the value of k for each.

(a)

q	1	2	3	4
$\mathrm{P}(Q = q)$	0.3	0.2	k	0.1

(b)

r	0	1	2
$\mathrm{P}(R = r)$	0.4	$2k$	k

(c)

s	5	10	15
$\mathrm{P}(S = s)$	$\frac{1}{10}$	$\frac{7}{10}$	k

(d)

t	1	2	3	4
$\mathrm{P}(T = t)$	$2k$	$4k$	$6k$	$8k$

3. The discrete random variable X has probability distribution:

x	1	2	3	4	5	6	7	8
$\mathrm{P}(X = x)$	0.15	0.2	0.1	0.05	c	0.1	0.08	0.02

Calculate:

(a) The value of c.
(b) $\mathrm{P}(X = 4)$.
(c) $\mathrm{P}(X < 4)$.
(d) $\mathrm{P}(X \leqslant 4)$.
(e) $\mathrm{P}(X > 4)$.
(f) $\mathrm{P}(X \geqslant 4)$.
(g) $\mathrm{P}(2 < X < 5)$.
(h) $\mathrm{P}(2 \leqslant X \leqslant 5)$.

4. A bag contains nine counters numbered 1 to 9. Two are selected at random, without replacement. Tabulate the probability distribution for the random variable E, representing the number of even numbers selected.

5. The discrete random variable X has probability distribution given by the function $\mathrm{P}(X = x) = k(x + 1)$, where k is a constant and X takes values $x = 0, 1, 2, 3$.
   (a) Tabulate the probability distribution for X.
   (b) Find $\mathrm{P}(X \geqslant 2)$.

6. The discrete random variable X has probability distribution given by the function $\mathrm{P}(X = x) = kx^2$, where k is a constant and X takes values $x = 1, 2, 3, 4$.
   (a) Tabulate the probability distribution for X.
   (b) Find $\mathrm{P}(X < 4)$.

7. Tabulate the probability distribution of each of the following discrete random variables:
   (a) F, the number of fives obtained when two cubical dice are rolled.
   (b) T, the number of tails obtained when three coins are tossed.
   (c) S, the sum obtained when two cubical dice are rolled.

6.2 The Expectation of a Discrete Random Variable

Whilst a single roll of a die is equally likely to land on any of the numbers from 1 to 6, if it is rolled many times and the *mean* of the observed values is calculated, it is very likely to be around 3.5. Letting random variable X represent the result of one roll of the die, the *long-run mean of successive observations of X will tend to 3.5*. This can be referred to as the *mean*, μ, of X, the *expectation* of X, or $\mathrm{E}(X)$.

The expectation μ of discrete random variable X:

$$\mathrm{E}(X) = \sum x_i P(X = x_i)$$

Example 1

Problem: The probability distribution of random variable X is tabulated as:

x	3	4	5	6
$P(X = x)$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$

Calculate $\mathrm{E}(X)$.

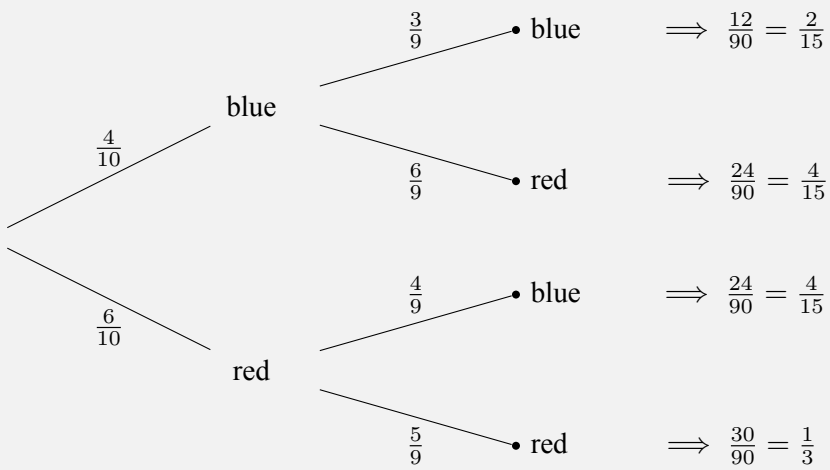
Solution:

$$\mathrm{E}(X) = 3 \times \frac{1}{8} + 4 \times \frac{1}{4} + 5 \times \frac{1}{2} + 6 \times \frac{1}{8} = \frac{37}{8}$$

Example 2

Problem: From a bag containing 4 blue and 6 red cubes one is drawn at random, followed by another without replacing the first. Letting discrete random variable X represent the number of red cubes drawn, calculate $\mathrm{E}(X)$.

Solution:



x	0	1	2
$P(X = x)$	$\frac{2}{15}$	$\frac{8}{15}$	$\frac{1}{3}$

$$\mathrm{E}(X) = 0 \times \frac{2}{15} + 1 \times \frac{8}{15} + 2 \times \frac{1}{3} = \frac{6}{5}$$

Exercise 6.2

1. Calculate the expected value for each of the probability distributions below.

(a)

x	2	4	6	8
$P(X = x)$	0.25	0.4	0.2	0.15

(b)

y	1	2	3	4	5
$P(Y = y)$	$\frac{1}{20}$	$\frac{7}{20}$	$\frac{3}{20}$	k	$\frac{1}{20}$

2. The random variable X represents the result of a roll of a fair, octahedral die with sides numbered 1 to 8.
   (a) Tabulate the probability distribution of X.
   (b) State $P(4 \leqslant X \leqslant 6)$.
   (c) Calculate $\mathrm{E}(X)$.

3. A discrete random variable X can assume values 10 and 20 only. Given $\mathrm{E}(X) = 16$, tabulate the probability distribution of X.

4. A box contains 3 red marbles and 5 green marbles. Two marbles are taken at random without replacement, and X is the number of green marbles obtained. Find the expectation of X.

5. A box contains 3 red marbles and 5 green marbles. Two marbles are taken at random with replacement, and Y is the number of green marbles obtained. Find the expectation of Y.

6. In the UKMT Senior Maths Challenge, 4 points are gained for a question answered correctly and 1 point is lost for a question answered incorrectly. A candidate guesses two questions, picking at random from the 5 choices for each question. Let random variable Y represent the points gained from those two questions.
   (a) Tabulate the probability distribution of Y.
   (b) Calculate $\mathrm{E}(Y)$.

7. A bag contains three £1 coins and two £2 coins. Two coins are taken at random, without replacement. Let discrete random variable M represent that amount of money that is taken from the bag, in pounds.
   (a) Tabulate the probability distribution of M.
   (b) Calculate μ, the mean value of M.

8. A game costs £4 to enter and consists of tossing a coin three times.
   - £8 is paid out if the coin comes up tails three times.
   - £5 is paid out if the coin comes up tails exactly twice.
   - £2 is paid out if the coin only comes up tails once.
   - Nothing is paid out if the coin shows heads three times.

   Let the random variable W represent the profit, in pounds, that a player makes if the game is played once.
   (a) Tabulate the probability distribution for W.
   (b) Calculate $\mathrm{E}(W)$.

9. The probability distribution of the discrete random variable X is tabulated below.

x	1	3	5	8	10
$P(X = x)$	0.1	a	0.3	b	0.2

   Given that $\mathrm{E}(X) = 6.3$, find the values of a and b.

6.3 The Law of Expectation

If discrete random variable X takes values $x_1, x_2...x_n$ with probabilities $P(X = x_i)$, then random variable $Y = 2X$ takes values $2x_1, 2x_2...2x_n$ with probabilities $P(X = x_i)$. For example, let X represent the roll of a die:

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

y	2	4	6	8	10	12
$P(Y = y)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$\mathrm{E}(X) = 1 \times \tfrac{1}{6} + ... + 6 \times \tfrac{1}{6} = 3.5 \qquad \mathrm{E}(2X) = \mathrm{E}(Y) = 2 \times \tfrac{1}{6} + ... + 12 \times \tfrac{1}{6} = 7$$

(Note that the distribution of $2X$ is *not* that of the sum of two *independent* dice rolls.)

It can be seen here that $\mathrm{E}(2X) = 2\mathrm{E}(X)$. The expectation of a random variable that is a *linear transformation* of another random variable for which the expectation is known can be calculated using the law:

For random variable X with constants a and b:

$$\mathrm{E}(aX \pm b) = a\mathrm{E}(X) \pm b$$

Proof:

$$\begin{aligned}
\mathrm{E}(aX + b) &= \sum (ax_i + b)\mathrm{P}(X = x_i) \\
&= \sum (ax_i)\mathrm{P}(X = x_i) + \sum b\mathrm{P}(X = x_i) \\
&= a\sum x_i\mathrm{P}(X = x_i) + b\sum \mathrm{P}(X = x_i) \\
&= a\mathrm{E}(X) + b
\end{aligned}$$

Example 1

Problem: Let X be the random variable such that $\mathrm{E}(X) = 3$. Calculate $\mathrm{E}(4X - 1)$.

Solution:

$$\begin{aligned}
\mathrm{E}(4X - 1) &= 4\mathrm{E}(X) - 1 \\
&= 4 \times 3 - 1 \\
&= 11
\end{aligned}$$

Example 2

Problem: Given tht $\mathrm{E}(10 - 2Y) = 50$, determine the value of $\mathrm{E}(Y)$.

Solution:

$$\begin{aligned}
\mathrm{E}(10 - 2Y) &= 50 \\
\mathrm{E}(-2Y + 10) &= 50 \\
-2\mathrm{E}(Y) + 10 &= 50 \\
-2\mathrm{E}(Y) &= 40 \\
\mathrm{E}(Y) &= -20
\end{aligned}$$

Exercise 6.2

1. Calculate:

   (a) $\mathrm{E}(X+4)$ given $\mathrm{E}(X)=3$ (b) $\mathrm{E}(Y-3)$ given $\mathrm{E}(Y)=0.4$ (c) $\mathrm{E}(2+Z)$ given $\mathrm{E}(Z)=-5$

2. Find:

   (a) $\mathrm{E}(3X)$ given $\mathrm{E}(X)=2.5$ (b) $\mathrm{E}(-\frac{1}{2}Y)$ given $\mathrm{E}(Y)=8$ (c) $\mathrm{E}(-W)$ given $\mathrm{E}(W)=-\frac{2}{3}$

3. Determine the value of:

   (a) $\mathrm{E}(2X-3)$ given $\mathrm{E}(X)=-4$ (b) $\mathrm{E}(-\frac{1}{3}Y+2)$ given $\mathrm{E}(Y)=12$ (c) $\mathrm{E}(3-4C)$ given $\mathrm{E}(C)=\frac{1}{2}$

4. Determine:

   (a) $\mathrm{E}(X)$ given $\mathrm{E}(4X+3)=15$ (b) $\mathrm{E}(Y)$ given $\mathrm{E}(4-Y)=2.4$ (c) $\mathrm{E}(Q)$ given $\mathrm{E}(5+\frac{2}{3}Q)=-1$

5. The probability distribution of X is tabulated as follows:

x	4	5	6	7
$P(X=x)$	0.3	0.5	k	0.05

   Determine the values of:

   (a) k (b) $\mathrm{P}(4<X\leqslant 6)$ (c) $\mathrm{E}(X)$ (d) $\mathrm{E}(4X+1)$

6. In a board game a fair, tetrahedral die with faces numbered from 1 to 4 is thrown, and the result is multiplied by 3 to give the number of places a player can move. Letting random variable M represent the number of places a player can move, determine the mean, μ, of M.

7. Random variables X and Y are defined such that the mean of X is $\frac{1}{4}$ and $Y=6X+4$. Determine the mean of Y.

8. A betting game involves a player rolling a fair, octagonal die with faces numbered from 1 to 8. Let X represent the result of a roll of the die.

   (a) Determine $E(X)$.

   Each game costs £10 to enter, with the number obtained of a single roll of the die being multiplied by 2 to give an amount in pounds that the player receives. Let Y represent the *profit* a player makes from playing the game a single time.

   (b) State an equation linking Y and X.

   (c) Determine the value of $E(Y)$.

   (d) Explain what this suggests will happen if a player plays the game repeatedly.

9. Random variables X and Y are defined such that the mean of Y is -5 and $Y=2X-10$. Determine the mean of X.

6.4 The Variance of a Random Variable

The *variance* of the values observed from successive rolls of a fair cubical die will tend towards a value in the long-run. Letting the random variable X represent the result of a roll of a die, the **variance**, σ^2, of X is $\frac{35}{12}$, or $\mathrm{V}(X) = \frac{35}{12}$.

The variance σ^2 of random variable X:

$$\mathrm{V}(X) \quad = \quad \mathrm{E}(X^2) \; - \; E^2(X)$$

Proof:

$$\begin{aligned}
\mathrm{V}(X) &= \mathrm{E}[(X-\mu)^2] \\
&= \mathrm{E}[(X-\mu)(X-\mu)] \\
&= \mathrm{E}[X^2 - 2\mu X + \mu^2] \\
&= \mathrm{E}(X^2) - 2\mu\mathrm{E}(X) + \mu^2 \\
&= \mathrm{E}(X^2) - 2(\mathrm{E}(X))^2 + (\mathrm{E}(X))^2 \\
&= \mathrm{E}(X^2) - \mathrm{E}^2(X)
\end{aligned}$$

Knowing the *expectation* and *variance* of random variables help statisticians better understand the *long-run* behaviour of repeated observations:

- $\mathrm{E}(X) = \mu$, the *long-run mean* of the distribution of X.
- $\mathrm{V}(X) = \sigma^2$, the *variance* of the distribution of X.

The standard deviation, σ, of a random variable X, notated $\mathrm{SD}(X)$, is the square root of the variance:

$$\mathrm{SD}(X) = \sqrt{\mathrm{V}(X)} \qquad \text{and} \qquad \mathrm{SD}(X), \mathrm{V}(X) \geqslant 0$$

Example

Problem: The probability distribution of random variable X is tabulated as:

x	0	1	2	3
$P(X=x)$	0.3	0.4	0.1	0.2

Calculate $\mathrm{E}(X)$ and $\mathrm{V}(X)$

Solution: *(Remember that $E(X^2)$ is not the variance.)*

$$\mathrm{E}(X) = 0 \times 0.3 + 1 \times 0.4 + 2 \times 0.1 + 3 \times 0.2 = 1.2$$

$$\mathrm{E}(X^2) = 0^2 \times 0.3 + 1^2 \times 0.4 + 2^2 \times 0.1 + 3^2 \times 0.2 = 2.6$$

$$\begin{aligned}
\mathrm{V}(X) &= \mathrm{E}(X^2) - E^2(X) \\
&= 2.6 - 1.2^2 \\
&= 1.16
\end{aligned}$$

Exercise 6.4

1. Calculate the mean and variance of each of the following probability distributions.

(a)

x	1	2	5	10
$P(X = x)$	0.4	0.3	0.2	0.1

(b)

y	0	1	2	3	4
$P(Y = y)$	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{4}$	k	$\frac{2}{5}$

2. The random variable X represents the result of a roll of a fair, tetrahedral die with sides numbered 1 to 4.

   (a) Tabulate the probability distribution of X.

   (b) Find $\mathrm{E}(X)$.

   (c) Calculate $\mathrm{V}(X)$.

3. A box contains 6 pink marbles and 4 orange marbles. Two marbles are taken at random without replacement, and X is the number of pink marbles obtained. Find the expectation and variance of X.

4. A box contains 2 red marbles and 6 green marbles. Two marbles are selected at random with replacement, and Y is the number of green marbles obtained. Find $\mathrm{E}(Y)$ and $\mathrm{V}(Y)$.

5. A discrete random variable X can assume values 4 and 12 only. Given $\mathrm{E}(X) = 10$, determine the variance of X.

6. A multiple-choice quiz gives four possible answers for each question, with 4 marks awarded for a correct answer and 1 mark taken away for an incorrect answer. Determine the expected gain in marks when the answer to a single questions is guessed, and the variance in the number of marks gained.

7. A game costs £10 to enter, and the player rolls two fair, cubical dice together. If both dice show a six the player wins £50, a single six means the player gets £10 back and no sixes means the player gets nothing. Determine the mean and variance in the amount of profit for playing the game once.

8. A fair coin is tossed three times. Let random variable X represent the number of tails shown.

   (a) Tabulate the probability distribution of X.

   (b) Find $\mathrm{E}(X)$ and $\mathrm{V}(X)$.

9. The random variable X takes values 7, 8 and 9. Given $\mathrm{P}(X = 8) = p$ and $\mathrm{P}(X = 7) = \mathrm{P}(X = 9)$:

   (a) Determine $\mathrm{E}(X)$.

   (b) Given $\mathrm{V}(X) = 0.8$, determine the value of p.

10. Random variable Y takes the value 1 with probability 0.9 and otherwise takes the value k. Determine the value of k if $\mathrm{V}(Y) = 0$.

6.5 The Law of Variance

Since variance is a measure of variability based on *squared* differences, $\mathrm{V}(2X) \neq 2\mathrm{V}(X)$.

For random variable X with constants a and b:

$$\mathrm{V}(aX \pm b) = a^2\mathrm{V}(X)$$

Proof:

$$\begin{aligned}
\mathrm{V}(aX+b) &= \mathrm{E}[((aX+b)-(\mathrm{E}(aX+b))^2] \\
&= \mathrm{E}[(aX+b-a\mathrm{E}(X)-b)^2] \\
&= \mathrm{E}[(aX-a\mathrm{E}(X))^2] \\
&= \mathrm{E}[(a(X-\mathrm{E}(X)))^2] \\
&= \mathrm{E}[a^2(X-\mathrm{E}(X))^2] \\
&= a^2\mathrm{E}[(X-\mathrm{E}(X))^2] \\
&= a^2\mathrm{V}(X)
\end{aligned}$$

Note that any laws relating to the *standard deviation* of a linear transformation ($\mathrm{SD}(aX \pm b)$) would not be able to be expressed so concisely, so such problems must be tackled by *first finding the variance* ($\mathrm{V}(aX \pm b)$).

Example 1

Problem: Let X be the random variable such that $\mathrm{E}(X) = -7$ and $\mathrm{V}(X) = 4$. Calculate $\mathrm{V}(2-3X)$.

Solution: *(It helps to rearrange into the form $aX \pm b$.)*

$$\begin{aligned}
\mathrm{V}(2-3X) &= \mathrm{V}(-3X+2) \\
&= (-3)^2\mathrm{V}(X) \\
&= 9 \times 4 \\
&= 36
\end{aligned}$$

Example 2

Problem: Random variable Y has mean 1.5 and standard deviation 0.8.
Determine the standard deviation of $10Y + 25$.

Solution:

$$\begin{aligned}
\mathrm{V}(10Y+25) &= 10^2\mathrm{V}(Y) \\
&= 100 \times 0.8^2 \\
&= 64 \\
\mathrm{SD} &= 8
\end{aligned}$$

Exercise 6.5

1. Given $\mathrm{E}(X) = 6$ and $\mathrm{V}(X) = 8$, determine the values of:

   (a) $\mathrm{E}(2X)$ (b) $\mathrm{V}(2X)$ (c) $\mathrm{SD}(2X)$

2. Given $\mathrm{E}(X) = -2$ and $\mathrm{V}(X) = 3$, determine the values of:

   (a) $\mathrm{E}(X + 5)$ (b) $\mathrm{V}(X + 5)$ (c) $\mathrm{SD}(X + 5)$

3. Given $\mathrm{E}(X) = 1.25$ and $\mathrm{SD}(X) = 0.64$, determine the values of:

   (a) $\mathrm{E}(3X - 2)$ (b) $\mathrm{V}(3X - 2)$ (c) $\mathrm{SD}(3X - 2)$

4. Given $\mathrm{E}(X) = 500$ and $\mathrm{V}(X) = 400$, determine the values of:

   (a) $\mathrm{E}(0.1X + 10)$ (b) $\mathrm{V}(0.1X + 10)$ (c) $\mathrm{SD}(0.1X + 10)$

5. The discrete random variable X has the following probability distribution.

x	1	2	4	8
$\mathrm{P}(X = x)$	0.1	0.5	0.3	0.1

   (a) Calculate $E(X)$ and $V(X)$. (b) Find $\mathrm{E}(2X + 1)$. (c) Find $\mathrm{V}(2X + 1)$.

6. Given $\mathrm{E}(Y) = -5$ and $\mathrm{V}(Y) = 3$ determine the values of:

   (a) $\mathrm{E}(-3Y - 2)$ (b) $\mathrm{V}(-3Y - 2)$ (c) $\mathrm{SD}(-3Y - 2)$

7. Random variable X has a mean of 2.4 and a standard deviation of 0.8. Find:

   (a) $\mathrm{E}(1 - 10X)$ (b) $\mathrm{V}(1 - 10X)$ (c) $\mathrm{SD}(1 - 10X)$

8. Given random variable X such that $\mathrm{E}(4X - 1) = 23$ and $\mathrm{V}(4X - 1) = 80$, determine the values of $\mathrm{E}(X)$ and $\mathrm{V}(X)$.

9. Given random variable Y such that $\mathrm{E}(3Y + 12) = 0$ and $\mathrm{V}(3Y + 12) = 81$, determine the mean and standard deviation of Y.

10. Given random variables X and Z such that $\mathrm{E}(Z) = 0$, $\mathrm{V}(Z) = 1$ and $X = a + bZ$:

    (a) State $\mathrm{SD}(Z)$.

    (b) Find $\mathrm{E}(X)$.

    (c) Determine the values of $\mathrm{V}(X)$ and $\mathrm{SD}(X)$.

6.6 Bivariate Random Variables

As well as linear transformations of a single random variable, new variables can be created by adding (or subtracting) random variables. Given random variables X and Y, for example, $W = X + Y$ is a *bivariate* random variable. When studying the long-run behaviour of bivariate random variables, by considering their expectation and variance, it is important to consider whether the variables are **independent**.

For **independent** random variables X and Y:

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y) \qquad \mathrm{V}(X+Y) = \mathrm{V}(X) + \mathrm{V}(Y)$$

$$\mathrm{E}(X-Y) = \mathrm{E}(X) - \mathrm{E}(Y) \quad \text{and} \quad \mathrm{V}(X-Y) = \mathrm{V}(X) + \mathrm{V}(Y)$$

$$\mathrm{E}(aX \pm bY) = a\mathrm{E}(X) \pm b\mathrm{E}(Y) \qquad \mathrm{V}(aX \pm bY) = a^2\mathrm{V}(X) + b^2\mathrm{V}(Y)$$

Example 1

Problem: Two independent random variables F and G have means of 15 and 12 and standard deviations of 3 and 4 respectively. Calculate $\mathrm{E}(3G - 2F)$ and $\mathrm{V}(F - G)$.

Solution: ($\mathrm{E}(X) = 15, \mathrm{E}(Y) = 12, \mathrm{SD}(X) = 3, \mathrm{SD}(Y) = 4$)

$$\begin{aligned}\mathrm{E}(3G - 2F) &= 3\mathrm{E}(G) - 2\mathrm{E}(F)\\ &= 3 \times 12 - 2 \times 15\\ &= 36 - 30\\ &= 6\end{aligned}$$

$$\begin{aligned}\mathrm{V}(F) &= 3^2 = 9\\ \mathrm{V}(G) &= 4^2 = 16\\ \mathrm{V}(F - G) &= \mathrm{V}(F) + \mathrm{V}(G)\\ &= 9 + 16\\ &= 25\end{aligned}$$

Example 2

Problem: Books in the fiction section of a library have a mean thickness of 3.1cm with standard deviation 0.2cm. If three books are randomly selected from the fiction section and stacked together, determine the mean and standard deviation of the collected, total thickness of the three books, stating one assumption required.

Solution:
Let random variable X represent the thickness of a book, so $\mathrm{E}(X) = 3.1$ and $\mathrm{V}(X) = 0.2^2 = 0.04$

Total thickness $T = X_1 + X_2 + X_3$. Assuming that the thickness of each book is independent:

$\mathrm{E}(T) = \mathrm{E}(X_1 + X_2 + X_3) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \mathrm{E}(X_3) = 3.1 + 3.1 + 3.1 = 9.3$

$\mathrm{V}(T) = \mathrm{V}(X_1 + X_2 + X_3) = \mathrm{V}(X_1) + \mathrm{V}(X_2) + \mathrm{V}(X_3) = 0.04 + 0.04 + 0.04 = 0.12$

$\mathrm{SD}(T) = \sqrt{\mathrm{V}(T)} = \sqrt{0.12} = 0.346$

Hence the mean total thickness of the three books is 9.3cm, and the standard deviation is 0.346cm.

Exercise 6.6

1. Independent random variables X and Y are defined such that $\mathrm{E}(X) = 5$, $\mathrm{V}(X) = 6$, $\mathrm{E}(Y) = 3$ and $\mathrm{V}(Y) = 8$. Determine the values of:

   (a) $\mathrm{E}(2X + 6Y)$
   (b) $\mathrm{V}(2X + 6Y)$
   (c) $\mathrm{E}(4X + 5Y)$
   (d) $\mathrm{V}(4X + 5Y)$
   (e) $\mathrm{E}(\frac{1}{2}Y - 2X)$
   (f) $\mathrm{V}(\frac{1}{2}Y - 2X)$

2. Random variables F and G have means of 3 and 2 respectively, and standard deviations of 5 and 6 respectively. State the information that is missing which means that it currently is not possible to determine the values of $\mathrm{E}(2F + G)$ and $\mathrm{V}(2F + G)$.

3. Independent random variables A and B have means of 1.5 and 1.2 respectively, and standard deviations of 1 and 0.8 respectively. Determine the values of $\mathrm{E}(A - B)$ and $\mathrm{V}(A - B)$.

4. Random variable X is defined such that $\mathrm{E}(X) = 10$ and $\mathrm{V}(X) = 3$. Random variables $\{X_1, X_2, X_3 \ldots\}$ represent independent observations of X. Determine the mean and variance of $X_1 + X_2 + X_3 + X_4$, representing the sum of four such observations of X.

5. A spreadsheet contains a large amount of cells, each containing numbers such that the mean of the numbers is 12 and the standard deviation is 1.2. Five numbers from the spreadsheet are selected at random, with replacement. Determine the mean and standard deviation of the total of the five numbers.

6. It is known that the mean volume of coffee poured per cup in a staffroom for teachers is 250ml, with a standard deviation of 10ml. Determine the mean and standard deviation of the total volume of coffee used for six cups, stating an assumption required.

7. Let random variable X represent the number obtained when a fair, cubical die numbered 1 to 6 is thrown, and random variable Y the number obtained when a fair, tetrahedral die numbered 1 to 4 is thrown.

   (a) Find $\mathrm{E}(X)$ and $\mathrm{E}(Y)$.

   (b) Find $\mathrm{V}(X)$ and $\mathrm{V}(Y)$.

   (c) Obtain the mean and variance of the result when both dice are thrown and the scores are added together.

8. The mean length of a car is 4.5 metres, with a standard deviation of 30 centimetres. Drivers arriving to queue for a ferry are told to line up one in front of another leaving as small a gap as possible between cars. When they do this, the mean gap they leave is 40 centimetres, with a standard deviation of 10 centimetres. Stating two assumptions required, determine the mean and standard deviation of the total length of a line of eight cars asked to line up in this way.

Review Exercise

1. A trial consists of tossing two unbiased coins. Let the random variable X represent the number of heads obtained.

   (a) Tabulate the probability distribution of X after one trial.

   (b) Calculate:

   i. $\mathrm{E}(X)$.

   ii. $\mathrm{V}(X)$.

   iii. $\mathrm{SD}(X)$.

2. Two independent random variables P and Q have means 5 and 10 and standard deviations 2 and 3 respectively. Find the mean and variance of each of the random variables:

   (a) $4P - 8$.

   (b) $3Q - P$.

3. A box contains 4 blue marbles and 5 red marbles. Three marbles are taken from the box without replacement. Let the random variable X represent the number of red marbles taken.

   (a) Tabulate the probability distribution of X.

   (b) Calculate the mean and variance of X.

4. Given that $\mathrm{E}(2X + 6) = 16$ and $\mathrm{V}(2X + 6) = 12$:

   (a) Calculate $\mathrm{E}(X)$.

   (b) Calculate $\mathrm{V}(X)$.

5. Two independent random variables A and B have means 53 and 21 and standard deviations 6 and 5 respectively. Find the mean and standard deviations of each of the following random variables:

   (a) $5B + 6$.

   (b) $2A - 3B$.

6. It is known that items of hand luggage carried by passengers on an airline have mean mass of 5.6kg and a standard deviation of 2.0kg, and that their hold luggage has a mean mass of 22.0kg and a standard deviation 5.0kg.

   (a) Calculate the mean and standard deviation of the total luggage mass for one passenger, stating an assumption you have made.

   The passengers themselves have masses with mean 75kg and standard deviation 12kg. A small plane carries 10 passengers.

   (b) Calculate the mean and standard deviation of the total load, including passengers and their luggage.

7

Discrete Distributions

"*All models are wrong, but some models are useful.*"

Models, Conditions and Assumptions

The above quote, attributed to the statistician George Box, highlights that no mathematical or statistical *model* can perfectly describe the behaviour of something in the real world. Nevertheless, if a model can get *close enough* then results obtained using it can be sufficiently accurate, and useful.

In this chapter, three *discrete probability distributions* that can model real-life data will be introduced:

- The Discrete Uniform Distribution
- The Binomial Distribution
- The Poisson Distribution

Studying a problem in context and identifying an appropriate model to use is an important skill for anyone working with data. The validity of the use of a model is based on a number of *conditions* being sufficiently met. When using a model it is vital to consider the conditions required for its use, and specifically whether they are each satisfied. In this chapter, and subsequent chapters introducing more probability distributions and hypothesis testing, the *conditions for the valid use* of models and tests are clearly indicated as they are introduced.

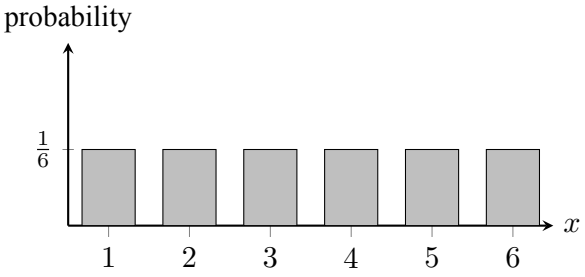
Sometimes it will be clear that a particular condition has been met, but there will be times when it is unclear. Whenever it is not clear that a condition has been met, statisticians must ask whether they can reasonably make an *assumption* that the condition is satisfied. It should always be recognised when an assumption is required, and it may be necessary to *justify* the use of an assumption in a context.

7.1 The Discrete Uniform Distribution

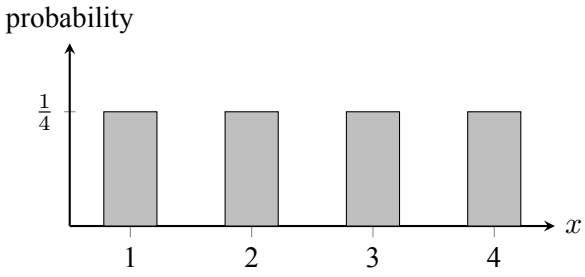
Consider the discrete random variables X and Y, representing the scores obtained from a cubical die numbered $\{1, 2, 3, 4, 5, 6\}$ and a tetrahedral die numbered $\{1, 2, 3, 4\}$ respectively. Their probability distributions have been tabulated and graphed below, and the expectation and variance calculated for each:

x	1	2	3	4	5	6
$\mathrm{P}(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

y	1	2	3	4
$\mathrm{P}(Y = y)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$



$\mathrm{E}(X) = 3.5$ and $\mathrm{V}(X) = \frac{35}{12}$



$\mathrm{E}(Y) = 2.5$ and $\mathrm{V}(Y) = \frac{5}{4}$

Whilst X and Y do not follow the exact same distribution, they share key characteristics, namely that they both take a *finite number of discrete values*, $\{1, 2, 3, \ldots, k\}$, and that *each of these values is equally likely*. It is these *shared characteristics* that define the *discrete uniform distribution*.

The Discrete Uniform Distribution:

- Notated $X \sim \mathrm{U}(k)$.
- Defined by one parameter, k.
- Takes k discrete values from 1 to k: $\{1, 2, 3, \ldots, k\}$.
- Each value is *equally likely*, with probability $\dfrac{1}{k}$.

Parameters define the precise distribution that random variables take within *"families"* of distributions. Here, we can say that X *"is distributed uniformly*" with parameter $k = 6$ (with values from 1 to 6). Using $\sim$, or *"tilde"*, to mean *"is distributed..."*, such statements can be notated more efficiently:

$$X \sim \mathrm{U}(6) \quad \text{and} \quad Y \sim \mathrm{U}(4)$$

The SQA Data Booklet gives the formulae to calculate *expectation* and *variance* for the discrete uniform distribution. For $X \sim \mathrm{U}(k)$:

$$\mathrm{E}(X) = \frac{k+1}{2} \quad \text{and} \quad \mathrm{V}(X) = \frac{k^2 - 1}{12}$$

The previously stated values for the expectation and variance of X and Y can be quickly obtained using these formulae.

Example

Problem: Given $T \sim \mathrm{U}(10)$ and $W = 3T + 1$, calculate $\mathrm{E}(W)$ and $\mathrm{V}(W)$.

Solution:

(a)
$$\begin{aligned}\mathrm{E}(T) &= \frac{10+1}{2} = 5.5\\ \mathrm{E}(W) &= \mathrm{E}(3T+1)\\ &= 3\mathrm{E}(T)+1\\ &= 3\times 5.5+1\\ &= 17.5\end{aligned}$$

(b)
$$\begin{aligned}\mathrm{V}(T) &= \frac{10^2-1}{12} = 8.25\\ \mathrm{V}(W) &= \mathrm{V}(3T+1)\\ &= 3^2\mathrm{V}(T)\\ &= 9\times 8.25\\ &= 74.25\end{aligned}$$

Exercise 7.1

1. Let $X \sim \mathrm{U}(5)$. Calculate the mean and variance of X.

2. Let $X \sim \mathrm{U}(15)$. Calculate the mean and variance of X.

3. A discrete random variable is uniformly distributed such that $\mathrm{P}(X = x) = \frac{1}{k}$, with $k = 8$.
   (a) Calculate $\mathrm{E}(X)$ and $\mathrm{V}(X)$.
   (b) Evaluate $\mathrm{E}(2X + 5)$ and $\mathrm{V}(2X + 5)$.

4. A regular, unbiased icosahedral die is rolled. Let the random variable X be the score on the top most face.
   (a) State the distribution of X.
   (b) Determine the values of $\mathrm{E}(X)$ and $\mathrm{V}(X)$.
   (c) Evaluate $\mathrm{E}(1 - 3X)$ and $\mathrm{V}(1 - 3X)$.

5. Two independent random variables X and Y are defined such that $X \sim \mathrm{U}(3)$ and $Y \sim \mathrm{U}(5)$.
   (a) Calculate $\mathrm{E}(X - Y)$.
   (b) Calculate $\mathrm{V}(X - Y)$.

6. Given $X \sim \mathrm{U}(k)$ such that $\mathrm{E}(X) = 5.5$:
   (a) Determine the value of k.
   (b) State $\mathrm{P}(2 \leqslant X < 6)$.
   (c) Find the standard deviation of X.

7. Given $Y \sim \mathrm{U}(k)$ such that the standard deviation of Y is $\frac{\sqrt{21}}{2}$:
   (a) Determine the value of k.
   (b) Find $\mathrm{E}(1 - 2Y)$.

8. Discrete, independent random variables X and Y are uniformly distributed such that:
   - $\mathrm{P}(X = 1) < \mathrm{P}(Y = 1)$
   - $\mathrm{E}(X - Y) = 2$ and $\mathrm{V}(X - Y) = 14$

   Determine the distributions of X and Y.

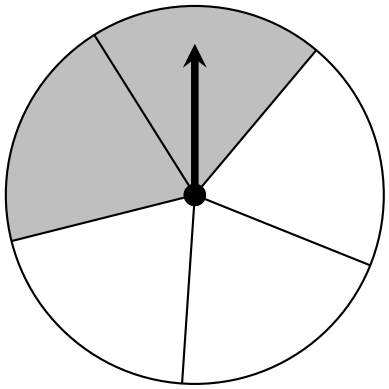
7.2 The Binomial Distribution

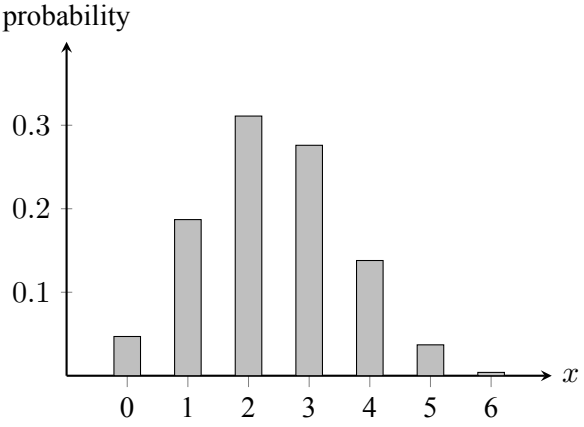
A *Bernoulli trial* is an experiment in which each trial has only two possible outcomes, such as a coin landing on either *heads* or *tails*, or asking a *yes or no* question in a survey. These responses can be more generally referred to as *successes* or *failures*. If random variable X represents the **number of successes** in ***n* trials**, each trial having **probability of success *p***, then X is **binomially distributed**.

The Binomial Distribution:

- Notated $X \sim \mathrm{B}(n, p)$.
- Defined by two parameters, n and p.
- Random variable X represents the number of successes in a **fixed** number of trials, n.
- Each trial has the **same** probability of success, p.
- Each trial must be **independent**.

Consider a spinning wheel that lands on a shaded sector with probability $\frac{2}{5}$, or 0.4. Let random variable X represent the number of times the wheel will land on the shaded sector *out of 6 trials*, with possible values from 0 to 6. Here X is *binomially distributed*, with $n = 6$ (number of trials) and $p = 0.4$ (probability of success), or: $X \sim \mathrm{B}(6, 0.4)$





To calculate the probability of *exactly two successes*, consider the probability of the first two spins resulting in success followed by four failures. Since each trial is independent:

$$\mathrm{P}(success, success, fail, fail, fail, fail) = 0.4 \times 0.4 \times 0.6 \times 0.6 \times 0.6 \times 0.6 = 0.4^2 \times 0.6^4 = 0.0207$$

However there are a number of other possible *combinations* of successes and failures that *also* give exactly two successes, *each* with probability 0.0207. The number of possible combinations from n trials resulting in x successes can be calculated using the *binomial coefficient*, ${}^nC_x$. For 6 trials of which 2 are successes, a calculator will provide the value of ${}^6C_2$.

$$\mathrm{P}(X = 2) = {}^6C_2 \times 0.4^2 \times 0.6^4 = 0.3110$$

The SQA Data Booklet gives the **formula** to calculate *individual* probabilities for the binomial distribution:

$$\text{Given } X \sim \mathrm{B}(n, p): \qquad \mathrm{P}(X = x) = {}^nC_x \times p^x \times q^{n-x}$$

Example 1

Problem: Given $X \sim \text{B}(7, \frac{1}{3})$, calculate $\text{P}(X = 1)$.

Solution:

$$\text{P}(X = 1) = {}^{7}C_{1} \times \left(\frac{1}{3}\right)^{1} \times \left(\frac{2}{3}\right)^{6} = 0.2048$$

Example 2

Problem: 12 students each take it in turns to select a playing card from a pack at random, before returning it. Calculate the probability that the card drawn is a Spade for fewer than 2 of the students.

Solution:
Let random variable X represent the number of times that the card drawn is spades from the 12 draws.

$$X \sim B\left(12, \frac{1}{4}\right)$$

$$\begin{aligned}
\text{P}(X < 2) &= \text{P}(X = 0) + \text{P}(X = 1) \\
&= {}^{12}C_{0} \times \left(\frac{1}{4}\right)^{0} \times \left(\frac{3}{4}\right)^{12} + {}^{12}C_{1} \times \left(\frac{1}{4}\right)^{1} \times \left(\frac{3}{4}\right)^{11} \\
&= 0.0317 + 0.1267 \\
&= 0.1584
\end{aligned}$$

Exercise 7.2

1. Given that $X \sim \text{B}(10, 0.25)$, calculate $\text{P}(X = 3)$.

2. Given that $X \sim \text{B}(8, 0.6)$, calculate $\text{P}(X = 5)$.

3. Given that $Y \sim \text{B}(13, 0.4)$, calculate $\text{P}(Y \leqslant 2)$.

4. 5% of bluebells have white flowers. The remainder have blue flowers. Determine the probability that a random sample of twelve bluebell plants includes exactly one with white flowers.

5. Amongst a group of four friends, let random variable X represent the number of them that were born on a Tuesday.
   (a) State the distribution of X.
   (b) Calculate the probability that only one of the four was born on a Tuesday.

6. In Bob's form class of 20 pupils, their form teacher assigns each pupil a number from 1 to 20 using the order listed on the register and, each morning, uses a random number generator to select which pupil should be responsible for read out notices. Calculate the probability that Bob is picked exactly twice in one week.

7.3 Binomial Calculations using Tables and Calculators

In practice, it will seldom be necessary to calculate binomial probabilities using the formula. Graphical calculators can be used to calculate both *individual* ($\mathrm{P}(X = x)$) and *cumulative* ($\mathrm{P}(X \leqslant x)$) probabilities. The Data Booklet provides **cumulative probability tables**, giving **"less than or equal to"** probabilities.

It is therefore important to be able to write any given probability that uses an inequality in terms of a *less than or equal to* probability.

Consider discrete random variable X taking values $x = \{0, 1, 2, 3, 4, 5, 6\}$:

$$\begin{aligned} \mathrm{P}(X < 5) &= \mathrm{P}(X \leqslant 4) && \{0, 1, 2, 3, 4, 5, 6\} \\ \mathrm{P}(X > 5) &= 1 - \mathrm{P}(X \leqslant 5) && \{0, 1, 2, 3, 4, 5, 6\} \\ \mathrm{P}(X = 5) &= \mathrm{P}(X \leqslant 5) - \mathrm{P}(X \leqslant 4) && \{0, 1, 2, 3, 4, 5, 6\} \\ \mathrm{P}(1 < X < 5) &= \mathrm{P}(X \leqslant 4) - \mathrm{P}(X \leqslant 1) && \{0, 1, 2, 3, 4, 5, 6\} \end{aligned}$$

Example

Problem: It is estimated that around 10% of people are left-handed. Taking this percentage to be accurate, calculate the probability that, in a random sample of 12 people, at least two of them will be left-handed.

Solution: *(Define a random variable, in context, and define the problem in probability terms.)*

Let random variable X represent the number of left-handed people out of the sample of 12.

$p = 0.1$ and $n = 12$

$X \sim \mathrm{B}(12, 0.1)$

$\mathrm{P}(X \geqslant 2) = 1 - \mathrm{P}(X \leqslant 1) = 1 - 0.6590 = 0.3410$

Exercise 7.3

1. Given that $X \sim \mathrm{B}(10, 0.3)$, calculate:

   (a) $\mathrm{P}(X \leqslant 6)$ (b) $\mathrm{P}(X = 6)$ (c) $\mathrm{P}(X < 6)$ (d) $\mathrm{P}(X > 6)$ (e) $\mathrm{P}(X \geqslant 6)$

2. Given that $X \sim \mathrm{B}(6, 0.15)$, calculate:

   (a) $\mathrm{P}(X \leqslant 2)$ (b) $\mathrm{P}(X \geqslant 2)$ (c) $\mathrm{P}(1 \leqslant X < 3)$

3. Given that $X \sim \mathrm{B}(14, 0.45)$, calculate:

   (a) $\mathrm{P}(X < 8)$ (b) $\mathrm{P}(X > 12)$ (c) $\mathrm{P}(3 < X \leqslant 6)$

4. Given that $X \sim \mathrm{B}(8, 0.8)$, calculate:

   (a) $\mathrm{P}(X = 5)$ (b) $\mathrm{P}(X \geqslant 7)$ (c) $\mathrm{P}(3 < X < 6)$

7.4 The Mean and Variance of the Binomial Distribution

The SQA Data Booklet gives the formulae for the *mean* and *variance* of binomial distribution.

Given random variable X such that $X \sim \mathrm{B}(n, p)$:

$$\mathrm{E}(X) = np \qquad \text{and} \qquad \mathrm{V}(X) = npq$$

Where $q = 1 - p$, representing the *probability of failure*.

Example

Problem: A multiple choice quiz is comprised of eight questions, each with four choices to pick from. Let random variable X represent the number of correct answers obtained out of eight from a student guessing completely at random. Calculate the mean and variance of X.

Solution:
$X \sim \mathrm{B}(8, 0.25)$

$\mathrm{E}(X) = np = 8 \times 0.25 = 2$

$\mathrm{V}(X) = npq = 8 \times 0.25 \times 0.75 = 1.5$

Exercise 7.4

1. Calculate the mean and variance for each random variable:

   (a) $X \sim \mathrm{B}(4, 0.2)$ (b) $Y \sim \mathrm{B}(6, \frac{1}{3})$ (c) $X \sim \mathrm{B}(4, 0.1)$ (d) $Y \sim \mathrm{B}(30, 0.4)$

2. An octahedral die numbered 1 to 8 is rolled ten times. Let the random variable S represent the number of times the die shows a square number from the ten rolls.

   (a) State the distribution of S.

   (b) Calculate $\mathrm{E}(S)$ and $\mathrm{V}(S)$.

   (c) Determine $\mathrm{P}(S > 5)$.

3. Given that $X \sim \mathrm{B}(6, p)$ such that $\mathrm{E}(X) = 1.2$:

   (a) Determine the value of p.

   (b) Calculate $\mathrm{SD}(X)$.

4. Given that $Y \sim \mathrm{B}(12, p)$ such that $\mathrm{V}(Y) = \frac{8}{3}$ and $p > 0.5$:

   (a) Determine the value of p.

   (b) Calculate $\mathrm{E}(Y)$.

5. Given that $X \sim \mathrm{B}(n, p)$ such that X has mean and variance of 16 and 3.2 respectively:

   (a) Determine the values of n and p.

   (b) Calculate $\mathrm{P}(5 < X < 10)$.

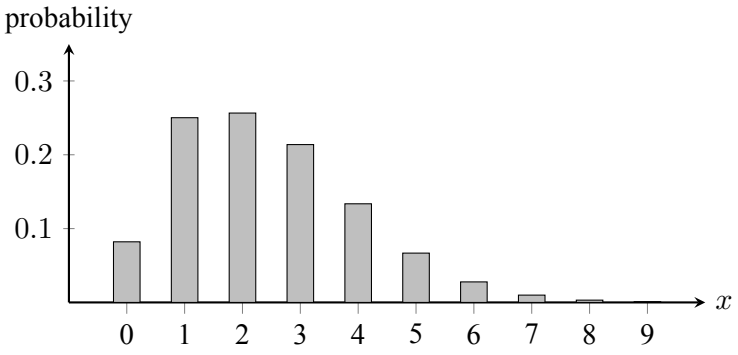
7.5 The Poisson Distribution

Consider a mobile network's customer support department looking at how many call agents to have on duty at any one time. The mean rate of calls may be well known, such as 2.5 per hour. However, in any one hour there may be 4 calls, 10 calls, no calls at all, and so on. Letting the random variable X represent the number of calls in any particular hour, if certain conditions are satisfied then X follows a **Poisson distribution** with mean rate $\lambda = 2.5$.

The Poisson Distribution:

- Notated $X \sim \text{Po}(\lambda)$.
- Defined by one parameter λ *("lambda")*.
- Random variable X represents the number of events occurring in a set amount of time or space.
- λ represents the **mean rate of events per unit of time or space**.
- Each event must be **independent** and **occur singly**.
- The mean rate must be **constant across the time period/space**.

Assume that the mean rate of 2.5 phone calls per hour is *constant* across the opening hours of the customer support department, and that there isn't some major network-wide event causing the calls to no longer be *independent*. The Poisson distribution can then be used to calculate, for example, the probability that there are exactly 4 calls in any one hour. Note that the graph below is truncated, and the Poisson distribution has non-zero probabilities for any non-negative integer $\{0, 1, 2, 3, ...\}$.



As with the binomial distribution, individual Poisson probabilities can calculated using a formula, provided in the SQA Data Booklet.

Given random variable $X \sim \text{Po}(\lambda)$:

$$\text{P}(X = x) = \frac{e^{-\lambda} \times \lambda^x}{x!}$$

For example $\text{P}(X = 4)$ in the probability distribution above can be calculated as:

$$\text{P}(X = 4) = \frac{e^{-2.5} \times 2.5^4}{4!} = 0.1336$$

Example

Problem: During a spell of rain, drops of rain fall on a lawn at a constant mean rate of 6 drops per square metre per second. Calculate the probability that exactly 10 drops fall in one square metres of the lawn in a second.

Solution: *(Define a random variable, in context, and define the problem in probability terms.)*

Let random variable X represent the number of drops of rain.

$X \sim \text{Po}(6)$

$$\text{P}(X = 10) = \frac{e^{-6} \times 6^{10}}{10!} = 0.0413$$

Exercise 7.5

1. Given random variable X such that $X \sim \text{Po}(2.5)$, use the formula to calculate:
   (a) $\text{P}(X = 4)$.
   (b) $\text{P}(X = 0)$.
   (c) $\text{P}(X < 2)$.
   (d) $\text{P}(X \geqslant 2)$.

2. Let random variable X represent the number of goals in a game of Sunday League football, known to follow a Poisson distribution with a mean of 4 goals.
   (a) State the distribution of X.
   (b) Use the formula to calculate the probability that a game finishes without any goals being scored.

7.5.1 Poisson Calculations using Tables and Calculators

As with the binomial distribution, in general it is far quicker to use the cumulative probability tables in the SQA Data Booklet, or a grapical calculator, to calculate Poisson probabilities.

Exercise 7.6

1. The random variable X follows a Poisson distribution with mean 6. Use the tables to calculate:
   (a) $\text{P}(X < 4)$. (b) $\text{P}(X > 5)$. (c) $\text{P}(X \leqslant 2)$. (d) $\text{P}(X \geqslant 7)$.

2. The random variable Y follows a Poisson distribution with mean 1.5. Use the tables to calculate:
   (a) $\text{P}(Y < 1)$. (b) $\text{P}(3 < Y < 6)$. (c) $\text{P}(1 \leqslant Y \leqslant 6)$. (d) $\text{P}(Y \geqslant 3)$.

3. Whilst watching a section of the night sky, the number of visible shooting stars is known to follow a Poisson distribution with a mean rate of one per hour. Let random variable X represent the number of shooting stars seen in an hour.
   (a) State the distribution of X.
   (b) Calculate the probability of more than three shooting stars being seen in any given hour.

7.6 The Mean and Variance of the Poisson Distribution

The mean and variance of a Poisson distribution are stated in the SQA Data Booklet. Given $X \sim \text{Po}(\lambda)$:

$$\text{E}(X) = \text{V}(X) = \lambda \qquad \text{and} \qquad \text{SD}(X) = \sqrt{\lambda}$$

Given two **independent** random variables X and Y such that $X \sim \text{Po}(\lambda_X)$ and $Y \sim \text{Po}(\lambda_Y)$:

$$X + Y \sim \text{Po}(\lambda_X + \lambda_Y)$$

The same result can be used to extend, or shorten, the time period of interest for a Poisson distribution, provided there is a *constant mean rate* throughout the entire time period.

Example

Problem: Betty has observed that the number of work emails she receives on weekdays between 10am and 11am follows a Poisson distribution with a mean rate of 7 emails. The number of personal emails she receives in this time is also Poisson distributed with a mean rate of 3 emails. Calculate the probability that she receives fewer than 8 emails in total on a weekday between 10am and 11am.

Solution:
Let random variable X represent the number of work emails, and Y the number of personal emails.

$\lambda_X = 7$ and $X \sim \text{Po}(7)$

$\lambda_Y = 3$ and $Y \sim \text{Po}(3)$

$\lambda_X + \lambda_Y = 7 + 3 = 10$, so $X + Y \sim \text{Po}(10)$

$\text{P}(X + Y < 8) = 0.2202$

Exercise 7.7

1. Given independent random variables X,Y such that $X \sim \text{Po}(1.6)$ and $Y \sim \text{Po}(0.4)$, calculate $\text{P}(X + Y \leqslant 4)$.

2. Given $F \sim \text{Po}(3.3)$ and $G \sim \text{Po}(1.2)$, with F and G independent, calculate $\text{P}(F + G \geqslant 2)$.

3. For a restaurant's Glasgow branch, the number of bottles of champagne sold on a Saturday night follows a Poisson distribution with a mean rate of 3 per night. The number sold for their Edinburgh branch is Poisson distributed with a mean of 5 per night. Stating an assumption required, calculate the probability at least 10 but fewer than 15 bottles of champagne are sold on a Saturday night between the two branches.

4. Shooting stars can be seen in a section of night sky at a mean rate of 1 per hour, and the number visible in an hour is Poisson distributed. Calculate the probability of watching the night sky for two hours and not seeing a single shooting star.

5. A data storage company using a large number of hard drives has seen that the number of drives failing per day follows a Poisson distribution with a constant mean rate of 1.5 per day. The company has staff monitoring the servers that contain the drives 24 hours per day, working in three separate 8 hour shifts each. Calculate the probability of more than 2 drives failing during one shift.

Review Exercise

1. Given that $X \sim \mathrm{U}(8)$, calculate:
   (a) $\mathrm{P}(X < 6)$.
   (b) The mean and variance of X.

2. Given that $X \sim \mathrm{B}(12, 0.35)$:
   (a) Find $\mathrm{P}(X = 3)$.
   (b) Find $\mathrm{P}(5 < X < 9)$.

3. X and Y are independent random variables such that $X \sim \mathrm{Po}(1.7)$ and $Y \sim \mathrm{Po}(6.8)$, find:
   (a) $\mathrm{P}(X + Y = 7)$.
   (b) $\mathrm{P}(X + Y \geqslant 9)$

4. The random variable X follows a Poisson distribution with standard deviation 2. Find $\mathrm{P}(X \leqslant 3)$.

5. A gardener plants seeds in batches of 20 and knows that only 85% of seeds actually germinate.
   (a) Find the chance of more that 15 seeds in the batch germinating.
   (b) Find the mean and standard deviation of the number of seeds that germinate.
   (c) State one assumption required for the valid use of the probability distribution used.

6. A restaurant has two food mixers, A and B. The number of times per week that mixer A breaks down has a Poisson distribution with mean 0.4, while independently the number of times that mixer B breaks down has a Poisson distribution with probability 0.1. Calculate the probability that in the next three weeks:
   (a) Mixer A will not break down at all.
   (b) There will be a total of 2 breakdowns.

7. A restaurant menu contains 20 dishes, consisting of starters, main courses and desserts. The table below breaks down the number of dishes for each course and by whether or not they are vegetarian:

	Starter	Main	Dessert
Vegetarian	2	2	6
Not Vegetarian	4	6	0

   Each week, the owner chooses one dish at random to check that the quality is up to standard. Each dish is equally likely to be selected, and a dish being chosen one week does not change the probability that is will be chosen in the following weeks.
   (a) Given that a dish selected to be checked is vegetarian, state the probability that it is a main course.

   Let random variable X represent the number of the dishes checked over a 8 week period that are desserts.
   (b) State the distribution of X.
   (c) Calculate the probability that more than half of the dishes checked in the 8 week period are dessert dishes.

8. For each random variable, calculate the probability of the value of an observation being more than one standard deviation above the mean.
   (a) $X \sim \mathrm{U}(20)$.
   (b) $Y \sim \mathrm{Po}(6)$.
   (c) $Q \sim \mathrm{B}(16, 0.2)$.

8

Continuous Distributions

Data that results from measuring is often continuous, such as the height of a person, the mass of a piece of fruit or the volume of tea in a cup. *Continuous* means that if there were no limits to the accuracy of the measuring device, there would be an infinite set of possible values in any given range. A **continuous random variable** takes a *range* of real numbers, instead of a list. For example, the time taken to complete a task could take *any real, positive number*, or $\{x \in R, x > 0\}$.

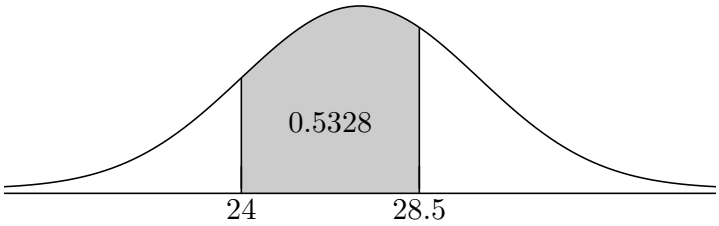
Probability and Continuous Distributions

Since a continuous random variable can take an infinite set of values, it is not possible for any *particular* value to be assigned a non-zero probability. This leads to the somewhat counter-intuitive result that, for any given value x, $\mathrm{P}(X = x) = 0$. Instead, probabilities are assigned to *ranges of values*. For example, letting random variable X represent the time taken to complete a task, measured in seconds, an example of probabilities for this continuous random variable could be:

$$\mathrm{P}(X = 26) = 0 \quad \textit{but} \quad \mathrm{P}(24 < X < 28.5) = 0.5328$$

This also means that for continuous random variables: $\mathrm{P}(X \leqslant a) = \mathrm{P}(X < a)$.

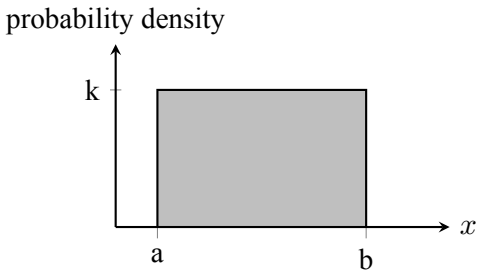
Probabilities are calculated not from the *height* of the graphed probability distributions as with discrete random variables, but instead from the **area under the curve** within the required interval. In this course, graphical calculators and tables of values mean that calculus will not be required to calculate these areas through integration.



The height of the curve is referred to as *probability density*, and indicates the "relative likelihood" of different values in the range. **The total area under the curve for any continuous distribution is 1.**

8.1 The Continuous Uniform Distribution

If random variable X can take *any real value* between a and b with a *constant probability density* then it follows a **continuous uniform distribution** with *two* parameters: a and b. This should not be confused with the *discrete uniform distribution*, which takes a *single* parameter only. Since graphing this distribution produces a rectangle, it is often also called a *rectangular distribution*.



The Continuous Uniform Distribution:

- Notated $X \sim \mathrm{U}(a, b)$.
- Defined by two parameters, a and b.
- Real, **continuous** values from a to b.
- Constant probability density of k.

The probability for any range of values can be calculated simply as the area of the rectangle enclosed. The value of k can be calculated using the formula given in the SQA Data Booklet, derived from the area under any "curve" for any continuous distribution being **equal to 1**. The formulae for the **expectation** and **variance** for the continuous uniform distribution are also given for $X \sim \mathrm{U}(a, b)$:

$$k = \frac{1}{b-a} \qquad \mathrm{E}(X) = \frac{a+b}{2} \qquad \mathrm{V}(X) = \frac{(b-a)^2}{12}$$

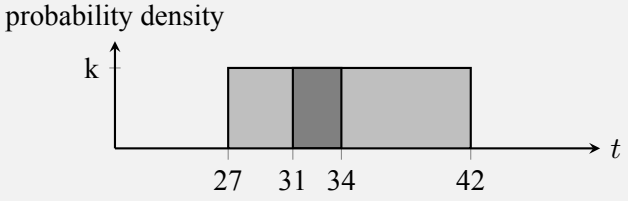
Example

Problem: Given random variable T such that $T \sim \mathrm{U}(27, 42)$, calculate $\mathrm{P}(31 < T < 34)$.

Solution: *(A sketch can help here.)*

$k = \frac{1}{b-a} = \frac{1}{42-27} = \frac{1}{15}$

$\mathrm{P}(31 < T < 34) = 3 \times \frac{1}{15} = \frac{3}{15} = \frac{1}{5}$



Exercise 8.1

1. The continuous random variable X is uniformly distributed over the interval $(6, 15)$. Find:
   (a) $\mathrm{P}(X < 12)$.
   (b) $\mathrm{P}(X > 8.1)$.
   (c) $\mathrm{P}(9.3 < X < 10.2)$.

2. The continuous random variable X is uniformly distributed over the interval $(24, 36)$. Find:
   (a) $\mathrm{P}(X \geqslant 32)$.
   (b) $\mathrm{P}(X = 29.5)$.
   (c) $\mathrm{P}(26.1 \leqslant X \leqslant 34)$.

3. The continuous random variable Y is uniformly distributed over the interval $(1.6, 4)$. Find:
   (a) $\mathrm{P}(Y < 2)$.
   (b) $\mathrm{P}(2 < Y < 5)$.
   (c) $\mathrm{P}(Y + 1 > 4)$.

4. The continuous random variable X is uniformly distributed over the interval $(-4, 6)$.
   (a) Evaluate the mean and standard deviation of X.
   (b) Calculate $\mathrm{E}(3X - 1)$ and $\mathrm{V}(5 - 2X)$.
   (c) Find $\mathrm{P}(X \leqslant 2.4)$.
   (d) Find $\mathrm{P}(-3 < X - 5 < 3)$.

5. The amount of time, in minutes, that a person must wait for a bus is uniformly distributed between 0 and 15 minutes, inclusive.
   (a) Find the probability that a person waits for less than 12.5 minutes.
   (b) Find the average amount of time that a person waits.
   (c) Find the standard deviation of the waiting time.

6. The average number of donuts a nine-year old child eats per month is uniformly distributed from 0.5 to 4 donuts. Find the probability that a randomly chosen nine-year old child eats more than two donuts on average per month.

7. A manufacturer produces sweets of length L millimetres, where L has a continuous uniform distribution with range $(15, 30)$.
   (a) Find the probability that a randomly selected sweet has a length greater than 24 millimetres.

   These sweets are randomly packed into bags containing 20 sweets.
   (b) Find the probability that a randomly selected bag will contain at least eight sweets of length greater than 24 millimetres.

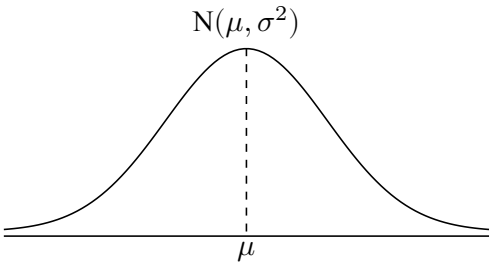
8. Given continuous random variable X such that $X \sim \mathrm{U}(1, 15)$:
   (a) Find μ, the mean of X.
   (b) Find σ, the standard deviation of X.
   (c) Determine $\mathrm{P}(X > \mu)$.
   (d) Find the probability that an observation from X falls within one standard deviation of the mean.

9. Given that continuous random variable X is uniformly distributed with a mean of 8 and a maximum value of 14:
   (a) State the distribution of X.
   (b) Find the standard deviation of X.
   (c) Determine $\mathrm{P}(X < 7)$.

10. Given that continuous random variable Y is uniformly distributed such that $\mathrm{E}(Y) = 19$ and $\mathrm{SD}(Y) = \sqrt{3}$:
    (a) Find the distribution of Y.
    (b) Determine $\mathrm{P}(\mathrm{E}(Y) - \mathrm{SD}(Y) < Y < \mathrm{E}(Y) + \mathrm{SD}(Y))$.

8.2 The Normal Distribution

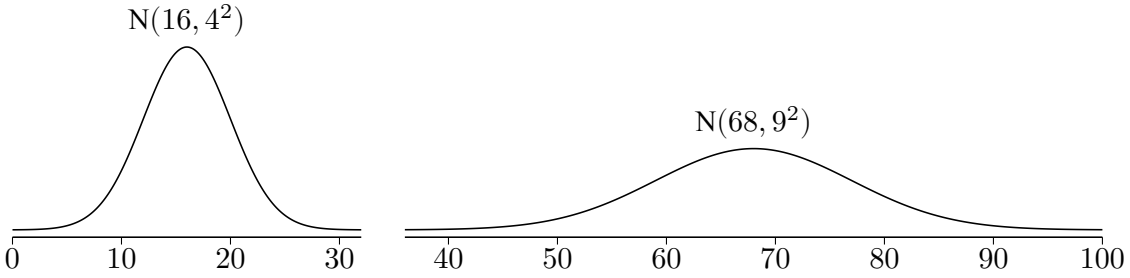
The **normal distribution**, also sometimes called the *Gaussian Distribution*, is one of the most important probability distributions in statistics. One of the reasons for this is that many continuous variables, from the mass of an animal to the time taken to complete a task, have been observed to be **normally distributed**.



pdf: $\dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

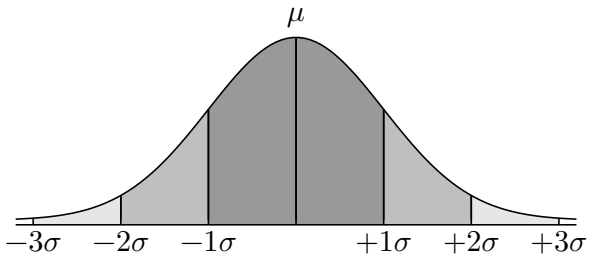
The Normal Distribution:

- Notated $X \sim \mathrm{N}(\mu, \sigma^2)$.
- Defined by two parameters, μ and σ^2.
- Often described as a *"bell-shaped curve"*.
- μ, or the **mean**, gives the *location* of the curve.
- The curve is *symmetrical about the mean*.
- σ^2, or the **variance**, gives the *spread* of the curve.



The shape of the curve illustrates that values closer to the mean are more likely to be observed, whilst those further from the mean are less likely. Whilst *any real value* is possible under the model, for **every normal distribution**:

- 68.27% of observed values lie within **one** standard deviation of the mean.
- 95.45% of observed values lie within **two** standard deviations of the mean.
- 99.73% of observed values lie within **three** standard deviations of the mean.



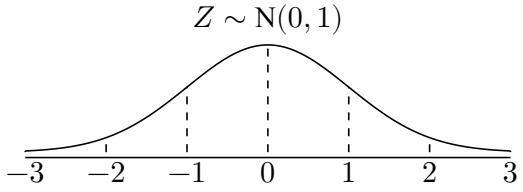
Exercise 8.2

For each, sketch the distribution, labelling the mean and two standard deviations either side of the mean.

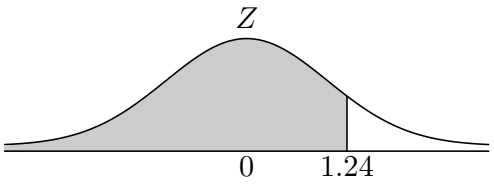
1. (a) $X \sim \mathrm{N}(60, 16)$ (b) $X \sim \mathrm{N}(243, 9)$
   (c) $X \sim \mathrm{N}(8, 1.3^2)$ (d) $X \sim \mathrm{N}(15, 1.44)$
2. The mass of a variety of pumpkin is normally distributed with mean 7kg and standard deviation 2kg.
3. The time taken to polish a spoon is normally distributed with mean 8 seconds and standard deviation 1.5 seconds.
4. The depth of water in a city's puddles after heavy rain is normally distributed with mean 32mm and standard deviation 9mm.

8.3 The Standard Normal Distribution *Z*

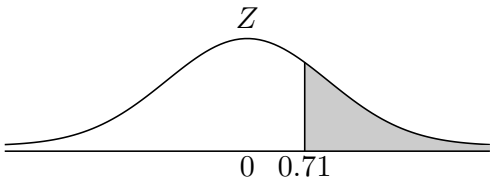
The random variable Z is specifically defined as normally distributed, with $\mu = 0$ and $\sigma = 1$:



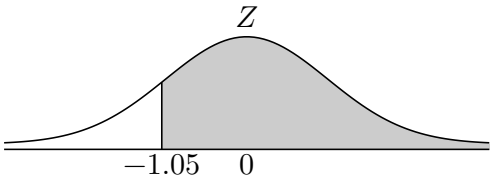
Z is called the **standard normal distribution**, and it is used to calculate probabilities for **any** normal distribution. Probabilities for the Z distribution can be obtained using the **cumulative probability table** on Page 11 of the SQA Data Booklet, or directly from a graphical calculator. Since the cumulative table only gives probabilities for **positive z-values**, negative z-values require consideration of the symmetry of the normal distribution.



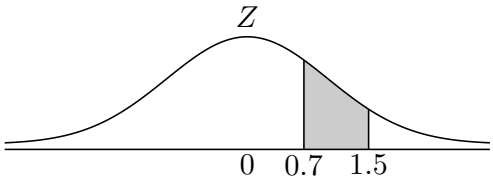
$$\mathrm{P}(Z < 1.24) = 0.8925$$



$$\begin{aligned}\mathrm{P}(Z > 0.71) &= 1 - \mathrm{P}(Z < 0.71)\\ &= 1 - 0.7611\\ &= 0.2389\end{aligned}$$



$$\begin{aligned}\mathrm{P}(Z > -1.05) &= \mathrm{P}(Z < 1.05)\\ &= 0.8531\end{aligned}$$



$$\begin{aligned}\mathrm{P}(0.7 < Z < 1.5) &= \mathrm{P}(Z < 1.5) - \mathrm{P}(Z < 0.7)\\ &= 0.9332 - 0.7580\\ &= 0.1752\end{aligned}$$

Exercise 8.3

1. Find:

(a) $\mathrm{P}(Z = 1.72)$	(b) $\mathrm{P}(Z < 1.72)$	(c) $\mathrm{P}(Z \leqslant 1.72)$	(d) $\mathrm{P}(Z < 0)$
(e) $\mathrm{P}(Z < 0.35)$	(f) $\mathrm{P}(Z < 2.03)$	(g) $\mathrm{P}(Z > 2.03)$	(h) $\mathrm{P}(Z > 1.19)$
(i) $\mathrm{P}(Z < -1.19)$	(j) $\mathrm{P}(Z > -0.56)$	(k) $\mathrm{P}(Z > -1.44)$	(l) $\mathrm{P}(Z < -0.57)$

2. Find:

(a) $\mathrm{P}(1.3 < Z < 2)$	(b) $\mathrm{P}(0.48 < Z < 1.93)$	(c) $\mathrm{P}(-2.3 < Z < -1.7)$
(d) $\mathrm{P}(-1.54 < Z < -0.76)$	(e) $\mathrm{P}(-0.7 < Z < 1.3)$	(f) $\mathrm{P}(-1 < Z < 1)$
(g) $\mathrm{P}(-2 < Z < 2)$	(h) $\mathrm{P}(-1.96 < Z < 1.96)$	(i) $\mathrm{P}(-1.64 < Z < 1.64)$

8.4 The *Z*-Transformation

A linear transformation of a normally distributed random variable will result in *another* normally distributed random variable, with the laws of expectation and variance applying as usual. In this way, any normal distributed can be related to the *standard normal distribution*, Z. Consider a random variable X with mean μ and variance σ^2:

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

Multiplying Z by σ and adding μ results in a normally distributed random variable:

$$\begin{aligned} \mathrm{E}(\sigma Z + \mu) &= \sigma \mathrm{E}(Z) + \mu & \mathrm{V}(\sigma Z + \mu) &= \sigma^2 \mathrm{V}(Z) \\ &= \sigma \times 0 + \mu & &= \sigma^2 \times 1 \\ &= \mu & &= \sigma^2 \end{aligned}$$

Hence $X = \sigma Z + \mu$, which can be rearranged to obtain the **z-transformation**:

The z-transformation:

$$Z = \frac{X - \mu}{\sigma}$$

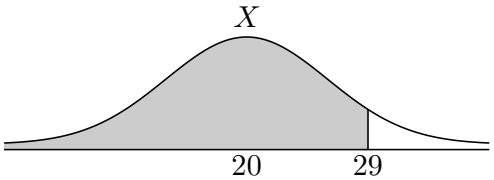
A useful way of thinking about the result of a z-transformation is that it is a measure of *how many standard deviations from the mean any given x-value is*, which then can be used to calculate a probability.

For example, consider normally distributed random variable X with a mean of 20 and standard deviation of 6. The probability that a random observation of X results in a value less than 29 can be sketched on the distribution of X, and compared to the *transformed value* of 1.5 on the distribution of Z. Note that 29 is 1.5 standard deviations over the mean of 20.

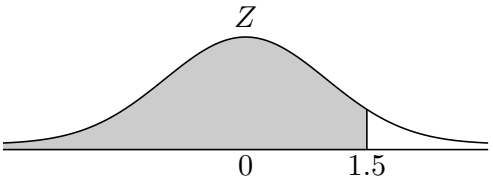
$$z = \frac{x - \mu}{\sigma} = \frac{29 - 20}{6} = 1.5$$

Hence, a probability statement about X can be rewritten as a probability statement about Z:

$$\mathrm{P}(X < 29) = \mathrm{P}\left(Z < \frac{X - \mu}{\sigma}\right) = \mathrm{P}\left(Z < \frac{29 - 20}{6}\right) = \mathrm{P}(Z < 1.5)$$



$\mathrm{P}(X < 29) = 0.9332$



$\mathrm{P}(Z < 1.5) = 0.9332$

Example

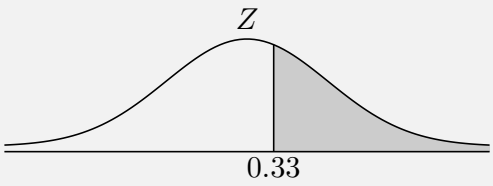
Problem: The mass of an adult red squirrel in an area of woodland is normally distributed with a mean of 295g and a standard deviation of 15g. Calculate the probability that a squirrel chosen at random has a mass greater than 300g.

Solution: *Defining the random variable clearly is important, and a sketch can be helpful.*

Let random variable X represent the mass of a squirrel.

$X \sim \mathrm{N}(295, 15^2)$

$\mathrm{P}(X > 300) = \mathrm{P}(Z > \frac{X-\mu}{\sigma}) = \mathrm{P}(Z > \frac{300-295}{15}) = \mathrm{P}(Z > 0.33) = 0.3707$



Exercise 8.4

1. Given $X \sim \mathrm{N}(24, 9)$, find:

   (a) $\mathrm{P}(X < 26)$ (b) $\mathrm{P}(X \leqslant 21)$ (c) $\mathrm{P}(X > 23)$ (d) $\mathrm{P}(21 < X < 25)$

2. Given that random variable X is normally distributed with mean 80 and standard deviation 4, find:

   (a) $\mathrm{P}(X > 83.1)$ (b) $\mathrm{P}(X < 79.3)$ (c) $\mathrm{P}(X > 72.7)$ (d) $\mathrm{P}(81 < X < 89)$

3. Let random variable X represent the tail length of a certain breed of mouse, which follows a normal distribution with a mean of 67.4mm and a standard deviation of 6mm.

   (a) State the distribution of X.

   (b) Find the probability that a mouse chosen at random will have a tail length:

   i. Less than 70mm. ii. More than 65mm. iii. Between 61mm and 66mm.

4. In a crop of 900 turnips harvested by a farmer, the mass of each is normally distributed with a mean of 150g and a standard deviation of 12g.

   (a) Calculate the probability that a turnip chosen at random will have a mass:

   i. Less than 165.36g. ii. Between 126.48g and 173.52g.

   (b) Estimate the number of turnips of the 900 that have a mass:

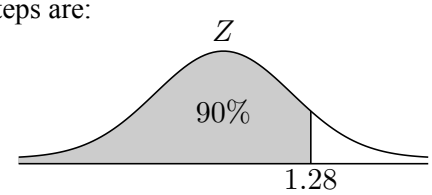
   i. Greater than 162g. ii. Less than 155g. iii. Between 140g and 160g.

8.5 Working Backwards

"Working backwards" means calculating a value for a normally distributed random variable when a **probability relating to this value is given**. For example, if random variable X is normally distributed with mean 280 and variance 16, it may be desired to calculate the value only exceeded by 10% of the observations. Typical steps are:

- Write a probability statement..
- Determine the z-value.
- Use the z-transformation.

$$\begin{aligned} \mathrm{P}(Z < 1.28) &= 0.9 \\ 1.28 &= \frac{x - 280}{4} \\ x &= 285.12 \end{aligned}$$



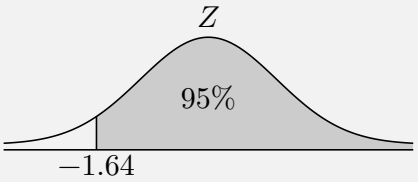
Hence, in the long run only 10% of observations exceed 285.12.

Example 1

Problem: The time before failure of a model of TV is normally distributed with mean 31 months and standard deviation 4 months. Calculate the minimum warranty length, in months, such that fewer than 5% of TVs will be expected to fail under warranty.

Solution:
Let random variable X represent the time taken for a TV to fail.

$$\begin{aligned} X &\sim \mathrm{N}(31, 4^2) \\ \mathrm{P}(Z < a) &= 0.05 \\ -1.64 &= \frac{x - 31}{4} \\ x &= 24.44 \end{aligned}$$



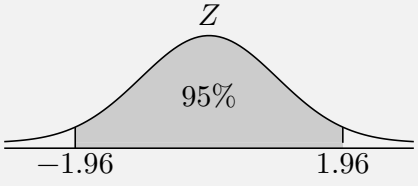
Hence, a 25 month warranty should expect to see less than 5% of TVs fail under warranty.

Example 2

Problem: The diameter of the apples grown in a garden are normally distributed with a mean of 7cm and a standard deviation of 0.25cm. Calculate the range of values, symmetrical about the mean, within which the diameters of 95% of the apples lie.

Solution:
Let random variable X represent the diameter of an apple.

$$\begin{aligned} X &\sim \mathrm{N}(7, 0.25^2) \\ \mathrm{P}(Z < a) &= 0.975 \\ 1.96 &= \frac{x - 7}{0.25} \\ x &= 7 + 1.96 \times 0.25 \\ x &= 7.49 \end{aligned}$$



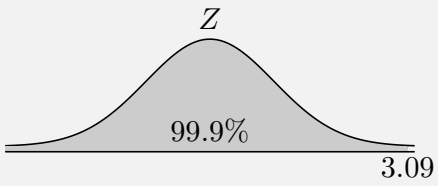
Whilst the higher value for the diameter of an apple is given by $7 + 1.96 \times 0.25 = 7.49$,
the lower value is given by $7 - 1.96 \times 0.25 = 6.51$.
Hence, 95% of the apples grown have a diameter between 6.51cm and 7.49cm.

Example 3

Problem: The masses of pumpkins grown in a farmer's field are normally distributed, with a mean of 6.8kg. Given that only one in every 1000 pumpkins has a mass greater than 10kg, calculate the standard deviation of the mass of a pumpkin.

Solution:
Let random variable X represent the mass of a pumpkin.

$$\begin{aligned} X \sim & \text{N}(6.8, \sigma^2) \\ \text{P}(Z < a) = & 0.999 \\ 3.09 = & \frac{10 - 6.8}{\sigma} \\ 3.09\sigma = & 3.2 \\ \sigma = & \frac{3.2}{3.09} \\ \sigma = & 1.04 \end{aligned}$$



Hence, the standard deviation in the mass of a pumpkin is 1.04kg.

Exercise 8.5

1. Given $Z \sim \text{N}(0, 1)$, find the value of a if:

   (a) $\text{P}(Z > a) = 0.1$ (b) $\text{P}(Z > a) = 0.05$ (c) $\text{P}(Z > a) = 0.001$ (d) $\text{P}(Z < a) = 0.99$
   (e) $\text{P}(Z < a) = 0.995$ (f) $\text{P}(Z < a) = 0.025$ (g) $\text{P}(Z < a) = 0.005$ (h) $\text{P}(Z > a) = 0.9$
   (i) $\text{P}(Z > a) = 0.975$ (j) $\text{P}(Z < a) = 0.999$ (k) $\text{P}(Z < a) = 0.791$ (l) $\text{P}(Z > a) = 0.003$

2. Let random variable X represent the total volume of petrol used by a motorist on their daily commute to and from work. The amount of petrol used is normally distributed with a mean of 2.4 litres and standard deviation 0.2 litres.

   (a) State the distribution of X.

   (b) Calculate the probability that less than 2 litres of petrol is used on a daily commute to and from work.

   (c) The motorist wishes to know the amount of petrol only exceeded on 1% of daily commutes. Find this value.

3. The time taken for a particular type of steak to be cooked until it is medium rare is normally distributed with a mean of 8.2 minutes and a standard deviation of 0.3 minutes.

   (a) Find the time needed, in minutes, such that 95% of such steaks will have reached medium rare.

   (b) Determine the range of times, symmetrical around the mean, between which 95% of all such steaks will reach medium rare.

4. In a Maths exam, the scores obtained by candidates are normally distributed with a mean of 57.3. Given that 5% of pupils fail to gain a grade from A to D, and that a score of 41 is required to gain a D, calculate the standard deviation.

5. In a Physics exam, the scores obtained by candidates are normally distributed. Given that 25% of pupils gain a grade A with a mark of 74 required, and 10% of pupils didn't get the 46 marks required for a grade C, calculate the mean and standard deviation.

8.6 Combining Normal Random Variables

Sums and differences of normally distributed random variables *are themselves normally distributed*, under the condition that **the normal random variables are independent**. For example, given **independent** normally distributed random variables X and Y with means of μ_X and μ_Y and standard deviations of σ_X and σ_Y respectively:

$$X + Y \sim \text{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \quad \text{and} \quad X - Y \sim \text{N}(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

The mean and variance of these *combined normal random variables* can be calculated using the laws of expectation and variance as usual.

It is important to remember that $2X$ is **not** the sum of two independent observations of X. Instead, the sum of n independent observations of X is $X_1 + X_2 + ... + X_n$.

$$\begin{aligned}\text{E}(X_1 + X_2 + ... + X_n) =& \text{E}(X_1) + \text{E}(X_2) + ... + \text{E}(X_n) = n\text{E}(X)\\ \text{V}(X_1 + X_2 + ... + X_n) =& \text{V}(X_1) + \text{V}(X_2) + ... + \text{V}(X_n) = n\text{V}(X)\end{aligned}$$

When calculating probabilities relating to combined normal random variables, it can be helpful to name the new random variable using some logical letter choice, such as T for *"Total"*, D for *"Difference"* and so on.

Example 1

Problem: The actual volume of paint contained within tins marked as containing 25 litres is normally distributed with a mean of 25.2 litres and a standard deviation of 0.3 litres. Calculate the probability that 4 such tins contain less than 100 litres in total.

Solution:

Let random variable X represent the volume of paint in a tin.

$X \sim \text{N}(25.2, 0.3^2)$

Let T represent the total volume of paint in the four tins.

Assuming the volume of paint in each tin is independent:

$T = X_1 + X_2 + X_3 + X_4$

$\text{E}(T) = \text{E}(X_1 + X_2 + X_3 + X_4) = \text{E}(X_1) + \text{E}(X_2) + \text{E}(X_3) + \text{E}(X_4) = 4 \times 25.2 = 100.8$

$\text{V}(T) = \text{V}(X_1 + X_2 + X_3 + X_4) = \text{V}(X_1) + \text{V}(X_2) + \text{V}(X_3) + \text{V}(X_4) = 4 \times 0.3^2 = 0.36$

$T \sim \text{N}(100.8, 0.36)$

$$\begin{aligned}\text{P}(T < 100) =& \text{P}\left(Z < \frac{100 - 100.8}{\sqrt{0.36}}\right)\\ =& \text{P}(Z < -1.33)\\ =& 0.0918\end{aligned}$$

Example 2

Problem: The time taken for a washing machine cycle to complete is normally distributed, with mean 57 minutes and standard deviation 2 minutes. The time taken for the owner of the washing machine to start the cycle, run 10 kilometres on their treadmill and get back to the washing machine is normally distributed with mean 58 minutes and standard deviation 3 minutes. Stating an assumption required, calculate the probability that the washing machine cycle is still running when the owner returns to it.

Solution: *Define the random variables used.*

Let random variable X represent the time taken for the washing machine cycle, and random variable Y represent the time taken for the owner to complete their run.

$X \sim \mathrm{N}(57, 2^2)$ and $Y \sim \mathrm{N}(58, 3^2)$

Assuming the time taken for the cycle to complete and the time taken for the owner to complete their run are independent:

$$\begin{aligned} \mathrm{P}(X > Y) &= \mathrm{P}(X - Y > 0) \\ &= \mathrm{P}\left(Z > \frac{0 - (-1)}{\sqrt{13}}\right) \\ &= \mathrm{P}(Z > 0.28) \\ &= 0.3897 \end{aligned}$$

$$\begin{aligned} \mathrm{E}(X - Y) &= \mathrm{E}(X) - \mathrm{E}(Y) = 57 - 58 = -1 \\ \mathrm{V}(X - Y) &= \mathrm{V}(X) + \mathrm{V}(Y) = 2^2 + 3^2 = 13 \\ X - Y &\sim \mathrm{N}(-1, 13) \end{aligned}$$

Exercise 8.6

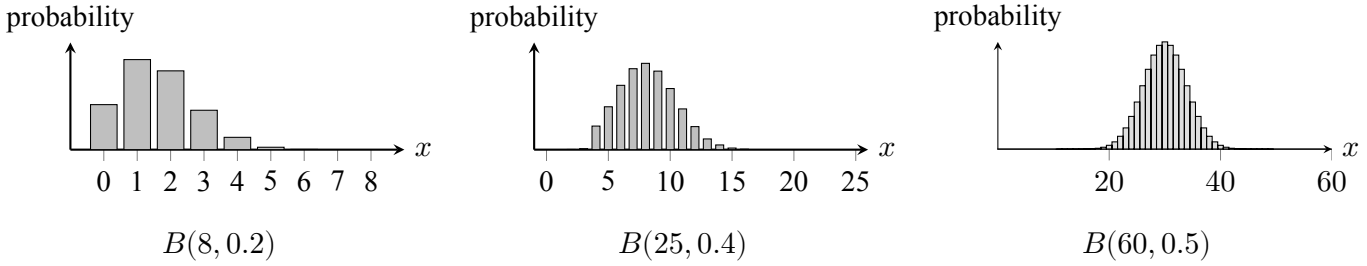
1. Bob observes that making his usual portion of mashed potatoes can be broken down into three distinct processes. The time taken to peel the potatoes is normally distributed with a mean of 5 minutes and a standard deviation of 1.2 minutes. The time taken to boil the potatoes is also normally distributed with a mean of 27 minutes and a standard deviation of 3.4 minutes. Finally, the mean time taken to mash and season the potatoes is 2 minutes with a standard deviation of 0.3 minutes.
   (a) Find the distribution of the time taken for Bob to make a portion of mash, stating two assumptions required.
   (b) Calculate the probability that it takes Bob under 35 minutes to make a portion.

2. The bar staff in a hotel ignore legal requirements to carefully measure any alcohol poured, and instead just estimate the amount of champagne to pour into each glass. The amount they pour is normally distributed with a mean of 127ml and a standard deviation of 5ml. Assuming that the amount of champagne poured in each of a series of glasses is independent, determine the probability that the total volume in five glasses is less than 625ml.

3. The amount of milk a school's Maths department uses per week is normally distributed, with a mean of 3.5 litres and a standard deviation of 0.4 litres. The Business department's milk use is also normally distributed, with a mean of 2.8 litres and a standard deviation of 0.5 litres. In any given week, calculate the probability that the Business department uses more milk than the Maths department, stating any assumptions required.

4. The length of a best-of-five match of tennis is normally distributed with a mean of 165 minutes and a standard deviation of 21 minutes, whilst the length of a best-of-three match has a mean of 95 minutes and a standard deviation of 10 minutes. A tournament has two best-of-five matches and one best-of-three match scheduled to be played on Court 7, beginning at 10am. Find the probability that the three matches will be completed by 4pm, stating two assumptions required.

8.7 Normal Approximation to the Binomial Distribution



Under certain conditions the *discrete* binomial distribution can begin to resemble a *continuous* normal distribution.. Given random variable $X \sim B(n, p)$, **a normal distribution can approximate the binomial distribution**, with *sufficiently accuracy*, when:

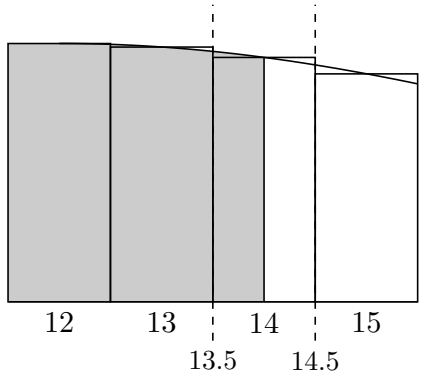
Conditions for normal approximation to binomial distributions :
$np > 5$ and $nq > 5$

The mean and variance of the *approximate* normal distribution are, as usual for a binomial distribution, $\mathrm{E}(X) = np$ and $\mathrm{V}(X) = npq$. Using '$\approx$' instead of '$\sim$' is important to show that the normal distribution is only an *approximation* to the actual binomial distribution.

$$\text{If } X \sim B(n, p) \text{ then } X \approx \mathrm{N}(np, npq)$$

Historically, the ability to approximate the binomial distribution using the normal distribution will have prevented the need for extensive tables of binomial probabilities. Additionally, binomial distribution calculations tend to be more computationally-intensive than normal distributions calculations. Whilst modern computing power may seem to render this argument null, it can still be desirable to speed up computation in return for a trivial loss in accuracy.

Approximating probabilities for a *discrete* distribution using a *continuous* distribution may lead to inaccuracy unless a **continuity correction** is applied. For example, consider random variable $X \sim B(24, 0.5)$, with its normal approximation $X \approx \mathrm{N}(12, 6)$ overlaid and the *continous* probability $\mathrm{P}(X < 14)$ shaded:



When calculating the discrete probability $\mathrm{P}(X \leqslant 14)$, the bar representing the discrete integer 14 ought to be included. The continuous probability $\mathrm{P}(X < 14.5)$, making a **continuity correction**, would provide a more accurate estimate of the desired probability. For the discrete probability $\mathrm{P}(X < 14)$, with the discrete integer value of 14 not to be included, the continuous probability $\mathrm{P}(X < 13.5)$ would be more accurate.

Example

Problem: A quiz contains 40 multiple choice questions, with four answers for every question. If a pupil picks answers entirely at random, use a suitable approximation to calculate the probability that they get at least half of the questions correct, and justify the suitability of the approximation used.

Solution: *"Suitable approximation" highlights that the normal approximation should be used, even if a graphical calculator can obtain the precise probability.*

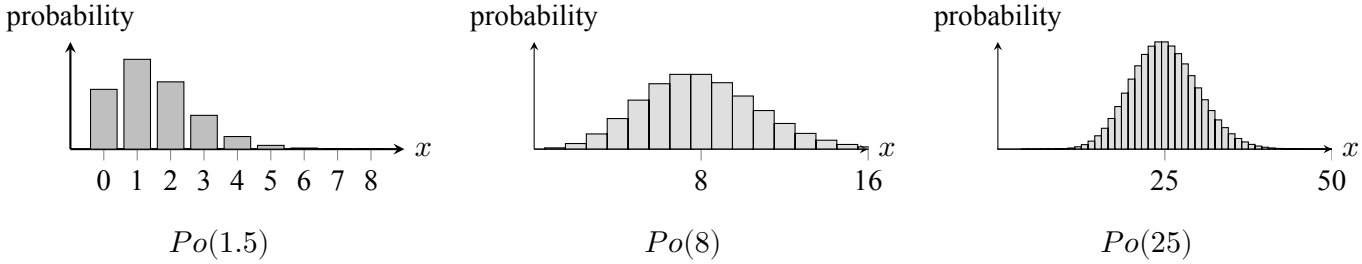
Let random variable X represent the number of questions answered correctly.
$X \sim B(40, 0.25)$
$np = 40 \times 0.25 = 10$, $nq = 40 \times 0.75 = 30$
$np > 5$ and $nq > 5$ therefore a normal approximation is appropriate.
$\mathrm{E}(X) = np = 10$ and $\mathrm{V}(X) = npq = 40 \times 0.25 \times 0.75 = 7.5$
$X \approx \mathrm{N}(10, 7.5)$

$$\begin{aligned}\mathrm{P}(X \geqslant 20) &= \mathrm{P}(X > 19.5) \\ &= \mathrm{P}\left(Z > \frac{19.5 - 10}{\sqrt{7.5}}\right) \\ &= \mathrm{P}(Z > 3.47) \\ &= 0.0003\end{aligned}$$

Exercise 8.7

1. Given $X \sim B(40, 0.4)$:
   (a) Show that a normal approximation is appropriate.
   (b) Find:
   i. $\mathrm{P}(X \leqslant 20)$ ii. $\mathrm{P}(25 \leqslant X \leqslant 28)$ iii. $\mathrm{P}(X > 31)$
2. Given $Y \sim B(81, \frac{2}{3})$:
   (a) Show that a normal approximation is appropriate.
   (b) Find:
   i. $\mathrm{P}(Y \geqslant 56)$ ii. $\mathrm{P}(Y = 48)$ iii. $\mathrm{P}(50 < Y < 60)$
3. To complete the mandatory online training required by a large corporation, 600 employees in an office each try to log on to the corporation's online training portal. The probability of an employee managing to successfully log in on their first attempt is 85%. Let the random variable X represent the number of employees that manage to log in successfully on their first attempt. State the distribution of X and use a suitable approximation to estimate the probability of over 100 of the 600 employees not managing to log in successfully on their first attempt, justifying its use.
4. A farmer believes that 60% of the apples that are brought into her grading shed are classified as Grade A. Assuming she is right, use a suitable approximation to estimate the probability that, in a random sample of 120 apples, there are more than 75 Grade A apples.
5. 200 people are randomly picked as part of a survey and asked whether or not they plan to vote for the party currently in power in an upcoming election. It is generally thought that 54% of people will vote for this party. Using a suitable approximation, calculate the probability that less than 45% of the people sampled indicate that they will vote for the party currently in power, and comment on what this might suggest should it happen.

8.8 Normal Approximation to the Poisson Distribution



$Po(1.5)$ $Po(8)$ $Po(25)$

As with the binomial distribution, under certain conditions the *discrete* Poisson distribution can also begin to resemble a *continuous* normal distribution. Given random variable $X \sim Po(\lambda)$, **a normal distribution can approximate the Poisson distribution**, with *sufficiently accuracy*, when:

Conditions for normal approximation to Poisson distributions :
$\lambda > 10$

The mean and variance of the *approximate* normal distribution are $\mathrm{E}(X) = \mathrm{V}(X) = \lambda$. .

$$\text{If } X \sim Po(\lambda) \text{ then } X \approx \mathrm{N}(\lambda, \lambda)$$

Since again this is a *continuous approximation* to a *discrete distribution*, any probabilities calculated will not be sufficiently accurate unless a **continuity correction** is applied. For a continuous approximation to a discrete distribution taking integers values, as is the case with both such approximations encountered in this chapter, continuity corrections will always require an addition or subtraction of 0.5.

Example

Problem: In a football league, goals occur at a constant mean rate of 2.5 goals per game. A full weekend of fixtures includes 10 games. Using a suitable approximation, and justifying its use, calculate the probability that there will be between 15 and 20 goals in total, inclusive, in a full weekend of fixtures.

Solution:

Let random variable T represent the number of goals across the 10 games, with $\lambda = 2.5 \times 10 = 25$: $T \sim Po(25)$

Since $\lambda > 10$, a normal approximation is appropriate.

$\mathrm{E}(T) = \mathrm{V}(T) = 25$

$T \approx \mathrm{N}(25, 25)$

$$\begin{aligned}
\mathrm{P}(15 \leqslant T \leqslant 20) &= \mathrm{P}(14.5 < T < 20.5) \\
&= \mathrm{P}\left(\frac{14.5 - 25}{\sqrt{25}} < Z < \frac{20.5 - 25}{\sqrt{25}}\right) \\
&= \mathrm{P}(-2.1 < Z < -0.9) \\
&= \mathrm{P}(Z < -0.9) - \mathrm{P}(Z < -2.1) \\
&= 0.1841 - 0.0179 \\
&= 0.1662
\end{aligned}$$

Exercise 8.8

1. Given $X \sim Po(24)$, use the normal approximation to find:

   (a) $\mathrm{P}(X \leqslant 25)$ (b) $\mathrm{P}(X = 21)$ (c) $\mathrm{P}(X > 23)$

2. Given $X \sim Po(60)$, use the normal approximation to find:

   (a) $\mathrm{P}(X \geqslant 57)$ (b) $\mathrm{P}(50 < X \leqslant 56)$ (c) $\mathrm{P}(62 < X < 66)$

3. The number of calls received by an office switchboard per hour follows a Poisson distribution with a parameter 30. Using the normal approximation to the Poisson distribution, find the probability that in one hour:

   (a) There are more than 33 calls.

   (b) There are between 25 and 28 calls, inclusive.

   (c) There are exactly 34 calls.

4. In a certain factory the number of incidents occurring in a month follows a Poisson distribution with mean 4. Find the probability that there will be at least 40 incidents during one year.

5. The number of bacteria on a plate viewed under a microscope follows a Poisson distribution with parameter 60.

   (a) Find the probability that there are between 55 and 75 bacteria, inclusive, on a plate.

   (b) A plate is rejected if fewer than 38 bacteria are found. If 2000 plates are viewed, determine the expected number of plates that will be rejected.

6. In an experiment with a radioactive substance, the number of particles reaching a counter over a given period of time follows a Poisson distribution with mean 22. Find the probability that the number of particles reaching the counter over the given period of time is:

   (a) Fewer than 22.

   (b) Between 25 and 30, not inclusive.

   (c) 18 or more.

7. The number of accidents on a certain railway line occurs at an average rate of one every 2 months. Justifying your method, estimate the probability that there are:

   (a) 28 or more accidents in 4 years.

   (b) Fewer than 30 accidents in 5 years.

8. The number of eggs laid by an insect follows a Poisson distribution with parameter 200. Estimate the probability that between 180 and 220 eggs, inclusive, are laid.

Review Exercise

1. Given $X \sim \mathrm{U}(1, 7)$, find:

   (a) $\mathrm{E}(X)$

   (b) $\mathrm{V}(3 - X)$

   (c) $\mathrm{P}(X < 3)$

2. Given $X \sim \mathrm{N}(20, 6)$, find:

   (a) $\mathrm{P}(X < 25)$

   (b) $\mathrm{P}(14 < X < 20)$

3. The height of an adult giraffe is normally distributed with a mean of 18.1 feet and a standard deviation of 0.7 feet. Let random variable H represent the height of an adult giraffe, in feet.

   (a) State the distribution of H.

   (b) Find $\mathrm{P}(H > 20)$.

   (c) A biologist wishes to know the height only exceeded by 5% of giraffes. Find this height.

4. A guitar player experiences broken strings at a constant mean rate of 2 per month.

   (a) Find the probability that no strings break in one month.

   (b) Use a suitable approximation to estimate the probability that exactly 26 strings break in one year.

5. For continuous random variable $X \sim \mathrm{N}(\mu, 3^2)$, $\mathrm{P}(X < 16) = 0.1$. Determine the value of μ for X.

6. A continuous uniform distribution is defined over the interval 10 to 20.

   (a) For a single observation from this distribution, calculate the probability of obtaining a value not greater than 13.

   (b) If a random sample of size 10 is taken from the distribution, calculate the probability of at least 3 of the observations resulting in value not greater than 13.

   (c) If a random sample of size 100 is taken from the distribution, use a suitable approximation to estimate the probability that over 30 of the observations result in a value not greater than 13.

7. A sports scientist is interested in the time taken for 800 metre runners to complete each of the two laps that make up the distance. They observe that the time it takes for an athlete they are studying to complete each lap is normally distributed, with a mean of 53.2 seconds with a standard deviation of 1.3 seconds.

   (a) Find the probability that the athlete completes the two laps in under 1 minute and 44 seconds.

   (b) An important assumption is required to calculate the answer for part (a). State the assumption and explain whether or not it is likely to have been met.

9

The Distribution of the Sample Mean

Previous chapters have investigated probabilities related to *individual* observations from random variables with known distributions. However, statisticians are rarely in the business of making a single observation from a population - instead, they are far more likely to draw a *sample* of observations and, commonly, calculate the *sample mean*. Given random variable X, the sample mean of X is also a random variable and is denoted $\overline{X}$.

Letting $X_1, X_2, ..., X_n$ represent a sample of n **independent and identically distributed (i.i.d.)** observations, where independence may be assumed through the use of *random sampling* methods, the sample mean may be expressed as:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + X_3 + ... + X_n)$$

The Expectation and Variance of the Sample Mean

The expected value of the sample mean is equal to the true value of the population mean, meaning it is an *unbiased estimator* of the population mean. That is, if a series of samples are taken from a distribution with population mean μ and the mean is calculated for each sample, in the long-run the mean of these sample means tends toward μ.

The *standard deviation of the sample mean* $\overline{X}$ is called the **standard error**. For an underlying population with standard deviation σ, the standard error of the sample mean for a sample of size n is $\frac{\sigma}{\sqrt{n}}$.

Proof: Given random variable X such that $\mathrm{E}(X) = \mu$ and $\mathrm{V}(X) = \sigma^2$, from which a sample of n independent observations $(X_1, X_2, ..., X_n)$ are drawn:

$$\begin{aligned}\mathrm{E}(\overline{X}) &= \mathrm{E}\left(\frac{1}{n}(X_1 + X_2 + ... + X_n)\right) \\ &= \frac{1}{n}(\mathrm{E}(X_1) + \mathrm{E}(X_2) + ... + \mathrm{E}(X_n)) \\ &= \frac{1}{n}(n\mu) \\ &= \mu\end{aligned}$$

$$\begin{aligned}\mathrm{V}(\overline{X}) &= \mathrm{V}\left(\frac{1}{n}(X_1 + X_2 + ... + X_n)\right) \\ &= \frac{1}{n^2}(\mathrm{V}(X_1) + \mathrm{V}(X_2) + ... + \mathrm{V}(X_n)) \\ &= \frac{1}{n^2}(n\sigma^2) \\ &= \frac{\sigma^2}{n}\end{aligned}$$

9.1 The Sample Mean of a Normally Distributed Random Variable

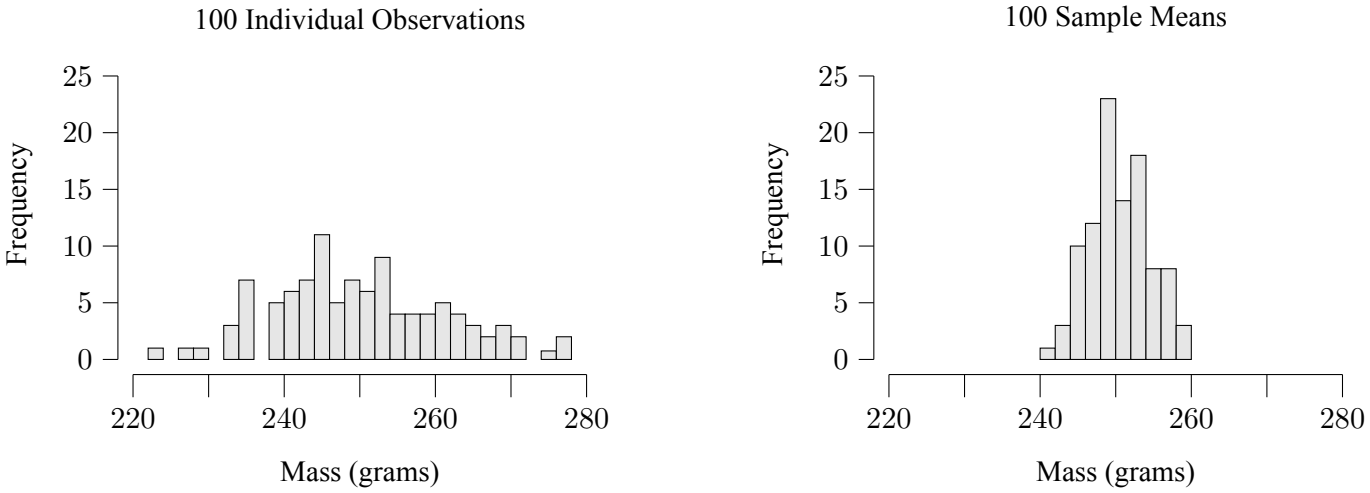
Since the sample mean $\overline{X}$ is a linear combination of independent observations from X, **if X is normally distributed then $\overline{X}$ is also normally distributed.**

Sample mean of normally distributed X:

$$\text{If } X \sim \text{N}(\mu, \sigma^2) \text{ then } \overline{X} \sim \text{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Due to the lower variance of the sample mean, *extreme results are less likely* when taking the mean of a sample in comparison to a single observation. For example, suppose the mass of adult red squirrels in a location is known to be normally distributed, with mean 250 grams and standard deviation 12 grams. If a single squirrel is randomly selected and its mass measured, it would not be that unlikely to encounter a squirrel lighter than 240 grams or heavier than 260 grams. However, if a random sample of three squirrels is taken, each weighed and the mean of the sample calculated then it would be very unlikely for the sample mean to be lower than 240 grams or heavier than 260 grams.

The histograms below show 100 simulated individual observations of squirrel masses compared to the means of 100 simulated samples, each of size 3. Note that the sample means also show a distribution that looks normal with a mean of 250 grams, but that the spread of the sample means is much less than that of the individual observations.



Given $X \sim \text{N}(\mu, \sigma^2)$:

The z-transformation for *single observations* from X has the *standard deviation* as the denominator: $Z = \dfrac{X - \mu}{\sigma}$

The z-transformation for the **sample mean** has the **standard error** as its denominator instead:

The z-transformation for the sample mean:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Example

Problem: The length of a mass-produced brand of fence panels is normally distributed with a mean of 180cm and a standard deviation of 0.4cm. A random sample of nine panels is taken and each is measured.

a) State the distribution of the length of a fence panel, and calculate the probability of a randomly picked fence panel having a length greater than 180.1cm.

b) State the distribution of the mean of a random sample of nine panels, and calculate the probability of the sample mean being greater than 180.1cm.

Solution: Let r.v. X represent the length of a fence panel.

a) $X \sim \text{N}(180, 0.4^2)$

$$\text{P}(X > 180.1) = \text{P}\left(Z > \frac{180.1 - 180}{0.4}\right) = \text{P}(Z > 0.25) = 0.4013$$

b) $\overline{X} \sim \text{N}\left(180, \frac{0.4^2}{9}\right)$

$$\text{P}(\overline{X} > 180.1) = \text{P}\left(Z > \frac{180.1 - 180}{\frac{0.4}{\sqrt{9}}}\right) = \text{P}(Z > 0.75) = 0.2266$$

Exercise 91

1. Given random variable $X \sim \text{N}(24, 4^2)$:
   (a) Calculate for a single observation:
      i. $\text{P}(X > 25.2)$
      ii. $\text{P}(X < 23.4)$
   (b) For a sample from X of size 3:
      i. State the distribution of $\overline{X}$
      ii. Calculate $\text{P}(\overline{X} > 25.2)$
      iii. Calculate $\text{P}(\overline{X} < 23.4)$

2. The tail length of a breed of mouse is normally distributed with a mean of 5.2cm and a standard deviation of 0.8cm. A random sample of 12 mice are captured, their tail lengths measured and then released. Calculate the probability that the mean tail length of the sample of mice is less than 5cm.

3. The time taken for a telecommunications company to interview candidates for a sales role is normally distributed with a mean of 28 minutes and a standard deviation of 4 minutes.
   (a) Determine the probability that an interview lasts over half an hour.
   (b) Stating an assumption required, calculate the probability that a sample of five interviews will have a mean length greater than half an hour.

4. The mass of a medium-sized whole chicken sold by a supermarket is normally distributed with a mean of 2.5kg and a standard deviation of 0.15kg. Calculate the probability that the mean mass of a random sample of 9 medium-sized chickens is less than 2.4kg.

5. The amount of food waste discarded by a family per day is normally distributed with a mean of 1.96kg and a standard deviation of 0.4kg.
   (a) Calculate the probability that, on any random day, less than 1kg is discarded.
   (b) Calculate the probability that over the course of a week the mean amount of daily food waste discarded is greater than 2kg.
   (c) Suggest a reason why the calculation in part (b) may not be reliable.

6. The distances recorded in long-jump attempts by a group of junior long-jump specialists at an athletics club are normally distributed with a mean of 6.5m and a standard deviation of 0.2m. The long-jump distances recorded by the junior heptathletes are also normally distributed, with mean 6.3m and standard deviation 0.4m. A coach randomly observes attempts by 4 long-jump specialists and 3 heptathletes. Calculate the probability that the heptathletes' mean distance is greater than that of the the long-jump specialists. *(Extension)*

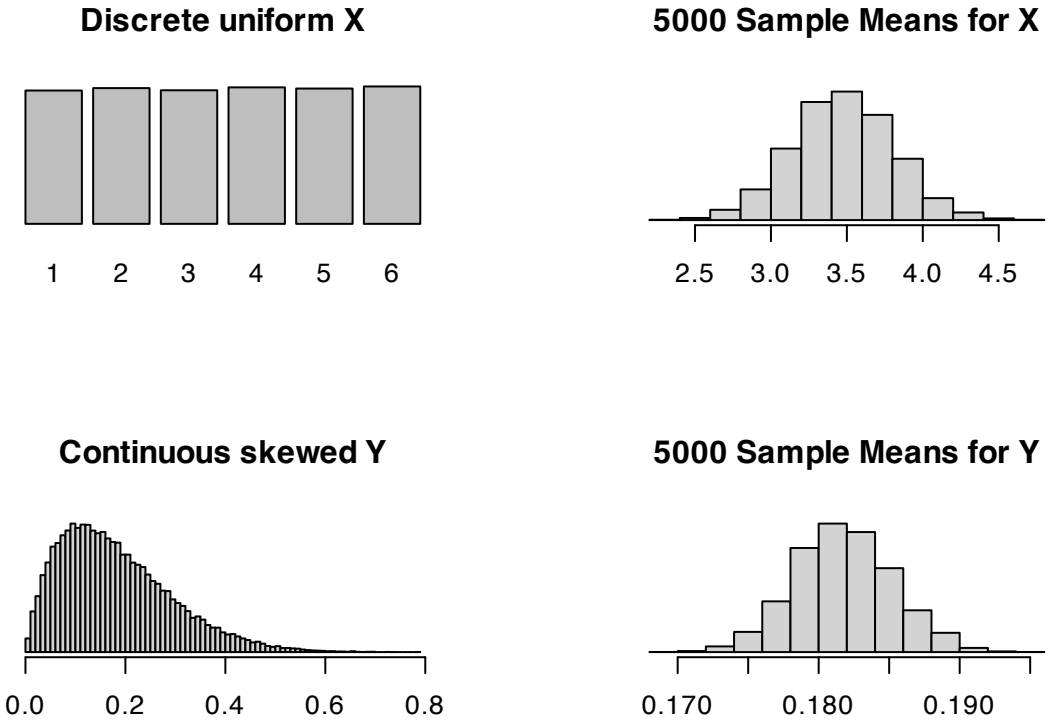
9.2 The Central Limit Theorem

The Central Limit Theorem (CLT) is one of the most important and useful concepts in statistics. It gives an insight into the prevalence of the normal distribution and, crucially for statisticians, allows the normal distribution to be used for statistical inference even when the data of interest is *not* normally distributed. The CLT says:

"When the underlying population is *not* normally distributed, for sufficiently large sample sizes
the sample mean will still be approximately normally distributed."

Note that a *"sufficiently large sample"* in this course is **greater than 20** ($n > 20$). Other statistical textbooks may use the value of 30, whilst in reality many statisticians appreciate that the nature of the underlying distribution may allow for smaller samples to be used or require larger samples in order for the distribution of the sample mean to be sufficiently approximately normal.

Proof of the CLT is beyond the scope of this course (and this textbook), but statistical software and/or experimentation can help give a sense of what this looks like in practice. The simulations below show random individual observations from distributions on the left, with corresponding distributions of sample means to the right. Every random sample in both cases was of size 25.



The **approximate** distribution of the sample mean can be stated provided the expectation and variance (or standard deviation) of the underlying distribution are known, or can be calculated. The notation '$\approx$' is used to highlight that this is an *approximate* distribution.

The Central Limit Theorem:

If $\mathrm{E}(X) = \mu$ and $\mathrm{V}(X) = \sigma^2$ then $\overline{X} \approx \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, provided $n > 20$.

Example

Problem: The mean mass of ceramic bowls produced by a potter is 400g and the standard deviation is 20g, and they can be ordered singly or in bulk. Bulk orders contain 36 mugs and, when such an order is placed, the potter selects the 36 mugs at random to be boxed up and shipped. Letting the random variable X represent the weight of a bowl, and $\overline{X}$ the mean bowl mass in a bulk order:

a. State the distribution of $\overline{X}$.

b. Calculate the probability that the mean bowl mass in a bulk order is greater than 402g.

Solution: *(Even through the underlying distribution of bowl masses is unknown, the CLT can be invoked since $n > 20$. The mean bowl mass in a sample of 36 bowls is approximately normally distributed.)*

a. $\overline{X} \approx \text{N}\left(400, \frac{20^2}{36}\right)$

b. $$\text{P}(\overline{X} > 402) = \text{P}\left(Z > \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = \text{P}\left(Z > \frac{402 - 400}{\frac{20}{\sqrt{36}}}\right) = \text{P}(Z > 0.6) = 0.2743$$

Exercise 9.2

1. For each sample from the random variable given, state the distribution of the sample mean where possible:
   (a) A sample of size 40 from X where $\text{E}(X) = 2.3$ and $\text{SD}(X) = 5$.
   (b) A sample of size 25 from normally distributed Y where $\text{E}(Y) = 83$ and $\text{V}(Y) = 9$.
   (c) A sample of size 7 from normally distributed D where $\text{E}(D) = 126$ and $\text{V}(D) = 10$.
   (d) A sample of size 12 from W where $\text{E}(W) = 14$ and $\text{SD}(W) = 2$.

2. The discrete random variable A has a mean of 18 and a standard deviation of 5. Calculate the probability that a sample of size 30 taken from A results in a mean less than 20.

3. The continuous random variable B has a mean of 248 and a variance of 16. Determine the probability of the mean of a sample of size 24 being greater than 250.

4. The mean height of the sunflowers grown in a field is 1.05m, with a standard deviation of 25cm. The grower measures the height of a random sample of 40 sunflowers. What is the probability that the mean height of the sample of sunflowers is less than a metre?

5. Eggs classified as *medium* by a farm have a mean mass of 78g and a standard deviation of 5g. Calculate the probability that, for a box of 24 randomly selected eggs, the mean egg mass is actually 78 when given correct to the nearest gram.

6. When *"40 litres"* of petrol is dispensed by a fuel pump design used by a supermarket, the mean actual amount dispensed is 40.1 litres with a standard deviation of 0.35 litres.
   (a) Explain why the known information is not sufficient to calculate the probability that a single *"40 litre"* fill actual gives less than 40 litres.

   As part of the supermarket's monitoring of the accuracy of its pumps, it takes 32 fills, each of size *"40 litres"*, from randomly chosen pumps at its stores nationwide.

   (b) Calculate the probability that the mean obtained from this random sample is less than 40 litres.
   (c) Explain why the Central Limit Theorem was required in order to obtain the answer for part (b).

Review Exercise

1. Given a normally distributed random variable X, with a mean of 15 and a standard deviation of 3, from which a sample of eight independent observations are taken:
   (a) State the distribution of the sample mean, $\overline{X}$.
   (b) Calculate $\text{P}(\overline{X} < 17)$.

2. A fair, tetrahedral (four-sided) die numbered 1 to 4 is rolled 25 times, and the score obtained for each roll recorded. Find the probability that the mean score for the 25 rolls is greater than 3.

3. At the start of a marathon, each runner is given a bottle containing 1 litre of water. The mean volume of water consumed by a runner during the marathon is 600 millilitres, whilst the standard deviation is 70 millilitres. At the end of the marathon, 40 runners are selected at random, and the volume of water consumed by each is measured. The mean volume of water consumed by the sample of 40 runners is then calculated.
   (a) In context, explain what the Central Limit Theorem says about the sample mean calculated.
   (b) State the distribution of the sample mean.
   (c) Find the probability that the sample mean calculated is between 610 millilitres and 620 millilitres.

4. The mass of mince pies sold individually at the counter at a bakery is normally distributed, with a mean of 60 grams and a standard deviation of 2 grams.
   (a) A customer asks for a mince pie, and it is selected at random by the bakery staff. Let random variable X represent the mass of the mince pie chosen.
      (i) State the distribution of X
      (ii) Find the probability that the mince pie chosen has a mass of less than 59 grams.
   (b) Another customer asks for eight mince pies, and they are selected at random by the bakery staff. Let $\overline{X}$ represent the mean mass for the sample of eight mince pies chosen.
      (i) State the distribution of $\overline{X}$.
      (ii) Find the probability that the sample mean is less than 59 grams.
   (c) The bakery changes its process, and now when a customer orders a number of mince pies the bakery staff places them on a set of weighing scales, one by one. Explain why this change is likely to mean the distribution of $\overline{X}$ is no longer as it previously was.

5. Let discrete random variable X represent the number of Video Assistant Refereem (VAR) reviews during the course of a game of football, known to follow a Poisson distribution with a mean of 4.5 reviews per game.
   (a) State the distribution of X.
   (b) Calculate the probability that a football game has no more than two VAR reviews.

   Thirty games are randomly chosen to be studied in greater detail by data analysts. They count the number of VAR reviews in each game, and calculate the mean number of reviews for this sample of thirty games.
   (c) State the distribution of $\overline{X}$, the calculated sample mean.
   (d) Find the probability that the sample mean calculated is greater than five.

10

An Introduction to Hypothesis Testing

In brief, hypothesis testing is an approach to assessing the statistical evidence for a claim through the analysis of data from a random sample. The claim will typically relate to some aspect of the distribution of a population of interest, often a *parameter* such as the *population mean*, μ, and the foundation of hypothesis testing is *probability*.

The Logic of Hypothesis Testing

Consider the following scenarios:

- A coin is tossed 50 times, and it lands showing *tails* 47 of those times.
- A cubical die is tossed 30 times, and a *six* is obtained 27 times.
- A random number generator gives the same number eight times in a row.

In each of these scenarios, there are two possible conclusions:

The coin, die and number generator are all fair, and a very unlikely event just happened...

or

The coin is biased, the die rigged, the random number generator flawed...

An *objective* process is required that allows a decision to be made between these two possibilities. This process, called **hypothesis testing**, will form a substantial part of the remainder of the course.

10.1 The Null Hypothesis and the Alternative Hypothesis

Most statistical studies begin with a claim, which is called the **alternative hypothesis**, or H_1. Initially in this chapter, these will consist of a claim about the value of the *population mean*, μ. A conservative stance in opposition to the alternative hypothesis will used as the basis for all calculations, assuming that claim is *not true*. This is called the **null hypothesis**, or H_0. It is *essential* that both hypotheses are formed *before* observing the data.

It should also first be decided whether evidence is being sought for the true value of the population mean being *less than* a specified value, *greater than* a specified value, or simply *different than* a specified value. When the alternative hypothesis is only interested in a *single direction*, the test being performed is described as ***one-tailed***. When the alternative hypothesis is interested in a change in *either direction*, it is referred to as a ***two-tailed*** test.

Example

Problem: Using suitable notation, state the null and alternative hypotheses for each of the following claims.

a. The mean volume of liquid in espressos sold by a coffee store is less than the 30ml stated on their menu.

b. The mean mass of red squirrels in a woodland population is different to the previously measured 240g.

Solution:

a. $H_0 : \mu = 30$ One-tailed.
$H_1 : \mu < 30$

b. $H_0 : \mu = 240$ Two-tailed.
$H_1 : \mu \neq 240$

Exercise 10.1

For each proposed hypothesis test, state the **null hypothesis** and the **alternative hypothesis** using appropriate notation.

1. A telecommunications provider wishes to test whether the mean amount of data used by its customers is more than 8.6 gigabytes per month.

2. A large school plans to test whether the mean amount its teachers spend on photocopying is different than the £12 per year they currently budget for.

3. A farmer wants to test whether the mean mass of the newborn lambs from his flock is less than 4.2 kilograms.

4. A sports statistician is planning a hypothesis test to assess the evidence that the mean actual playing time during a game of football has changed from 58 minutes.

5. A kettle manufacturer wishes to test whether the mean time for its kettles to boil one litre of water is more than the one minute and 5 seconds stated on the packaging.

6. The mean volume of petrol a driver uses driving from Glasgow to Blackburn is 15.7 litres. The driver wishes to test whether the mean volume of petrol used has changed.

Evidence to Suggest

The *true value* of a population parameter cannot be known by studying a sample. This means that, without a census of the entire population, a statistician will never say that they have *proved* or *disproved* a claim. Instead they will say that they have **evidence to suggest** or **insufficient evidence to suggest** a particular claim is true.

The Significance Level and the *p*-value

Once the null and alternative hypotheses have been stated, a conservative viewpoint is taken - it is *assumed that the null hypothesis is true*. Only an observed value that is then *very unlikely to have occurred by random chance* will lead to the decision to **reject the null hypothesis** in favour of the alternative hypothesis. The decision to *reject or not reject* is made in relation to the *null hypothesis*, whilst the *evidence (to suggest)*, or lack of, relates to the *alternative hypothesis*:

The probability of such an extreme sample purely by chance is called the ***p*-value**. A decision should be made in advance of collecting the data regarding the **significance level**, α, or *"alpha"*. When the p-value is less than α this is referred to as a *statistically-significant* result.

The decision rule for the p-value and the significance level α:

If p-value $< \alpha$ then **reject $\mathbf{H_0}$** and conclude that
there is *evidence to suggest* that the alternative hypothesis is true.

If p-value $> \alpha$ then **do not reject $\mathbf{H_0}$** and conclude that
there is *insufficient evidence to suggest* that the alternative hypothesis is true.

A significance level of 5%, or $\alpha = 0.05$, is common, and it is recommended to use this value if a question does not already specify one. In some fields a smaller value is chosen to reduce the risk of making a *Type I Error* - incorrectly rejecting the null hypothesis when it is in fact true. Failing to reject the null hypothesis error when it is in fact false is called a *Type II Error*.

	Do Not Reject H_0	Reject H_0
H_0 is true	*Correct decision*	*Type I error*
H_1 is true	*Type II error*	*Correct decision*

When performing *any* significance test the following procedure is generally performed:

1. State H_0 and H_1, and whether the test will be one-tailed or two-tailed.

2. Consider the appropriate distribution under H_0.

3. Decide on the significance level of the test, α, and the rejection criteria.

4. Calculate the test statistic and/or p-value.

5. Decide whether to reject or not reject H_0 and write a conclusion in statistical and in layman's terms, using context.

10.2 One-Sample *z*-Test for the Population Mean using a *p*-value

The *one-sample z-test for the population mean* is a hypothesis test used when a sample of data is taken from a population so that a claim about the *true value of the population mean*, μ, can be assessed. Since z-tests are based on the normal distribution, they are called **parametric tests**.

Conditions for valid use of the one-sample z-test for the population mean:

- The *sample mean* is **normally distributed**.
- The *population standard deviation*, σ, is **known**.
- The data was obtained through a **random sample**.

If one or more of these conditions is not known to be satisfied, then it is required to make an **assumption** that it *is* satisfied for the test to be valid. In any statistical study assumptions that have been made should be recognised and consideration should be given as to whether they could be justified.

Note that when the sample size is greater than 20 **the CLT can be invoked**, so the sample mean is at least *approximately* normally distributed regardless of the distribution of the underlying population. Also, again for sample sizes greater than 20, the *sample* standard deviation, s, is considered to provide a *sufficiently accurate estimate* for the *population* standard deviation, σ. These concepts remove the need for the first two assumptions to be made for *sufficiently large samples*.

Example 1

Problem: The designers of an electric car state that the distance that can be covered on a full charge is normally distributed, with a mean of 260 miles and a standard deviation of 10 miles. A car magazine randomly selects 14 of them from showrooms to be tested, with a sample mean of 257 miles obtained. Stating one assumption required, perform a hypothesis test to assess, at the 1% level of significance, the claim by the car magazine that the mean distance is actually less than the designers' claim.

Solution: $\mu = 260$, $\bar{x} = 257$, $n = 14$, $\sigma = 10$

Assume that the standard deviation of distance covered on a full charge is unchanged from the 10 miles stated.

$$\left.\begin{array}{ll} H_0: & \mu = 260 \\ H_1: & \mu < 260 \end{array}\right] \begin{array}{l} \alpha = 0.01 \\ \text{One-tailed} \end{array}$$

Under H_0:

$$P(\bar{X} < 257) = P\left(Z < \frac{257 - 260}{\frac{10}{\sqrt{14}}}\right) = P(Z < -1.12) = 0.1314$$

Since $0.1314 > 0.01$, do not reject H_0 at the 1% level of significance. There is insufficient evidence to suggest that the mean distance the car can cover on a full charge is less than the stated 260 miles, and so no evidence to support the car magazine's claim.

Example 2

Problem: The mass of crisps dispensed into a sharing-sized bag by a machine is normally distributed with a mean of 150g and a standard deviation of 2g. After being serviced, a random sample of 30 bags gives a mean of 150.8g. Test at the 5% level of significance the claim that the machine is now dispensing too much.

Solution: $\mu = 150$, $\bar{x} = 150.8$, $n = 30$, $\sigma = 2$

$$\left.\begin{array}{ll} H_0: & \mu = 150 \\ H_1: & \mu > 150 \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{One-tailed} \end{array}$$

Under H_0:

$$P(\bar{X} > 150.8) = P\left(Z > \frac{150.8 - 150}{\frac{2}{\sqrt{30}}}\right) = P(Z > 2.19) = 0.0143$$

Since $0.0143 < 0.05$, reject H_0 at the 5% level of significance.
There is evidence to suggest that the mean mass of crisps dispensed per bag is more than 150g.

For a *two-tailed* test, the probability of such an extreme result *in either direction* must be considered. Given a symmetrical distribution of the test statistic, such as with a z-test, a p-value for a two-tailed test is *double* the initial probability obtained.

Example 3

Problem: A tomato grower has been monitoring the mass of the tomatoes he grows, with the mean mass of a tomato shown to be 82 grams. Following a change in the soil composition in which they are grown, the grower wants to see whether the mean mass of their tomatoes has been affected. A sample of 36 tomatoes gives a mean mass of 83 grams and a standard deviation of 3 grams. Perform a hypothesis test to determine whether there is evidence to suggest that the mean mass of their tomatoes has changed, stating one assumption required and justifying the validity of the test with reference to the sample size.

Solution: $\mu = 82$, $\bar{x} = 83$, $n = 36$, $s = 3$

Assume that the sample of tomatoes was random. Since $n > 20$, the distribution of the sample mean is approximately normally distributed regardless of the underlying distribution of the mass of a tomato, and the sample standard deviation, s, provides a sufficiently reliable estimate for the population standard deviation, σ.

$\sigma \approx 3$

$$\left.\begin{array}{ll} H_0: & \mu = 82 \\ H_1: & \mu \neq 82 \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{Two-tailed} \end{array}$$

Under H_0:

$$P(\bar{X} > 83) = P\left(Z > \frac{83 - 82}{\frac{3}{\sqrt{36}}}\right) = P(Z > 2) = 0.0228$$

p-value $= 2 \times 0.0228 = 0.0456$ and $\alpha = 0.05$

Since $0.0.0456 < 0.5$, reject H_0 at the 5% level of significance.
There is evidence to suggest that the mean mass of the grower's tomatoes has changed from 82 grams.

Exercise 10.2

1. The length of cookery videos uploaded to a video sharing platform has been historically found to be normally distributed, with a mean of 11.7 minutes and a standard deviation of 2.3 minutes. A random sample of 30 videos uploaded this year tagged as 'cookery' on the platform gives a mean of 10.8 minutes. Assuming that the standard deviation of video length has not changed, test at the 5% level of significant the claim that the mean length of cookery videos on the platform is now less than 11.7 minutes.

2. The time taken for a teacher to mark an exam paper is normally distributed with a mean of 6.3 minutes and a standard deviation of 0.8 minutes. After more detailed marking instructions are issued, a random sample of 25 papers result in a mean marking time of 6.5 minutes per paper. Assuming that the standard deviation of time taken remains unchanged, assess at the 1% significant level the claim that the mean marking time has increased since the new marking instructions were introduced, using a hypothesis test.

3. The figure often quoted as the mean time spent with the ball being in-play during a game of football 58 minutes, with the standard deviation being 4 minutes, and it is assumed that the time for which the ball is in-play is normally distributed. In the face of increasing complaints by fans about the impact of timewasting on actual playing time, a sports statistician is planning a hypothesis test to assess the evidence that the mean time during which the ball is in-play has changed from 58 minutes. A random sample of 15 games gives a mean in-play time of 61 minutes.

   (a) State one assumption required for the valid use of a z-test for the mean time spent with the ball in-play.

   (b) Perform a parametric test at the 1% level of significance to assess the evidence that the mean actual playing time has changed from 58 minutes.

4. A farmer wants to test whether the mean mass of the newborn lambs from his flock is less than 4.2 kilograms. The mass of his newborn lambs is normally distributed, and the standard deviation has been previously established to be 0.3 kilograms. A random sample of 8 newborn lambs are weighed, and the sample mean is found to be 4.0 kilograms.

   (a) State one assumption required for the valid use of a z-test.

   (b) Perform a hypothesis test to assess the evidence at the 1% level of significance that the mean mass of his newborn lambs is less than 4.2 kilograms.

5. Last year, the value of houses in an area were normally distributed, with a mean of £184 000 and a standard deviation of £7000. A random sample of 15 of the houses this year gives a mean value of £187 000. Perform a hypothesis test at the 10% level of significance to assess the claim that the mean house price in the area has increased.

6. The heights of mature oak trees in some woodland are normally distributed with a mean of 21 metres and a standard deviation of 1.5 metres. A random sample of 30 of these trees yields a sample mean of 20.4 metres.

   (a) Test at the 10% level the claim that the mean height of the trees has changed.

   (b) Explain why the sample size meant that a parametric test could have been used in part (a) even if it were not known that the heights of the mature oak trees are normally distributed.

These questions match those from page 117, and the conclusions reached should correspond.

7. The mass of a pine marten in Scotland is historically known to be normally distributed, with a mean of 1.31kg and a standard deviation of 0.23kg. A sample of 22 pine martens captured, weighed and released in Scotland in 2023 gives a mean mass of 1.38kg. Stating two assumptions required, test at the 5% level of significance the claim that the mean mass of a pine marten in Scotland has increased.

8. The time taken by runners to complete a 5km run in a weekly race in a local park is normally distributed with a mean time of 26.3 minutes and a standard deviation of 1.4 minutes. Following some development works completed in the park, the route is unchanged but some of the running surfaces have changed. A random sample of 40 runners completing the course after the development work gives a mean of 25.9 minutes. Stating one assumption required, test the claim at the 0.1% significance level that the mean run time is no longer 26.3 minutes.

9. A telecommunications provider wishes to test whether the mean amount of data used by its customer is more than 8.6 gigabytes per month. The amount of data used by a customer is known to be normally distributed with standard deviation of 1.7 gigabytes. A random sample of 32 customers gives a mean of 9.1 gigabytes. Perform a hypothesis test at the 5% level of significance to assess the evidence that the mean amount of data used by a customer is greater than 8.6 gigabytes.

10. A large school plans to test whether the mean amount its teachers spend on photocopying is different than the £12 per year they currently budget for. A random sample of 23 teachers gives a mean of £10.90 and a standard deviation of £2.30.

    (a) Explain why a z-test for the mean amount spent can still be performed despite the population standard deviation spend is not known.

    (b) Perform a hypothesis test at the 10% level of significant to assess the evidence that the mean spend per teacher is different than £12.

11. A kettle manufacturer wishes to test whether the mean time for its kettles to boil one litre of water is more than the one minute and 5 seconds stated on the packaging. The standard deviation in the boiling time is 4 seconds, and the time taken to boil is normally distributed. A sample of 32 kettles gives a mean time of one minute and 7 seconds. Assuming that the standard deviation remains unchanged, perform a hypothesis test at the 5% level of significance to assess the evidence that the mean time taken to boil one litre is more than the packaging states.

12. The volume of petrol a driver uses driving from Glasgow to Blackburn is normally distributed, with previous studies showing the mean to be 15.7 litres and the standard deviation to be 0.3 litres. The driver believes their new driving style may have caused the mean volume of petrol used to change. Subsequently, recording the fuel consumption for 5 journeys gives a mean volume of petrol for a journey of 15.4 litres.

    (a) State two assumptions required for the valid use of a z-test.

    (b) Perform a z-test at the 1% level of significance to assess the evidence for their claim.

10.3 One-Sample z-Test for the Population Mean using a Test Statistic

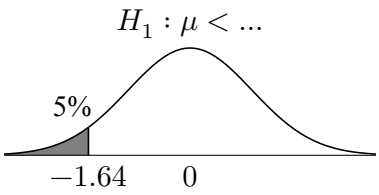
Every hypothesis test performed so far has involved calculating the probability of the observed sample mean, $\bar{x}$, by first using the *z-transformation for a sample mean* to obtain what is called a *z-score*. Instead of comparing a p-value to the significance level, another approach is to instead compare the z-score, referred to as the **test statistic**, to separately calculated cut-off values, called **critical values**.

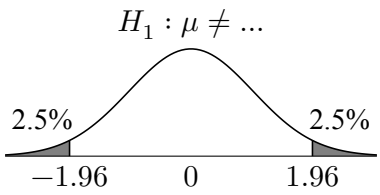
For a one-sample z-test for the population mean:

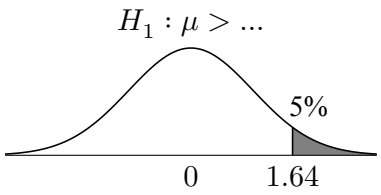
$$\text{Test statistic: } z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

A *z-test statistic* of 0 occurs when the sample mean of the observed data is equal to the value for μ claimed by the null hypothesis. The further the test statistic is from 0, the more extreme the observed data is.

Critical values can be found by considering *where the 5% least-likely z-scores lie that, if observed, would cause the null hypothesis to be rejected*. For a two-tailed test, the 5% least-likely observed values for the sample mean are split between 2.5% in the upper tail of the distribution and 2.5% in the lower tail. The symmetry of the distribution means that the values to the left of the mean are the *negative* of the equivalent values on the right.







The relevant areas of the z-distribution *beyond* the critical values, shown shaded on in the diagrams above, are often referred to as **critical regions**, and a test statistic within a critical region leads to the null hypothesis being rejected in favour of the alternative hypothesis.

The decision rule for test statistics and critical values for parametric tests:

If the **test statistic is further from zero than the critical value(s)** then **reject $\mathbf{H_0}$** and conclude that there is *evidence to suggest* that the alternative hypothesis is true.

If the **test statistic is not further zero than the critical value(s)** then **do not reject $\mathbf{H_0}$** and conclude that there is *insufficient evidence to suggest* that the alternative hypothesis is true.

One key need for this approach is that it can be difficult to calculate p-values for some distributions. **Page 12 of the SQA Data Booklet** can be used to find common critical values for z-tests without the need for more extensive probability tables or a graphical calculator. It is expected that a statistician is able to both calculate and interpret a p-value when required, as well as use test statistics and critical values. For the remainder of this textbook, the predominant approach used in examples will be **test statistics and critical values**.

Note that the following examples approach problems already covered in the examples on Pages 120-121, which used a *p*-value approach, and lead to identical conclusions to the hypothesis tests conducted.

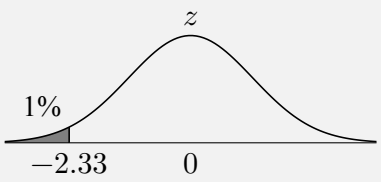
Example 1

Problem: The designers of an electric car state that the distance that can be covered on a full charge is normally distributed, with a mean of 260 miles and a standard deviation of 10 miles. A car magazine randomly selects 14 of them from showrooms to be tested, with a sample mean of 257 miles obtained. Perform a hypothesis test to assess, at the 1% level of significance, the claim by the car magazine that the mean distance is actually less than the designers' claim.

Solution: $\mu = 260$, $\bar{x} = 257$, $n = 14$, $\sigma = 10$

$$\left.\begin{array}{ll} H_0: & \mu = 260 \\ H_1: & \mu < 260 \end{array}\right] \begin{array}{l} \alpha = 0.01 \\ \text{One-tailed} \end{array}$$

Test statistic: $z = \dfrac{257 - 260}{\dfrac{10}{\sqrt{14}}} = -1.12$



Critical value: -2.33

Since $-1.12 > -2.33$, do not reject H_0 at the 1% level of significance. There is insufficient evidence to suggest that the mean distance the car can cover on a full charge is less than the stated 260 miles, and so no evidence to support the car magazine's claim.

Example 2

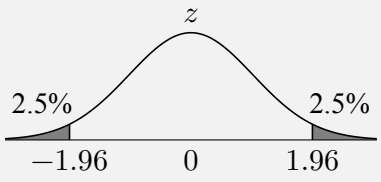
Problem: A tomato grower has been monitoring the mass of the tomatoes he grows, with the mean mass of a tomato shown to be 82 grams. Following a change in the soil composition in which they are grown, the grower wants to see whether the mean mass of their tomatoes has been affected. A sample of 36 tomatoes gives a mean mass of 83 grams and a standard deviation of 3 grams. Perform a hypothesis test to determine whether there is evidence to suggest that the mean mass of their tomatoes has changed.

Solution: $\mu = 82$, $\bar{x} = 83$, $n = 36$, $s = 3$

Since $n > 20$, take $\sigma \approx 3$.

$$\left.\begin{array}{ll} H_0: & \mu = 82 \\ H_1: & \mu \neq 82 \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{Two-tailed} \end{array}$$

Test statistic: $z = \dfrac{83 - 82}{\dfrac{3}{\sqrt{36}}} = 2$



Critical value: 1.96

Since $2 > 1.96$, reject H_0 at the 5% level of significance.
There is evidence to suggest that the mean mass of the grower's tomatoes has changed from 82 grams.

Exercise 10.3

1. A packing process packages ground coffee into bags. The mass of each bag is normally distributed, with standard deviation 15g. To comply with the technical standards stipulated for the process, the mean mass of coffee per bag should be 500g. After a recent adjustment of some of the machinery involved in the process, a random sample of 25 bags is taken and the sample mean is found to be 512g. Test the claim that there has been a change in the mean mass of coffee per bag at the 5% level of significance.

2. The lengths of metal bars produced by a particular machine in a factory are normally distributed with mean length 420cm and standard deviation 12cm. The machine is serviced, after which a random sample of 80 bars gives a mean length of 423cm. Perform a parametric test to assess whether there is evidence at the 1% level of significance of an increase in the mean length of the bars produced, assuming the standard deviation of bar length remains the same.

3. In 1980 the mass of a herring in the North Sea was found to be normally distributed with a mean of 160g and a standard deviation of 10g. In 2020 a sample of 15 herrings yielded a mean mass of 158g. Assuming that the standard deviation is unchanged, test at the 1% level of significance the hypothesis that the mean mass has decreased since 1980.

4. The mass of a pack of salt is normally distributed with a mean of 200g and a standard deviation of 12g. After servicing, it is suspected that the salt dispensing machine is not dispensing the same mean amount of salt. A sample of 23 bags yields a mean mass of 205g per pack. Stating two assumptions required, test the hypothesis that the mean amount of salt dispensed by the machine is no longer 200g.

5. Last season, the distances covered in a 90-minute game by the players of a football team were normally distributed with a mean of 7.1km. A random sample, of size 26, of distances covered by players in the new season gave a sample mean distance covered of 7.3km and a standard deviation of 0.7km.

   (a) Perform a hypothesis test to assess the evidence, at the 5% level of significance, that the mean distance players are covering in a game this season has changed.

   (b) Explain why a z-test could still be performed in part (a) despite the population standard deviation in running distance being unknown.

6. A popular type of loaf sold by a regional chain of bakeries has been previously observed to have a mean mass of 500g, with a standard deviation of 20g. Following a change in the recipe used for the loaves, a random sample of 25 loaves yielded a mean mass of 483g.

   (a) State the condition for a sample size which implies that a sample mean will be approximately normally distributed even whilst the underlying distribution is not normally distributed.

   (b) State one assumption required for the valid use of a z-test for the population mean loaf mass.

   (c) If the assumption from part (b) was to be considered unreliable, explain what could be done to allow a z-test to still be performed.

   (d) Test at the 5% significance level the claim that the mean mass of a loaf has decreased.

These questions match those from page 113, and the conclusions reached should correspond.

7. The mass of a pine marten in Scotland is historically known to be normally distributed, with a mean of 1.31kg and a standard deviation of 0.23kg. A sample of 22 pine martens captured, weighed and released in Scotland in 2023 gives a mean mass of 1.38kg. Stating two assumptions required, test at the 5% level of significance the claim that the mean mass of a pine marten in Scotland has increased.

8. The time taken by runners to complete a 5km run in a weekly race in a local park is normally distributed with a mean time of 26.3 minutes and a standard deviation of 1.4 minutes. Following some development works completed in the park, the route is unchanged but some of the running surfaces have changed. A random sample of 40 runners completing the course after the development work gives a mean of 25.9 minutes. Stating one assumption required, test the claim at the 0.1% significance level that the mean run time is no longer 26.3 minutes.

9. A telecommunications provider wishes to test whether the mean amount of data used by its customer is more than 8.6 gigabytes per month. The amount of data used by a customer is known to be normally distributed with standard deviation of 1.7 gigabytes. A random sample of 32 customers gives a mean of 9.1 gigabytes. Perform a hypothesis test at the 5% level of significance to assess the evidence that the mean amount of data used by a customer is greater than 8.6 gigabytes.

10. A large school plans to test whether the mean amount its teachers spend on photocopying is different than the £12 per year they currently budget for. A random sample of 23 teachers gives a mean of £10.90 and a standard deviation of £2.30.

    (a) Explain why a z-test for the mean amount spent can still be performed despite the population standard deviation spend is not known.

    (b) Perform a hypothesis test at the 10% level of significant to assess the evidence that the mean spend per teacher is different than £12.

11. A kettle manufacturer wishes to test whether the mean time for its kettles to boil one litre of water is more than the one minute and 5 seconds stated on the packaging. The standard deviation in the boiling time is 4 seconds, and the time taken to boil is normally distributed. A sample of 32 kettles gives a mean time of one minute and 7 seconds. Assuming that the standard deviation remains unchanged, perform a hypothesis test at the 5% level of significance to assess the evidence that the mean time taken to boil one litre is more than the packaging states.

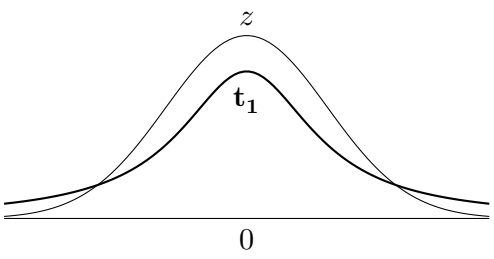
12. The volume of petrol a driver uses driving from Glasgow to Blackburn is normally distributed, with previous studies showing the mean to be 15.7 litres and the standard deviation to be 0.3 litres. The driver believes their new driving style may have caused the mean volume of petrol used to change. Subsequently, recording the fuel consumption for 5 journeys gives a mean volume of petrol for a journey of 15.4 litres.

    (a) State two assumptions required for the valid use of a z-test.

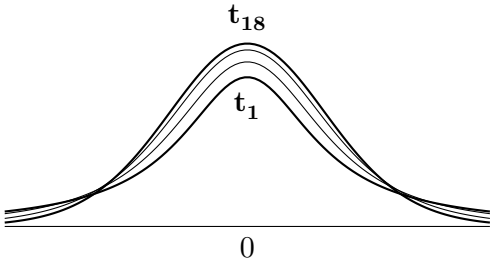
    (b) Perform a z-test at the 1% level of significance to assess the evidence for their claim.

10.4 One-Sample *t*-Test for the Population Mean

In the early 20th century William Sealy Gosset, statistician and Head Experimental Brewer at the Guinness factory in Dublin, described the distribution of the sample mean when the **population standard deviation is unknown**. This was necessary since he was working with **small sample sizes**, for which the sample standard deviation could not be used as a reliable estimator for the population standard deviation. Since Guinness did not want rival companies to know that they were employing a statistician to improve their brewing, he published his work under the pen-name *Student*, hence the distribution he described becoming known as **Student's *t*-distribution**, or *the t-distribution*.



t-distribution, t_1, and normal distribution, z



t-distributions with $\nu = 1, 2, 5, 18$

Not knowing the population standard deviation, σ, creates additional uncertainty about the distribution of the sample mean, with the **sample standard deviation**, s, used instead. This results in the shape of a t-distribution resembling that of a normal distribution but with *fatter tails*. The t-distribution takes one parameter called **degrees of freedom**, notated using the Greek letter ν, which is the *number of data points from the sample that are free to vary* under the null hypothesis. For the one-sample t-test for population mean, the degrees of freedom can be calculated for a sample size n as:

$$\nu = n - 1$$

The one-sample t-test for population mean is a **parametric test** that should be used, in this course, in place of the equivalent z-test when the population standard deviation, σ, is not known **and** the sample size is less than 20.

Conditions for valid use of the one-sample t-test for the population mean:

- The *underlying population* is **normally distributed**.
- The data was obtained through a **random sample**.

The t-test statistic is identical to that of the equivalent z-test, except the use of the sample standard deviation, s, instead of the population standard deviation, σ.

For a one-sample t-test for the population mean:

$$\text{Test statistic: } t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Critical values for the t-test can be found on **Page 13 of the SQA Data Booklet**, referring to the significance level, whether the test is one-tailed or two-tailed and the degrees of freedom.

Example 1

Problem: A supermarket sells sharing-sized bags of a particular brand of crisps. A consumer watchdog is asked to investigate a claim that the mean mass of crisps contained in a bag is less than the stated contents of 150 grams. A random sample of bags gives the following results, in grams:

148 151 149 152 145 150 146 151

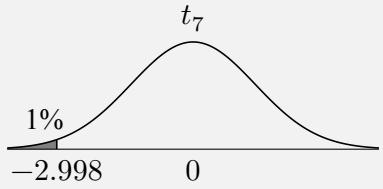
Stating a necessary assumption, perform a parametric test to assess the claim at the 1% level of significance.

Solution: $\mu = 150$, $\bar{x} = 149$, $n = 8$, $s = 2.51$

Assume that the mass of crisps in a bag is normally distributed.

$$\left.\begin{array}{ll} H_0: & \mu = 150 \\ H_1: & \mu < 150 \end{array}\right] \begin{array}{l} \alpha = 0.01 \\ \text{One-tailed} \end{array}$$

Test statistic: $t = \dfrac{149 - 150}{\dfrac{2.51}{\sqrt{8}}} = -1.127$



Degrees of freedom: $\nu = n - 1 = 8 - 1 = 7$

Critical value: -2.998

Since $-1.127 > -2.998$, do not reject H_0 at the 1% level of significance. There is insufficient evidence to suggest that the mean mass of crisps in a bag is less than the stated value of 150g.

Example 2

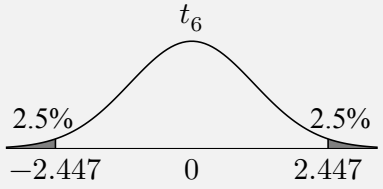
Problem: Marks scored in an exam are known to be approximately normally distributed, and historically the mean mark achieved by students is 72. After some minor tweaks to the exam, marks gained by a random sample of 7 students generated the summary data below. Perform a hypothesis test at the 5% level of significance to determine whether there is evidence that the difficulty of the exam has changed for students.

$$\Sigma x = 571 \qquad \Sigma x^2 = 46851 \qquad n = 7$$

Solution: $\mu = 72$, $\bar{x} = 81.57$, $n = 7$, $s = 6.754$

$$\left.\begin{array}{ll} H_0: & \mu = 72 \\ H_1: & \mu \neq 72 \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{Two-tailed} \end{array}$$

Test statistic: $t = \dfrac{81.57 - 72}{\dfrac{6.754}{\sqrt{7}}} = 3.749$



Degrees of freedom: $\nu = n - 1 = 7 - 1 = 6$

Critical value: 2.447

Since $3.749 > 2.447$, reject H_0 at the 5% level of significance. There is evidence to suggest that the mean mark students will achieve is no longer 72, and that the difficulty of the exam has changed.

Exercise 10.4

1. The amount of sugar contained in frozen pizzas, per 100g, is normally distributed. A random sample of 8 brands of frozen pizza gives the following sugar content per 100g, measuring in grams:

   3.5 2.1 5.3 4.4 3.4 2.2 4.1 5.0

   Test at the 5% significance level the claim that the mean sugar content per 100g in a frozen pizza is less than 5 grams.

2. In a game of football, the amount of time between a goal being scored and the game resuming is normally distributed. The time taken to resume, in seconds, is recorded for a random sample of 6 goals, with results:

   49 25 31 23 38 29

   Test at the 10% level of significance the claim that the mean time taken for a game to resume after a goal is greater than the 30 seconds that is generally considered to be typical.

3. A camera tripod can supposedly safely hold masses of up to 4.5kg. A photography shop disputes this figure, and randomly selects 10 tripods. Each is loaded progressively with heavier masses until it fails. The results, in kg are summarised as follows:

   $$n = 10 \quad \Sigma x = 40.5 \quad \Sigma x^2 = 165.49$$

   Stating an assumption required, test at the 1% level the claim that the mean mass the tripods can withstand before failure is actually less than 4.5kg.

4. *Vinho Verde* is a type of Portuguese wine that originated in the Minho Province in the far north of the country. In 2009, the University of Minho supported a study which examined the extent to which wine quality can be judged through measuring its physiochemical properties. As part of the study, the acidity level of a random sample of 11 white *Vinho Verde* wines all made using using the *alvarinho* grape were measured. The summary data obtained from this sample was as follows (measured in pH):

   $$\Sigma x = 34.89 \qquad \Sigma x^2 = 110.81$$

   Worldwide, the mean acidity level of white wines is 3.4 pH.

   Stating one assumption required, perform a parametric test at the 5% level of significance to determine whether the white *Vinho Verde* wines made from the *alvarinho* grape differ in acidity level from those made worldwide.

5. A *soporific drug* is one which has been shown to induce sleep, and increase drowsiness. An early 20th century study into possibly soporific drugs involved giving ten patients a dose of one such drug and recording whether they reported sleeping *more* than usual that night (recorded as a positive number) or *less* than usual (recorded as a negative number). The results recorded, in hours, were:

   0.7 -1.6 -0.2 -1.2 -0.1 3.4 3.7 0.8 0.0 2.0

   The study wished to determine whether there was evidence from the study of the drug leading to increased sleep. Perform a hypothesis test to assess the evidence at the 2.5% level of significance.

6. A particular model of bass guitar made by a company is known to have historically been produced with a mean mass of 9.4 pounds, where the mass of a bass guitar is normally distributed. Now under new ownership, a bass guitar magazine is looking into claims that the bass guitars of that model now being produced are heavier, possibly due to an undeclared change in the wood used. A random sample of 14 such bass guitars gives the following masses, in pounds:

9.4	9.6	9.3	9.5	9.7	9.6	9.6
9.5	9.4	9.7	9.6	9.3	9.8	9.7

Perform a hypothesis test at the 1% level of significance to assess evidence for the claims.

7. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. As part of this, 14 newly hatched chicks were randomly selected to be given a soybean-based feed supplement. Their masses in grams after six weeks were recorded, and the following summary statistics were obtained:

variable	n	min	max	mean	sd
weight (grams)	14	158.0	329.0	246.4	54.1

It has historically been observed that the mean mass of a chick after six weeks is 260 grams. Stating one assumption required, perform a hypothesis test to determine whether there is evidence at the 5% level of significance that giving a soybean-based supplement affects the mass a chick will reach by six weeks.

8. The second-most common species of tree in the United States is the *loblolly pine*, regarded as one of the most commercially important species of tree in the Southeastern US for its timber. A survey based on a sample of 15 of the loblolly trees in one region revealed a sample mean height of 32.4 feet and a standard deviation of 2.07 feet. The heights of the trees are known to be normally distributed. Stating one assumption required, test at the 1% significance level whether the data provides evidence to suggest the mean height of loblolly trees in the region is greater than 30 feet.

9. A school is looking to figure out a course for a the annual school cross-country, with the goal of choosing a course such that the mean time it takes a student at the school to finish is 45 minutes. The first ten students to respond to a request by the PE department for volunteers to test the course have their time taken to complete the course recorded. The results of the t-test used to assess whether the course meets their requirements is as follows:

```
t-test for the population mean
n = 10    sample mean = 53.1    sample SD = 11.2492
alternative hypothesis: true mean is not equal to 45
t = ****    p-value = 0.0488
```

   (a) State the null and alternative hypotheses for the test conducted.

   (b) Determine the missing value of the t-test statistic, indicated by **** in the computer output.

   (c) Find the critical value for the hypothesis test, at the 5% level of significance.

   (d) Use the p-value to write a conclusion to the hypothesis test.

   (e) Name the sampling method used and comment on the appropriateness.

10. An experiment into the cold tolerance of a species of grass plant looked at the effect of low temperature on its carbon dioxide (CO2) uptake. It it known that, under normal conditions, the CO2 uptake rate for the species is known to be 33 $\mu mol/m^2$, whilst the sample of 7 areas exposed to chilling resulted in a sample mean uptake of 29.97 $\mu mol/m^2$ and a standard deviation of 8.335 $\mu mol/m^2$. Perform a parametric test to assess the evidence for CO2 uptake being affected by low temperature for the grass plant, at the 10% level of significance.

10.5 One-Sample *z*-Test for the Population Proportion

If X of the elements from a random sample of size n from a population of interest satisfy some condition, then the **sample proportion**, $\hat{P}$, can be calculated as:

$$\hat{P} = \frac{X}{n}$$

For example, if a survey of registered voters reveals 114 out of 250 intend to vote for the incumbent candidate, the *sample proportion*, $\hat{p}$, is $\frac{114}{250} = 0.456$. The true value of the **population proportion**, $\boldsymbol{p}$, may be something different entirely. The **one-sample z-test for proportion** assesses whether an *observed sample proportion* provides statistically significant evidence against a *null-hypothesised population proportion*. The test statistic can be derived from the distributions of the random variables X and hence $\hat{P}$:

$$X \sim B(n, p)$$

When $np > 5$ and $nq > 5$, X is *approximately* normally distributed:

$$\begin{aligned} E(X) &= np \\ V(X) &= npq \\ X &\approx N(np, npq) \end{aligned}$$

The distribution of $\hat{P}$ can now be obtained since $\hat{P} = \frac{X}{n}$:

$$E\left(\hat{P}\right) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n} \times np = p$$

$$V\left(\hat{P}\right) = V\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right)^2 V(X) = \frac{1}{n^2} \times npq = \frac{pq}{n}$$

Therefore:

$$\hat{P} \approx N\left(p, \frac{pq}{n}\right)$$

The z-transformation leads to the required test statistic:

For a one-sample z-test for the population proportion:

$$\text{Test statistic: } z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Note that $\sqrt{\frac{pq}{n}}$ is called the **standard error of the sample proportion**.

Conditions for valid use of the one-sample z-test for the population proportion:

- $\boldsymbol{np > 5}$ **and** $\boldsymbol{nq > 5}$
- The data was obtained through a **random sample**.

Example 1

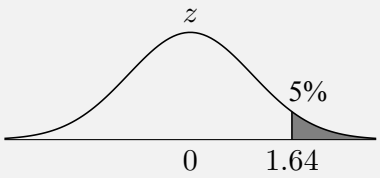
Problem: A data storage company knows that 10% of the hard drives they use will fail within the first year of their use. They have decided to trial the use of a cheaper model of hard drive, but are concerned that it may have a higher failure rate. During this trial, 60 of the new hard drives are randomly selected to be monitored for a year once installed. Of those 60, 11 fail.

Use a parametric hypothesis test to assess the evidence, at the 5% level of significance, that the new drives have a higher first-year failure rate than the previous model.

Solution: $p = 0.1$, $q = 0.9$, $n = 60$, $\hat{p} = \frac{11}{60} = 0.183$

$$\left.\begin{array}{ll} H_0: & p = 0.1 \\ H_1: & p > 0.1 \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{One-tailed} \end{array}$$

Test statistic: $z = \dfrac{0.183 - 0.1}{\sqrt{\dfrac{0.1 \times 0.9}{60}}} = 2.14$



Critical value: 1.64

Since $2.14 > 1.64$, reject H_0 at the 5% level of significance.
There is evidence to suggest that the proportion of the new hard drives that fail within their first year of use is higher than 10%.

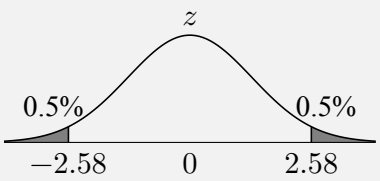
Example 2

Problem: A candidate in a mayoral election claims that 85% of voters intend to vote for them. Out of a random sample of 80 people eligible to vote, 56 say that they intend to vote for that candidate. Perform a hypothesis test to assess the evidence at the 1% level of significance that their claim may be untrue.

Solution: $p = 0.85$, $q = 0.15$, $n = 80$, $\hat{p} = \frac{56}{80} = 0.7$

$$\left.\begin{array}{ll} H_0: & p = 0.85 \\ H_1: & p \neq 0.85 \end{array}\right] \begin{array}{l} \alpha = 0.1 \\ \text{Two-tailed} \end{array}$$

Test statistic: $z = \dfrac{0.7 - 0.85}{\sqrt{\dfrac{0.85 \times 0.15}{80}}} = 3.76$



Critical value: 2.58

Since $3.76 > 2.58$, reject H_0 at the 1% level of significance.
There is evidence to suggest that the true proportion of voters intending to vote for that candidate is not 85%, and that the candidates claim may be untrue.

Exercise 10.5

1. A coin is thrown 500 times and 267 heads are obtained. Test whether the coin is biased, using a 10% significance level.

2. A newspaper claims that over 60% of its readers are car owners. In a random sample of 312 readers there are 208 car owners. Test whether there is evidence at the 1% level of significance to support their claim.

3. In a public opinion poll, 1000 randomly chosen electors were asked whether they would vote for the 'Purple Party' at the next election, and 357 replied 'yes'. The leader of the 'Purple Party' believes the true proportion is 0.4. Test at the 5% significance level whether she is overestimating her support.

4. The proportion of left-handed people in the population is 10%. A random sample of 400 professional artists found that 47 were left-handed. Perform a hypothesis test at the 5% level of significance to assess the evidence for the claim that the proportion of left-handed artists is not 10%.

5. A seed company sells pansy seeds in mixed packets and claims that 20% of the resulting plants will have red flowers. A packet of seeds is sown by a gardener who finds that only 9 out of 82 plants have red flowers. Perform a hypothesis test to assess the claim that the seed company's claim is not correct, using a 1% significance level.

6. A teacher says that the majority of pupils in a school prefer cats to dogs. A sample of 25 pupils are asked which they prefer, and only 7 prefer cats. Perform a parametric test at the 5% level of significance to assess the claim that the teacher is wrong, stating an assumption required.

7. Scotland's 2022 census revealed that of those in employment, 7% were working in construction. In 2024, a random sample of 200 adults living in Scotland revealed that, of the 189 who were in some form of employment, only 8 were working in construction. Perform a parametric test to assess whether the data gives evidence at the 10% level of significance that the proportion of working adults employed in construction has changed since the census.

8. A football manager claims that the *Video Assistant Referee* (VAR) system used in all games played in the top league leads to the on-field referee being instructed to watch a replay on a pitchside monitor in at least three-quarters of games. A journalist looks back at the 60 games played up to that point in the season and finds that the VAR instructed the referee to watch a replay in only 38 of the games. Assess at the 1% significance level whether the data supports the journalist's concerns that the football manager was incorrect in their statement.

9. In a factory, ready-sliced potatoes are cooked to make crisps, which are then placed into bags before flavouring is added and the bags sealed. Occasionally a packet of crisps is sealed without flavouring having been added. The crisp manufacturers have decided that they are content with this process provided no more than 1.8% of bags are flavourless.

   (a) A random sample of 50 bags is taken. Explain why performing a z-test for the proportion of flavourless crisps obtained from this sample would not be appropriate.

   (b) Determine the minimum sample size required such that a z-test would be appropriate.

   (c) From a much larger sample of 500 bags of crisps, 14 are found to be flavourless. Test at the 5% level of significance whether the manufacturers should be concerned with the process.

10.6 Choosing the Right Hypothesis Test

Selecting the appropriate hypothesis test to be used in any given context is an important skill for a statistican, and warrants careful practice. The following questions contain a mix of *z-tests for the population mean, t-tests for the population mean and z-tests for the population proportion.*

Exercise 10.6

1. A keen cyclist often takes a route which consists of a circuit, travelling through a forest and around a hill before returning to her starting point, timing herself on each occasion. The time it takes her to complete a circuit is normally distributed, with mean 42.7 minutes and standard deviation 3.8 minutes. After buying a new bike, she records the time is takes her to complete a circuit on 7 randomly selected occasions, obtaining a sample mean of 41.3 minutes. Assuming the standard deviation in her time taken remains unchanged, perform a hypothesis test to assess her data for evidence that her mean time taken to complete the route has reduced.

2. A popular chain of fast food restaurants states on its website that a vegan alternative to its most popular menu item is available in over 80% of its restaurants. A researcher for a website that publishes articles on veganism and vegetarianism selects 60 of its restaurants at random and asks one reader from nearby each restaurant to visit and see whether the vegan alternative it available. Out of the 60 restaurants visited, the vegan alternative was only available in 42 of them. Perform a hypothesis test at the 5% level of significance to assess whether the data suggests the chain is exaggerating the availability of the vegan alternative.

3. The mass of beef mince contained within packs that state they contain 1kg of mince is measured, for a random sample of 8 packs. The results obtained, in kg, are:

   0.97 1.01 0.99 0.97 1.00 0.98 0.99 0.96

   Perform a parametric test to assess the data for evidence that the mean mass of mince in a pack is less than that stated on the packaging.

4. A report by the *Energy Saving Trust* states that households in the UK use a mean of 349 litres of water each day. A family living in an area in which water usage is metered and billed for monitors its daily volume of water used on 8 randomly selected days. They find that the sample yields a mean volume of water used of 357 litres and a standard deviation of 12 litres. Stating one assumption required, perform a test at the 10% level of significance to assess whether the family has a higher mean daily water consumption than the figure quotes by the report.

5. *Screen time* refers to the amount of time a person spends in front of some kind of digital screen, such as a smartphone, tablet or computer monitor. Researchers exploring screen time in the UK reported that the mean daily screen time for an adult in the UK is 6 hours and 9 minutes, or 6.15 hours, with a standard deviation of 1 hour and 18 minutes, or 1.3 hours. Interested in regional variations, researchers in located in Scotland contacted 20 randomly selected adults living in Scotland and asked them to report their screen time for the previous day. 16 of them responded, with a sample mean of 4 hours and 51 minutes, or 4.85 hours, obtained.

   (a) Assuming the standard deviation for adults in Scotland is the same as that for the UK as a whole, perform a hypothesis test at the 5% level of significance to assess whether the data supports any suggestion that mean daily screen time differs between adults in Scotland and the UK as a whole.

   (b) Give one reason why the method of gathering data used may call into question the reliability of the hypothesis test.

   A report by another group of researchers wish to test the claim made by a newspaper that ”a majority of adults in Scotland are taking steps to reduce their screen time”. A survey of 120 randomly selected adults are asked whether they are taking steps to reduce their screen time, with 71 stating that they are, and the remaining 49 that they are not.

   (c) Perform a parametric test to assess, at the 5% significance level, whether there is evidence to support the newspaper’s claim.

Review Exercise

1. A pharmaceutical company has developed a drug designed to reduce the recovery time needed for a specific type of operation from the well-established mean time of 11 days. Following trials involving treating 24 randomly selected patients admitted to hospital for the operation with the drug, it is found that the mean recovery time for these patients is 9.6 days, with a standard deviation of 2.7 days.

   (a) Explain why a z-test can be used for this data despite the population standard deviation, σ, not being known.

   (b) Perform a hypothesis test at the 1% level of significance to assess the evidence that the drug may reduce the mean recovery time for the operation.

2. A car company wishes to check that a new machine at their factory producing front bumpers is still maintaining the mean mass of 2.8kg for the part as the old machine it replaced. Eight bumpers produced by the machine are randomly selected and weighed. The results, in kilograms, are:

   2.9 2.6 3.3 3.1 3.0 2.9 3.1 2.8

   Perform a parametric test at the 5% level of significance to determine whether there is evidence to suggest that the mean mass of the part produced by the new machine has changed, stating an assumption required.

3. The speed of cars passing through a village is being studied in an effort to improve safety for pedestrians, and is known to be normally distributed. The mean speed has previously been monitored over a number of years and found to be consistently at 33 mph, with a standard deviation of 4 mph. Following the installation of speed bumps, the speed of 15 randomly selected cars is measured at the same point as before, with a sample mean of 31 mph observed.

   (a) Perform a parametric test to determine whether the data provides evidence to support the council's claim that the speed bumps have been successful in reducing the mean speed of cars passing through the village.

   (b) State an assumption required for the test conducted in part (a).

4. A driving instructor posts an advert online stating nine out of ten of their students pass their driving test first time. A random sample of 60 students taught by the instructor reveals that only 47 of them passed their driving test first time. Perform a hypothesis test at the 1% level of significance to assess the evidence for the claim that the proportion of students who pass first time is less than the instructor claims.

5. A coach working with a junior athletics club has observed that for many years the mean reaction time in 100m races for the sprinters at the club has been consistently stable at 185 milliseconds. Claiming that this figure may no longer be valid, she randomly selects 8 sprinters and records their reaction time in their next races. None of the athletes raced together in their next race. The summary data, where times were measured in milliseconds, is below:

   $$n = 8 \quad \Sigma x = 1372 \quad \Sigma x^2 = 237466$$

   (a) Stating an assumption required, perform a t-test to test the coach's claim at the 10% level of significance.

   (b) The coach initially debated recording the reaction times of the eight athletes running in the gold medal race of club's annual 100m championships. Suggest one reason why she was correct to reject this approach.

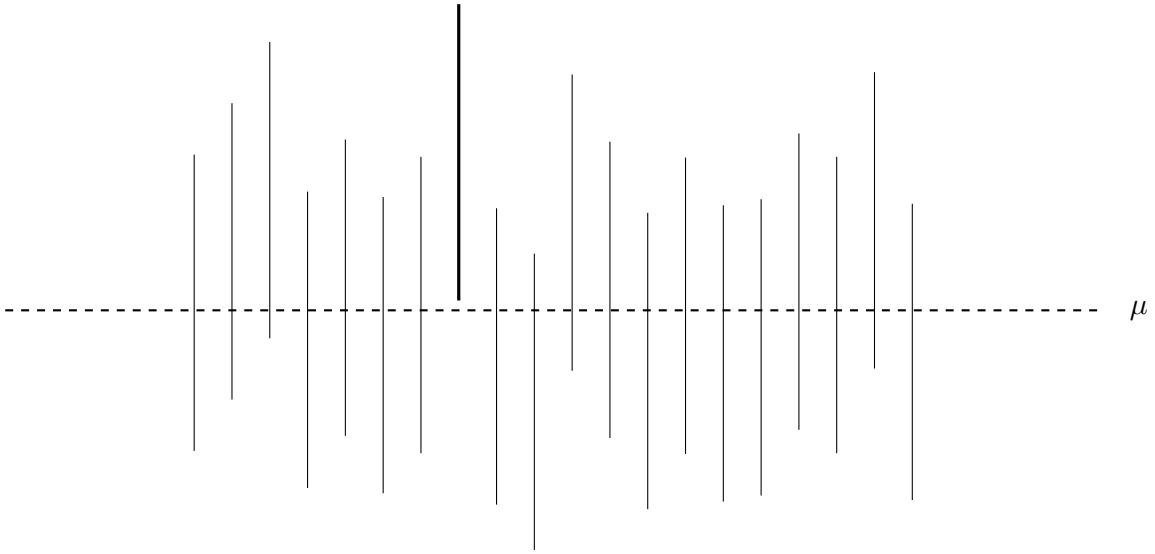
11

Confidence Intervals

From a random sample of data from a population, the sample mean could be calculated and used as a *point estimate* for the true, unknown value of the population mean. If the sample was very small, or the variability of the data is very high, then it would not be appropriate to have a great deal of confidence in the *precision* of this value. To avoid the misleading sense of precision that point estimates on their own can give, many institutions will require *confidence intervals* to routinely be stated as part of any statistical analysis. Statistical software will often also report a confidence interval whenever a hypothesis test is performed.

"A 95% confidence interval for the mean weight of an adult pine marten, in kg, is (1.42 , 1.58)"

A common misconception is that such a confidence interval can be interpreted as *"There is a 95% probability that the true mean weight is between 1.42 and 1.58."* Care should be taken to *avoid* interpreting the significance level used as a probability relating to any particular generated confidence interval and the parameter it relates to. Instead, *"95% confidence"* refers to the **long-run success rate of the procedure** used to generate the interval. If random samples are repeatedly taken from a distribution with population mean μ, and a confidence interval produced from each, in the long-run 95% of these intervals will contain μ. The figure below shows twenty 95% confidence intervals generated using simulated data, one of which does not contain the true value of the population mean, μ, in the interval.



11.1 A z-Confidence Interval for the Population Mean

Given a *random sample* from an distribution that is *normally distributed* with *known population standard deviation* σ, a $(1-\alpha)\%$ *z-confidence interval* for the population mean, μ, can be calculated:

A z-confidence interval for the population mean:

A $(1-\alpha)\%$ CI for μ is: $\bar{x} \pm z_{1-\alpha/2} \times \dfrac{\sigma}{\sqrt{n}}$

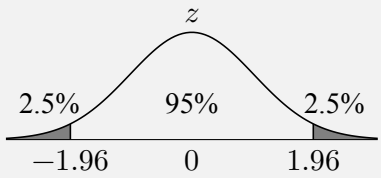
Note that the conditions required match those for a z-test for the population mean.

A sketch can be used to help find the required z-value. In this course, only *two-tailed* confidence intervals will calculated.

Example 1

Problem: The masses of ceramic dishes sold by a kitchenware shop are known to be normally distributed with a standard deviation of 12g. A random sample of 15 such ceramic dishes are weighed and the sample mean is 430 grams. Obtain a 95% confidence interval for the mean mass of a ceramic dish.

Solution: $\bar{x} = 430, n = 15, \sigma = 12$



A 95% CI for μ is: $\quad 430 \pm 1.96 \times \dfrac{12}{\sqrt{15}} \quad = \quad 430 \pm 6.07$

Therefore a 95% confidence interval for the population mean dish mass, in grams, is $(423.93, 436.07)$.

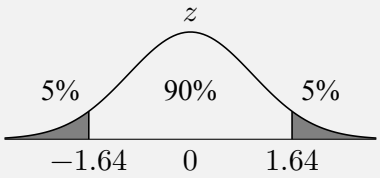
It should be observed that the z-confidence interval is symmetrical with the **sample mean at its centre**. The width of the interval is determined by the **standard error** and the confidence level required.

Example 2

Problem: An airline picks a random sample of 25 of their flights and records how long it takes for all the passengers to depart the plane from the time the cabin doors are opened. The sample mean of 6.5 minutes is obtained, and a sample standard deviation of 0.8 minutes. Determine a 90% confidence interval for the mean time taken for all passengers to depart one of the airline's planes, and interpret this confidence interval.

Solution: $\bar{x} = 6.5, n = 25, s = 0.8$

Since $n > 20$ a z-confidence interval is appropriate, and $\sigma \approx 0.8$.



A 90% CI for μ is: $\quad 6.5 \pm 1.64 \times \dfrac{0.8}{\sqrt{25}} \quad = \quad 6.5 \pm 0.26 \quad \therefore \quad (6.24, 6.76)$

90% of all confidence intervals generated by this procedure will contain the value of the population mean.

Exercise 11.1

1. The height of a red panda is known to be normally distributed with a standard deviation of 2cm. A random sample of 8 red pandas have their heights measured, with a sample mean of 57cm obtained. Obtain a 99% confidence interval for the mean height of a red panda.

2. An airline knows from historical data that the mass of the hold baggage for a passenger is normally distributed with a standard deviation of 2.1kg. The mass of the hold baggage is measured for a random sample of 11 passengers, with a sample mean of 13.8kg calculated. Assuming that the standard deviation remains unchanged, obtain a 90% confidence interval for the mean mass of the hold luggage for the airline's passengers.

3. The time taken for a chef to make a particular dish from their restaurant's main course menu is normally distributed, with the standard deviation known to be 0.9 minutes. On 7 randomly chosen occasions, the time it takes to make the dish is measured, with a sample mean of 15.2 minutes obtained. Calculate a 95% confidence interval for the mean amount of time it takes for the chef to make the dish.

4. The speeds of heavy goods vehicles (HGVs) crossing an old, rural bridge are normally distributed, with a historically observed standard deviation of 2.6 mph. Following improvements to the surface of the road on the bridge, the local council wishes to know what the mean speed of HGVs crossing the bridge is. The speeds of 17 randomly selected HGVs are recorded, with a sample mean of 39 mph. Obtain a 99% confidence interval for the mean speed of HGVs crossing the bridge, stating one assumption required.

5. The masses of the eggs laid by the chickens on a farm are normally distributed. The farmer weighs 25 randomly chosen eggs, in grams, obtaining the following summary statistics:

$$n = 25 \quad \Sigma x = 1254 \quad \Sigma x^2 = 63186$$

   Obtain a 90% confidence interval for the mean mass of the chicken eggs laid by the chickens at the farm.

6. A restaurant manager is concerned about the time a table has to wait to have a drinks order taken once they have been seated. On 30 random occasions throughout a week, the time taken is recorded, in minutes. The manager puts the data into some statistical software and obtains the following summary data:

```
x -> drinks order wait data
n = 30    sum of x = 171    x bar = 5.7    SD of x = 0.8
```

   Obtain a 95% confidence interval for the mean wait time.

7. A workers' rights organisation wishes to understand the number of hours worked per week by full-time employees of the local council. A random sample of 24 employees gives a sample mean of 38.2 hours and a standard deviation of 1.4 hours.
   (a) Obtain a 95% confidence interval for the mean number of hours worked by full-time employees of the local council.
   (b) The organisation would like a narrower confidence interval, and one organisation member suggests that obtaining instead a 99% confidence interval will provide this. Calculate the width of a 99% confidence interval for the mean hours worked and comment on the member's suggestion.
   (c) The organisation finally decides to stick with a 95% confidence interval, but to extend the sample size to obtain a confidence interval with width no greater than 1. Determine the minimum number of **additional** full-time workers that must be included in the sample in order to obtain this, assuming the sample standard deviation remains at 1.4 hours.

8. For a 90% z-confidence interval for the population mean to have a width of 12.4, based on a sample of size 7, calculate the value of σ.

11.2 Confidence Intervals and Hypothesis Testing

The calculations involved in calculating a *two-tailed* z-confidence interval for the population mean correspond with those for a *two-tailed* z-test for the population mean. Such a confidence interval could therefore allow inferences to be made about the population mean, μ.

Inference using confidence intervals:

If a given value for μ is **outwith a calculated confidence interval** then this would *support a claim* that it is *not the true value of the population mean*.

If a given value for μ is **within a calculated confidence interval** then this would *not support a claim* that it is not the true value of the population mean.

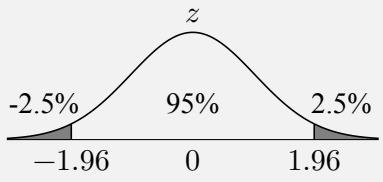
In this course, it should be noted that an instruction to perform a *test* requires a hypothesis test to be conducted using either a p-value or a test statistic, rather than a confidence interval.

Example

Problem: The mass of food eaten each day by a family's cat is normally distributed with a mean of 280 grams and a standard deviation of 13 grams. Recently, they notice that their cat doesn't seem to eat the same amount of food as it used to. On 14 randomly chosen days they measure the mass of food that their cat eats, obtaining a mean of 273 grams per day. Stating an assumption required, obtain a 95% confidence interval for the mass of food eaten daily by their cat, and comment on their concern that their cat doesn't eat the same amount of the food as it did previously.

Solution: $\bar{x} = 273, n = 14, \sigma = 13$

Assume that the standard deviation of mass of food eaten per day is unchanged from 13 grams.



A 95% CI for μ is: $\quad 273 \pm 1.96 \times \dfrac{13}{\sqrt{14}} \quad = \quad 273 \pm 6.81$

Therefore a 95% confidence interval for the mean mass of food now being eaten daily by their cat, in grams, is $(266.19, 279.81)$.

Since 280 lies outwith the confidence interval, this supports their concern that the cat is not eating the same mean daily mass of 280 grams as it was before.

Exercise 11.2

1. A packing process packages ground coffee into bags. The mass of each bag is normally distributed with standard deviation 15g. A random sample of 25 bags is taken and the sample mean is found to be 505g. Obtain a 95% confidence interval for the mean mass of ground coffee in a bag, and comment on the claim that the mean mass being dispensed is not the 500g that it is supposed to be.

2. The lengths of metal bars produced by a particular machine are normally distributed with standard deviation 12cm. The machine is serviced, after which a random sample of 80 bars gives a mean length of 423cm.

   (a) Obtain a 99% confidence interval for the mean length of the bars now produced, assuming the standard deviation remains the same.

   (b) Comment on the concern that that the mean bar length produced by the machine is no longer the 420cm that it was prior to it being serviced.

3. A large school is studying the amount of money its teachers spend on photocopying as it prepares its budget for the next academic year. A random sample of 23 teachers gives a mean spend of £10.90 per year and a standard deviation of £2.30. Obtain a 90% confidence interval for the mean spend per teacher each year, and comment on the suggestion that the figure of £12 used previously is still accurate.

4. A sports statistician is analysing the number of goals scored in games of football. A random sample of 35 games gives a sample mean of 2.7 goals and a sample standard deviation of 0.6 goals. Obtain a 99% confidence interval for the mean number of goals scored in a game, and comment on their claim that the number of goals scored has changed from the historically observed mean of 2.5 goals per game.

5. Extensive records of the heights of mature oak trees throughout the 1970s and 1980s in some woodland show that they are normally distributed with a mean of 21 metres and a standard deviation of 1.5 metres. Recently, a random sample of 30 mature oak trees in the woodland trees yielded a sample mean of 20.4 metres.

   (a) Stating one assumption required, obtain a 90% confidence interval for the mean height of mature oak trees in the woodland now.

   (b) Comment on the suggestion that the mean height of mature oak trees in the woodland has changed.

6. A researcher is examining a statistical report on the salaries of delivery drivers in a city. She sees that a random sample of full-time delivery drivers in the city were asked their annual salaries. The report includes computer output related to a z-test on the hypothesis that the mean salaries for the delivery drivers is different to the national mean of £30 000. The report concludes that there is not sufficient evidence to suggest that the mean salary of delivery drivers in the city is different to the that seen nationally.

```
z-test for the population mean
n = 22    sample mean = 29300    sample SD = 1720
alternative hypothesis: true mean is not equal to 30000
z = -1.91    p-value = 0.0563
```

   (a) Obtain a 95% confidence interval for the mean annual salary of delivery drivers in the city.

   (b) With reference to the confidence interval, comment on whether this confidence interval supports the conclusion made by the report.

11.3 A *t*-Confidence Interval for the Population Mean

Given a random sample of data from an underlying population that is *normally distributed*, a t-confidence interval can be calculated using the *sample standard deviation*, s. The conditions for valid use of the a t-confidence interval match those for a one-sample t-test, with the degrees of freedom calculated as $\nu = n - 1$.

A t-confidence interval for the population mean:

A $(1-\alpha)\%$ CI for μ is: $\bar{x} \pm t_{n-1,1-\alpha/2} \times \dfrac{s}{\sqrt{n}}$

As with a z-confidence interval, a *t-confidence interval* can be used in a similar way to ***two-tailed*** *one sample t-test for the population mean*.

Example

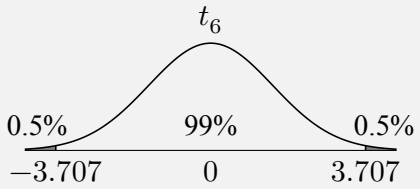
Problem: Seven mid-terrace houses in Fife are randomly chosen to have their electricity usage for 24 hours recorded. The results, in kWh, are:

7.8 10.1 8.4 9.8 8.3 9.4 7.3

Stating any assumption required, determine a 99% confidence interval for the mean daily electricity usage of mid-terrace houses in Fife, and comment on the claim that their electricity consumption is not the same as the mean of 9.7 kWh used by *semi-detached* houses in Fife.

Solution: $\bar{x} = 8.729$, $n = 7$, $s = 1.055$

Assume the underlying population of daily energy usage is normally distributed.



A 99% CI for μ is: $\quad 8.729 \pm 3.707 \times \dfrac{1.055}{\sqrt{7}} \quad = \quad 8.729 \pm 1.478$

Therefore a 99% CI for the mean daily electricity usage of mid-terrace houses in Fife, in kWh, is $(7.251, 10.207)$.

Since 9.7 kWh is within the confidence interval, there is insufficient evidence here to support the claim that mid-terrace houses and semi-detached houses in Fife use different mean amounts of electricity.

Exercise 11.3

1. As part of a project to better understand acidity levels in lakes in England, a random sample of 8 lakes are selected and the pH level of each measured. The pH levels observed were:

   $$6.7 \quad 7.1 \quad 8.3 \quad 7.4 \quad 7.6 \quad 6.8 \quad 7.0 \quad 8.1$$

   Obtain a 95% confidence interval for the mean pH level of lakes in England.

2. EnergyPower makes replacement phone batteries. The capacity of the batteries is normally distributed with a mean of 4500 mAH. A phone repair store is concerned that a recent large delivery of EnergyPower batteries may not be genuine, and measures the capacity of a random sample of 7 batteries. The results, in mAH, are:

   $$4780 \quad 4120 \quad 4340 \quad 4360 \quad 4110 \quad 4570 \quad 4390$$

   Obtain a 99% confidence interval for the mean capacity of the batteries delivered and comment on their concern that they are not genuine EnergyPower batteries.

3. The lifespan of a type of light bulb made by a brand is normally distributed. A consumer watchdog records the lifespan of a random sample of ten such bulbs and obtains the following summary statistics, in hours:

   $$n = 10 \quad \Sigma x = 9920 \quad \Sigma x^2 = 9868100$$

   Obtain a 90% confidence interval for the mean lifespan of the bulbs.

4. A camera tripod can supposedly safely hold masses of up to 4.5kg. A photography shop disputes this figure, and randomly selects 10 tripods. Each is loaded progressively with heavier masses until it fails. The results, in kg are summarised in the following computer output:

   ```
   summary(tripod mass data)
   n = 10  median = 4.05  mean = 4.05
   sum of x = 40.5  sum of x squared = 165.49
   min = 3.40  max = 4.70
   ```

   (a) State an assumption required in order to calculate a confidence interval for the mean mass a tripod can withstand before failure.

   (b) Obtain a 99% confidence interval for the mean mass the tripods can support, and comment on the claim that the 4.5kg figure stated is not correct.

5. The amount of sugar contained in frozen pizzas, per 100g, is normally distributed. A random sample of 8 brands of frozen pizza gives the following sugar content per 100g, measuring in grams:

   $$3.5 \quad 2.1 \quad 5.3 \quad 4.4 \quad 3.4 \quad 2.2 \quad 4.1 \quad 5.0$$

   (a) Obtain a 95% confidence interval for the mean sugar content of frozen pizzas, and comment on a statement on a nutrition website that the mean amount is 5 grams.

   (b) Determine the minimum sample size required for an interval of width 0.5 grams, if the sample standard deviation remains the same.

11.4 A z-Confidence Interval for the Population Proportion

In Chapter 9 a *normal approximation* to the distribution of the sample proportion, $\hat{P}$, was introduced, with a standard error of $\sqrt{\frac{pq}{n}}$, where p is the null-hypothesised *population proportion*. When calculating a z-confidence interval for the population proportion, it is necessary to instead use the **sample proportion**, $\hat{p}$, to *estimate* the standard error as $\sqrt{\frac{\hat{p}\hat{q}}{n}}$.

A z-confidence interval for the population proportion:

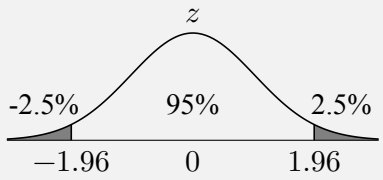
A $(1-\alpha)\%$ CI for p is: $\hat{p} \pm z_{1-\alpha/2} \times \sqrt{\frac{\hat{p}\hat{q}}{n}}$

As for a z-test for proportion, the conditions for validity are that the data is from a *random sample* and that $n\hat{p}, n\hat{q} > 5$. Since a *normal approximation to the binomial distribution* is being used (see Chapter 10, Section 10.8), it is often referred to as an *approximate confidence interval*.

Example

Problem: A tech website surveys a random selection of its subscribers and finds that 41 out of 50 of them have changed the password for their WiFi router from the default one as recommended by experts. Obtain an approximate 95% confidence interval for the proportion of its subscribers that have changed their router's password.

Solution: $\hat{p} = \frac{41}{50} = 0.82$, $\hat{q} = 0.18$, $n = 50$



A 95% CI for p is: $= 0.82 \pm 1.96 \times \sqrt{\frac{0.82 \times 0.18}{50}} = 0.82 \pm 0.060$

Therefore an approximate 95% confidence interval for the proportion of the website's readers that have changed their router's password is $(0.760, 0.880)$

Given the potential for confusion between the population proportion p and the sample proportion $\hat{p}$, as well as q and $\hat{q}$, it is recommended to check all working carefully to ensure the correct notation is being used.

Exercise 11.4

1. A newspaper conducts a survey to explore the proportion of its readers that are car owners. In a random sample of 312 readers there are 208 car owners. Obtain an approximate 99% confidence interval for the proportion of car owners amongst the newspaper's readers, justifying the method used.

2. A sample of 25 pupils in a school are asked whether they prefer cats or dogs, and only 7 prefer cats. Obtain a 90% confidence interval for the proportion of pupils in the school that prefer cats, stating an assumption required.

3. In a public opinion poll, 1000 randomly chosen electors were asked whether they would vote for the 'Purple Party' at the next election, and 357 replied 'yes'. Obtain an approximate 95% confidence interval for the true proportion of electors that intend to vote for the 'Purple Party' at the next election.

4. The proportion of left-handed people in the general population is 10%. A random sample of 400 professional artists found that 47 were left-handed. Obtain a 95% confidence interval for the proportion of professional artists that are left-handed, and comment on the claim that the proportion of left-handed artists is not the same as that seen in the general population.

5. A seed company sells pansy seeds in mixed packets and claims that 20% of the resulting plants will have red flowers. A packet of seeds is sown by a gardener who finds that only 9 out of 82 plants have red flowers. Obtain an approximate 99% confidence interval for the proportion of plants that have red flowers, and comment on whether this supports the seed company's claim.

6. A particular gene is historically thought to be present in the DNA of 4% of people. Taking advantage of the increased affordability and availability of genetic testing, a charity randomly selects 240 people to be tested for the gene. Of the 240 people in the sample, 13 are found to have the gene. Obtain an approximate 95% confidence interval for the proportion of people that have the gene, and comment on the suggestion that the 4% figure is not correct.

7. The proportion of students at a university that subscribe to a music streaming service is thought to be between 45% and 75%. A random sample of n students gives a sample proportion of 58%. Determine the minimum sample size n such that a 95% confidence interval may be argued to support a claim that a majority of students at the university subscribe to a music streaming service.

8. A guitar website hosts a forum in which its users can advertise their guitars as for sale. The website claims that one in every three of the guitars listed on the forum are vintage instruments from the 1950s, 1960s and 1970s. A forum user selects at random one of the ten most recently listed guitars, then every tenth guitar after that until they have looked back over the last three months of for-sale guitars. The decade of manufacture of every guitar in the sample is recorded. The results are:

1950s	1960s	1970s	1980s	1990s	2000s	2010s	2020s
7	14	17	9	7	15	31	7

   (a) State the type of sampling used by the forum user.

   (b) Obtain an approximate 95% confidence interval for the proportion of guitars for sale that are vintage, manufactured in the 1950s, 1960s and 1970s.

   (c) Comment on whether the confidence interval supports the website's claim.

Review Exercise

1. A road safety researcher randomly selects eight cars passing by on a motorway and records their speed, in mph. The following results are obtained:

   63 74 57 84 80 68 70 91

   Obtain a 95% confidence interval for the mean speed of the cars passing by on the motorway, stating an assumption required.

2. A random sample of 50 residents of a town are asked whether or not they are aware that a proposal for a large retail park on the outskirts of the town is currently being considered by the local council. 14 of the 50 residents were aware of the proposals, whilst the rest were not. Obtain an approximate 99% confidence interval for the proportion of the town's residents that are aware of the proposals.

3. The volume of wine served when a restaurant's customer orders a small glass of wine is normally distributed with a standard deviation of 1.7ml. A random sample of 32 small glasses of wine served by the restaurant gives a mean of 126.3ml per glass.

   (a) Obtain a 90% confidence interval for the mean volume of wine served when a small glass of wine is ordered by a customer in the restaurant, stating an assumption required.

   (b) Comment on the suggestion that the mean volume served is not the correct measure of 125ml, stated on the restaurant's wine list.

4. A pharmaceutical company has developed a drug designed to reduce the recovery time needed for a specific type of operation. Following trials involving treating 25 randomly selected patients admitted to hospital for the operation with the drug, it is found that the mean recovery time for these patients is 11.4 days. Computer output summarising this data, measured in days, is as follows:

   ```
   summary(recovery time in days)
   n = 25  mean = 11.4
   A 95% z-confidence interval for the mean is:
   (10.3 , ******)
   ```

   (a) State the value from the computer output that has been replaced by ******.

   (b) The sample standard deviation was considered a reliable estimator for the population standard deviation recovery time, since the sample was sufficiently large. Determine the value of the sample standard deviation used to calculate the confidence interval.

5. A driving instructor posts an advert online stating nine out of ten of their students pass their driving test first time. A random sample of 60 students taught by the instructor reveals that only 47 of them passed their driving test first time. Obtain an approximate 95% confidence interval for the proportion of the instructor's students that pass first time, and comment on the claim that the true proportion is not the nine out of ten claimed by the instructor.

12

Chi-Squared Goodness-Of-Fit Test

The **chi-squared distribution**, χ^2_ν, describes the distribution of *sums* of *scaled, squared random variables*. χ^2 hypothesis tests offer a way to analyse categorical data, and are a type of **non-parametric** test.

12.1 χ^2 Goodness-of-Fit Test

A *goodness-of-fit* test assesses how well observed data, which can be broken down into categories, fits a hypothesised model. In this course, the data examined will be either *categorical* or, as in the following example, *discrete*. Suppose a statistics teacher's cubical die is suspected of being not *fair* and their students roll the die 48 times to check for any bias. The number of times each outcome occurs, or the **observed frequencies**, are recorded in a frequency table:

Outcome	1	2	3	4	5	6
Observed Frequencies (O_i)	7	8	5	10	7	11

With the null hypothesis being that the die *is* fair, a third row of **expected frequencies** can be added to the table by dividing the total count of 48 *observations* by the 6 equally-likely outcomes:

Outcome	1	2	3	4	5	6
Observed Frequencies (O_i)	7	8	5	10	7	11
Expected Frequencies (E_i)	8	8	8	8	8	8

The χ^2 goodness-of-fit test asks whether the **observed values** are so *significantly* different than the **expected values**, under the model described in the null hypothesis, that there is evidence to suggest *the model is in fact incorrect*.

H_0 : *The die is fair / the data follows a* $\mathrm{U}(6)$ *distribution*
H_1 : *The die is not fair / the data is distributed differently.*

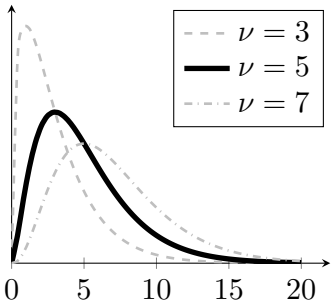
The theory underpinning the test statistic used for a goodness-of-fit test is beyond the scope of the Advanced Higher Statistics course. However, a glimpse into the rationale may be found by noting that the test statistic, on the next page, takes the *sum* of *squared differences* between the observed and expected frequencies, *scaled* by the expected frequencies.

For a χ^2 goodness-of-fit test:

Test statistic: $X^2 = \dfrac{(O_1 - E_1)^2}{E_1} + \dfrac{(O_2 - E_2)^2}{E_2} + ... + \dfrac{(O_n - E_n)^2}{E_n} = \sum \dfrac{(O_i - E_i)^2}{E_i}$

The test statistic follows a χ^2_ν distribution with degrees of freedom $\nu = c - 1 - m$, where:

- c is the number of *categories*.
- m is the number of *model parameters* estimated using the observed data.



χ^2_ν distributions with ν degrees of freedom

The decision rule for χ^2 tests:

If **test statistic** $>$ **critical value**, *reject* H_0.

If **test statistic** $<$ **critical value**, *do not reject* H_0.

For the *"fair"* die, the test statistic can now be calculated:

$$X^2 = \frac{(7-8)^2}{8} + \frac{(8-8)^2}{8} + \frac{(5-8)^2}{8} + \frac{(10-8)^2}{8} + \frac{(7-8)^2}{8} + \frac{(11-8)^2}{8} = 3$$

There are six categories, so $c = 6$, and no parameters of the model were calculated using the data, so $m = 0$. The degrees of freedom are therefore $\nu = 6 - 1 - 0 = 5$, and the test statistic can be compared against the χ^2 critical values from **page 14** of the data booklet. With $\nu = 5$ and taking $\alpha = 0.05$, the critical value of $\chi^2_{5,0.95} = 11.070$ can be found.

The smallest possible observed test statistic is 0, representing data that perfectly matches the model, and for $\nu = 5$ less than 5% of test statistics exceed 11.070 by random chance. Note that chi-squared tests are always *one-tailed*.

Since $3 < 11.070$ the students should *not* reject the null hypothesis at the 5% level of significance.

There is *insufficient evidence* to suggest that the model is not correct, and that the die is not fair.

Conditions for valid use of a goodness-of-fit test:

- No **expected** values should be *less than 1*.
- At least 80% of the **expected** values should be *at least 5*.

The data for teacher's die met the criteria for the valid use of a goodness-of-fit test, which is given on **page 6 of the data booklet**. Both conditions need to be verified before test statistics, degrees of freedom and critical values can be calculated.

Example

Problem: A chocolatier says that their boxes are randomly filled with caramel, coffee and strawberry flavoured chocolates in the ratio $2 : 2 : 1$. Suspecting that this is not true, a chocolate-fan counts the number of chocolates of each flavour in a box and finds 8 are caramel, 13 are coffee and 11 are strawberry. Perform a non-parametric hypothesis test to assess the evidence at the 10% significance level that the flavours are not distributed as claimed by the chocolatier.

Solution:

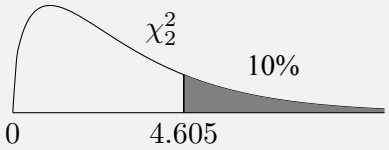
H_0 : the chocolates are distributed in the ratio 2:2:1
H_1 : the chocolates are distributed differently
] $\alpha = 0.1$, One-tailed

Flavour	caramel	chocolate	strawberry
Observed Frequencies (O_i)	8	13	11
Expected Frequencies (E_i)	12.8	12.8	6.4

$\nu = 3 - 1 - 0 = 2$

Test statistic: $X^2 = \frac{(8-12.8)^2}{12.8} + \frac{(13-12.8)^2}{12.8} + \frac{(11-6.4)^2}{6.4} = 5.109$

Critical value: $\chi^2_{2,0.9} = 4.605$



Since $5.109 > 4.605$, reject H_0 at the 10% level of significance.
There is evidence to suggest that the flavours are not distributed in the ratio stated by the chocolatier.

Exercise 12.1

1. A fair, tetrahedral die with faces numbered 1 to 4 is rolled 24 times. The results obtained are:

Outcome	1	2	3	4
Observed Frequencies	6	1	8	9

   Perform a goodness-of-fit test at the 5% level of significance to assess whether there is evidence to support a claim that the die is biased.

2. A question in an old textbook states that a person is equally likely to be born on any day of the week. A student asks a random sample of 49 people which day of the week they were born on, and records the results.

Day	Mon	Tue	Wed	Thurs	Fri	Sat	Sun
Frequency	11	9	4	9	7	3	6

   Perform a non-parametric test at the 10% significance level to test whether the data supports the textbook's statement.

3. A local council has previously monitored the numbers of cars entering a city, classifying each as either internal combustion engine (ICE), hybrid or fully electric. The ratio of cars in each category was found to be to be 5:2:1 respectively. Interested in whether this has since changed, a random sample is taken of 80 cars entering the city, with the results: 43 ICE, 16 hybrid and the remaining cars fully electric. Perform a hypothesis test at the 1% level of significance to test the claim the the ratio has changed from what it was previously.

12.2 Combining Columns

If the expected categories do not meet the conditions stated on Page 151 then two or more categories should be *"combined"* (or *"collapsed"*) until the conditions are satisfied. The degrees of freedom for the test should be calculated based on the *final combined frequency table* rather than the original table. Whilst it is recommended to combine those categories with the smallest expected values, it may at times be more intuitive to instead collapse those categories that more naturally can be grouped.

Example 1

Problem: An video game allows players to buy in-game "loot boxes" with real money, each containing one in-game item. The game's publisher states that 60% of loot boxes contain a "Regular" item, 25% contain a "Premium" item, 10% contain a "Rare' item and the remaining 5% contain an "Ultra Rare" item. A player records the types of item contained in each of a series of loot boxes and records the results in the frequency table below.

Type	Regular	Premium	Rare	Ultra Rare
Observed Frequencies (O_i)	33	16	1	0

Perform a χ^2 goodness-of-fit test to determine whether there is evidence at the 1% significance level to suggest the publisher's claim is not correct.

Solution:

$\left.\begin{array}{ll} \mathrm{H}_0: & \text{the items types are distributed as claimed} \\ \mathrm{H}_1: & \text{they are distributed differently} \end{array}\right] \begin{array}{l} \alpha = 0.01 \\ \text{One-tailed} \end{array}$

Type	Regular	Premium	Rare	Ultra Rare
Observed Frequencies (O_i)	33	16	1	0
Expected Frequencies (E_i)	30	12.5	5	2.5

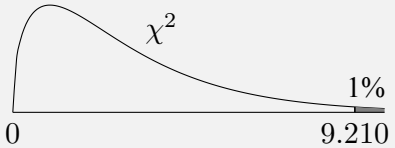
*(Since only 75% of **expected** values are 5 or greater, which is less than the required 80%, the last two columns must be combined.)*

Type	Regular	Premium	Rare/Ultra Rare
Observed Frequencies (O_i)	33	16	1
Expected Frequencies (E_i)	30	12.5	7.5

$\nu = 3 - 1 - 0 = 2$

Test statistic: $X^2 = \frac{(33-30)^2}{30} + \frac{(16-12.5)^2}{12.5} + \frac{(1-7.5)^2}{7.5} = 6.913$

Critical value: $\chi^2_{2,0.99} = 9.210$



Since $6.913 < 9.210$, do not reject H_0 at the 1% level of significance.
There is insufficient evidence to suggest that the item types are not distributed as stated by the game's publisher.

Exercise 12.2

1. Historically, grades attained (*A*, *B*, *C*, *D* and *No Award*) by pupils in a school course's end of year exam have been distributed following the ratio $4 : 5 : 3 : 2 : 1$. Following a change in the teaching resources used, the results of 40 students are recorded as follows:

Grade	*A*	*B*	*C*	*D*	*No Award*
Frequency	13	19	6	0	2

   Perform a chi-squared test to assess, at the 10% level of significance, whether the data provides evidence to suggest the distribution of grades attained has changed.

2. The proportion of people in the USA with each eye colour is as follows:

Eye Colour	*Brown*	*Blue*	*Hazel*	*Green*	*Other*
Proportion	0.45	0.27	0.18	0.09	0.01

   A random sample of 58 people living in San Francisco have their eye colour recorded. The results are shown in the table below:

Eye Colour	*Brown*	*Blue*	*Hazel*	*Green*	*Other*
Frequency	29	13	10	2	4

   Perform a chi-squared goodness-of-fit test to assess at the 1% significance level whether there is evidence of a difference in distribution of eye colours in San Francisco.

3. A coffee shop sells five of types of coffee. Historically, the five types of coffee have been sold in the following proportions:

   - 34% cappuccinos
   - 27% lattes
   - 18% espressos
   - 11% flat whites
   - 10% americanos

   Suspecting that these proportions are may have changed, the coffee shop records the type of coffee sold for a random sample of 40 individual orders:

Cappuccinos	Lattes	Espressos	Flat whites	Americanos
7	8	7	9	9

   Perform a non-parametric test to assess at the 5% level of significance the claim that the historical proportions of each type of coffee sold are no longer accurate.

12.3 Goodness-of-fit and Binomial Distributions

A goodness-of-fit test can also be performed to assess whether observed data fits a *binomial* distribution, in which the **total observed frequencies** will be split in proportion to the *calculated binomial probabilities*.

Example

Problem: A supermarket randomly samples boxes of duck eggs delivered by a supplier, each containing four eggs, and counts how many of the four eggs are broken for each box. The results of this are:

Eggs broken	0	1	2	3	4
Observed Frequencies (O_i)	122	10	0	4	8

To better predict the occurrences of damaged eggs, the supermarket decides that the number of eggs broken in each box of four follows a binomial distribution, with $p = 0.1$. Perform a hypothesis test to assess this claim and comment on the findings with reference to the assumptions required for a binomial model.

Solution:

$\left.\begin{array}{ll} \text{H}_0: & \text{No. of eggs broken per box follows a B}(4, 0.1)\text{ distribution.} \\ \text{H}_1: & \text{No. of eggs broken per box follows a different distribution.} \end{array}\right]$ $\alpha = 0.05$, One-tailed

Total observations $= \mathbf{144}$. Letting random variable X represent the number of eggs broken in a box:

$$\text{P}(X = 0) = 0.6561 \qquad \text{P}(X = 1) = 0.2916 \qquad \text{P}(X = 2) = 0.0486 \qquad \cdots$$

$$0.6561 \times \mathbf{144} = 94.48 \qquad 0.2916 \times \mathbf{144} = 41.99 \qquad 0.0486 \times \mathbf{144} = 7.00 \qquad \cdots$$

Eggs broken	0	1	2	3	4
Observed Frequencies (O_i)	122	10	0	4	8
Expected Frequencies (E_i)	94.48	41.99	7.00	0.52	0.01

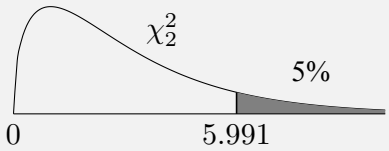
(Combining the final three columns is necessary here to satisfy the conditions for expected frequencies.)

Eggs broken	0	1	2 or more
Observed Frequencies (O_i)	122	10	12
Expected Frequencies (E_i)	94.48	41.99	7.53

$\nu = 3 - 1 - 0 = 2$

Test statistic: $X^2 = \frac{(122-94.48)^2}{94.48} + \frac{(10-41.99)^2}{41.99} + \frac{(12-7.53)^2}{7.53} = 35.041$

Critical value: $\chi^2_{2,0.95} = 5.991$



Since $35.041 > 5.991$, reject H_0 at the 5% level of significance.
There is evidence to suggest that the number of eggs broken per box does not follow a $\text{B}(4, 0.1)$ distribution.

A condition for a binomial model is that each trial is independent. Since broken eggs most likely occur together, this condition may not be satisfied and a binomial model is unlikely to be suitable.

Note that estimating the parameter p from *the data itself* would require an additional degree of freedom to be subtracted (since $m = 1$). However this is rarely assessed in this course.

Exercise 12.3

1. A manufacturer of festive products sells Christmas crackers, in boxes of four. Since 20% of all crackers they produce are known to be faulty, failing to produce the required "crack" when pulled, they wish to test whether the number of faulty crackers in a box could be modelled using a binomial distribution.

   (a) State the distribution of X, representing the number of faulty crackers in a box.

   (b) Calculate:

   i. $\mathrm{P}(X = 0)$ ii. $\mathrm{P}(X = 1)$ iii. $\mathrm{P}(X = 2)$ iv. $\mathrm{P}(X = 3)$ v. $\mathrm{P}(X = 4)$

   The manufacturer tests 60 boxes of crackers, and records the number of faulty crackets in each box, with the results shown in the table below:

No of Faulty Crackers	0	1	2	3	4
Frequency	34	17	4	2	3

   (c) Use a suitable hypothesis test to assess at the 10% level of significance whether the use of a binomial distribution is supported by the data.

2. A school's mathematics department runs a short multiple choice festive revision quiz, which requires students to choose the correct answer from one of four options. The quiz contains five questions in total.

   Teachers in the department suspect that, instead of carefully working using their knowledge to work out the answers to the mathematical questions, the students are simply guessing each question at random.

   (a) State the distribution of X, the number of correct answers out of 5, assuming that a student is guessing at random for every question.

   To investigate these, the number of correct answers out of five for each of the 80 students is first set out in a table:

No. of Correct Answers	0	1	2	3	4	5
Frequency	9	1	29	25	10	6

   (b) Perform a chi-squared test to determine if the teachers' suspicious are backed by the data.

3. A car mechanic has established that the proportion of tyres with tread that is below the legal minimum, for cars coming in for a service, is 15%. They are interested in whether the number of the four tyres on a car brought in for a service that have less than the legal minimum tread, and thus need to be replaced, may be modelled using a binomial distribution: $X \sim \mathrm{B}(4, 0.15)$.

   The number of tyres needing replaced for low tread for a random sample of cars brought in for a service are recorded:

No. of Tyres	0	1	2	3	4
Frequency	41	20	8	1	2

   (a) Perform a chi-squared test to assess whether the binomial distribution may be suitable, at the 5% significance level.

   (b) If the proportion of faulty tyres had not been known, and an estimate had instead been obtained using the data, explain how this would affect the calculation of the degrees of freedom,ν , for the test.

12.4 Goodness-of-fit and Poisson Distributions

Just as with a binomial distribution, using a Poisson distribution as a model for a goodness-of-fit test requires the total observations to be split according to calculated probabilities. Should the only parameter for a Poisson distribution, λ, be required to be *estimated from the data*, an additional degree of freedom will be *subtracted*, since $m = 1$.

Example

Problem: It is generally said that the number of goals scored in a game of football follows a Poisson distribution. A football fan decides to see whether this statement is still reasonable, and records the number of goals scored in every game across a weekend in each of the top four tiers of English football:

Goals	0	1	2	3	4	5	6	7 or more
Observed Frequencies (O_i)	4	8	13	9	8	2	2	0

Perform a non-parametric tests at the 1% level of significance to assess whether there is evidence to suggest the model is longer valid.

Solution: $\lambda = \frac{\text{total no. of goals}}{\text{total no. of games}} = \frac{115}{46} = 2.5$

$\left.\begin{array}{ll} \text{H}_0: & \text{No. of goals per game follows a Poisson distribution (with } \lambda = 2.5) \\ \text{H}_1: & \text{No. of goals per game follows a different distribution.} \end{array}\right]$ $\alpha = 0.01$, One-tailed

Total observations$= 46$.
P$(X = 0) = 0.0821$
Expected frequency for 0 goals $= 0.0821 \times 46 = 3.777$

Goals	0	1	2	3	4	5	6	7 or more
Observed Frequencies (O_i)	4	8	13	9	8	2	2	0
Expected Frequencies (E_i)	3.777	9.439	11.799	9.835	6.146	3.073	1.279	0.652

(Combining the final three columns is necessary here to satisfy the conditions for expected frequencies.)

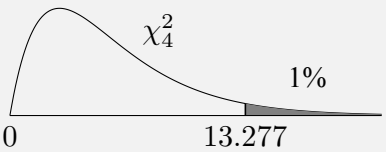
Goals	0	1	2	3	4	5 or more
Observed Frequencies (O_i)	4	8	13	9	8	4
Expected Frequencies (E_i)	3.777	9.439	11.799	9.835	6.146	5.004

(Since λ was calculated using the data, subtract an additional degree of freedom.)

$\nu = 6 - 1 - 1 = 4$

Test statistic: $X^2 = \frac{(4-3.777)^2}{3.777} + ... + \frac{(4-5.004)^2}{5.004} = 1.186$

Critical value: $\chi^2_{4,0.99} = 13.277$



Since $1.186 < 13.277$, do not reject H$_0$ at the 1% level of significance.
There is insufficient evidence to suggest that the number of goals scored per football game does not follow a Poisson distribution, so the model still seems reasonable.

Note that whilst the χ^2 test statistic is a measure of how well the data fits a model, the goodness-of-fit test does not show evidence *for* the model - only that the data does or does not *contradict* it.

Exercise 12.4

1. A keen amateur astronomer recalls hearing during a statistics class that shooting stars can be seen in a local section of night sky at a mean rate of 1 per hour, and the number visible in an hour is Poisson distributed. Over the course of a year they carefully watch the night sky on clear nights, and record the number of shooting starts they see within each hour. The data they obtain for the total of 70 hours spent watching for shooting stars is shown in the table below:

Number of Shooting Stars	0	1	2	3 or more
Frequency	23	20	15	12

   (a) If the number of shooting stars visible in a section of sky in an hour is distributed as they were told in the statistics class, calculate the probability of seeing, in any given hour:

      i. No shooting stars.

      ii. Exactly one shooting star.

      iii. Exactly two shooting stars.

      iv. Three or more shooting stars.

   (b) Use a non-parametric test to determine whether there is evidence at the 5% significance level that the distribution stated during the statistics class is not correct.

2. A restaurant had generally understood that the number of bottles of wine sold on any given night could be modelled by a Poisson distribution with parameter λ. Suspecting this is no longer the case, the restaurant recorded the number of bottles sold on 50 randomly selected nights. The table below shows the data obtained:

Number of Bottles	0	1	2	3	4	5 or more
Frequency	2	13	19	15	1	0

   (a) Calculate a suitable estimate for the value of λ.

   (b) Use a hypothesis test to show that there is evidence at the 1% level of significance to suggest that a Poisson distribution is no longer an appropriate model for the number of bottles of wine sold.

3. It is often stated that the number of goals scored in a game of football follows a Poisson distribution with a parameter of 2.5. To explore this, a sports statistician obtains the relevant data for all 380 games of the 2023-24 English Premier League season:

Number of Goals	0	1	2	3	4	5	6	7	8	9 or more
Frequency	11	42	81	80	81	48	24	10	3	0

   (a) Perform a chi-squared test to assess whether the data suggests a Poisson model with a parameter of 2.5 is suitable.

   It is subsequently claimed that a Poisson model is still appropriate, but that the value of the parameter has changed.

   (b) Calculate an estimated value for λ using the data, to 2 decimal places.

Review Exercise

1. The population of the United Kingdom can be approximately split between those living in England, Scotland, Wales and Northern Ireland in the ratio $30 : 3 : 2 : 1$. Out of a random sample of 117 people who applied to take part in a UK-wide television show, 102 lived in England, 4 lived in Scotland, and 1 lived in Northern Ireland.

   Perform a non-parametric test to assess whether there is evidence to suggest the distribution of people applying for the TV show differs from that of the wider UK population, at the 5% level of significance.

2. A football team practises for the possibility of a penalty shootout each training session by having their five chosen penalty takers take one penalty each. They know that 15% of all penalties taken as part of this training will result in a failure to score. Over the course of a season, the number of *failed attempts* for each of the 60 sets of five penalties is recorded:

Number of Failed Attempts	0	1	2	3	4	5
Frequency	33	12	7	5	3	0

   (a) Perform a test at the 10% significance level to assess the evidence that the data fits a binomial distribution where $p = 0.15$.

   (b) A coach comments that some of the five penalty takes are clearly better at taking penalties than others. Explain why this would mean that a binomial would be inappropriate.

3. A supermarket chain is monitoring concerns regarding shoplifting in one of its stores. It is proposed that a Poisson distribution may be used to model the number of incidents of shoplifting recorded in any given day.

   The table below shows *observed* frequencies of the numbers of incidents recorded per day across a period of time, with some *expected* frequencies calculated using a Poisson distribution estimating its parameter from the observed data.

Incidents	0	1	2	3	4	5	6 or more
Observed	19	37	27	10	6	1	0
Expected	22.31	33.47		12.55	4.71	1.41	

   (a) Calculate the mean of the Poisson distribution used.

   (b) Determine the missing expected values.

   (c) Carry out a non-parametric test at the 1% significance level to determine whether a Poisson distribution is an appropriate model for the data.

   It is suggested that the same mean number of incidents of shoplifting per day obtained in part (a) was already known to be the appropriate value to use for the Poisson distribution.

   (d) Had this been known before the test was conducted, state what the test statistic and critical value would have been.

Exercise 12.1

1. $6.333 < 7.815$, do not reject H_0
2. $7.143 < 10.645$, do not reject H_0
3. $13.88 > 9.210$, reject H_0

Exercise 12.2

1. $7.779 < 9.488$, do not reject H_0
2. $3.049 < 11.345$, do not reject H_0
3. $14.906 > 7.815$, reject H_0

Exercise 12.3

1. (a) $X \sim B(4, 0.2)$
   (b) i. 0.4096
   ii. 0.4096
   iii. 0.1536
   iv. 0.0256
   v. 0.0016
   (c) $6.266 > 4.605$, reject H_0
2. (a) $X \sim B(5, 0.25)$
   (b) $257.862 > 9.488$, reject H_0 at the 5% level
3. (a) $3,144 < 5.991$, do not reject H_0
   (b) An additional 1 must be subtracted from the number of categories to obtain the degrees of freedom.

Exercise 12.4

1. (a) i. 0.3679
   ii. 0.3679
   iii. 0.1839
   iv. 0.0803
   (b) $9.173 > 7.815$, reject H_0
2. (a) $\lambda = 2$
   (b) $14.839 > 11.345$, reject H_0
3. (a) $98.443 > 14.067$, reject H_0 at the 5% level
   (b) $\lambda = 3.28$

Review Exercise

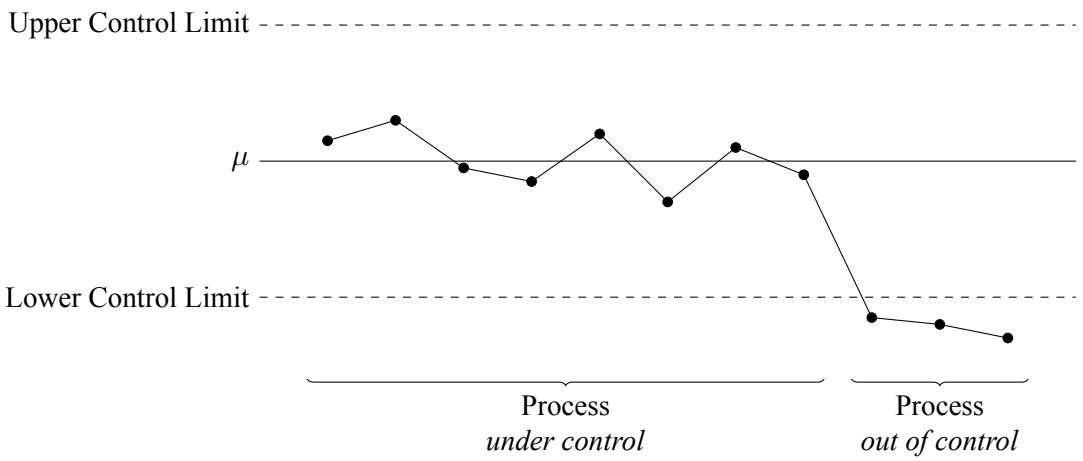
1. $3.759 < 5.991$, do not reject H_0
2. (a) $9.790 > 4.605$, reject H_0
   (b) Some players being better penalty takers than others would mean the probability of success for each trial (penalty attempt) would not be equal, and hence a binomial distribution would not be appropriate.
3. (a) $\lambda = 1.5$
   (b) $25.10, 0.45$
   (c) $1.553 > 11.345$, reject H_0
   (d) The test statistic would still be 1.553, whilst the critical value would instead be 13.277

13

Control Charts

In any manufacturing process, performance must be monitored to ensure that the products being manufactured are not faulty or that the process is not flawed in some way. A common tool used to monitor performance is a *control chart*, in which a series of samples are taken, measured and plotted against calculated limits to observe for signs of a problem.

A simplified control chart:



Variation naturally occurs as a result of random causes in any manufacturing process, referred to as *common cause variation*, and often there is no need for this to be investigated or addressed. This could be, for example, because the variation is within tolerable limits, or the cost of reducing the variation would outweigh the potential benefit of doing so. A process performing in this manner could be described as *under control*.

However, if something occurs within the manufacturing process that results in a significant change in the output, a system, such as a control chart, needs to be in place to raise the alarm that the process may be *out of control*. The process can then be reviewed and any possible *special cause variation* investigated, such as machine wear and tear, poor raw materials, poorly trained operatives, or an accident or malfunction.

The Principle of a Control Chart

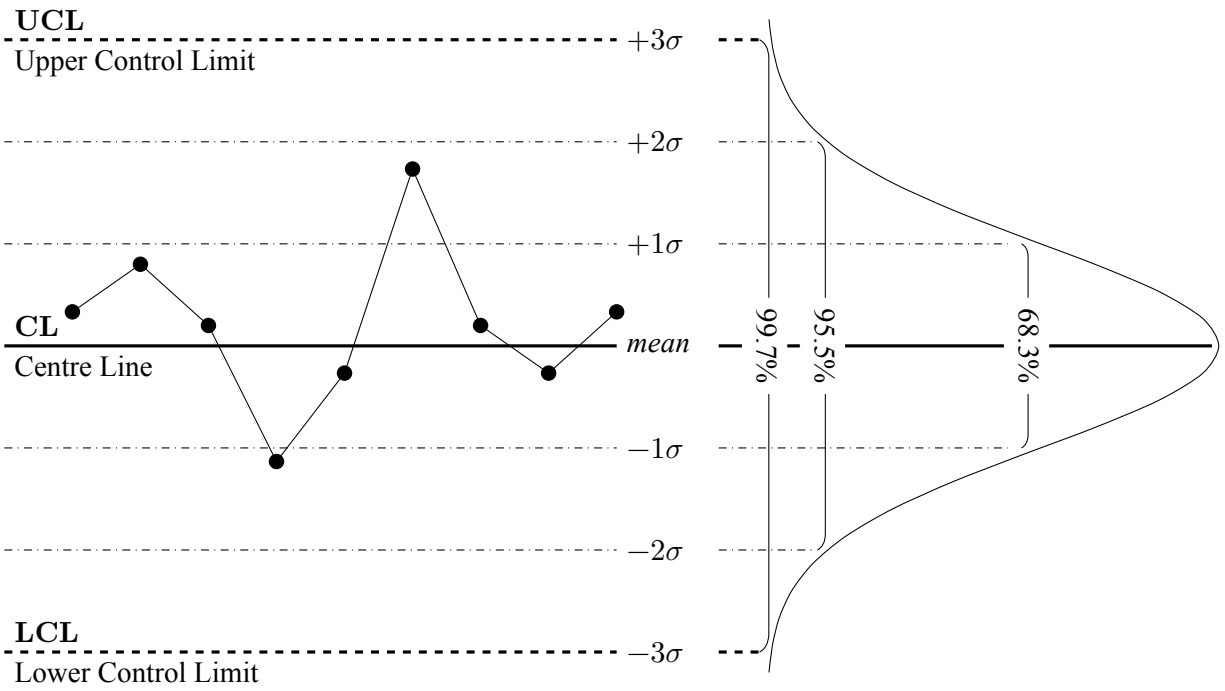
The control chart was invented in the 1920's by Walter Shewart, who was working for a communications company whose engineers wanted to improve the reliability of their telephone transmission systems. His theory was based around the normal distribution bell curve.

The Advanced Higher Statistics course covers two types of control charts:

- A Shewart x-bar chart for the sample **mean**.
- A Shewart p chart for the sample **proportion**.

In either case, random samples of the product are taken at regular intervals and the sample means or sample proportions are plotted onto the control chart. A line on the chart is drawn to indicate the target value for the population mean or population proportion, as well as 1-sigma , 2-sigma and 3-sigma lines either side of the target value.

The 3-sigma (3σ) limits are often referred to as the *upper* and *lower control limits*, with the other limits called *warning lines*.



The Western Electric Company Rules (*WECO Rules*) are commonly used to determine whether the production process is *in or out of statistical control*, and are provided on Page 4 of the SQA booklet of Formulae and Tables.

A process maybe out of statistical control when:

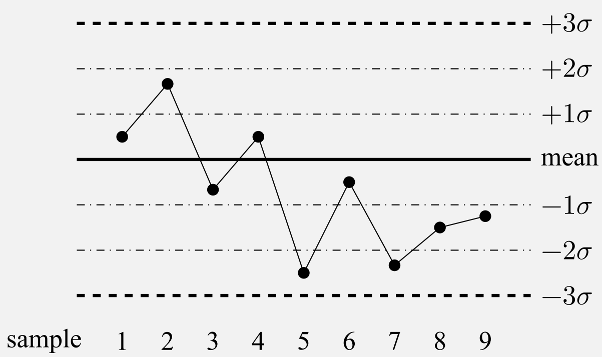
- *Any* single data point falls outside a 3σ limit.
- *Two out of three* consecutive points fall beyond the *same* 2σ limit.
- *Four out of five* consecutive points fall beyond the *same* 1σ limit.
- *Eight consecutive* points fall on the *same side of the centre line*.

13.1 Applying the WECO Rules

Each rule must be carefully checked for to determine whether a process is *in statistical control*. If it is *not* in statistical control, it is often necessary to identify the relevant WECO rule and at which point this could be declared. The following examples focus on context-free control charts for the purpose of practising *applying the WECO rules*.

Example

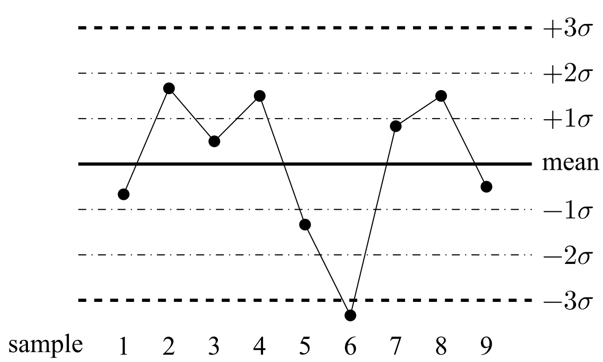
Problem: For the control chart below, determine whether the process is in control. If it is out of control, identify at which sample this could be determined, with justification.

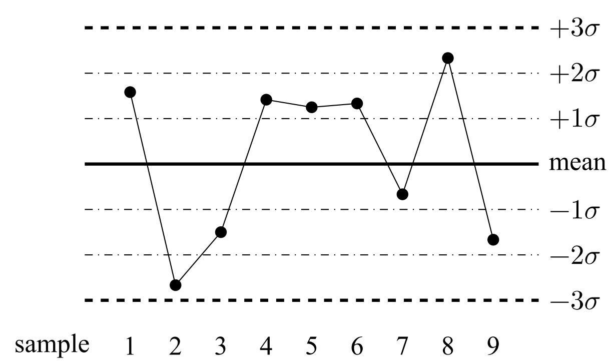


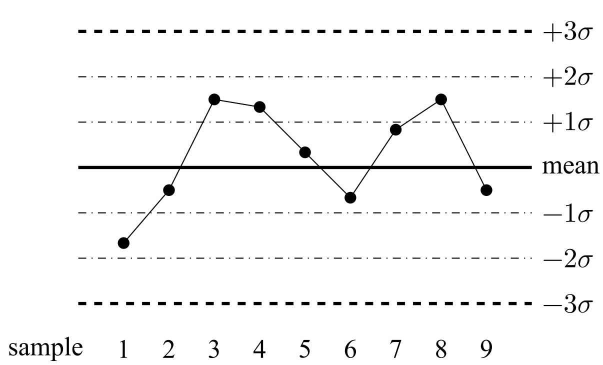
Solution: The process is out of control, as two out of the three values from samples 5 to 7 lie below the same 2σ limit (-2σ). The process could have been identified as out of control at sample 7.

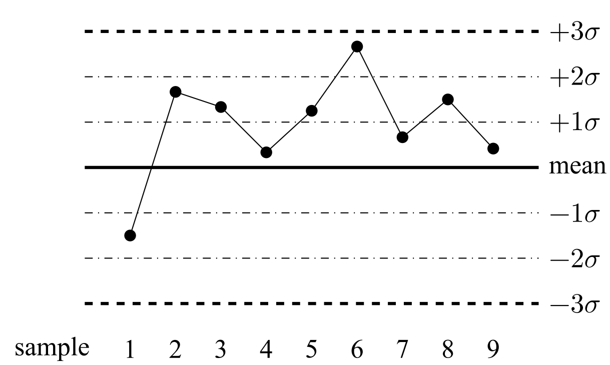
Exercise 13.1

For each control charts below, determine whether the process is in control. If it is out of control, identify at which sample this could be determined, with justification.









13.2 Shewart *x*-bar Charts with a Known Population Mean

Suppose that a manufacturing process is working properly and its performance is normally distributed, with mean μ and standard deviation σ. If samples of size n are taken regularly from the product, and the mean obtained for each sample, then:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Sigma limits can be calculated as:

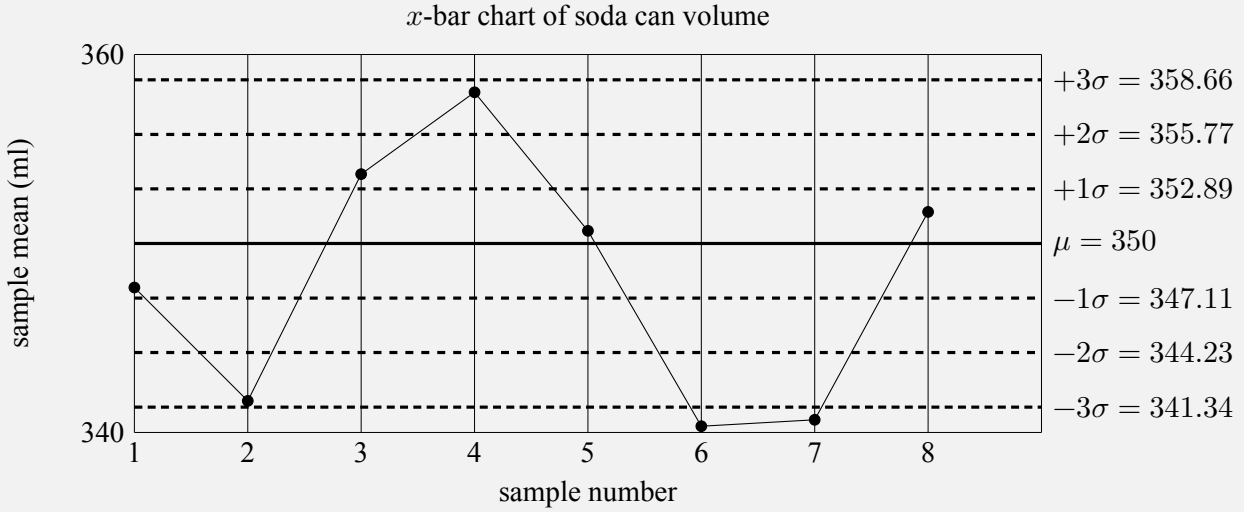
$$\underbrace{1\sigma \text{ limits: } \mu \pm \frac{\sigma}{\sqrt{n}} \qquad 2\sigma \text{ limits: } \mu \pm \frac{2\sigma}{\sqrt{n}}}_{\textit{Warning lines}} \qquad \underbrace{3\sigma \text{ limits: } \mu \pm \frac{3\sigma}{\sqrt{n}}}_{\textit{UCL and LCL}}$$

A control chart is created with horizontal lines indicating the population mean and each of the 6 sigma limits calculated. The sample mean for each of the samples drawn at regular intervals is plotted on the chart, and the *WECO rules* applied to determine if or when the production process is out of control.

Example

Problem: A machine fills cans of soda with mean volume 350ml and standard deviation 5ml. Every hour a sample of 3 cans is taken, their volumes measured, and the sample mean calculated before being added to an x-bar chart.

Sample number	1	2	3	4	5	6	7	8
Sample mean	347.67	341.67	353.67	358	350.67	340.33	340.67	351.67



(a) Confirm the 2-sigma limits on the control chart.

(b) Determine when the process is out of control.

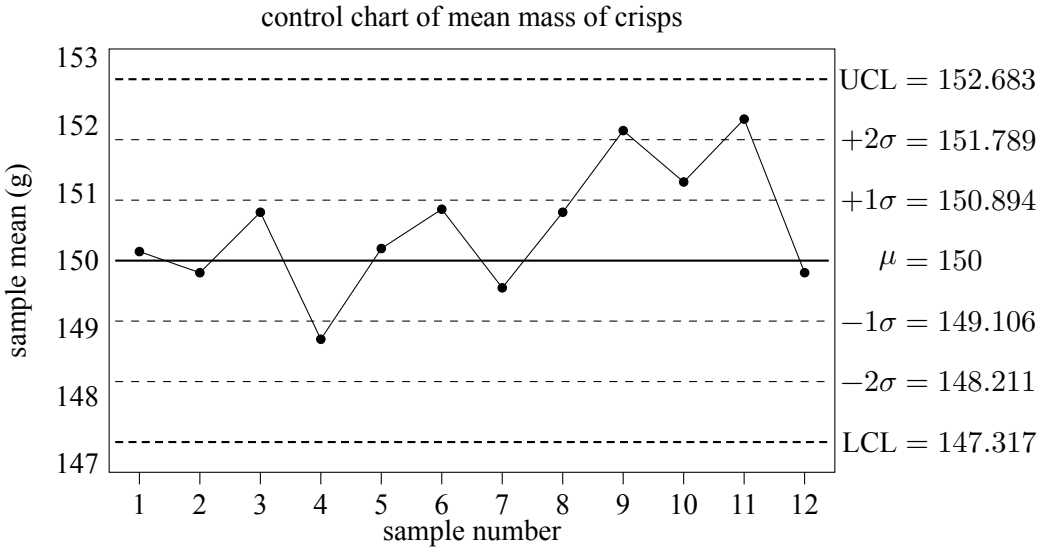
Solution:

(a) 2σ limits: $\mu \pm \frac{2\sigma}{\sqrt{n}} = 350 \pm \frac{2 \times 5}{\sqrt{3}} = 350 \pm 5.77$, so 344.23 and 355.77.

(b) The process is out of control for samples 6 and 7 as the sample means both lie below the lower control limit (-3σ limit). The process needs to be investigated.

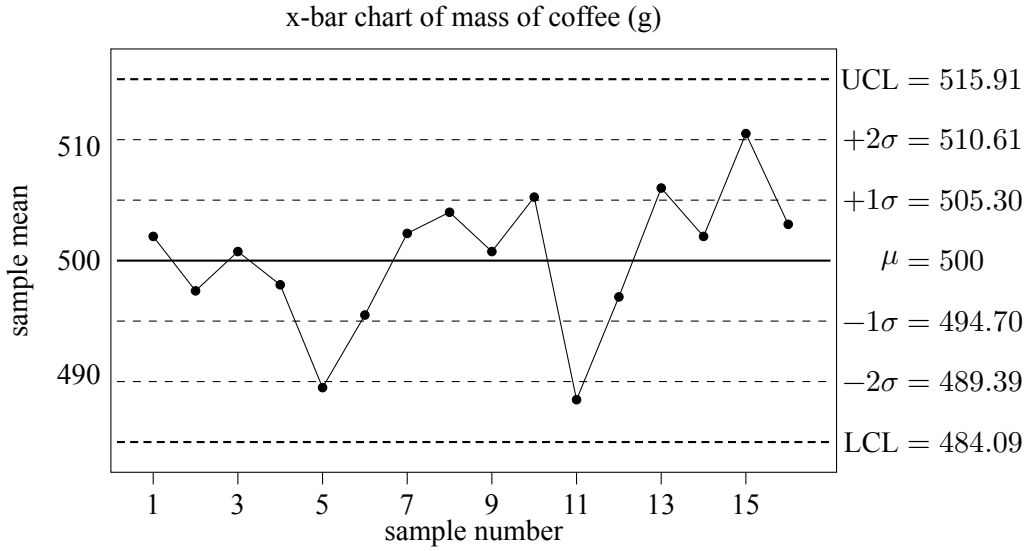
Exercise 13.2

1. A machine dispenses crisps into sharing-sized bags with a mean mass of 150g and a standard deviation of 2g. It is known that the mass of the crisps is normally distributed. A random sample of 5 of the sharing-sized bags of crisps is taken every hour to check that the machine is in statistical control. A control chart showing 12 consecutive samples from a particular shift is shown.



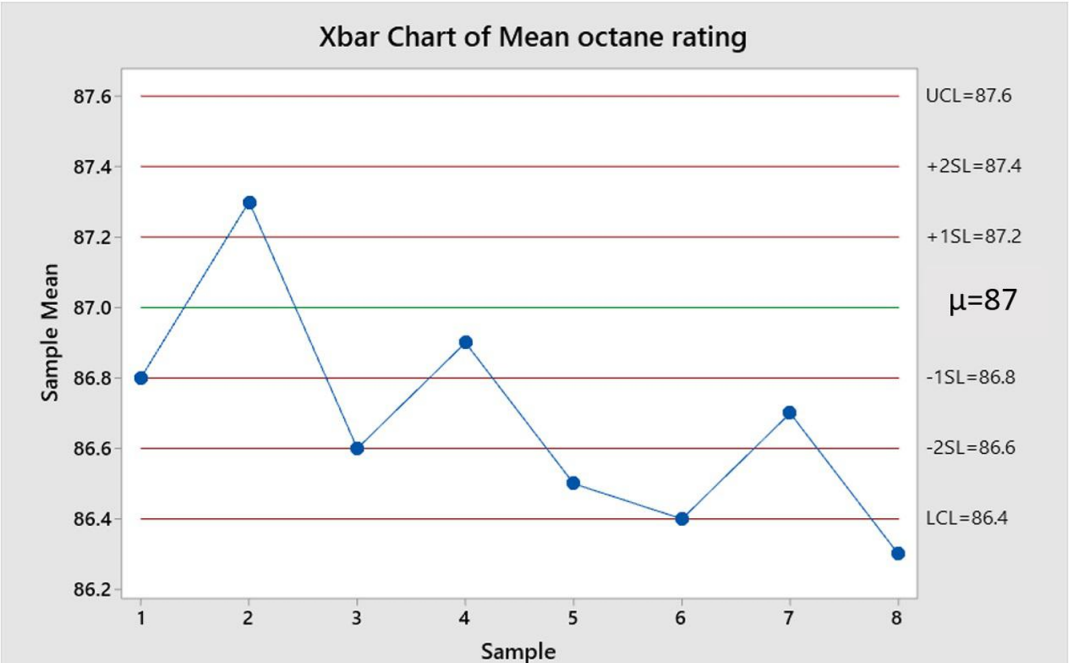
Confirm the 2σ limits shown on the chart and determine if the machine was in control for the whole shift.

2. A packing process packages ground coffee into bags. The mass of each bag is normally distributed with mean 500g and standard deviation 15g. Random samples of 8 coffee bags are taken at regular intervals and weighed. The control chart shown gives 16 consecutive sample means.



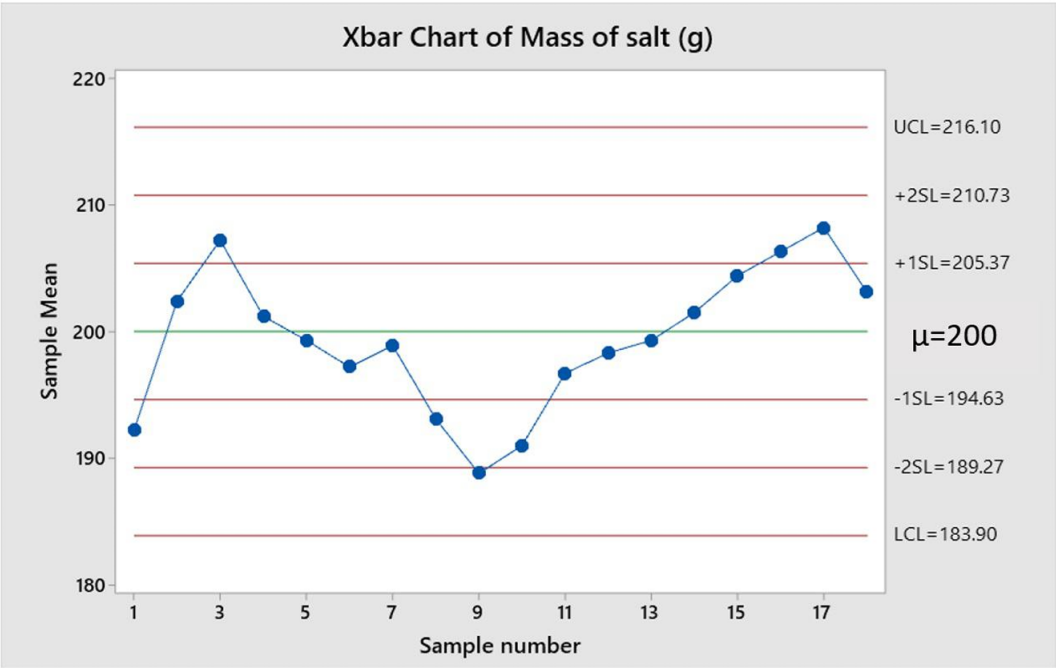
Confirm the 1σ limits shown on the chart and determine if the packing process was in control for all the samples given.

3. When a process for cracking crude oil is working to specification it yields petrol for which random samples give octane ratings which are normally distributed with mean 87.0 and standard deviation 0.40.



Given that each sample is of size 4, confirm the upper and lower control limits shown on the chart and determine if the process was in control for all the samples given.

4. A machine dispenses sea salt flakes into bags. The mass of salt in the bags is known to be normally distributed with mean 200 grams and standard deviation 12g. Samples of 5 bags are taken every 30 minutes and the sample mean plotted on a Shewart x-bar chart.



Confirm the 1 and 2 sigma limits on the chart and determine whether the process was in statistical control for the samples shown. State an assumption that has been made.

5. When a medical manufacturing process is in a state of statistical control, it produces tablets with masses which are normally distributed with mean 6.0mg and standard deviation 0.13mg. Random samples of 5 tablets are taken at regular intervals and the mean mass recorded. The sample means obtained during a day's production, used to produce a control chart, are shown in the table below.

Sample number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Mass (mg)	6.04	5.89	5.99	6.06	5.96	6.08	5.87	6.25
Sample number	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Mass (mg)	6.05	5.86	5.96	6.08	6.19	6.05	5.92	5.90

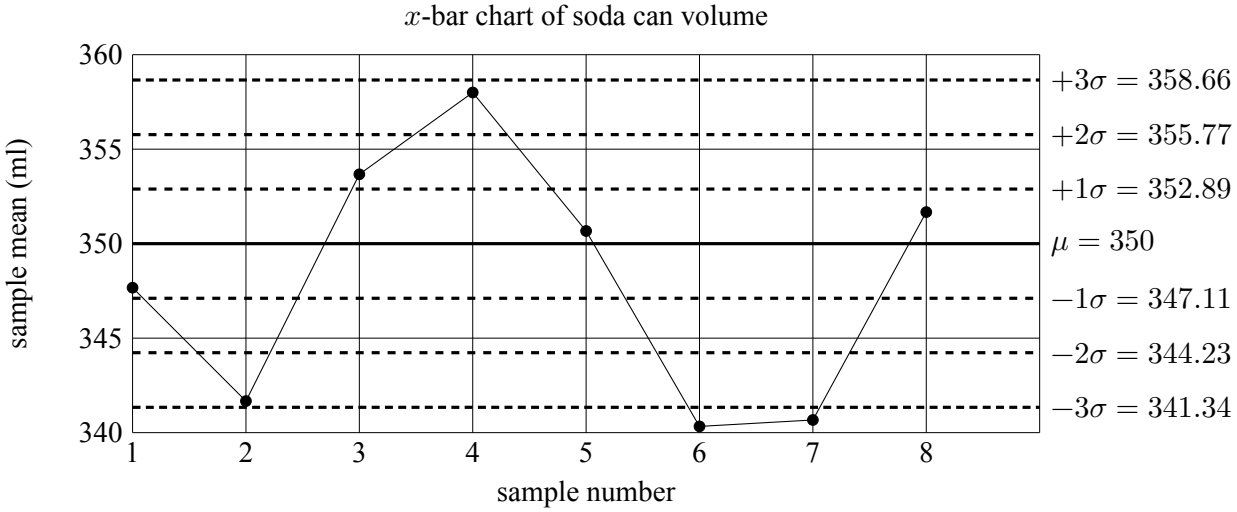
Calculate the upper and lower control limits and determine if the process is in statistical control for the day sampled.

Shewart *x*-bar Charts with an Unknown Population Mean

If the population mean is unknown, the value for μ must be estimated from the *mean of the mean* of the first few sample measurements, known as $\bar{\bar{x}}$. This situation would be common when a control chart for a manufacturing process is being set up for the first time. When μ is unknown, often σ is also unknown. It goes beyond the syllabus for this course to estimate the value of σ so it will be stated in the question.

6. A new medical manufacturing process is set up to produce a new tablet. In order to ensure the process is in statistical control, a control chart is created from random samples of size 5. The first ten sample means were used to create the target value $\bar{\bar{x}}$ for the chart, and it is known that $\sigma = 0.06$ mg.

Sample number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Mass (mg)	4.92	4.94	5.04	5.01	4.98	5.07	4.97	4.99	5.06	5.02
Sample number	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Mass (mg)	4.99	4.98	4.92	4.95	4.93	4.97	4.91	5.01	5.04	4.98



(a) Confirm the value $\bar{\bar{x}} = 5$ mg using the first 10 sample means.

(b) Determine if the process is in statistical control for the initial 20 samples.

13.3 Shewart *p*-Charts for a Proportion

For some quality control processes the attribute of interest will be a *quality* rather than a numerical value, meaning calculating a mean is not possible. For such cases, a control chart for x-bar is not appropriate. Such a *quality* may typically take the form of a binary outcome, such as *defective* or *not defective*. The *proportion* of products within each sample that take a particular outcome (such as *defective*) may instead be calculated, and a control chart for proportion, or p-chart, created.

For $np, nq > 5$ the distribution of the sample proportion may be described by $\hat{P} \approx N\left(p, \frac{pq}{n}\right)$.

This leads to the 3-sigma control limits for a p-chart being given, for population proportion p, by $p \pm 3\sqrt{\frac{pq}{n}}$.

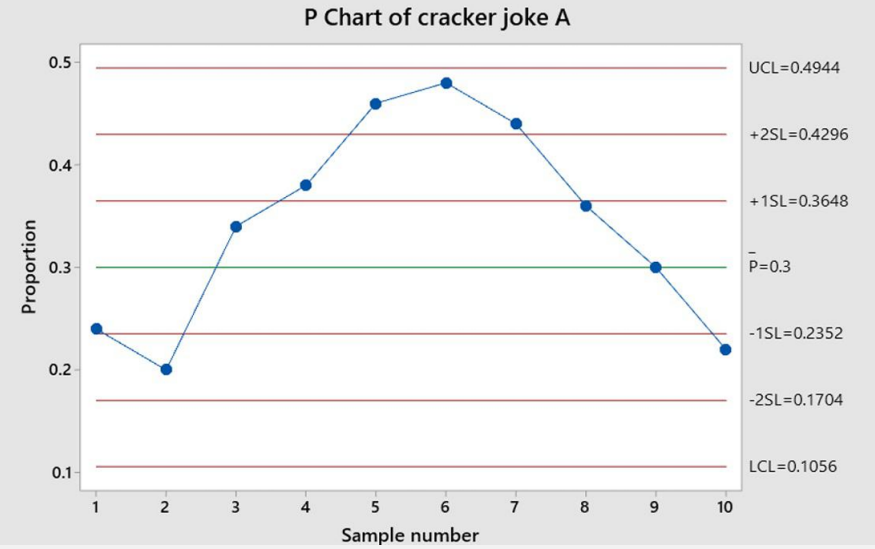
Similarly, the 1σ and 2σ limits be be calculated respectively as $p \pm \sqrt{\frac{pq}{n}}$ and $p \pm 2\sqrt{\frac{pq}{n}}$.

It should be noted that, depending on the value of p, it is not unusual for the lower control limit to be negative or the upper control limit to be greater than 1. When this happens the lower control limit should be given as 0 or the upper control limit given as 1.

If no historical value of the population proportion has been suggested, then p must be estimated from the sample data provided, and $\hat{p}$ used.

Example

Problem: A machine fills Christmas crackers. 30% of the crackers are supposed to contain joke A. Every 30 minutes a sample of 50 crackers is taken and the number containing joke A is found. The proportion of each sample that has joke A is plotted on a p-chart.



Verify the 3σ limits on the chart and determine if the p-chart is in statistical control for the samples given.

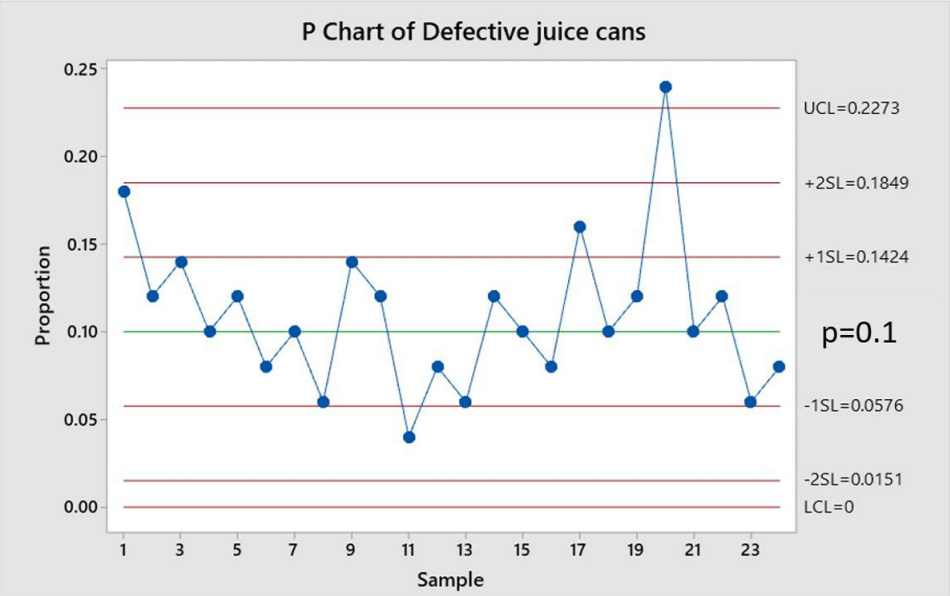
Solution:

3σ limits: $p \pm 3 \times \sqrt{\frac{pq}{n}} = 0.3 \pm 3 \times \sqrt{\frac{0.3 \times 0.7}{50}} = 0.3 \pm 0.1944$, so 0.1056 and 0.4944.

The process is out of statistical control at sample 6 as two out of three consecutive sample means lie above the upper 2σ limit. The process needs to be investigated.

Exercise 13.3

1. A manufacturer of orange juice cans claims that, historically, 10% of the cans produced are defective. 24 samples, each of 50 cans were taken from the production line at daily intervals, and the number of defective cans in the samples were plotted on a p-chart.

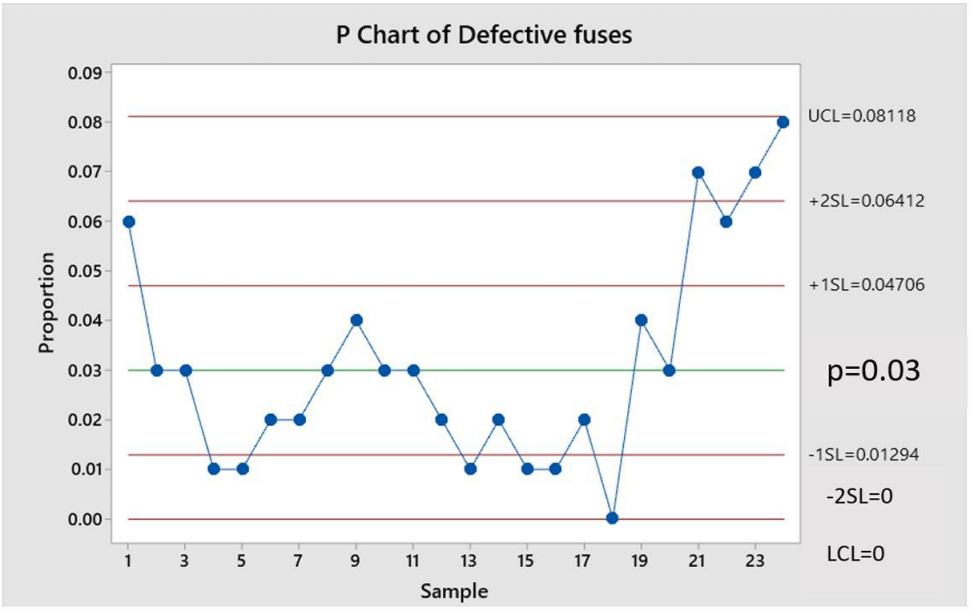


   (a) Confirm the 1 sigma limits given on the p-chart above.

   (b) Explain the value of the lower control limit.

   (c) Determine if the p-chart was in statistical control for all 24 days.

2. A factory machine produces fuses. Historically 3% of these fuses are defective. Random samples of 100 are taken hourly and the number of defective fuses counted. The p-chart below was created for 24 samples.

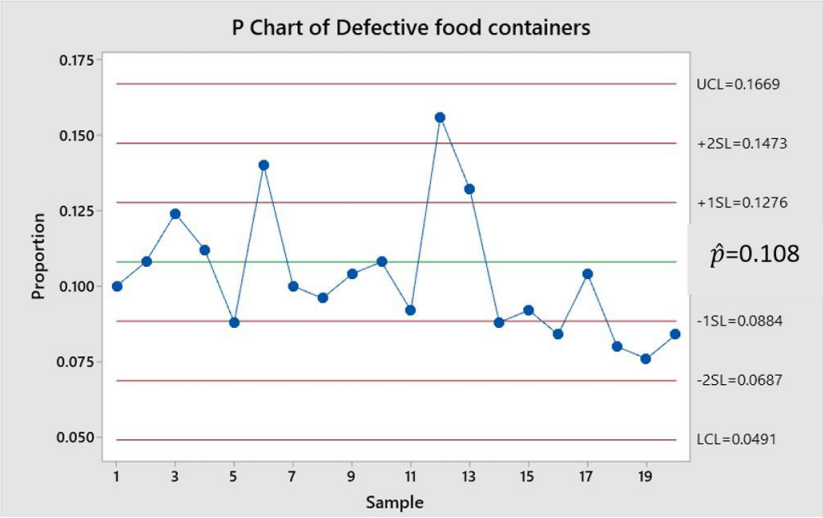


   (a) Confirm the upper and lower control limits.

   (b) Determine if the p-chart was in statistical control.

3. A food retailer insists that the packaging for foodstuffs is produced to a very high-level standard. One of the main supermarkets has asked for evidence of this claim and has sent a manager to the food retailer to inspect the packaging production process. It was decided to create a p-chart using random samples of 250 of a particular type of cardboard food container. The target value was estimated using the first 10 samples.

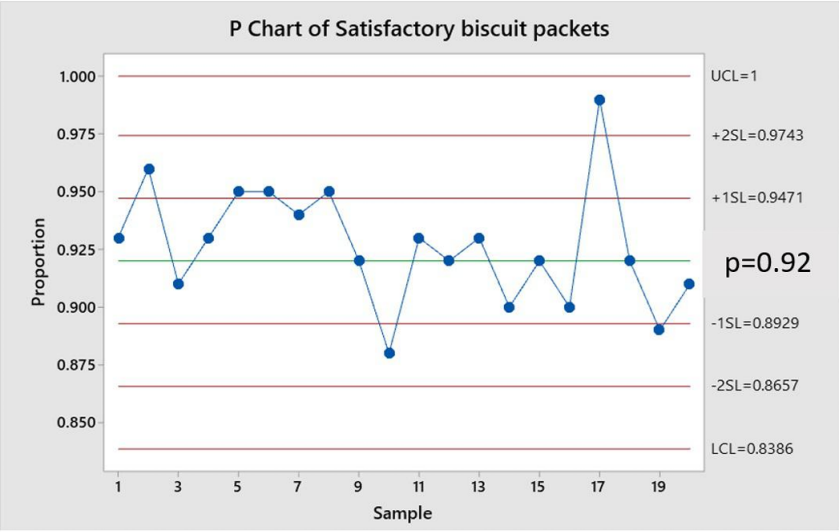
Sample no.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
No. of defective containers	25	27	31	28	22	35	25	24	26	27
Sample no.	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
No. of defective containers	23	39	33	22	23	24	26	20	19	21



(a) Confirm the estimated target value of $\hat{p} = 0.108$ using the first 10 samples.

(b) Determine if the supermarket manager will be satisfied with the results of his inspection.

4. Every hour, 100 small packets of biscuits are randomly sampled from a production line and weighed. A packet is deemed satisfactory if its mass of 25 grams or more, and it is known historically that 92% of biscuit packets are classed as satisfactory. The proportion of satisfactory packets in 20 successive samples are plotted on the p-chart below.



(a) Comment on the upper control limit given on the p-chart.

(b) Determine if the process is in control during throughout the period sampled.

14

Chi-Squared Test for Association

Categorical variables can be described as *independent* if knowledge about one of the variables reveals nothing about the other. Another way of saying this is that there is **no association** between the variables. For example, knowing a person's *favourite colour* reveals nothing about whether or not they are *left-handed or right-handed*, and so there is no association between the someone's favourite colour and their handedness.

On the other hand, knowing whether or not someone *enjoys camping* is likely to help in predicting whether or not they *own a tent*, and so these variables are *not independent*. Another way of saying this is that there **is an association** between enjoying camping and owning a tent.

Suppose a researcher wishes to investigate whether there is a link between someone's *handedness* and whether they have colour vision deficiency (*CVD*), often known as *colour-blindness*. Data is obtained for a random sample of 800 people, and the data is displayed in a *contingency table*:

Observed	CVD	No CVD
Left-handed	4	76
Right-handed	58	662

The contingency table shows that 5% (4 out of 80) of left-handed people had colour-blindness, whilst slightly over 8% (58 out of 720) of right-handed people had colour-blindness. However, this difference in proportions may have occurred due to an association between colour-blindness and handedness, or through *random chance*, since this is a *random sample* of individuals.

A hypothesis test is required to assess whether the data in the contingency table gives evidence to suggest that there is in fact an association between handedness and colour-blindness.

The hypotheses for this test will be as follows:

H_0 : There is **no association** between *handedness* and *colour-blindness*.

H_1 : There is an **association** between *handedness* and *colour-blindness*.

14.1 χ^2 Test for Association

The χ^2 test for association is used to explore a possible *association* between *two categorical variables*, the data for which are typically displayed in a **contingency table**.

The hypotheses for this **non-parametric** test should be stated *in context*.

χ^2 Test for Association hypotheses:

H_0 : There is **no association** between the variables.

H_1 : There is an **association** between the variables.

The test uses the *same test statistic and conditions for validity* as the χ^2 goodness-of-fit test, given on page 6 of the SQA Data Booklet. The decision rule also matches that for the goodness-of-fit test, stated in Chapter 13.

Returning to the data on handedness and colour-blindness, a χ^2 test for association can now be performed. The first step is to include *marginal totals* for each row and column of the contingency table, as well as an total number of observations. The expected value for each cell can then be calculated by multiplying marginal totals and dividing by the total frequency, as shown below:

Observed	CVD	No CVD	
Left-handed	4	76	*80*
Right-handed	58	662	*720*
	62	*738*	*800*

Expected	CVD	No CVD	
Left-handed	6.2	73.8	*80*
Right-handed	55.8	664.2	*720*
	62	*738*	*800*

$\frac{80 \times 738}{800}$

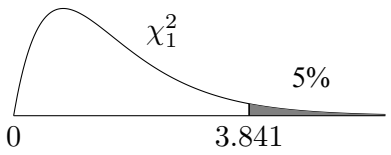
The expected values produced meet the required conditions, in that all are at least 1, and at least 80% are at least 5. Note that these expected values can also be produced quickly using some graphical calculators, as can the χ^2 test statistic:

$$X^2 = \frac{(4-6.2)^2}{6.2} + \frac{(76-73.8)^2}{73.8} + \frac{(58-55.8)^2}{55.8} + \frac{(662-664.2)^2}{664.2} = 0.940$$

The number of *degrees of freedom* for the test statistic is calculated as $\nu = (r-1)(c-1)$, where r is the number of rows and c is the number of columns of data in the contingency table, *not including marginal rows and columns*.

Here there are *two rows* ($r = 2$) and *two columns* ($c = 2$), so $\nu = (2-1)(2-1) = 1$. The χ^2_1 critical value, taking a 5% level of significance, can be obtained from Page 14 of the Data Booklet:

Critical value: $\chi^2_{1,0.95} = 3.841$



Since $0.940 < 3.841$, H_0 is not rejected at the 5% level of significance.

There is insufficient evidence to suggest that there is an association between handedness and colour-blindness.

Example 1

Problem: Researchers working for an educational organisation wish to explore a possible connection between playing a musical instrument and learning a foreign language in school-age teenagers. They randomly select three schools within the region and then, within each school, select a number of S4 pupils to interview.

During the interview, each pupil is asked whether or not they play at least one musical instrument, and whether or not they are studying at least one modern language at National 5 level. The results are summarised in the contingency table below:

	Modern language	No modern language
Plays an instrument	32	18
Doesn't play an instrument	10	20

Perform a χ^2 test at the 1% level of significance to assess whether there is evidence to suggest that there is an association between playing a musical instrument and learning a language.

Solution:

H_0 : There is no association between playing a musical instrument and learning a language.
H_1 : There is an association between playing a musical instrument and learning a language.
$\alpha = 0.01$, One-tailed

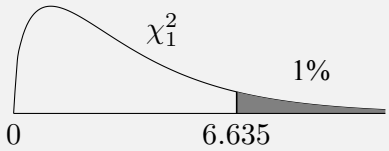
Observed	Language	No language	
Instrument	32	18	*50*
No instrument	10	20	*30*
	42	*38*	*80*

Expected	Language	No language	
Instrument	26.25	23.75	*50*
No instrument	15.75	14.25	*30*
	42	*38*	*80*

$\nu = (r-1)(c-1) = (2-1)(2-1) = 1$

Test statistic: $X^2 = \dfrac{(32-26.24)^2}{26.25} + ... + \dfrac{(20-14.25)^2}{14.25} = 7.071$

Critical value: $\chi^2_{1,0.99} = 6.635$



Since $7.017 > 6.635$, reject H_0 at the 1% level of significance. There is evidence to suggest that there is an association between playing a musical instrument and learning a language.

A vital check carried out before calculating degrees of freedom and the test statistic was that the conditions for validity were met. Here, none of the expected values were less than 1 and at least 80% were at least 5.

Marginal totals are helpful when manually calculating expected values. When a graphical calculator is available, expected values can be quickly produced using matrices.

It should be noted that since the test's purpose of looking for evidence for an *association* is the same thing as looking for a *lack of independence*, the wording of a question may be framed in that manner, as below.

In such cases, reworded yet equivalent hypothesis should be used for the test for consistency.

Example 2

Problem: The outcomes for a random sample of 80 driving tests sat in three different test centres located across Lancashire are recorded. The organisation responsible for overseeing driving test standards is seeking to explore a claim made that the centres do not use the same standards, and that the outcome of driver's test is not independent to location chosen.

		Test Centre		
		Blackburn	Preston	Chorley
Outcome	Pass	19	15	2
	Fail	13	21	10

Use a suitable non-parametric test to assess the data in the contingency table at the 10% level of significance, and hence comment on whether the data supports the claim.

Solution:

H_0 : The test outcome and test centre used are independent.
H_1 : The test outcome and test centre used are not independent.
$\alpha = 0.1$, One-tailed

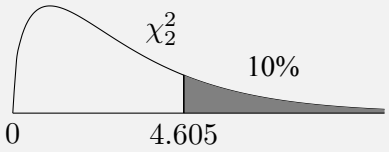
Observed	Blackburn	Preston	Chorley
Pass	19	15	2
Fail	13	21	10

Expected	Blackburn	Preston	Chorley
Pass	14.4	16.2	5.4
Fail	17.6	19.8	6.6

$\nu = (r-1)(c-1) = (2-1)(3-1) = 2$

Test statistic: $X^2 = \dfrac{(17-14.4)^2}{14.4} + ... + \dfrac{(9-6.6)^2}{6.6} = 6.726$

Critical value: $\chi^2_{2,0.9} = 4.605$



Since $6.726 < 4.605$, reject H_0 at the 10% level of significance.
There is evidence to suggest the driving centre used and the outcome of the test are not independent. This suggests the claim may be true (although there could be other reasons apart from a difference in standards applied by the centres).

Note that an equivalent conclusion would be that there is evidence to suggest that there is an *association* between driving centre used and the outcome of the test. However, it is generally preferable for a solution to mirror the language used in the question.

Exercise 14A

1. For all the games in the last 12 months, an ice hockey team records whether they won or lost, as well as whether the game was played at the weekend or on a weekday. The results were collated using a contingency table:

	Won	Lost
Weekend	13	8
Weekday	7	14

   Perform a χ^2 test, at the 10% level of significance, to assess whether there is evidence of an association between when a game is played and the result for the team.

2. A police service gives motorists caught speeding a choice between attending a speed awareness course and a fine together with points added to their licence. A random sample of such motorists has their records checked to see whether they committed a further offence in the following 12 months. The data is recorded in a contingency table:

	Awareness course	Fine and points
Further offence	12	16
No further offence	7	5

   Carry out a non-parametric test, at the 5% level of significance, to determine if there is an association between the choice of speeding penalty and whether a further speeding offence is committed.

3. A sports scientist is studying the effects of caffeine on the performance of professional darts players. 40 players are each randomly allocated to be given either a *caffeinated* or *caffeine-free* drink, each identical in appearance and taste. They are then given one attempt to hit a bullseye on a dart board. The results are as follows:

	Caffeinated	Caffeine-free
Hit	7	15
Miss	17	11

   The sports scientist claims that this shows that caffeine has an effect on the performance of darts players. Perform a χ^2 test of association to determine whether the data supports the claim, at the 10% level of significance.

4. Some modern cameras allow greater control over the sound the shutter makes when taking a photograph; however, some photographers claim that for one particular model there is an association between the shutter mode used and autofocus performance. 90 photos are taken with this model using the three different shutter modes, and each photograph is carefully checked to see whether it was in focus or not.

	In focus	Not in focus
Normal mode	11	4
Quiet mode	13	5
Silent mode	19	8

   Assess whether the data in the contingency table shows evidence for the claim, at the 5% level of significance.

5. Over the course of a month, a number of employees working at a company's office are selected at random each day to have their arrival time and their mode of transport noted. No employee is asked more than once. The results are shown in the contingency table below:

	Car	Bus	Cycle	Walk
On time	48	37	12	23
Late	11	9	3	7

   Perform a chi-squared test at the 1% level of significance to test whether there is an association between the mode of transport an employee uses and whether they make it into work on time.

14.2 Combining Columns or Rows

Where the conditions for validity for a χ^2 test of association are not met, it may be still possible to carry out a test by *combining rows and/or columns*. It is generally advised to collapse those which most naturally may be combined.

Example

Problem: A pharmaceutical company arranges for a random sample of 120 patients to take part in a trial of a new pain relief medication. Some will be given the new medication, whilst others will instead be given a placebo treatment that is identical in appearance and packaging. At the end of the week, each patient will be asked whether they experienced a reduction, an increase in pain, or no change. The data collected is shown in the table below:

		Pain reported		
		Reduction	No change	Increase
Treatment	Medication	45	17	4
	Placebo	28	23	3

Assess at the 5% level of significance whether there is evidence of an association between the treatment given and the pain reported by the patient.

Solution:

H_0 : There is no association between the treatment given and pain reported.
H_1 : There is an association between the treatment given and pain reported.
$\alpha = 0.05$, One-tailed

Observed	Reduction	No change	Increase
Medication	45	17	4
Placebo	28	23	3

Expected	Reduction	No change	Increase
Medication	40.15	22	3.85
Placebo	32.85	18	3.15

Since less than 80% of expected values are at least 5, the last two columns should be combined.

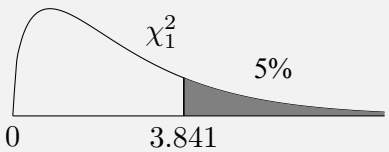
Observed	Reduction	No reduction
Medication	45	21
Placebo	28	26

Expected	Reduction	No reduction
Medication	40.15	25.85
Placebo	32.85	21.15

$\nu = (r-1)(c-1) = (2-1)(2-1) = 1$

Test statistic: $X^2 = \dfrac{(45-40)^2}{40} + ... + \dfrac{(26-21.15)^2}{21.15} = 3.324$

Critical value: $\chi^2_{1,0.95} = 3.841$



Since $3.324 < 3.841$, do not reject H_0 at the 5% level of significance.
There is insufficient evidence to suggest that there is an association between the treatment given and the pain reported by patients.

As with all all hypothesis tests, a decision to not reject H_0 does not mean there is evidence that it is true - only that there is insufficient evidence to suggest that it is *not* true in favour of H_1.

Exercise 14B

1. A random sample of residents listed on the electoral register in four Scottish cities are selected and contacted to find out whether they voted in the most recent UK General Election, and the results shown in a contingency table. A *spoiled vote* means they deliberately left their ballot paper blank or otherwise voided it, often seen as a 'protest vote'.

Observed	Perth	Stirling	Glasgow	Dundee
Voted	51	43	27	40
Didn't vote	17	11	18	22
Spoiled	3	7	2	1

Expected	Perth	Stirling	Glasgow	Dundee
Voted	47.24	40.58	31.27	41.91
Didn't vote	19.95	17.14	13.21	17.7
Spoiled	3.81	3.28	2.52	3.38

Perform a test at the 5% level of significance to assess whether there a difference in voting habits between the cities.

2. Data is collected from a random sample of people living across the United Kingdom on how many teaspoons (tsp) of sugar they typically add to a cup of tea. The contingency table below shows the results:

Observed	0 tsp	1 tsp	2 tsp	3+ tsp
England	132	34	11	7
Scotland	52	27	13	4
Wales	36	11	10	3
N. Ireland	27	16	5	0

Expected	0 tsp	1 tsp	2 tsp	3+ tsp
England	132	34	11	7
Scotland	52	27	13	4
Wales	36	11	10	3
N. Ireland	27	16	5	0

Use a chi-squared test to assess, at the 5% level of significance, whether there is a difference in the amount of sugar the residents of the various parts of the United Kingdom prefer in their tea.

3. A researcher studying a medical condition claims there is a link between having the condition and someone's blood type. A random sample of 620 patients registered with a local doctor have their blood tested, with the test revealing both which of the four major blood types they belong to and whether they test positive or negative for the condition.

	O	A	B	AB
Positive	7	24	10	2
Negative	318	188	50	21

Perform a non-parametric test at the 10% level of significance to assess whether there is evidence to support a claim that there is an association between a patient's blood type and testing positive for the medical condition.

Review Exercise

1. In statistician Ronald Fisher's 1935 book, '*A Design of Experiments*', he details setting up a test to judge a claim made by the scientist Muriel Bristol: that she could tell whether milk was added to a cup *before* tea, or *after*. A statistics student, whose friend claims to only like tea when the milk is added first, decides to set up a similar experiment for herself, disbelieving that her friend can really tell the difference.

   Over the course of a term, she makes 24 cups of tea for her friend, each time randomly selecting whether the milk is added before or after. Her friend then states whether he likes it or not, without knowing which method was used. The results are shown in a contingency table:

		Verdict	
		Likes the tea	Dislikes the tea
Method	Milk first	9	4
	Milk second	2	9

   Perform a non-parametric hypothesis test to assess at the 10% level of significance the evidence for an association between the order the milk is added in and whether her friend likes the tea.

2. Two bird feeders are placed in a garden, with a different type of bird food in each: *sunflower hearts* and *mealworms*. The garden's owner observes the number of each species of bird that visit each of the two feeders over the course of a week:

	Robins	Blue tits	Goldfinches	Bullfinches
Sunflower hearts	5	34	7	1
Mealworms	23	11	2	3

   Perform a non-parametric hypothesis test to assess at the 2.5% level of significance the evidence for an association between the species of the bird and the bird feeder it visits.

3. Two types of nutritional supplement, *A* and *B*, are given to newly hatched chicks. The growth over the next six months for each chick is categorised *high*, *medium* or *low*, with the results displayed in a contingency table set up as below:

	High	Medium	Low
Feed A	17	11	7
Feed B	13	18	5

   Computer output for a hypothesis test conducted on this data is shown below:

```
chi-squared test for association
n = 71        df = 2
alternative hypothesis: there is an asociation between the variables
test statistic = 2.543    critical value = 4.605    p-value = 0.2804
```

   (a) Use the degrees of freedom (*df*) to explain whether any columns or rows were combined.

   (b) State the significance level used.

   (c) State a suitable null hypothesis for the test.

   (d) Use the *p*-value to write a suitable conclusion to the test.

Exercise 14A

1. $3.436 > 2.706$, reject H_0
2. $0.807 < 3.841$, do not reject H_0
3. $4.121 > 2.706$, reject H_0
4. $0.0465 < 5.991$, do not reject H_0
5. $0.282 < 11.345$, do not reject H_0

Exercise 14B

1. $3.328 > 7.815$, do not reject H_0
2. $15.737 > 12.592$, reject H_0
3. $25.106 > 4.605$, reject H_0

Review Exercise

1. $6.254 > 2.706$, reject H_0
2. $23.478 > 7.378$, reject H_0
3. (a) No, with justification
   (b) 10%
   (c) There is no association between the type of feed used and the growth of a chick.
   (d) $0.2804 > 0.1$, do not reject H_0, with context

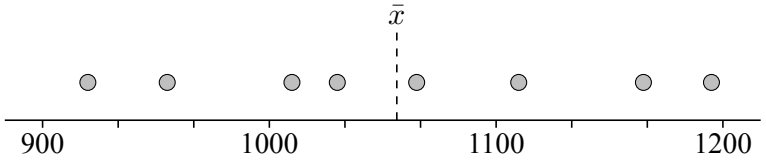
15

Two-Sample Parametric Tests

Chapter 10 introduced three hypothesis tests which use a single sample of data to assess the evidence for claims relating to the population mean. For example, a study to assess the longevity of a brand's light bulbs might check a random sample of such bulbs, to assess whether the mean lifespan is greater than 1000 hours. The data from the sample is shown below, with the sample mean $\bar{x}$ indicated:

$H_0: \quad \mu = 1000$

$H_1: \quad \mu > 1000$

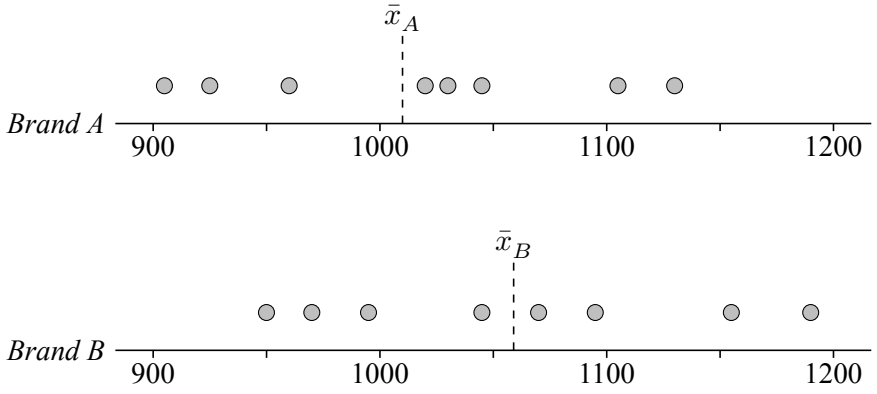


This use of a *single sample* of data, assessing a *single* population parameter, is an example of a **one-sample test**.

Should a study instead wish to compare the longevity of *two different brands' light bulbs* they would instead check *two samples* of data: one sample from *Brand A* and one sample from *Brand B*. In such a case, a typical set of hypotheses might instead ask whether there is evidence to suggest that there is a difference between the population mean lifespan for Brand A, μ_A, and the population mean lifespan for Brand B, μ_B. The data from the samples are shown below, with the sample means $\bar{x}_A$ and $\bar{x}_B$ indicated:

$H_0: \quad \mu_A = \mu_B$

$H_1: \quad \mu_A \neq \mu_B$



Such a test, comparing the parameters of *two* populations by using *two samples*, is an example of a **two-sample test**.

15.1 Two-Sample *z*-test for Population Means

If random samples of data are obtained from each of two populations then, given certain conditions are met, a *two-sample z-test* may be used to explore evidence for a difference in the population means.

Conditions for valid use of the two-sample z-test for population means:

- The *populations* are **normally distributed**.
- The *population standard deviations*, σ_1 and σ_2, are **known**.
- The data was obtained through **random** and **independent samples**.

When the sample sizes n_1 and n_2 are greater than 20 the *Central Limit Theorem* can be invoked, so the sample means are at least approximately normally distributed regardless of the distribution of the underlying population. Also for sample sizes greater than 20, the sample standard deviations s_1 and s_2 provide *sufficiently accurate estimates* for the *population* standard deviations σ_1 and σ_2.

Given the conditions are satisfied, then the sample means can be described as independent random variables $\overline{X}_1$ and $\overline{X}_2$:

$$\overline{X}_1 \sim \mathrm{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \qquad \text{and} \qquad \overline{X}_2 \sim \mathrm{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

The *difference* between the sample means is also therefore normally distributed:

$$\overline{X}_1 - \overline{X}_2 \sim \mathrm{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

The distribution of the test statistic is provided in the SQA Data Booklet as:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

Taking a null hypothesis that *the difference between the population means is zero*, or $\mu_1 - \mu_2 = 0$, a more streamlined version can be obtained:

For a two-sample z-test for population means:

$$\text{Test statistic: } z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

The null hypothesis above can be rewritten as $\mu_1 = \mu_2$, and either one-tailed or two-tailed tests can be performed:

$\mathrm{H}_0 : \mu_1 = \mu_2$	$\mathrm{H}_0 : \mu_1 = \mu_2$	$\mathrm{H}_0 : \mu_1 = \mu_2$
$\mathrm{H}_1 : \mu_1 \neq \mu_2$	$\mathrm{H}_1 : \mu_1 > \mu_2$	$\mathrm{H}_1 : \mu_1 < \mu_2$

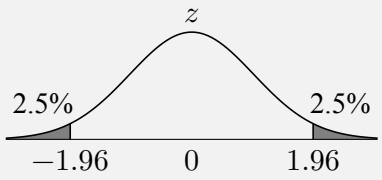
Example 1

Problem: To compare the longevity of two brands' light bulbs, a random sample of eight bulbs from *AlwaysBrite* and eight bulbs from *BulbCity* are tested. The sample of AlwaysBrite bulbs had have a mean lifespan 1015 hours, with a sample mean of 1059 hours before failure for BulbCity.

Assuming that the lifespan for both populations of bulbs is known to be normally distributed, with a standard deviation of 84 hours, use a parametric test to determine whether there is evidence for a difference in longevity between the brands.

Solution: $\bar{x}_A = 1015$, $\bar{x}_B = 1059$, $\sigma_A = 84$, $\sigma_B = 84$ $n_A = 8$, $n_B = 8$

$$\left.\begin{array}{ll} H_0: & \mu_A = \mu_B \\ H_1: & \mu_A \neq \mu_B \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{Two-tailed} \end{array}$$



Test statistic: $z = \dfrac{1015 - 1059}{\sqrt{\dfrac{84^2}{8} + \dfrac{84^2}{8}}} = -1.05$

Critical value: $z_{0.025} = -1.96$

Since $-1.05 > -1.96$, do not reject H_0 at the 5% level of significance. There is insufficient evidence to suggest that there is a difference in the mean lifespans of the two brands of light bulbs.

Example 2

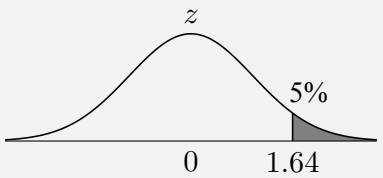
Problem: A survey found that the mean hotel room rate in Edinburgh is £88.42 and that the mean room rate in Glasgow is £80.61. The data was obtained from samples of 50 hotels each and the sample standard deviations were £5.62 and £4.83 respectively. Assess the evidence at the 5% level of significance that mean hotel room rates are more expensive in Edinburgh, stating any assumption made.

Solution: $\bar{x}_e = 88.42$, $\bar{x}_g = 80.61$, $s_e = 5.62$, $s_g = 4.83$ $n_e = 50$, $n_g = 50$

Since both samples sizes are greater than 20, both sample means are approximately normally distributed due to the Central Limit Theorem, and both sample standard deviations of hotel room rate may be considered sufficiently accurate estimators for their respective population standard deviations.

Assume that the samples of hotels for each city were random.

$$\left.\begin{array}{ll} H_0: & \mu_e = \mu_g \\ H_1: & \mu_e > \mu_g \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{One-tailed} \end{array}$$



Test statistic: $z = \dfrac{88.42 - 80.61}{\sqrt{\dfrac{5.62^2}{50} + \dfrac{4.83^2}{50}}} = 7.45$

Critical value: $z_{0.95} = 1.64$

Since $7.45 > 1.64$, reject H_0 at the 5% level of significance. There is evidence to suggest that the mean hotel room rate in Edinburgh is more expensive than in Glasgow.

Exercise 15A

1. The alkalinity, in milligrams per litre, of water in the upper reaches of rivers in a particular region is known to be normally distributed with a standard deviation of 10 mg/l. Alkalinity readings in the lower reaches of rivers in the same region are also known to be normally distributed, but with a standard deviation of 25 mg/l. Ten alkalinity readings are made in the upper reaches of rivers in the region and fifteen in the lower reaches with sample means of 80.5 mg/l and 99.0 mg/l respectively. Test, at the 1% level of significance, the claim that the true mean alkalinity of water in the lower reaches of this river is greater than that in the upper reaches, stating an assumption made.

2. The mass of crisps delivered into bags by a machine is known to be normally distributed with a standard deviation of 0.5 g. Prior to a minor overhaul of the machine, the contents, in grams, of a random sample of six bags are as follows:

   151.7 152.6 150.8 151.9 152.3 151.5

   After the overhaul, which from past experience is known not to affect the standard deviation, the contents of a random sample of twelve bags were measured with the results as follows:

   151.1 150.7 149.0 150.3 151.3 151.4
   150.8 149.5 150.2 150.6 150.9 151.3

   Test the hypothesis that the minor overhaul has had no effect on the mean mass of crisps delivered by the machine.

3. The same test was given to randomly selected groups of 100 scouts and 144 guides. The mean score for the scouts was 27.53 and the mean score for the guides was 26.81. Assuming a common population standard deviation of 3.48, test, using a 5% level of significance, whether the scouts' performance in the test was better than that of the guides.

4. The manager of a lemonade bottling plant is interested in comparing the performances of two production lines, one of which has only recently been installed. For each line she selects 10 one-hour periods at random and records the number of crates completed in each hour. The table below gives the results:

Production Line	Number of crates completed per hour									
1 (Old)	74	77	78	70	87	83	76	78	81	76
2 (New)	78	87	79	82	87	81	85	80	82	83

   From past experience with this kind of equipment, it is known that the variance in these figures will be 25 for Line 1 and 10 for Line 2. Assuming that these samples came from normal populations with these variances, test the hypothesis that the two populations have the same mean.

5. A medical researcher wishes to see whether the pulse rates of smokers are higher than the pulse rates of non-smokers. Random samples of 100 smokers and 100 non-smokers are selected and the results shown below.

Smokers	Non-smokers
$\bar{x}_1 = 90$	$\bar{x}_2$
$s_1 = 5$	$s_2 = 6$
$n_1 = 100$	$n_2 = 100$

   Determine if the researcher has evidence at the 1% level that smokers have higher pulse rates than non-smokers.

6. It is believed that the lengths of new-born babies are normally distributed. The lengths of a random sample of 44 new-born boys at the maternity hospital in Glasgow were measured with sample mean 51.97cm and standard deviation 2.02cm. A random sample of 41 baby girls was also taken and the measurements of their lengths gave a sample mean of 50.21cm and standard deviation of 1.88cm. Test at the 5% level the hypothesis that the mean length of the baby boys was longer than that of the baby girls.

15.2 Two-Sample z-test for Population Proportions

If random, independent samples of data are obtained from each of *two populations*, and the two **sample proportions** $\hat{p}_1$ and $\hat{p}_2$ are obtained, then a *two-sample z-test for population proportions* may be used to explore evidence for a difference in the **population proportions**, p_1 and p_2.

Conditions for valid use of the two-sample z-test for population proportions:

- Each of $n_1\hat{p}_1, n_1\hat{q}_1, n_2\hat{p}_2$ and $n_2\hat{q}_2$ are **greater than 5**.
- The data was obtained through **random** and **independent samples**.

This parametric test can take one-tailed or two-tailed hypotheses which relation to the population proportions, p_1 and p_2:

$$\begin{array}{lll} \mathrm{H}_0 : p_1 = p_2 & \mathrm{H}_0 : p_1 = p_2 & \mathrm{H}_0 : p_1 = p_2 \\ \mathrm{H}_1 : p_1 \neq p_2 & \mathrm{H}_1 : p_1 > p_2 & \mathrm{H}_1 : p_1 < p_2 \end{array}$$

The test statistic is provided in the SQA Data Booklet, and it uses an *estimate of the pooled population proportion*, $\hat{p}$:

For a two-sample z-test for population proportions:

Test statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ where $\hat{p} = \dfrac{\hat{p}_1 n_1 + n_2 \hat{p}_2}{n_1 + n_2}$

Example

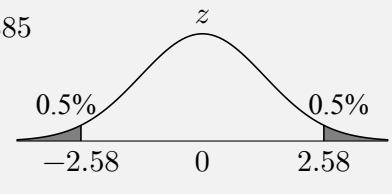
Problem: A sample of 50 randomly selected men with high triglyceride levels consumed 2 tablespoons of oat bran daily for 6 weeks. After 6 weeks, 60% of the men had lowered their triglyceride levels. A sample of 80 men consumed 2 tablespoons of wheat bran for 6 weeks. After 6 weeks 25% had lower triglyceride levels. Assess the evidence for a difference in the proportions at the 1% significance level.

Solution: $\hat{p}_{oat} = 0.6, \hat{p}_{wheat} = 0.25, n_{oat} = 50, n_{wheat} = 80$

$$\left.\begin{array}{ll} H_0 : & p_{oat} = p_{wheat} \\ H_1 : & p_{oat} \neq p_{wheat} \end{array}\right] \begin{array}{l} \alpha = 0.01 \\ \text{Two-tailed} \end{array}$$

Estimate of pooled population proportion: $\hat{p} = \dfrac{0.6 \times 50 + 0.25 \times 80}{50 + 80} = 0.385$

Test statistic: $z = \dfrac{0.6 - 0.25}{\sqrt{0.385 \times 0.615\left(\dfrac{1}{50} + \dfrac{1}{80}\right)}} = 3.99$



Critical value: $z_{0.995} = 2.58$

Since $3.99 > 2.58$, reject H_0 at the 1% level of significance. There is evidence to suggest that the proportions of men who experience a reduction in triglyceride levels by consuming oat bran or wheat bran daily are not the same.

Exercise 15B

1. In a random sample of 70 students, 45 are able to identify a certain brand of cola. In a random sample of 80 teachers, 35 are able to identify the same brand of cola. Test at the 5% significance level the hypothesis that the students are better at identifying the cola and confirm the results given below.

2. A coffee shop chain has launched a new blend of coffee and wishes to investigate whether their younger customers prefer the new blend of coffee more than their older customers. In a random sample of 60 customers aged under 25, 16 preferred the old blend over the new blend, whereas in a random sample of 80 customers aged over 50, 35 preferred the old blend. Test at the 5% significance level the hypothesis that there is a difference in preference between the age groups.

3. A sample of 150 people from a certain industrial community showed that 80 people suffered from a lung disease. A sample of 100 people from a rural community showed that 30 suffered from the same lung disease. Test at the 1% significance level if there is a difference between the proportions of people who suffer from the disease in the two communities, stating an assumption made.

4. In a random sample of 100 store customers, 46 used a Mastercard. In another random sample of 100, 58 used a Visa card. Assess the data for evidence of a difference in the proportion of people who use each type of credit card.

5. A recent survey of people living in Glasgow showed that in a sample of 120 people, 35% had visited Disney World. In another sample of 140 people, 25% had visited Disneyland. Test at the 5% significance level the hypothesis that more people visit Disney World stating any assumptions made.

6. The Trial Urban District Assessment (TUDA) is a study sponsored by the US government of student achievement in large urban school district. In 2009, 1311 of a random sample of 1900 eighth-graders from Houston performed at or above the basic level in mathematics. In 2011, 1440 of a random sample of 2000 eighth-graders from Houston performed at or above the basic level. Assess the evidence for an increase in the proportion of eighth-graders who performed at or above the basic level in mathematics from 2009 to 2011 at the 5% significance level.

15.3 Two-Sample *t*-test for Population Means

When sample sizes are fewer than 20 and population standard deviations are unknown, a one-sample z-test for population means is not appropriate. Instead, a *two-sample t-test for population means* may be used, given certain conditions are met:

Conditions for valid use of the two-sample t-test for population means:

- The *populations* are **normally distributed**.
- The populations have **equal variance**.
- The data was obtained through **random** and **independent samples**.

The test statistic is provided in the SQA Data Booklet, and follows a t-distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. It requires calculation of a *pooled estimate of the variance*, s^2, and as with the two-sample z-test, a more streamline version of the test statistic under a null hypothesis of $\mu_1 = \mu_2$ is as follows:

For a two-sample t-test for population means:

Test statistic: $t_{n_1+n_2-2} = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ where $s^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$

Example

Problem: It is known that the volume of hot drinks dispensed by a type of machine is normally distributed. A random sample of 10 hot drinks from Dispenser A has a mean volume of 203 ml and a standard deviation of 3 ml. A random sample of 15 hot drinks from Dispenser B has corresponding values of 206 ml and 5 ml. Test, at the 5% significance level, the hypothesis that there is no difference in the mean volume dispensed by the two machines.

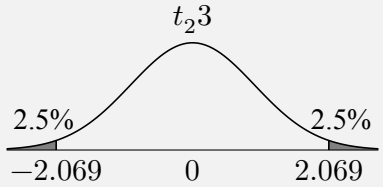
Solution: $\bar{x}_A = 203$, $\bar{x}_B = 206$, $s_A = 3$, $s_B = 5$, $n_A = 10$, $n_B = 15$

$$\left.\begin{array}{ll} H_0: & \mu_A = \mu_B \\ H_1: & \mu_A \neq \mu_B \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{Two-tailed} \end{array}$$

Estimate of pooled variance: $s^2 = \dfrac{(10-1)\times 3^2 + (15-1)\times 5^2}{10+15-2} = 18.739$

$s = \sqrt{18.739} = 4.329$

Test statistic: $t = \dfrac{203 - 206}{4.329 \times \sqrt{\dfrac{1}{10} + \dfrac{1}{15}}} = -1.697$



Critical value: $t_{23,0.975} = -2.069$

Since $-1.697 > -2.069$, do not reject H_0 at the 5% level of significance. There is insufficient evidence to suggest that the mean volume dispensed by the two machines is different.

Exercise 15C

1. A random sample of 10 yellow grapefruit is weighed and the mean mass is found to be 201.4g. The value of an unbiased estimate for the population variance is 234.1g^2. The corresponding figures for a random sample of 8 pink grapefruit are 221.8g and 281.9g^2.

   Determine, using a 1% significance level whether there is evidence that pink grapefruit weigh more than yellow grapefruit. State any assumptions made.

2. The times (in minutes) it took six randomly selected field mice to learn to run a simple maze and the times it took six randomly selected house mice to learn to run the same maze are given.

Field Mice	18	24	12	20	13	15
House Mice	25	16	10	19	14	16

   Use a parametric test to determine if there is a difference in learning rate between the two types of mice.

3. Mr Brown is the owner of a small bakery in a large town. He believes that the smell of fresh baking will encourage customers to purchase goods from his bakery. To investigate this belief, he records the daily sales for 10 days when all the bakery's windows are open and the daily sales for another 10 days when all the windows are closed. The following sales, in £s, are recorded.

Windows open	202.0	204.5	207.0	215.5	190.8	215.6	208.8	187.8	204.1	185.7
Windows closed	193.5	192.2	199.4	177.6	205.4	200.6	181.6	169.2	172.2	192.8

   Assuming that these data may be deemed to be random samples from normal populations with the same variance, investigate the baker's belief.

4. A microbiologist wishes to determine whether there is any difference in the time it takes to make yoghurt from the two different starters: lactobacillus acidophilus (A) and bulgarius (B). Seven batches of yoghurt were made with each of the starters. The table below shows the time taken, in hours, to make each batch along with corresponding summary data.

Starter A	6.8	6.3	7.4	6.1	8.2	7.3	6.9
Starter B	6.1	6.4	5.7	5.5	6.9	6.3	6.7

Variable	N	Mean	SE Mean	StDev
Starter A	7	7.000	0.269	0.712
Starter B	7	6.229	0.191	0.506

   Assuming that both sets of times may be considered to be random samples from normal populations with the same variance, test at the 5% significance level the hypothesis that the mean time taken to make yoghurt is the same for both starters. Comment on any different to the conclusion if a 1% significance level had been used instead.

5. An investigation was conducted into the dust content in the flue gases of two types of solid-fuel boilers. Thirteen boilers of type A and nine boilers of type B were used under identical fuelling and extraction conditions. Over a similar period, the following quantities, in grams, of dust were deposited in similar traps inserted in each of the twenty-two flues.

Type A	73.1	56.4	82.1	67.2	55.2	78.7	75.1	48.0	60.6	63.1	53.3	55.5	61.5
Type B	53.0	39.3	55.8	46.0	56.4	58.8	41.2	66.6	58.9				

A test for equality of population means was performed at the 5% level of significance.

```
two sample t-test
alternative hypothesis: true difference in means is not equal to 0
t = 2.52    df=20    p-value = 0.020
sample estimates:
mean of A = 63.8    mean of B = 52.89
sd of A = 10.6    sd of B = 9.00
```

Confirm the test statistic by calculating the pooled estimate of variance and draw a suitable conclusion for the test, stating any assumption required.

6. As part of a research study into pattern recognition, subjects were asked to examine a picture and see if they could distinguish a word. The picture contained the word 'technology' written backwards and camouflaged by an elaborate pattern. Of the 23 randomly selected librarians who took part, 11 succeeded in recognising the word whilst of 19 randomly selected designers, 13 succeeded. The times, in seconds, for the successful subjects to recognise the word were as follows:

Librarians	55	18	99	54	87	11	62	68	27	90	57		
Designers	23	69	34	27	51	29	45	42	48	74	31	30	31

   (a) Stating any necessary assumptions, investigate whether the mean time for the librarians is longer than for the designers.

   (b) Suggest another parametric hypothesis test that could be carried out using this data and perform this test.

7. A council official wants to compare two brands of paint used for painting stripes on local roads. Twenty locations are selected for the road stripes. The first brand is used on ten randomly selected locations and the second brand on the remaining 10. The number of months each brand lasts is shown below.

Brand A	35.6	36.1	37.0	35.8	34.9	34.9	36.0	37.8	36.6	36.5
Brand B	37.2	36.4	39.7	37.5	37.2	40.5	38.8	38.2	37.7	36.6

Carry out a hypothesis test making clear any assumptions made.

8. The U.S. Bureau of Labour Statistics conducts monthly surveys to estimate hourly earnings of nonsupervisory employees in various industry groups. Results of the surveys can be found in Employment and Earnings. Independent random samples of 14 mine workers and 17 construction workers yielded the following statistics.

Industry	Mean	Standard Deviation
Mining	\$13.93	\$2.25
Construction	\$14.42	\$2.36

Assuming that the hourly earnings of the employees are normally distributed, determine at the 5% significance level if the data provide sufficient evidence to conclude that mine workers earn less on average than construction workers.

15.4 Paired *t*-test for the Population Mean Difference

Suppose six randomly selected athletes have their times to complete a 100 metres sprint recorded, then they each repeat the sprint a month later after some training on technique. Here the two sets of data, *time before* and *time after*, are *not independent*. Instead, each value from the first set of data pairs with a value from the second set; this is called *paired data*. In such a case, the previous two-sample tests in this chapter would be inappropriate. Instead, a single sample of *differences*, d_i, for each pair should be calculated and the mean difference assessed using a *paired t-test*. Possible hypotheses would then related to the population mean difference, μ_d:

$$\begin{array}{lll} \mathrm{H}_0 : \mu_d = 0 & \mathrm{H}_0 : \mu_d = 0 & \mathrm{H}_0 : \mu_d = 0 \\ \mathrm{H}_1 : \mu_d \neq 0 & \mathrm{H}_1 : \mu_d > 0 & \mathrm{H}_1 : \mu_d < 0 \end{array}$$

The test statistic derives from the one-sample t-test covered in Chapter 10, with $\bar{d}$ (or $\bar{x}_d$) representing the sample mean difference and s_d the sample standard deviation in differences.

For a paired t-test:

Test statistic: $t_{n-1} = \dfrac{\bar{d} - \mu_d}{\dfrac{s_d}{\sqrt{n}}}$

Conditions for valid use of a paired t-test:

- The *populations of differences* is **normally distributed**.
- The data was obtained through **random sampling**.

Careful attention to how the sample differences was created is needed for one-tailed hypotheses.

Example

Problem: Six randomly selected children have their IQ measured before and after starting a regular programme of exercise.

IQ before programme	132	98	112	99	168	125
IQ after programme	134	94	117	103	165	129

Assess whether there evidence that exercise affects IQ measurement, at the 5% level of significance.

Solution:

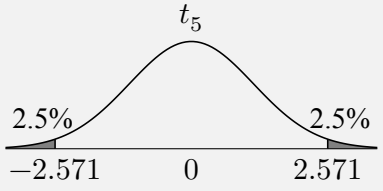
Differences (after−before)	2	-4	5	4	-3	4

$\bar{d} = 1.33$, $s_d = 3.88$, $n = 6$

$$\left.\begin{array}{ll} H_0 : & \mu_d = 0 \\ H_1 : & \mu_d \neq 0 \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{Two-tailed} \end{array}$$

$\nu = 6 - 1 = 5$

Test statistic: $t = \dfrac{1.33 - 0}{\dfrac{3.88}{\sqrt{6}}} = 0.840$



Critical value: $t_{5,0.975} = 2.571$

Since $0.840 < 2.571$, do not reject H_0 at the 5% level of significance. There is insufficient evidence to suggest that the mean difference in IQ measurement is not zero, and so no evidence that exercise affects IQ measurement.

Exercise 15D

1. Eight amateur runners are selected to study the effects of consuming energy gels before running on runners' performances at middle-distances. Each has their finish time recorded for a 5000m run without consuming energy gels beforehand, then a month later their time is recorded again for a 5000m run before which each consumed two energy gels. The data is shown in the table below, with times recorded to the nearest minute.

Runners	A	B	C	D	E	F	G	H
Without gels	24	19	21	17	26	32	18	20
With gels	22	19	20	18	23	28	15	20

Perform a hypothesis test at the 5% level of significance to assess whether the data supports a conclusion that consuming energy gels improves performances for runners.

2. A group of 8 patients with a particular illness were given a special diet and it was desired to test the mass gain in kg at the end of a 2-week period. Their masses are shown below.

Person	A	B	C	D	E	F	G	H
Before	82.27	78.18	86.36	85.00	95.45	75.45	83.18	83.64
After	82.87	79.54	87.36	86.10	94.99	75.48	83.54	82.15

A hypothesis test is conducted to assess the data, with the computer output produced shown below.

```
paired t-test
alternative hypothesis: true difference in means is greater than 0
t = 0.94    df=7    p-value = 0.189
sample estimates:
mean of differences = 0.313    standard deviation of differences = 0.941
```

Confirm the test statistic given and determine if there is evidence at the 5% level that the diet has resulted in significant mass gain, stating any assumptions made.

3. 2. The percentage body fat of 25 randomly selected eleven-year-old boys was measured before and after an exercise program. The difference for each boy was calculated and the summary statistics were calculated:

$$\Sigma d = -30.2 \quad \Sigma d^2 = 286.92$$

(a) Test the hypothesis, at the 1% significance level, that the exercise program has made no difference to the percentage of body fat, stating any assumption made.

(b) Calculate a 99% confidence interval for the mean decrease in percentage body fat and comment on your answer with respect to part a).

Review Exercise

1. An estate agent compares the selling prices of homes in two neighbouring villages to see if there is a difference in price. The results of the study are shown below, measured in pounds.

Village 1	Village 2
$\bar{x}_1 = 281904$	$\bar{x}_2 = 265872$
$s_1 = 30267$	$s_2 = 41092$
$n_1 = 35$	$n_2 = 40$

   Determine if there is evidence at the 1% level to reject the claim that the mean cost of homes in both villages is the same, stating any assumptions made.

2. A random sample of eleven students sat a Chemistry examination consisting of one theory paper and one practical paper. Their marks out of 100 are given in the table below:

Student	A	B	C	D	E	F	G	H	I	J	K
Theory Mark	30	42	49	50	63	38	43	36	54	42	26
Practical Mark	52	58	42	67	94	68	22	34	55	48	17

   Assuming the differences in pairs to be normally distributed, test, at the 5% level of significance the hypothesis of no difference in the mean mark on the two papers.

3. Blood pressure levels were measured in a random sample of 19 men with a particular illness, aged 50 to 59 years. The sample mean systolic blood pressure was 146.6 mmHg and the sample standard deviation was 17.9 mmHG. In a second random sample of 15 men, aged 50 to 59 years, in good health, the mean systolic blood pressure reading was 137.4 mmHg and the standard deviation was 15.1 mmHG.

   Assess the evidence at the 5% level that the mean systolic blood pressure in men with the illness is greater than that of men in good health.

4. In a random sample of 120 workers from a factory in city A, it was found that 5% were unable to read, while in a random sample of 200 workers from a similar factory in city B, it was found that 11% were unable to read. Determine whether it be concluded that there is a difference in the proportions of non-readers in the two cities. Test at the 10% significance level.

   Re-test at the 5% significance level and comment.

5. The standard deviation of the scores obtained on a particular test of mathematical ability is known to be 15. A school experiments with a new method of teaching which is supposed to increase general mathematical awareness. A group of 36 students are randomly assigned to one of two classes. The 20 students in Class A are given the new method of teaching while the other 16 students in Class B are taught in the standard way.

   At the end of the year, the two classes are given the same test of mathematical ability. The mean for Class A is 120.4 and the mean for Class B is 113.1. Test at the 5% significance level the hypothesis that the new method leads to a higher mean performance.

16

Wilcoxon Signed Rank Test

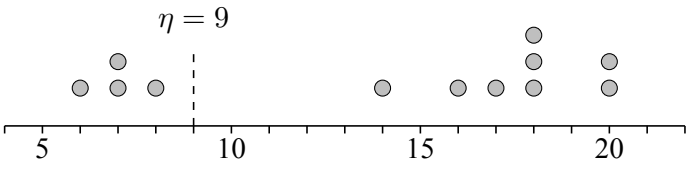
The z and t hypothesis tests from chapters 10 and 15 which explore the *locations* of distributions have all focused on the *population mean*, μ. Those *parametric* tests each rely on either normally distributed populations or sample sizes greater than 20 such that the sample mean is approximately normally distributed.

Given small samples ($n < 20$) from populations whose distributions may not be assumed to be normally distributed, a number of **non-parametric** tests may be used which instead explore the locations of distributions by considering the **population median**, η (*'eta'*).

16.1 Sign Test

The Sign Test is a simple non-parametric test that, whilst *not required for the Advanced Higher Statistics course*, helps to lay the groundwork for the tests covered later in this chapter. It uses a useful property of the population median, η, namely that 50% of observations sampled from the population can be expected to be greater than η and 50% can be expected to be less than η. This means that the number of observations on one side of the population median in a sample of size n should follow a $B(n, 0.5)$ distribution. For example, consider the following random sample of values ($n = 12$) and a claim that the population median is less than 9:

$$\left.\begin{array}{ll} \mathrm{H}_0: & \eta = 9 \\ \mathrm{H}_1: & \eta < 9 \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{One-tailed} \end{array}$$



With 4 out of the 12 observed values being less than 9, the p-value for this test can be calculated as the probability of observing 4 or fewer values less than $\eta = 9$. Letting random variable X represent the number of such values, which follows a $B(12, 0.5)$ distribution under the null hypothesis:

$$p\text{-value} = P(X \leqslant 4) = 0.1938$$

Since $0.1938 > 0.05$, do not reject H_0 at the 5% level of significance.

There is insufficient evidence to suggest that the population median is less than 9.

16.2 Wilcoxon Signed Rank Test

Whilst the Sign Test is only concerned with the number of values that fall on either side of the hypothesised population median, η, the Wilcoxon Signed Rank test also considers how far from the median these values lie. More specifically, it applies weighting to the values by assigning *rank* 1 to the value closest to η, *rank* 2 to the second closest, and so on.

Zeros: Any values that are exactly equal to the median, η, are ignored and not given a rank, with the sample size n reduced to reflect the number of *remaining values to be ranked*.

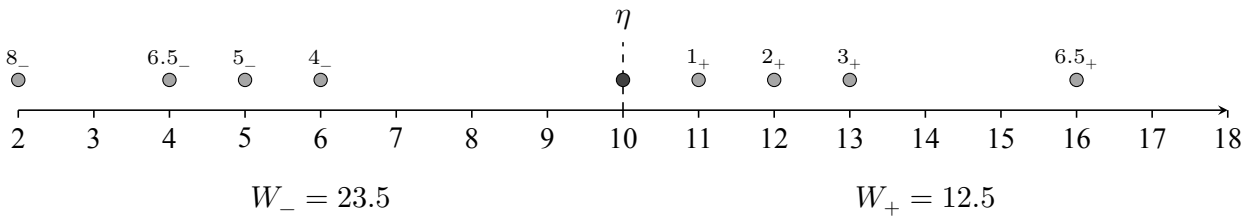
Tied ranks: Values that are equally distant from the median, η, share ranks. For example, if two values are equally far from η when looking to assign rank 4, then they are each ranked 4.5, since $\frac{4+5}{2} = 4.5$. If three values are equally distant from η when assigning rank 7, then they are each ranked 8 since $\frac{7+8+9}{3} = 8$.

- Ranks of values greater than the median are denoted as *positive ranks*, and their total as W_+.

- Ranks of values greater than the median are denoted as *negative ranks*, and their total as W_-.

Consider a random sample of nine values from a distribution with median $\eta = 10$: $\quad 2, 4, 5, 6, 10, 11, 12, 13, 16$

$$\left.\begin{array}{ll} \mathrm{H}_0: & \eta = 10 \\ \mathrm{H}_1: & \eta \neq 10 \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{Two-tailed} \end{array}$$



Note that the sum of W_+ and W_- is $12.5 + 23.5 = 36$, which will always be the total when ranks are assigned to $n = 8$ values, since $1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 = 36$. It is recommended to check that the sum of W_+ and W_- does give the appropriate total for the sample size, which can be calculated for a sample of size n as $\frac{1}{2}n(n+1)$.

For a Wilcoxon Signed Rank test:

- The **test statistic**, denoted W, is the *lower* of W_+ and W_-.
- The **critical value** can be obtained from Page 15 of the Data Booklet.
- H_0 is rejected if the test statistic W is **less than or equal to** the critical value.

Test statistic: $\quad W = 12.5$

Critical value: $\quad CV = 3 \qquad (\alpha = 0.05, n = 8$ and two-tailed.)

Since $12.5 > 3$, do not reject H_0 at the 5% level of significance.
There is insufficient evidence to suggest that the population **median** is not 10.

The Wilcoxon Signed Rank test will typically be used for small samples for which the Central Limit Theorem cannot be invoked, from populations which cannot be reasonably assumed to be normally distributed in order to use a t-test. The table of critical values only extends to samples up to size $n = 20$, once zeros have been discarded.

Conditions for valid use:

- The underlying distribution is **symmetrical**.
- The data was obtained through a **random sample**.

A suggested methodical approach to obtaining the signed ranks required without a diagram is:

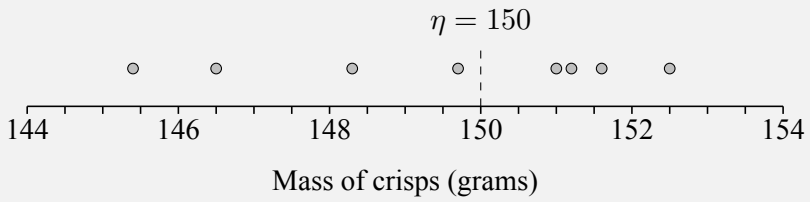
- **Subtract** η from every value in the sample.
- Ignoring any **zeros**, assign ranks to the remaining n values in ascending order by **absolute** value.

The following example should be compared to Example 1 from Section 10.7, Chapter 10, in which a t-test was used.

Example 1

Problem: A supermarket sells sharing-sized bags of a particular brand of crisps. A consumer watchdog is asked to investigate a claim that the median mass of crisps contained in a bag is less than the stated contents of 150 grams. A random sample of bags gives the following results, in grams:

148.3 151.2 149.7 152.5 145.4 151.0 146.5 151.6



Stating a necessary assumption, perform a non-parametric test to assess the claim at the 1% level of significance.

Solution:
Assume that the mass of crisps in a bag is symmetrically distributed.

$$\left.\begin{array}{ll} \mathrm{H}_0: & \eta = 150 \\ \mathrm{H}_1: & \eta < 150 \end{array}\right] \begin{array}{l} \alpha = 0.01 \\ \text{One-tailed} \end{array}$$

148.3	151.2	149.7	152.5	145.4	151.0	146.5	151.6	
-1.7	1.2	-0.3	2.5	-4.6	1.0	-3.5	1.6	(*subtract* $\eta = 150$)
5_-	3_+	1_-	6_+	8_-	2_+	7_-	4_+	(*assign signed ranks*)

$W_+ = 15$ $\qquad \frac{1}{2} \times 8 \times 9 = 36$

$W_- = 21$ $\qquad$ *Check:* $15 + 21 = 36$

Test statistic: $W = 15$

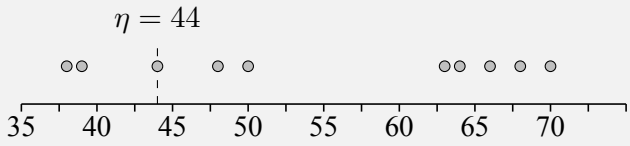
Critical value: $CV = 1$ $\qquad$ ($\alpha = 0.01$, $n = 8$ and one-tailed.)

Since $15 > 1$, do not reject H_0 at the 1% level of significance. There is insufficient evidence to suggest that the median mass of crisps in the bags is less than the stated value of 150g.

Example 2

Problem: Before learner drivers can sit their practical test, they must sit a theory test. Part of this is a hazard perception test, in which a maximum of 75 points are available. Prior to some changes to the hazard perception test, the median score was 44. The points scored by ten randomly selected learner drivers in the updated hazard perception test are recorded:

63 38 48 44 50 70 68 39 64 66



Hazard perception test scores

Test whether there is evidence, at the 10% level of significance, that the median score has changed from 44.

Solution:

$$\left.\begin{array}{ll} \mathrm{H}_0: & \eta = 44 \\ \mathrm{H}_1: & \eta \neq 44 \end{array}\right] \begin{array}{l} \alpha = 0.1 \\ \text{Two-tailed} \end{array}$$

63	38	48	44	50	70	68	39	64	66	
19	-6	4	0	6	26	24	-5	20	22	(*subtract* $\eta = 44$)
5_+	3.5_-	1_+		3.5_+	9_+	8_+	2_-	6_+	7_+	(*assign signed ranks*)

$W_+ = 39.5$ $\qquad \frac{1}{2} \times 9 \times 10 = 45$

$W_- = 5.5$ $\qquad$ *Check:* $39.5 + 5.5 = 45$

Test statistic: $W = 5.5$

Critical value: $CV = 8$ $\qquad$ ($\alpha = 0.1$, $n = 9$ and two-tailed.)

Since $5.5 < 8$, reject H_0 at the 10% level of significance. There is evidence to suggest that the median score achieved in the hazard perception tests has changed from 44 points.

Exercise 16A

1. It is known that a certain tree produces apples with a median mass of 150g. Eight apples are chosen at random from a crate containing apples picked from the tree. The masses, in grams, are:

   147 138 171 142 152 145 141 143

   Use a non-parametric test, at the 5% significance level, to test whether the median mass of the apples is different from 150g.

2. For number of social media posts made in the last week are recorded for a random sample of eight UK teenagers, with the results:

   5 2 18 11 0 6 1 0

   Test at the 5% significance level the claim that the median number of social media posts made per week by a UK teenager is 10.

3. A recycling centre allows residents to deposit their metal waste into a skip set aside for it. The decision as to how often the skip will need to be emptied is based on an understanding that the median daily mass of metal waste deposited into the skip is 15kg, based on previous records. It has been suggested that this figure is no longer accurate. The recycling centre recorded the amount of metal waste deposited on eight randomly selected days. The results, in kg, were:

   11.6 14.7 15.1 14.3 16.2 9.7 15.2 10.7

   Perform a non-parametric test to assess, at the 10% level of significance, whether this sample provides evidence to support the suggestion that the median daily mass of metal waste deposited has changed.

4. The lifetimes of a random sample of a particular brand of candle, measured in minutes, are:

   372 352 335 364 345 360 354 358 348 341

   Determine if there is evidence at the 5% significance level that the median lifetime of this brand of candles is different from 6 hours.

5. A company's complaints department has a policy that states that the median length of time its customers should have to wait before receiving a response to any complaint should not exceed 14 days. A sample of ten customers who have received responses to complaints were asked how long they had to wait. The results, in days, were:

   15 10 17 13 18 12 23 20 19 21

   (a) State two assumptions required in order for a Wilcoxon Signed Rank Test to be performed.

   (b) Assess whether there is evidence to support a claim that the complaint department's policy is being breached, at the 5% level of significance.

6. A company is concerned about the length of time customers have to wait on a helpline before being connected to an agent. Previous records shows that the median waiting time was 34 minutes, and following additional training for the agents the company wishes to gauge the impact it may have had. Customer waiting times, in minutes, for a random sample of ten calls are recorded:

   55 20 31 12 18 32 28 16 14 33

   Stating an assumption required, perform a non-parametric test at the 5% level of significance to assess whether there is evidence the median waiting time has reduced from 34 minutes.

7. A medical centre schedules doctors appointments for patients on the basis of a median appointment duration of 8 minutes. Wishing to determine whether this figure is reasonable, the length of ten randomly chosen appointments are recorded, and displayed below (in minutes).

   7 8 6 5 9 5 7 5 10 4

   Assess at the 5% level whether the data suggests the median appointment duration is not 8 minutes.

16.3 Wilcoxon Signed Rank Test for Paired Data

In Chapter 15, the *paired t-test* was introduced. It is used for assessing the *mean of the differences between paired values* from small samples, and a key assumption for the test is that the *population of differences is normally distributed*. Where this assumption cannot reasonably be made, a **Wilcoxon Signed Rank Test for Paired Data** can instead be used to assesses the **median of the differences**, η_d.

Conditions for valid use of the Wilcoxon Signed Rank Test for Paired Data:

- The **population of differences are symmetrical**.
- The data was obtained through **random sampling**.

Just as with the paired t-test, the paired samples are considered as a *single sample of differences*, allowing a similar process to the one-sample Wilcoxon Signed Rank Test.

Example

Problem: An athletics coach wishes to assess the value to athletes of an intensive period of weight training. Twelve 400-metre runners are selected at random and their times to complete this distance, in seconds, are recorded. Following a programme of weight training they run the distance again, with their times in seconds again recorded. The table below summarises the results.

Athlete	A	B	C	D	E	F	G	H	I	J	K	L
Before	51.0	49.8	49.5	50.1	51.6	48.9	52.4	50.6	53.1	48.6	52.9	53.4
After	50.6	50.4	48.9	49.1	51.6	47.6	53.5	49.9	51.0	48.5	50.6	51.7

Stating an assumption required, perform a non-parametric test to assess the evidence that the training programme will improve athletes' times for the 400 metre distance.

Solution:
(Note that $d = \textit{Before} - \textit{After}$ will give differences such that a positive number represents an improvement in running time, and η_d represents the population median of differences).

$$\left.\begin{array}{ll} H_0: & \eta_d = 0 \\ H_1: & \eta_d > 0 \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{One-tailed} \end{array}$$

0.4	-0.6	0.6	1.0	0.0	1.3	-1.1	0.7	2.1	0.1	2.3	1.7	
2_+	3.5_-	3.5_+	6_+		8_+	7_-	5_+	10_+	1_+	11_+	9_+	(*assign signed ranks*)

$$W_+ = 55.5 \qquad \tfrac{1}{2} \times 11 \times 12 = 66$$
$$W_- = 10.5 \qquad \textit{Check: } 55.5 + 10.5 = 66$$

Test statistic: $W = 10.5$
Critical value: $CV = 13$ ($\alpha = 0.05$, $n = 11$ and one-tailed.)

Since $10.5 < 13$, reject H_0 at the 5% level of significance. There is evidence to suggest that the median difference in 400 metres times is greater than zero, meaning times have improved after the weight training programme.

Exercise 16B

1. To measure the effectiveness of a drug for asthmatic relief, twelve subjects, all susceptible to asthma, were each randomly administered either the drug or a placebo during two separate asthma attacks. After one hour an asthmatic

index was obtained on each subject with the following results:

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Drug	28	31	17	18	31	12	33	24	18	25	19	17
Placebo	32	33	23	26	34	17	30	24	19	23	21	24

Investigate the claim that the drug reduces the asthmatic index.

2. A group of 8 patients with a particular illness were given a special diet and it was desired to test the assess the impact at the end of a 2-week period. Their masses, measured before and after the diet, are shown below, in kg.

Patient	A	B	C	D	E	F	G	H
Before	82.27	78.18	86.36	85.00	95.45	75.45	83.18	83.64
After	82.87	79.54	87.36	86.10	94.99	75.48	83.54	82.15

Use a non-parametric test to determine if there is evidence at the 5% level that the diet increases patient mass, stating any assumptions made.

3. As part of her research into the behaviour of the human memory, a psychologist asked 15 randomly selected school pupils to talk for five minutes on ‘my day at school’. Each pupil was then asked to estimate how many times they thought that they had used the word “like” during the five minutes. The table below gives their estimates together with the true values.

Pupil	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
True value	12	20	1	8	0	12	12	17	6	5	24	23	10	18	16
Estimated value	9	19	3	14	4	12	16	14	5	9	20	16	11	17	19

Use a non-parametric test to investigate whether pupils can remember accurately the frequency with which they use a particular word in a verbal description.

4. Eight amateur runners are selected to study the effects of consuming energy gels before running on runners’ performances at middle-distances. Each has their finish time recorded for a 5000m run without consuming energy gels beforehand, then a month later their time is recorded again for a 5000m run before which each consumed two energy gels. The data is shown in the table below, with times recorded to the nearest minute.

Runners	A	B	C	D	E	F	G	H
Without gels	24	19	21	17	26	32	18	20
With gels	22	19	20	18	23	28	15	20

Perform a hypothesis test at the 5% level of significance to assess whether the data supports a conclusion that consuming energy gels improves performances for runners.

5. The delay in sound transmission through an audio interface is called *latency*. A music technology website wishes to investigate whether a company’s new audio interface, really offers low latency than its older one. A random sample of ten computers each have their audio latency measured, first with the older interface and then with the new interface. The results are as follows, measured in milliseconds:

Computer	1	2	3	4	5	6	7	8	9	10
Old interface	5.7	3.6	7.3	6.1	2.8	4.8	5.3	2.9	6.7	11.7
New interface	5.4	3.5	7.5	5.4	2.7	4.0	5.3	2.6	6.3	11.6

Stating an assumption required, perform a non-parametric test to assess whether the newer interface offers lower latency, at the 5% level of significance.

16.4 Normal Approximation to the Wilcoxon Signed Rank Test

The Data Booklet only provides Wilcoxon critical values for effective sample sizes where $n \leqslant 20$, due to the complexity of obtaining exact critical values as n increases. When dealing with a larger sample, where $n > 20$ after zeros have been removed, a Wilcoxon Signed Rank Test may still be performed using a **normal approximation**. Page 15 of the Data Booklet provides the required properties of the distribution of the test statistic, W:

$$E(W) = \frac{1}{4}n(n+1) \quad \text{and} \quad V(W) = \frac{1}{24}n(n+1)(2n+1)$$

Since here a *discrete* distribution is being approximated by the *continuous* normal distribution, a **continuity correction** is required. If the test statistic W is taken as the lower of W_+ and W_-, a continuity correction of $+0.5$ should be used.

Example

Problem: It is claimed by a local resident that more than 50% of all the vehicles on an urban road exceed the 30 mph speed limit. The speed of each of a random sample of 24 vehicles is recorded with the following results:

22	24	26	28	29	30	30	32	33	34	35	35
37	38	38	39	40	41	42	45	48	56	62	72

Use a non-parametric test to investigate the resident's claim.

Solution:

$H_0 : \ \eta = 30$
$H_1 : \ \eta > 30$
$\alpha = 0.05$, One-tailed

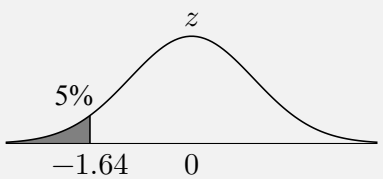
22	24	26	28	29	30	30	32	33	34	35	35	
37	38	38	39	40	41	42	45	48	56	62	72	
-8	-6	-4	-2	-1	0	0	2	3	4	5	5	(*subtract* $\eta = 30$)
7	8	8	9	10	11	12	15	18	26	32	42	
12_-	9_-	5.5_-	2.5_-	1_-			2.5_+	4_+	5.5_+	7.5_+	7.5_+	(*assign signed ranks*)
10_+	12_+	12_+	13_+	14_+	15_+	16_+	17_+	18_+	19_+	20_+	21_+	22_+

$W_+ = 223$ $\qquad \frac{1}{2} \times 22 \times 23 = 253$
$W_- = 30$ $\qquad$ *Check:* $223 + 30 = 253$

$$E(W) = \frac{1}{4} \times 22 \times 23 = 126.5 \quad \text{and} \quad V(W) = \frac{1}{24} \times 22 \times 23 \times 45 = 948.75$$

Test statistic: $z = \dfrac{30.5 - 126.5}{\sqrt{948.75}} = -3.12$

Critical value: -1.64



Since $-3.12 < -1.64$, reject H_0 at the 5% level of significance. There is evidence to suggest that the median speed of the vehicles on the road is greater than 30 mph, and so more than 50% of the vehicles are speeding..

Exercise 16C

Note: sums of signed ranks are often provided in questions that require a normal approximation to the Wilcoxon test. The data must still be inspected to check whether any values have been be discarded, to determine the effective sample size

1. Twenty six apples are chosen at random from a crate containing a large number of apples. The masses (to the nearest gram) are:

136	137	138	139	141	141	142	142	143	145	146	146	147
148	149	150	152	152	156	157	159	161	162	167	171	173

$$W_+ = 149.5 \qquad W_- = 175.5$$

   Use a Wilcoxon Signed Rank test, at the 5% significance level, to test whether the median mass of the apples is different than 150 grams, stating any assumptions made.

2. The lifetimes of a random sample of a particular type of candle, measured to the nearest minute, are:

268	335	339	341	341	345	346	348	352	354	355	356
357	358	359	361	362	363	364	367	371	373	374	374

$$W_+ = 91 \qquad W_- = 209$$

   The manufacturer claims that the average lifetime is at least 6 hours. Use a non-parametric test to assess whether the sample gives evidence to suggest the manufacturer's claim is not correct.

3. A local politician stated that the median hourly wage paid to young people aged 16-17 in the area is £9.20. Believing that the true median hourly wage paid is lower than this, a newspaper randomly selected 28 young people currently doing paid work and recorded their hourly wage (in £):

8.65	9.10	8.71	9.80	8.60	8.95	10.35	8.60	8.86	9.25	8.60	8.70	9.20	8.92
8.82	8.60	9.83	13.95	9.00	8.60	8.63	9.35	8.90	9.70	8.70	9.00	10.10	9.50

$$W_+ = 152 \qquad W_- = 173$$

   Use a non-parametric test to determine whether the sample supports the newspaper's suspicions.

4. A hillwalking website publishes details of popular hiking routes around Scotland. One particular route states that it takes 1 hour and 50 minutes to complete. Following upgrades to some of the paths which the route follows, the time taken to complete the route is recorded for a random sample of 32 hikers. The results of a non-parametric test conducted on the resulting data is shown below, with times recorded in minutes:

```
Wilcoxon test for the population median
data: hike times
alternative hypothesis: true location is less than 110
n = 32    sample median = 102
W = 148    p-value = 0.0154
```

   (a) State the null and alternative hypotheses for the test conducted.

   (b) Use the value of W provided in the computer output to perform a test of the above hypotheses at the 5% level of significance.

   (c) State an assumption required for the test.

Review Exercise

1. A random sample of nine primary school pupils are selected to test the number of times tables questions they can answer correctly in one minute. The results are are follows:

   24 33 36 19 37 21 16 47 20

   Perform a test to determine whether the data supports a claim that the median number of times tables questions a primary school pupil can answer in a minute is greater than 20.

2. To test whether displaying the speeds of cars travelling up a school drive on a digital display is effective in reducing driving speeds, ten cars that regularly drive up the school drive are randomly selected. For each, a speed is recorded on an occasion in which their speed is not displayed on a digital sign, and a speed is recorded when their speed is displayed. The results are as follows, in mph:

Vehicle	A	B	C	D	E	F	G	H	I	J
Speed not displayed	8	9	11	9	11	13	9	16	7	21
Speed is displayed	7	10	10	9	10	14	8	10	9	14

   Use a non-parametric tests to assess whether this suggests displaying the driving speeds of cars travelling up the school drive is effective in reducing speeds, stating an assumption required.

3. A national newspaper states that the median amount of fuel motorists put into their vehicles' tanks when stopping at a petrol station is 20 litres. To investigate this claim, a motoring magazine stops thirty people at random whilst leaving a petrol station and asks how much fuel they just bought. The results, to the nearest litre, are:

   32 45 17 54 20 9 15 34 38 26
   40 24 61 6 18 21 34 48 28 25
   9 10 13 10 25 22 38 41 23 29

   $$W_+ = 106 \qquad W_- = 329$$

   Perform a non-parametric test to assess whether there is evidence to suggest that the figure claimed by the national newspaper is incorrect, at the 5% level of significance.

4. A car company wishes to check that a new machine at their factory producing front bumpers is still maintaining the median mass of 2.8kg for the part as the old machine it replaced. Eight bumpers produced by the machine are randomly selected and weighed, in kilograms. Computer output from a test performed on the data is as follows:

```
Wilcoxon test for the population median
data: bumper mass
alternative hypothesis: true location is not equal to 2.8
W = 25    n = 7 ranks assigned    sample median = 2.95
```

   By first calculating the required test statistic, perform a non-parametric test at the 5% level of significance to determine whether there is evidence to suggest that the median mass of the part produced by the new machine has changed, stating an assumption required.

Chapter 16 Answers

For any questions for which a level of significance is not specified, a 5% level of significance has been used. Full answers should include, of course, conclusions in context

Exercise 16A

1. $9 > 3 \implies$ do not reject H_0
2. $5.5 < 3 \implies$ reject H_0
3. $8 > 5 \implies$ do not reject H_0
4. $7.5 > 5 \implies$ do not reject H_0
5. $10 \leqslant 10 \implies$ reject H_0
6. $9 < 10 \implies$ reject H_0
7. $6.5 > 5 \implies$ do not reject H_0

Exercise 16B

1. $8.5 < 13 \implies$ reject H_0
2. $11 > 5 \implies$ do not reject H_0
3. $46 > 21 \implies$ do not reject H_0
4. $1.5 < 2 \implies$ reject H_0
5. $4 < 8 \implies$ reject H_0

Exercise 16C

1. $-0.34 > -1.96 \implies$ do not reject H_0
2. $-1.67 < -1.64 \implies$ reject H_0
3. $-0.27 > -1.64 \implies$ do not reject H_0
4. (a) $H_0 : \eta = 110, \quad H_1 : \eta < 110$
   (b) $-2.16 > -1.64 \implies$ reject H_0
   (c) Assume that time taken to complete the hike is symmetrically distributed.

Review Exercise

1. $5 \leqslant 5 \implies$ reject H_0
2. $14 > 8 \implies$ do not reject H_0
3. $-2.40 > -1.96 \implies$ reject H_0
4. $3 > 2 \implies$ do not reject H_0

17

Mann-Whitney Rank Sum Test

The Mann-Whitney Rank Sum test is a **two-sample** test, which is **non-parametric**. It is an example of a *permutation test*, which are those that can be approached by considering every possible equally-likely arrangement of sample data. Consider random samples of two mice of species A and four of species B, which will ordered from lightest to heaviest, and then ranks from 1 to 6 assigned. Below are each of the 15 possible arrangements of ranks for the six mice, and alongside each arrangement is the *sum of the ranks* obtained by the two mice of species A:

Arrangement	Rank Sum	Arrangement	Rank Sum	Arrangement	Rank Sum
$AABBBB$	$1+2=3$	$ABBBAB$	$1+5=6$	$BBABAB$	$3+5=8$
$ABABBB$	$1+3=4$	$ABBBBA$	$1+6=7$	$BBABBA$	$3+6=9$
$ABBABB$	$1+4=5$	$BABBAB$	$2+5=7$	$BBBAAB$	$4+5=9$
$BAABBB$	$2+3=5$	$BBAABB$	$3+4=7$	$BBBABA$	$4+6=10$
$BABABB$	$2+4=6$	$BABBBA$	$2+6=8$	$BBBBAA$	$5+6=11$

If the population distributions of mass for mice from species A and mice from species B are *identical* then the probability of obtaining for A a rank sum less than or equal to 4 (for example) by random chance can be calculated:

$$\text{P(rank sum obtained} \leqslant 4) = \frac{2}{15} = 0.1333$$

If it is assumed that the two population distributions of mice masses have the *same shape and variance* then hypotheses may be formed which focus on a possible difference in the *medians*, η_A and η_B, of the distributions. Hypotheses for this two-sample test may be one-tailed or two-tailed. Here, a one-tailed example is given:

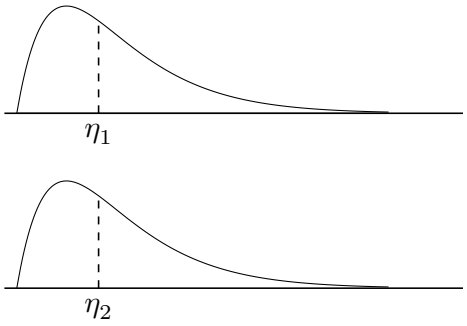
$$\left.\begin{array}{ll} \text{H}_0: & \eta_A = \eta_B \\ \text{H}_1: & \eta_A < \eta_B \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{One-tailed} \end{array}$$

The p-value of 0.1333 obtained about would not provide evidence to suggest, at the 5% level of significance, that the median mass of a mouse of species A is less than the median mass of a mouse of species B (since $0.1333 > 0.05$).

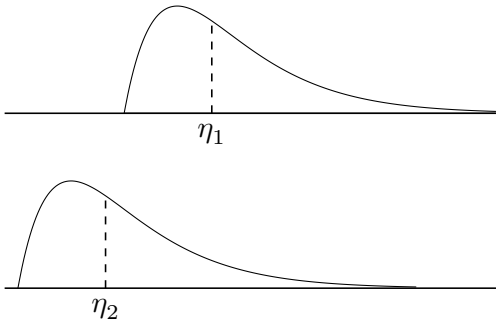
Since the process of listing possible arrangements would become excessively time-consuming for anything other than small samples, Mann-Whitney tests will be typically conducted using *critical values* from the SQA Data Booklet, and *test statistics*. However it is also necessary to be able to perform the test *from first principles* as above, covered on Page 200.

17.1 Mann-Whitney Rank Sum Test

Given two independent samples, the non-parametric **Mann-Whitney Rank Sum test** can be used to assess whether the two population distributions have different *medians*. It is typically used in place of a two-sample t-test where the two distributions are not *normal* and sample sizes are small. Instead a required assumption is that the population distributions have the *same shape and variance*, to test whether there is evidence of a difference in *locations*:



Same locations: $\eta_1 = \eta_2$



Different locations: $\eta_1 \neq \eta_2$

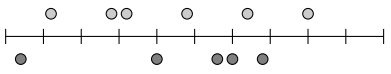
Conditions for valid use of the Mann-Whitney Rank Sum Test:

- The two distributions of the underlying populations have **the same shape and variance**.
- The data were obtained through **random** and **independent samples**

Consider two independent samples taken from populations A and B and the following hypotheses:

$$\left.\begin{array}{l} \mathrm{H}_0 : \eta_A = \eta_B \\ \mathrm{H}_0 : \eta_A \neq \eta_B \end{array}\right\} \begin{array}{l} \alpha = 0.05 \\ \text{Two-tailed} \end{array}$$

A	2.8	3.2	3.3	3.7	4.1	4.5
B	2.6	3.5	3.9	4.0	4.2	



To perform the test using a test statistic and critical value approach, first the samples are combined, ordered and ranked, with **rank sums** W_A and W_B obtained:

Values	2.6	2.8	3.2	3.3	3.5	3.7	3.8	4.0	4.1	4.2	4.5
Sample	B	A	A	A	B	A	B	B	A	B	A
Ranks	1	2	3	4	5	6	7	8	9	10	11

$$W_A = 35$$
$$W_B = 31$$

The test statistic, W, is the *smallest possible rank sum that can be obtained for the smaller sample*, here B. A time-consuming approach would be to reorder the values and rerank them to check if this produces a smaller rank sum W_B.

Values	4.5	4.2	4.1	4.0	3.8	3.7	3.5	3.3	3.2	2.8	2.6
Sample	A	B	A	B	B	A	B	A	A	A	B
Ranks	1	2	3	4	5	6	7	8	9	10	11

$$W_A = 37$$
$$W_B = 29$$

Hence, the test statistic can now be confirmed as $W = 29$. However this value can be obtained without reordering using the formula $m(m + n + 1) - W_m$, where m and n represent the sizes of the smaller and larger samples respectively:

$$5(5 + 6 + 1) - 31 = 29$$

For a Mann-Whitney Rank Sum test:

- The **test statistic**, denoted W, is the *lower* of W_m and $m(m+n+1) - W_m$.
- The **critical value** can be obtained from Page 16 of the Data Booklet.
- H_0 is rejected if the test statistic W is **less than or equal to** the critical value.

Taking as the test statistic the value $W = 29$, the test can now be concluded.

From page 16 of the SQA Data Booklet the critical value of 18 can be obtained, using $m = 5$, $n = 6$ and a two-tailed test at the 5% significance level.

Since $29 > 18$, H_0 can not be rejected at the 5% level of significance. There is insufficient evidence to suggest that the two distributions have different population medians.

Note that whilst only the rank sum for the *smaller sample* is required, to minimise errors it is recommended to calculate rank sums for both samples and check that they add to the required total of $\frac{1}{2}(m+n)(m+n+1)$. Here $\frac{1}{2} \times 11 \times 12 = 66$, and it should be observed that for each pair of rank sums the total is 66.

Example

Problem: Times taken to run 5km for a random sample of 7 hockey players and a random sample of 5 rugby players are recorded. The table below gives the results, to the nearest minute:

Hockey	18	19	20	22	23	24	27
Rugby	23	28	29	30	33		

Stating an assumption required, perform a non-parametric test to assess the data for evidence of a difference in the median time taken for hockey and rugby players to run 5km.

Solution: *Assume the distributions of time taken to run 5km for hockey players and rugbey players have the same shape and variance.*

$H_0: \eta_{\text{hockey}} = \eta_{\text{rugby}}$
$H_1: \eta_{\text{hockey}} \neq \eta_{\text{rugby}}$
$\alpha = 0.05$, Two-tailed

Values	18	19	20	22	23	23	24	27	28	29	30	33
Sample	H	H	H	H	R	H	H	H	R	R	R	R
Ranks	1	2	3	4	5.5	5.5	7	8	9	10	11	12

$n = 7$ $\quad W_H = 30.5$ $\quad \frac{1}{2} \times 12 \times 13 = 78$

$m = 5$ $\quad W_R = 47.5$ $\quad$ *Check:* $39.5 + 5.5 = 45$

Check: $5 \times 13 - 47.5 = 17.5$

Test statistic: $W = 17.5$

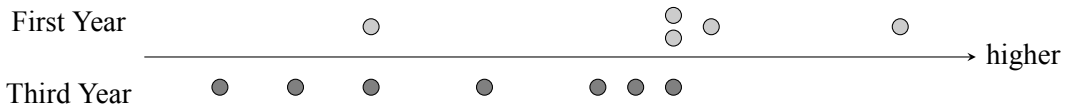
Critical value: $CV = 20$ $\quad$ ($\alpha = 0.05, m = 5, n = 7$ and two-tailed.)

Since $17.5 < 20$, reject H_0 at the 5% level of significance. There is evidence to suggest that the median time taken for rugby players to run 5km is different than the median time taken for hockey players.

Exercise 17A

1. One year on World Maths Day (14th March) pupils across all the year groups in a school attempt a times tables speed challenge. A teacher checks the scores obtained by five randomly selected First Year pupils and seven randomly selected Third Year pupils. Their results are shown below.

First Year	14	6	14	15	20		
Third Year	12	9	6	13	2	14	4



A Mann-Whitney test is performed to assess whether the median score for First Year pupils is greater than that of Third Year pupils.

(a) Use the data provided in the table to verify the test statistic of 20.5.

(b) Conclude the test at the 5% level of significance.

2. A highways official wants to compare two brands of paint used for road markings. Twenty similar locations are selected and at ten of them brand A is used and at the others brand B. The number of months the markings last for were recorded and the results shown below.

Brand A	35.6	36.1	37.0	35.8	34.9	34.9	36.0	37.8	36.6	36.5
Brand B	37.2	36.4	39.7	37.5	37.2	40.5	38.8	38.2	37.7	36.6

Perform a Mann-Whitney test, using the rank sum for Brand A of 65.5, to determine whether there is a difference between the two brands' durability.

3. The durability of two brands of power lawn mowers are to be compared. For independent random samples from each brand, the number of years each lawn mower lasted before breaking down beyond repair is recorded:

Brand X	2.3	3.7	1.9	6.8	3.5	4.4
Brand Y	5.9	3.8	6.4	5.6	4.9	5.9

Test at the 10% level of significance whether the data indicates that there is a difference in median number of year that power lawn mowers from each brand last.

4. The contents of a random sample of six pots of 'Extrafrute' jam were measured. The results, in grams, were:

342 354 349 350 347 356

The results for a random sample of eight pots of 'Jambo' jam were:

340 341 348 342 346 347 342 344

Use a non-parametric test to assess the evidence that the pots of the 'Extrafrute' brand have more jam in them.

5. A farmer decides to test a new diet for egg production. A group of twelve hens is randomly divided into two groups of six. One group is fed the new diet and the other the old. The egg yields recorded in the first year of the test are displayed in the stem-and-leaf diagram below.

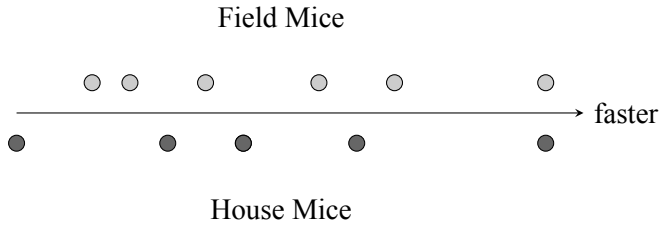
Old Diet			New Diet
	7	22	
	0	23	
4	2	24	5 8
	0	25	5 8 9
	1	26	0

24 | 8 represents 248 grams

Use a Mann-Whitney test at the 5% level to test the hypothesis that the new diet leads to improved egg production.

6. The times (in minutes) it took six randomly selected field mice to learn to run a simple maze and the times it took six randomly selected house mice to learn to run the same maze are given.

Field Mice	House Mice
18	24
24	16
12	10
20	19
13	14
15	16



Stating an assumption required, use a non-parametric test to determine if there is a difference in learning rate between the two types of mice.

7. An investigation was conducted into the dust content in the flue gases of two types of solid-fuel boilers. Thirteen boilers of type A and nine boilers of type B were used under identical fuelling and extraction conditions. Over a similar period, the following quantities, in grams, of dust were deposited in similar traps inserted in each of the twenty-two flues.

Type A	73.1	56.4	82.1	67.2	55.2	78.7	75.1	48.0	60.6	63.1	53.3	55.5	61.5
Type B	53.0	39.3	55.8	46.0	56.4	58.8	41.2	66.6	58.9				

A non-parametric test for a difference in population medians was performed at the 5% level of significance.

```
Mann-Whitney rank sum test
data: TypeA and TypeB
alternative hypothesis: true location shift is not equal to 0
W for TypeB = 134.5    p-value = 0.0416
```

(a) State the suitable null and alternative hypotheses the the test conducted.

(b) By performing the usual check and the rank sum for Type B given in the computer output, obtain the required test statistic.

(c) Use the answer from (b) to complete the hypothesis test and interpret the result.

(d) State three assumptions required for the test performed to be valid.

17.2 Normal Approximation to the Mann-Whitney Rank Sum Test

The Data Booklet only provides Mann-Whitney critical values for sample sizes where $n \leqslant 20$ and $m \leqslant 20$. When dealing with larger samples, where $m, n > 20$, a Mann-Whitney Rank Sum Test may still be performed using a **normal approximation**. Page 16 of the Data Booklet provides the required properties of the distribution of the test statistic, W:

$$\mathrm{E}(W) = \frac{1}{2}m(m+n+1) \quad \text{and} \quad \mathrm{V}(W) = \frac{1}{12}mn(m+n+1)$$

Since here a *discrete* distribution is being approximated by the *continuous* normal distribution, a **continuity correction** is required. If the test statistic W is taken as the lowest possible rank sum W_m *for the smaller sample*, a continuity correction of $+0.5$ should be used.

Example

Problem: 26 Atlantic herrings and 24 North Sea herrings were caught and their lengths measured in cm. The 50 lengths were ordered from smallest to largest and the ranks summed:

$$W_{\text{Atlantic}} = 503 \qquad W_{\text{North Sea}} = 772$$

Assess at the 5% significance level whether Atlantic and North Sea herrings differ in median length.

Solution:

$$\left.\begin{array}{ll} H_0: & \eta_{\text{Atlantic}} = \eta_{\text{North Sea}} \\ H_1: & \eta_{\text{Atlantic}} \neq \eta_{\text{North Sea}} \end{array}\right] \begin{array}{l} \alpha = 0.05 \\ \text{Two-tailed} \end{array}$$

Check that the rank sums for the smaller sample:

$m = 24, n = 26, W_m = 772$

$24 \times (24 + 26 + 1) - 772 = 452$

Note that $452 < 772$

Hence $W = 452$

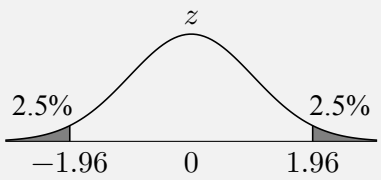
$\mathrm{E}(W) = \frac{1}{2} \times 24 \times (24 + 26 + 1) = 612$

$\mathrm{V}(W) = \frac{1}{12} \times 24 \times 26 \times (24 + 26 + 1) = 2652$

$W \approx \mathrm{N}(612, 2652)$

Test statistic: $z = \dfrac{452.5 - 612}{\sqrt{2652}} = -3.10$

Critical value: -1.96



Since $-3.10 < -1.95$, reject H_0 at the 5% level of significance. There is evidence to suggest that the median length of an Atlantic herring differs from that of a North Sea herring.

Exercise 17B

1. A commuter made a number of car journeys, on weekday mornings, from Perth to Glasgow. She believed that journeys on wet days would take longer, on average, than journeys on dry days. Journey length (in minutes) and weather conditions (either *dry*, D, or *wet*, W) for 48 randomly chosen days throughout the Autumn months, of which 21 were classified as *dry*. The results are shown below.

40	D	50	D	55	D	59	D	61	W	70	W
44	D	51	W	56	W	59	W	62	W	72	W
45	W	51	W	56	D	59	W	63	W	73	W
45	D	52	D	56	D	59	D	63	W	74	D
47	D	52	D	57	D	60	W	63	W	77	W
47	W	54	W	57	W	61	W	68	W	77	W
48	D	54	W	58	W	61	W	68	W	82	D
50	D	55	D	58	W	61	D	70	W	86	W

The data were ranked with the shortest time given rank 1, and the rank sum for the *dry* days was 399.

(a) Perform a Mann-Whitney test at the 1% significance level to determine whether or not her belief is supported by the data.

(b) State an assumption required for the Mann-Whitney test conducted in part (a) to be valid.

2. Archaeologists determined the silver content (%Ag) of random samples of 30 coins discovered in Cyprus from each of the first and second mintages of the reign of King Manuel I. The data were ranked and ranks sums obtained:

$$\text{First mintage rank sum} = 1015 \qquad \text{Second mintage rank sum} = 815$$

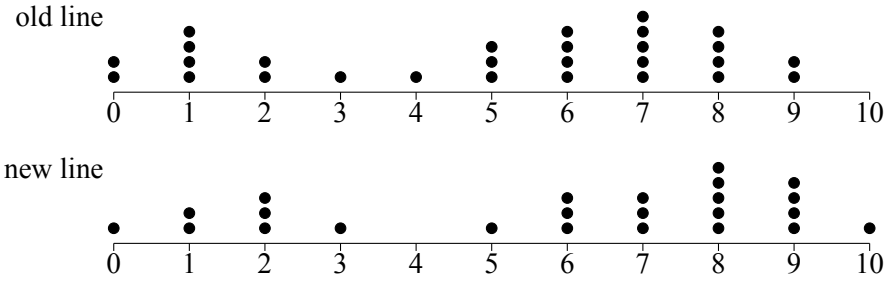
Carry out a hypothesis test to compare the medians of the silver content for the two mintages.

3. During the development of a new adhesive, a research team tested the tensile strength (N/m^2) of random samples of 21 bonds made with Formulation A and 25 bonds made with Formulation B. The results were ranked and ranks sums were:

$$W_A = 583 \qquad W_B = 498$$

Perform a non-parametric hypothesis test to determine whether or not the data provide evidence that Formulation A makes stronger bands than Formulation B.

4. A factory which makes electric heaters has two manufacturing lines in operation, one older and one newer. A quality control inspector wishing to compare the performance of each line selects a random sample of 28 heaters from the old line and 24 heaters from the new line. A quality rating is given for each heater from 0 (terrible) to 10 (flawless), and the results displayed in the dot plots below.



(a) Explain why the dot plots suggest a Mann-Whitney test would be more appropriate for this data than a t-test.

The rank sum for the new line is found to be 562.5.

(b) Carry out a Mann-Whitney test to determine if there is evidence, at the 10% level of significance, that the median quality rating for heaters produced by the new line is higher than that of the old line.

17.3 Mann-Whitney from First Principles

In Chapter 10, hypothesis tests were initially introduced using p-values, which can be compared directly to the chosen significance level to conclude a test. Exact p-values can also be calculated directly for Mann-Whitney tests using an approach introduced on the first page of this chapter. The conditions required for a Mann-Whitney test using the *first principles* method are the same as when using critical values, but with the addition that the data must contain *no tied values*, and hence no shared ranks.

The standard steps involved in performing a one-tailed test in this manner are, for samples of size m and n (where $m \leqslant n$):

- Use ${}^{m+n}C_m$ to calculate the number of possible arrangements of the $m+n$ values.
- Use the observed data to obtain the rank sum for the smaller sample (of size m) as usual.
- Typically by listing all relevant *subsets* of ranks for the smaller sample, determine the number of possible such subsets that results in a rank sum equal to or less than the rank sum obtained from the data.
- The p-value for the test can be calculated using the values from the previous steps as $\dfrac{\text{relevant arrangements}}{\text{possible arrangements}}$.

Example

Problem: Three randomly selected cameras from Brand A and seven from Brand B are tested for reliability, with the number of *shutter actuations* before failure recorded. When ranked with 10 as the most reliable (greatest number of actuations) to 1 as the least reliable (least number of actuations), a rank sum of 9 is obtained for Brand A.

By considering possible arrangements of rankings, and stating any assumptions required, perform a Mann-Whitney test at the 10% significance level to assess whether the data suggests Brand A's cameras are less reliable.

Solution:

Assume that the population distributions of number of actuations for each brand's cameras have the same shape and variance.

$$\left.\begin{array}{ll} H_0: & \eta_{\text{A}} = \eta_{\text{B}} \\ H_1: & \eta_{\text{A}} < \eta_{\text{B}} \end{array}\right] \begin{array}{l} \alpha = 0.1 \\ \text{One-tailed} \end{array}$$

Number of possible arrangements $= {}^{10}C_3 = 120$.

Subsets of relevant arrangements with a rank sum for A$\leqslant 9$:

Ranks for A	Rank Sum
$1, 2, 3$	6
$1, 2, 4$	7
$1, 2, 5$	8
$1, 2, 6$	9
$1, 3, 4$	8
$1, 3, 5$	9
$2, 3, 4$	9

p-value $= \text{P}(\text{rank sum} \leqslant 9) = \dfrac{7}{120} = 0.0583$

Since $0.0583 < 0.1$, reject H_0 at the 10% level of significance. There is evidence to suggest that the median number of shutter actuations before failure for Brand A is less than that for Brand B, and so Brand A's cameras are less reliable.

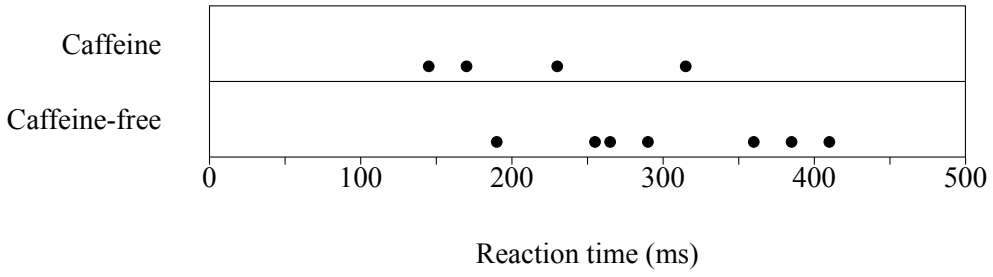
Exercise 17C

1. A Mann-Whitney test is to be performed using two independent samples of sizes 3 and 5. Once data for the samples have been collected and ordered, a rank sum will be obtained for the smaller sample.
   (a) Determine the number of possible arrangements of the 8 values.
   (b) List all the possible subsets of rankings for the smaller sample which would result in a rank sum of 8 or lower.
   (c) Hence determine $\text{P}(\text{rank sum} \leqslant 8)$.

2. A Mann-Whitney test is to be performed using two independent samples of sizes 4 and 5. Once data for the samples have been collected and ordered, a rank sum will be obtained for the smaller sample.
   (a) Determine the number of possible arrangements of the 9 values.
   (b) List all the possible subsets of rankings for the smaller sample which would result in a rank sum of 12 or lower.
   (c) Hence determine $\text{P}(\text{rank sum} \leqslant 12)$.

3. To investigate the effects of caffeine on reaction times, four students from a group of eleven are randomly chosen to be given a soft drink containing caffeine, whilst the remaining seven are given the caffeine-free version of the soft drink. The students will be unaware which they have consumed.

   The reaction time of each student is then tested using an online test, with the results shown on the dot plot below.



   (a) Show the calculation required to obtain the rank sum of 15 for the caffeine group.
   (b) List all of the subsets of ranks for the caffeine group which would produce a value equal to or less than 15.

   A non-parametric hypothesis test is to be conducted under the following hypotheses:

$$\text{H}_0 : \eta_{\text{caffeine}} = \eta_{\text{caffeine-free}}$$
$$\text{H}_0 : \eta_{\text{caffeine}} < \eta_{\text{caffeine-free}}$$

   (c) State an assumption required for the test.
   (d) Under the null hypothesis, calculate the probability of the rank sum for the caffeine group being equal to or less than 15.
   (e) Use your answer to part (c) to conclude the test at the 5% level of significance.

4. A 800m race between four hockey players and eight rugby players results in the hockey players finishing first, second, fourth and sixth. Assume that the population distributions of times for each group have the same shape and are equal in variance.
   (a) By considering possible arrangements and relevant subsets, calculate the p-value for a Mann-Whitney test with an alternative hypothesis that hockey players are faster over 800m than rugby platers.
   (b) Use your answer from (a) to conclude the test at the 10% level of significance.
   (c) State an assumption required for the test to be valid.

Review Exercise

1. A wildlife enthusiast sets up a remote camera which is configured to take a photo when movement is detected. Sometimes it is placed in some local woods, whilst at other times it is positioned beside a pond. The photos are later reviewed and the number of animals spotted each day by the camera is recorded. The number of animal sightings at each location over the last two weeks is shown below.

Woods	7	3	6	2	10	8		
Pond	6	4	1	7	3	1	4	3

Perform a non-parametric hypothesis test at the 5% level of significance to determine whether there is a difference in the median number of wildlife sightings between the two camera locations.

2. 23 red wines and 23 white wines in stock at a large supermarket are randomly chosen to have their alcohol content (%ABV) recorded. The results are displayed in the stem-and-leaf diagram below.

white									red							
							15	0	0							
						0	14	0	2	5	5	8				
			5	5	3	0	0	13	0	0	5	5	5	5	5	8
9	8	5	5	5	5	0	0	12	0	5	5	8	8			
		8	5	5	3	0	0	11	0	5						
					5	5	2	10	5							

$12 \mid 8$ represents 12.8%

(a) Comment on what the stem-and-leaf diagram shows.

The data are ranked from lowest to highest, and a rank sum for the white wines of 401.5 is obtained.

(b) Perform a Mann-Whitney test to assess at the 10% level of significance whether there is evidence to suggest the median alcohol content differs between white and red wines.

A t-test was also considered for a comparison of the alcohol content of the two types of wine.

(c) Comment on what the stem-and-leaf diagram may suggest about the suitability of both a t-test and a Mann-Whitney test.

3. Independent random samples, of sizes 4 and 9 respectively, are taken from populations A and B. A non-parametric hypothesis test is to be performed at the 1% significance level with the hypotheses as follows:

$$\begin{aligned} \mathrm{H}_0 &: \eta_a = \eta_B \\ \mathrm{H}_1 &: \eta_A < \eta_B \end{aligned}$$

When the data from the two samples are ranked from 1 (smallest) to 13 (largest), a rank sum of 12 is obtained for A, with no values tied.

(a) State an assumption required for the test to be performed.

(b) Find the number of possible arrangements of rankings for the 13 values.

(c) List all relevant subsets which provide a rank sum for A of 12 or less.

(d) Hence find the p-value for the test, and state an appropriate conclusion.

Chapter 17 Answers

For any questions for which a level of significance is not specified, a 5% level of significance has been used. Full answers should include, of course, conclusions in context

Exercise 17A

1. (a) Rank sum of 20.5, or rank sum of 44.5 leading to smaller value of 20.5.
   (b) $20.5 < 21 \implies$ reject H_0
2. $65.5 < 78 \implies$ reject H_0
3. $28 \leqslant 28 \implies$ reject H_0
4. $27.5 < 31 \implies$ reject H_0
5. $29 > 28 \implies$ do not reject H_0
6. Assume distributions of learning times for each type of mouse have the same shape and variance.
   $38.5 > 26 \implies$ do not reject H_0
7. (a) $\mathrm{H}_0 : \eta_A = \eta_B$ and $\mathrm{H}_1 : \eta_A \neq \eta_B$
   (b) $72.5 > 73 \implies$ reject H_0
   (c) Assume the distributions of mass of dust deposited for each boiler type have the same shape and variance. Assume the samples of boilers are random. Assume the samples are independent.

Exercise 17B

1. (a) $-2.39 < -2.33 \implies$ reject H_0
   (b) Assume that the distributions of driving times for each weather condition have the same shape and variance.
2. $-1.47 > -1.96 \implies$ do not reject H_0
3. $-1.96 > -1.64 \implies$ do not reject H_0
4. (a) The distributions of gradings for each line appear to have the same shape and variance, to a Mann-Whitney test is appropriate. The distributions do not appear to be normal, so a t-test would not be appropriate.
   (b) $-1.34 < -1.28 \implies$ reject H_0

Exercise 17C

1. (a) 56
   (b) $\{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}$
   (c) $\frac{4}{56} = \frac{1}{14}$ or 0.0714
   (a) 126
   (b) $\{1,2,3,4\}, \{1,2,3,5\}, \{1,2,3,6\}, \{1,2,4,5\}$
   (c) $\frac{4}{126} = \frac{2}{63}$ or 0.0317
2. (a) $1 + 2 + 4 + 8 = 15$
   (b) $\{1,2,3,4\}, \{1,2,3,5\}, \{1,2,3,6\}, \{1,2,3,7\}, \{1,2,3,8\}$
   $\{1,2,3,9\}, \{1,2,4,5\}, \{1,2,4,6\}, \{1,2,4,7\}, \{1,2,4,8\}$
   $\{1,2,5,6\}, \{1,2,5,7\}, \{1,3,4,5\}, \{1,3,4,6\}, \{1,3,4,7\}$
   $\{1,3,5,6\}, \{2,3,4,5\}, \{2,3,4,6\}$
   (c) Assume that the population distributions of reactions times for each group have the same shape and variance.

(d) $\frac{18}{330} = \frac{3}{55}$ or 0.0545

(e) $0.0545 > 0.05 \implies$ do not reject H_0

3. (a) $\frac{7}{495} = 0.0141$

(b) $0.0141 < 0.1 \implies$ reject H_0

(c) Assume that the samples of hockey and rugby players are random.

Review Exercise

1. $34 > 29 \implies$ do not reject H_0

2. (a) The stem-and-leaf diagram shows that red wines seem to contain more alcohol on average than white wines.

(b) $-3.04 < -1.64 \implies$ reject H_0

(c) Since the stem-and-leaf diagram seems to suggest that the distributions of alcohol content for the two groups have the same shape and variance, a Mann-Whitney test would be appropriate. Since the distributions both appear to be normal-like, a t-test would also be appropriate.

(d) i. Assume that the distributions of A and B have the same shape and variance.

ii. 715

iii. $\{1, 2, 3, 4\}, \{1, 2, 3, 5\}, \{1, 2, 3, 6\}, \{1, 2, 4, 5\}$

iv. $0.0056 < 0.01 \implies$ reject H_0

18

Further Linear Regression

Linear regression, first introduced in Chapter 5, will be explored in much more detail in this chapter.

18.1 The Method of Least Squares Regression

A straight-line relationship can be valuable in many situations in establishing and summarising the dependence of one variable on another, and in making predictions.

In most experiments, it is assumed that x is under the control of the experimenter and can be known exactly. In such cases x is called the *predictor* (or *independent*) variable. Y is known as the *response* (or *dependent*) variable, and since it is conditional on x, Y is a random variable.

The (conditional) expected value and variance of Y is commonly of interest.

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

ε is a random variables called the **error** that describes how far Y is from its expected value. α and β are fixed parameters whose value cannot be known exactly without examining all possible occurrences of Y and x. However the sample observations (x_i, y_i) can be used to give estimates a and b of α and β. This leads to the relationship $\hat{Y}_i = a + bx_i$, where $\hat{Y}_i$ denotes the predicted value of Y_i for a given x_i.

The method of least squares gives formulae for a and b derived by considering how $\sum \varepsilon_i^2$ might be minimised:

$$b = \frac{S_{xy}}{S_{xx}} \qquad \text{and} \qquad a = \bar{y} - b\bar{x}$$

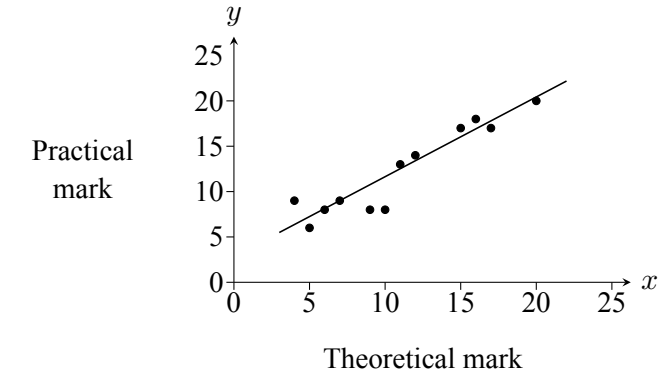
The above estimates, given on Page 5 of the Data Booklet, rely on several assumptions about the errors, ε_i:

- ε_i, are independent.
- $\mathrm{E}(\varepsilon_i) = 0$
- $\mathrm{V}(\varepsilon_i) = \sigma^2$ where $\mathrm{V}(Y) = \sigma^2$

18.2 Contexts for Exercises

The following example and contexts 1 to 5 were all introduced in Chapter 5, and will be used throughout this chapter. For convenience, key summary data and sample statistics previously calculated are provided.

Example: A biology exam consists of two papers: theoretical (x) and practical (y). The marks scored on each test is obtained for a random sample of 12 students who sat the exam, and a scatterplot is constructed to display the results.



Summary data for sample:

$\Sigma x = 132$ $\quad S_{xx} = 290$
$\Sigma x^2 = 1742$ $\quad S_{yy} = 256.25$
$\Sigma y = 147$ $\quad S_{xy} = 255$
$\Sigma y^2 = 2057$
$\Sigma xy = 1872$ $\quad n = 12$

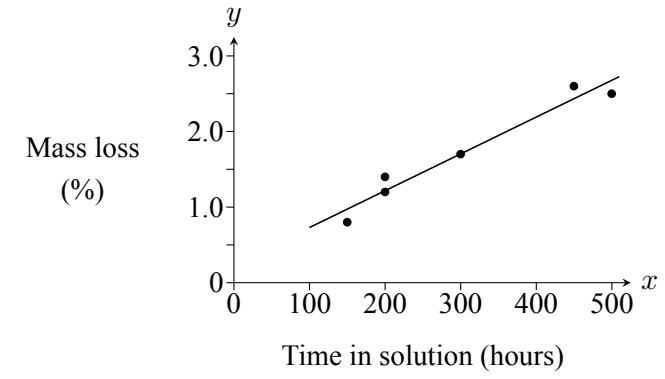
Sample correlation coefficient

$r = 0.935$

Equation of y on x regression line:

$\hat{Y} = 2.581 + 0.879x$

Context 1: Six metal plates were immersed in a weak acid solution for various lengths of time. Their percentage losses in mass were then measured and a scatterplot drawn of time in solution (x hours) against mass loss (y%).



Summary data for sample:

$\Sigma x = 1800$ $\quad S_{xx} = 105000$
$\Sigma x^2 = 645000$ $\quad S_{yy} = 2.6$
$\Sigma y = 10.2$ $\quad S_{xy} = 510$
$\Sigma y^2 = 19.94$
$\Sigma xy = 3570$ $\quad n = 6$

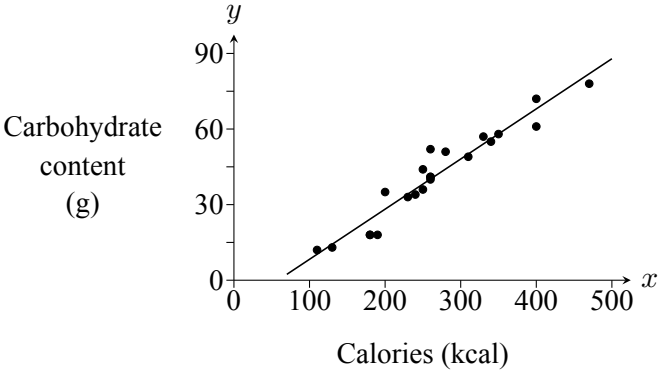
Sample correlation coefficient

$r = 0.976$

Equation of y on x regression line:

$\hat{Y} = 0.243 + 0.00486x$

Context 2: A researcher is wanting to know if there is a link between the number of calories (x) and the carbohydrate content (y) of different types of caffeinated drinks available at coffee shops. The summary statistics below are from a random sample of 22 caffeinated drink options from a range of coffee shops.



Summary data for sample:

$\Sigma x = 5880$ $\quad S_{xx} = 166636$
$\Sigma x^2 = 1738200$ $\quad S_{yy} = 7143.1$
$\Sigma y = 916$ $\quad S_{xy} = 33218$
$\Sigma y^2 = 45282$
$\Sigma xy = 278040$ $\quad n = 22$

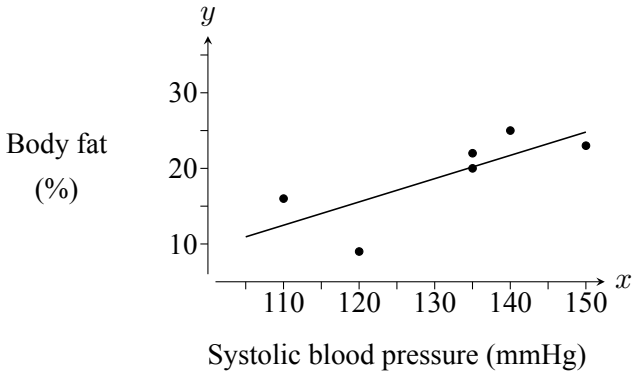
Sample correlation coefficient

$r = 0.963$

Equation of y on x regression line:

$\hat{Y} = -11.6 + 0.199x$

Context 3: The systolic blood pressure in mmHg (x) and percentage body fat (y) of six randomly selected patients with the same consultant were recorded, and the data shown in the scatterplot below.



Summary data for sample:

$\Sigma x = 790$ $\quad S_{xx} = 1033.3$
$\Sigma x^2 = 105050$ $\quad S_{yy} = 170.83$
$\Sigma y = 115$ $\quad S_{xy} = 318.33$
$\Sigma y^2 = 2375$
$\Sigma xy = 15460$ $\quad n = 6$

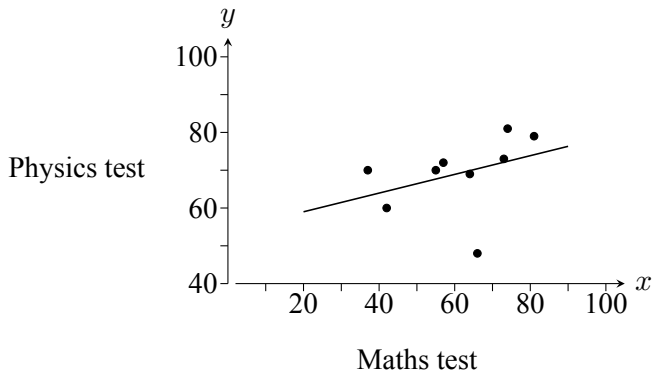
Sample correlation coefficient

$r = 0.758$

Equation of y on x regression line:

$\hat{Y} = -21.4 + 0.308x$

Context 4: Researchers collect marks for 9 students randomly selected from a group of students taking both a maths test (x) and a physics test (y). The results are shown in the scatterplot below.



Summary data for sample:

$\Sigma x = 549$ $\quad S_{xx} = 1736$
$\Sigma x^2 = 35225$ $\quad S_{yy} = 792.89$
$\Sigma y = 622$ $\quad S_{xy} = 428$
$\Sigma y^2 = 43780$
$\Sigma xy = 38370$ $\quad n = 9$

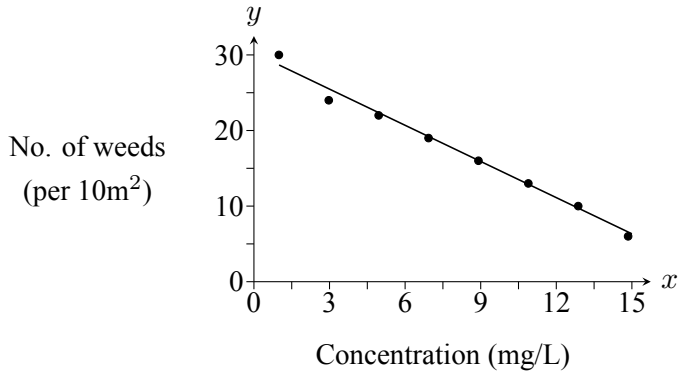
Sample correlation coefficient

$r = 0.365$

Equation of y on x regression line:

$\hat{Y} = 54.1 + 0.247x$

Context 5: A scientist is investigating the effectiveness of a particular weed killer. She uses different concentrations of weed killer (mg/litre, x) and counts the number of surviving weeds (per 10 square metres, y) in a fixed area. The results were as follows.



Summary data for sample:

$\Sigma x = 64$ $\quad S_{xx} = 168$
$\Sigma x^2 = 680$ $\quad S_{yy} = 432$
$\Sigma y = 140$ $\quad S_{xy} = -268$
$\Sigma y^2 = 2882$
$\Sigma xy = 852$ $\quad n = 8$

Sample correlation coefficient

$r = -0.995$

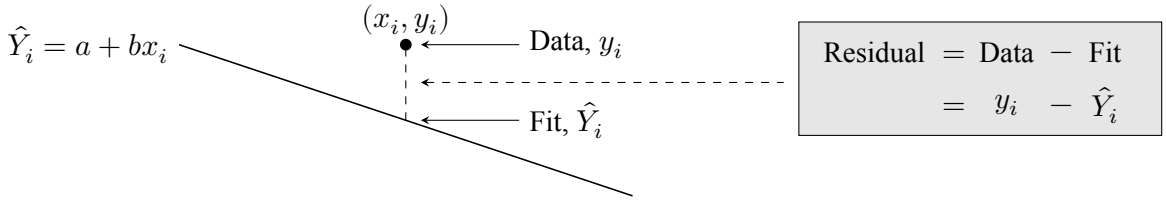
Equation of y on x regression line:

$\hat{Y} = 30.26 - 1.595x$

18.3 Calculating the Value of a Residual

For any given x_i, the equation of the least squares regression line $\hat{Y}_i = a + bx_i$ gives a *predicted* corresponding value $\hat{Y}_i$, referred to as a *fitted* value, or **fit**. For any observed data point (x_i, y_i), the value of y_i may be called the **data** value.

The difference between an observed data value y_i and a corresponding fitted value $\hat{Y}_i$ is called a **residual**:



For data points which lie *above* the regression line, their residual is *positive*.

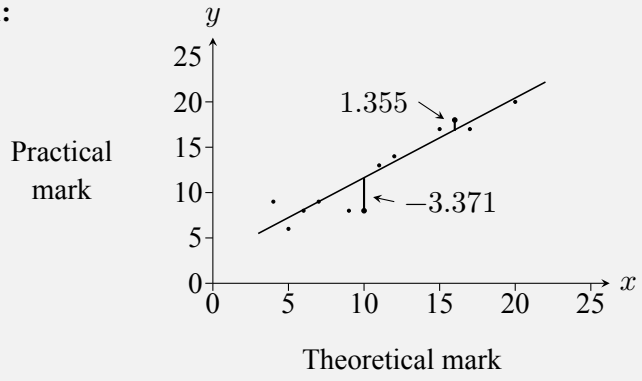
For those which lie *below* the line, their residual is *negative*.

Example

Problem: Students sitting a biology exam are given a mark for each of the two papers: theoretical (x) and practical (y). From a random sample of 12 students, the y on x regression line is calculated as $\hat{Y} = 2.581 + 0.879x$. Calculate the residual for a student who gains:

(a) 16 in the theoretical test and 18 in the practical test.
(b) 10 in the theoretical test and 8 in the practical test.

Solution:



(a)
$\hat{Y} = 2.581 + 0.879 \times 16 = 16.645$
Residual $= 18 - 16.645 = 1.355$

(b)
$\hat{Y} = 2.581 + 0.879 \times 10 = 11.371$
Residual $= 8 - 11.371 = -3.371$

Exercise 18A

Referring to Contexts 1 to 5, calculate the residual for each of the following:

1. From Context 1: A metal plate in the acid solution for 200 hours with 1.4% mass loss.

2. From Context 2: A caffeinated drink with 400 calories and a carbohydrate content of 61 grams.

3. From Context 3: A patient with systolic blood pressure of 110 and body fat % of 16.

4. From Context 4: A student who gained a maths mark of 74 and a physics mark of 81.

5. From Context 5: A weed killer concentration of 3 and 24 surviving weeds counted.

18.4 Estimating the Error Variance

One of the assumptions made for the least squares regression model is that the variance of the errors is a constant for all x_i (and Y_i), or:

$$\mathrm{V}(\varepsilon_i) = \sigma^2 \qquad (\text{where } \mathrm{V}(Y_i) = \sigma^2)$$

The error variance σ^2 is estimated by:

$$s^2 = \frac{SSR}{n-2}$$

Where the (always positive) *sum of squared residuals*, SSR, is calculated as:

$$SSR = \sum \left(y_i - \hat{Y}_i\right)^2 \qquad \text{or} \qquad SSR = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

Both the formula for estimating the error variance σ^2 and the more convenient second formula for the sum of squared residuals SSR are given on page 5 of the SQA Data Booklet.

Example

Problem: Students sitting a biology exam are given a mark for each of the two papers: theoretical (x) and practical (y). From a random sample of 12 students: $S_{xx} = 290, S_{yy} = 256.25, S_{xy} = 255$.

(a) Calculate the sum of squared residuals for this data set.
(b) Obtain an estimate of the error variance.

Solution:

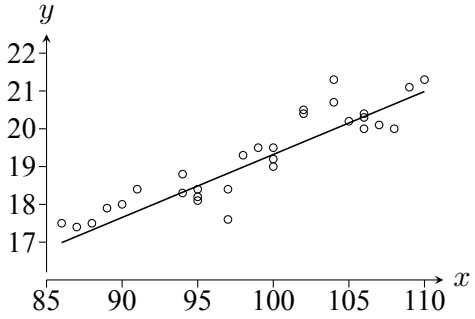
(a) $SSR = S_{yy} - \dfrac{(S_{xy})^2}{S_{xx}} = 256.25 - \dfrac{255^2}{290} = 32.026$

(b) $s^2 = \dfrac{SRR}{n-2} = \dfrac{32.026}{12-2} = 3.203$

Exercise 18B

1. Calculate the sum of squared residuals and hence estimate the error variance for each of Contexts 1 to 5.

2. The age in months of 30 randomly selected primary school pupils (x) and the length of their right foot in centimetres (y) were recorded, with the following results:



Foot length (cm)

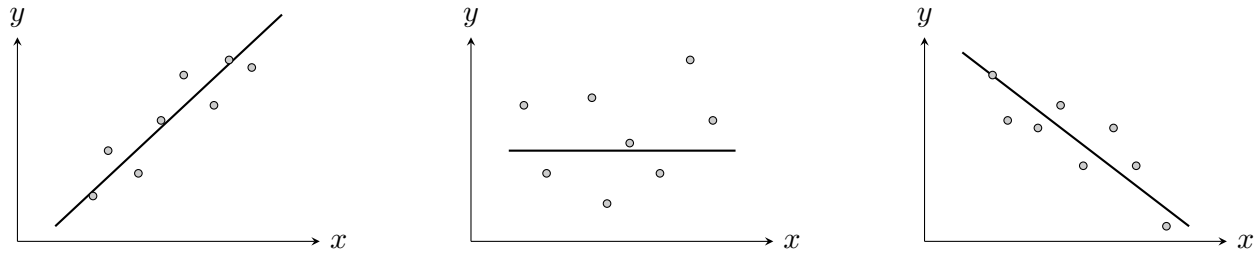
Age (months)

Summary data for sample:
$\Sigma x = 2964$
$\Sigma x^2 = 294248$
$\Sigma y = 577.3$
$\Sigma y^2 = 11152.71$
$\Sigma xy = 57262.7$
$n = 30$

(a) Find the equation of the least squares regression line of y on x.
(b) Estimate the error variance.

18.5 Test for the Slope Parameter β

There are three possibilities for the slope parameter β:



$\beta < 0$ $\beta = 0$ $\beta > 0$

In the case where $\beta = 0$ it can be seen that knowing x_i has no value in predicting y_i. In other words, a linear regression model is only **useful for prediction** in cases where the slope parameter is *non-zero*, or $\beta \neq 0$. The *parametric* **Test for Beta** assesses whether data from a sample provide evidence to suggest that this is the case. The test statistic below follows a t-distribution with $n-2$ degrees of freedom, and in this course the test will always be *two-tailed*.

Test for β test statistic:

$$t_{n-2} = \frac{b\sqrt{S_{xx}}}{s}$$

Conditions for valid use of a test for β:

- ε_i are independent and identically distributed as $\mathrm{N}(0, \sigma^2)$.
- The data was obtained through random sampling.

The conclusion must comment on the evidence for the model being useful for prediction in the given context.

Example

Problem: A bivariate sample has summary statistics:

$$\sum x = 117 \qquad \sum x^2 = 1869 \qquad \sum y = 660 \qquad \sum y^2 = 54638 \qquad \sum xy = 9519 \qquad n = 8$$

Test at the 1% significance level for evidence that a linear regression model would be useful for prediction.

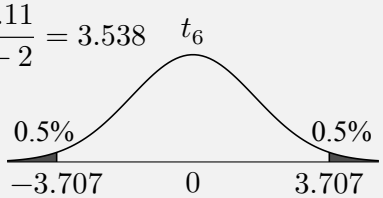
Solution:

$$\left.\begin{array}{ll} H_0: & \beta = 0 \\ H_1: & \beta \neq 0 \end{array}\right] \begin{array}{l} \alpha = 0.01 \\ \text{Two-tailed} \end{array}$$

$$S_{xx} = 157.875 \qquad S_{yy} = 188 \qquad S_{xy} = -133.5 \qquad b = \frac{S_{xy}}{S_{xx}} = \frac{-133.5}{157.875} = -0.8456$$

$$SSR = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = 188 - \frac{(-133.5)^2}{157.875} = 75.11, \quad s = \sqrt{\frac{SSR}{n-2}} = \sqrt{\frac{75.11}{8-2}} = 3.538$$

Test statistic: $t = \dfrac{b\sqrt{S_{xx}}}{s} = \dfrac{-0.8456 \times \sqrt{157.875}}{3.538} = -3.003$



Critical value: $t_{6,0.995} = -3.707$

Since $-3.003 > -3.707$, do not reject H_0 at the 1% level of significance. There is insufficient evidence to suggest that the slope parameter is not zero, and so no evidence that a linear model would be useful for prediction.

Exercise 18C

1. Six metal plates were immersed in a weak acid solution for various lengths of time. Their percentage losses in mass were then measured and a scatterplot drawn of time in solution (x hours) against mass loss (y%). The equation of the least squares regression line for y on x was obtained.

   (See Context 1)

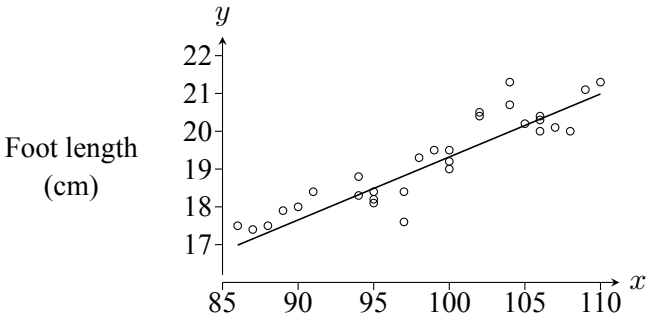
   Perform a test at the 5% level of significance to assess whether there is evidence to suggest that a linear model is useful for prediction.

2. A scientist is investigating the effectiveness of a particular weed killer. She uses different concentrations of weed killer (mg/litre, x) and counts the number of surviving weeds (per 10 square metres, y) in a fixed area. The equation of the least squares regression line for y on x was obtained.

   (See Context 5)

   Assess at the 10% level of significance whether the linear model is useful for prediction.

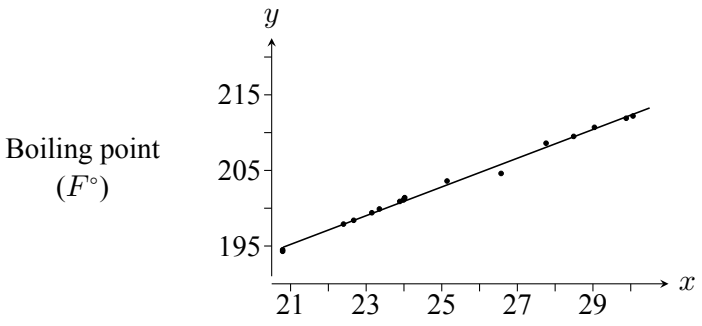
3. The age in months of 30 randomly selected primary school pupils (x) and the length of their right foot in centimetres (y) were recorded, with the following results:



Summary data for sample:
$\Sigma x = 2964$
$\Sigma x^2 = 294248$
$\Sigma y = 577.3$
$\Sigma y^2 = 11152.71$
$\Sigma xy = 57262.7$
$n = 30$

   Test at the 1% significance level for evidence that the model is useful for prediction, stating any assumption made.

4. James Forbes was a Scottish physicist who, amongst other things, was interested in finding a way to estimate altitude from measurements of the boiling temperature of water. Some of his observations on boiling point (y, in F°) and barometric pressure (x, in inches of mercury) are summarised below and shown in the scatterplot.



Summary data for sample:

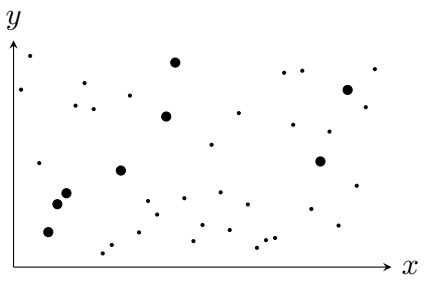
$S_{xx} = 145.94$
$S_{yy} = 530.78$
$S_{xy} = 277.54$

$n = 17$

   The least squares regression line of $\hat{Y} = 155.29 + 1.9018x$ was fitted. Test at the 1% significance level if there is any evidence that the model is useful for prediction, stating any assumption made.

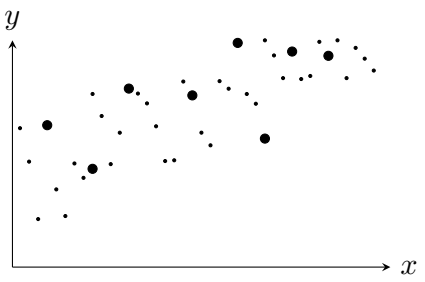
18.6 Test for the Product Moment Correlation Coefficient ρ

Pearson's Product Moment Correlation Coefficient, r, is a sample statistic which measures the strength of the linear association between two variables. It is an estimate of the population parameter ρ (*"rho"*), where $-1 \leqslant \rho \leqslant 1$. If $\rho = 0$ then there is no *linear* relationship between the variables. Given a sample of bivariate data for which there exists *no* linear relationship between the variables, it is regardless common for the sample to give a *non-zero* value for r:



Population:
$\rho = 0$

Sample:
$r = 0.639$



Population:
$\rho = 0.791$

Sample:
$r = 0.634$

The *parametric* **Test for Rho** assesses a sample of bivariate data for evidence of a **linear association** between the variables, or $\rho \neq 0$. The test statistic below follows a t_{n-2} distribution, and in this course the test will always be *two-tailed*.

Test for ρ test statistic:	Conditions for valid use of a test for ρ:
$t_{n-2} = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	• ε_i are independent and identically distributed as $\mathrm{N}(0, \sigma^2)$. • The data was obtained through random sampling.

The conclusion must comment on the evidence for a linear association between the two variables in the given context.

Example

Problem: A bivariate sample has summary statistics:

$$\sum x = 117 \qquad \sum x^2 = 1869 \qquad \sum y = 660 \qquad \sum y^2 = 54638 \qquad \sum xy = 9519 \qquad n = 8$$

Test at the 1% significance level for evidence of a linear association between x and y.

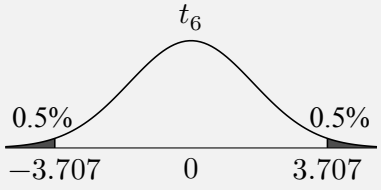
Solution:

$$\left.\begin{array}{ll} H_0: & \rho = 0 \\ H_1: & \rho \neq 0 \end{array}\right] \begin{array}{l} \alpha = 0.01 \\ \text{Two-tailed} \end{array}$$

$S_{xx} = 157.875 \qquad S_{yy} = 188 \qquad S_{xy} = -133.5$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-133.5}{\sqrt{157.875 \times 188}} = -0.7749$$

Test statistic: $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{-0.7749\sqrt{8-2}}{\sqrt{1-(-0.7749)^2}} = -3.003$



Critical value: $t_{6,0.995} = -3.707$

Since $-3.003 > -3.707$, do not reject H_0 at the 1% level of significance. There is insufficient evidence to suggest that there is a linear relationship between thw two variables.

Exercise 18D

1. A researcher is wanting to know if there is a link between the number of calories (x) and the carbohydrate content (y) of different types of caffeinated drinks available at coffee shops. Summary statistics are obtained from a random sample of 22 caffeinated drink options from a range of coffee shops.

   (See Context 2)

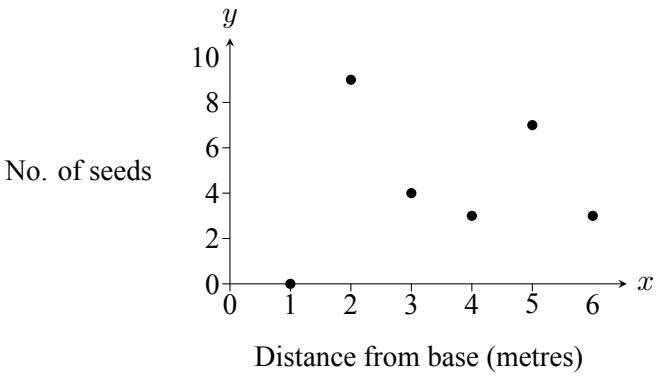
   Perform a test at the 5% level of significance to assess whether there is evidence to suggest that there is a linear relationship between the number of calories a caffeinated drink contains and its carbohydrate content.

2. Researchers collect marks for 9 students randomly selected from a group of students taking both a maths test (x) and a physics test (y).

   (See Context 4)

   Perform a test at the 10% level of significance to assess the linear association.

3. A researcher laid a transect from the base of a tree and placed a quadrat at one metre intervals along the transect line. The number of seed dropped by the tree contained in each quadrat was recorded.



Summary data for sample:

$\Sigma x = 21$

$\Sigma x^2 = 91$

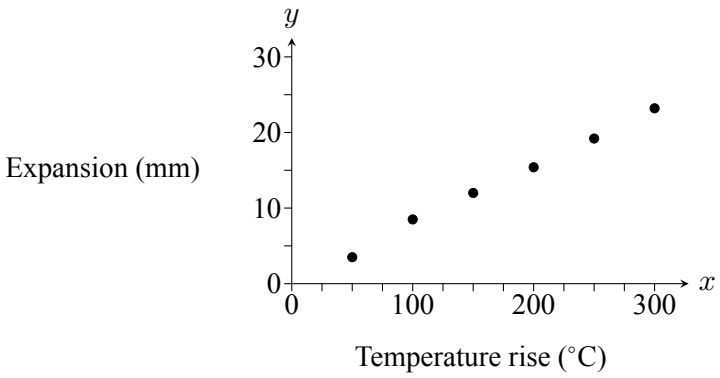
$S_{yy} = 54.833$

$S_{xy} = 3.5$

$n = 30$

   Calculate the sample product moment correlation coefficient and test at the 10% significance level for a linear association between the variables.

4. When a type of metal bar is heated it expands. The amount by which it expands (in mm) and the increase in temperature (in °C) for a random sample of 6 such bars heated to specific temperatures is recorded.



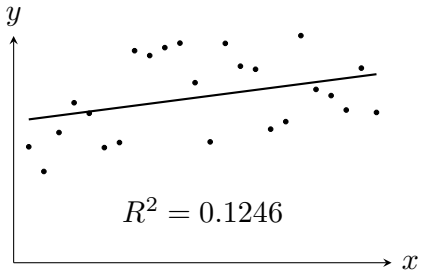
Summary data for sample:

$r = 0.981$

   Use a parametric hypothesis test to assess, at the 0.1% level of significance, the evidence for a linear association between temperature rise and expansion.

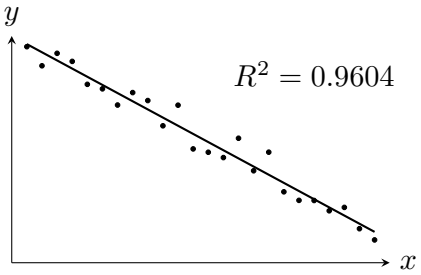
18.7 The Coefficient of Determination

The *coefficient of determination*, R^2,is the proportion of the total variation in the response variable, Y, that is explained by the linear regression model.



$$R^2 = \frac{\text{explained variation}}{\text{total variantion}}$$

$$\left(R^2 = \frac{b^2 S_{xx}}{S_{yy}} = \frac{(S_{xy})^2}{S_{xx} S_{yy}} \right)$$



For the linear regression models covered in this course, the coefficient of determination is connected to the *correlation coefficient*, r, by a simple relationship:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \qquad \Longrightarrow \qquad \boxed{r^2 = R^2}$$

Everything above implies that $0 \leqslant R^2 \leqslant 1$.

Example

Problem: Students sitting a biology exam are given a mark for each of the two papers: theoretical (x) and practical (y). From a random sample of 12 students, the following summary statistics are obtained:

$$S_{xx} = 290 \quad S_{yy} = 256.25 \quad S_{xy} = 255 \quad n = 12 \quad r = 0.935$$

A linear regression model is proposed, and an equation for the regression line of y on x is obtained:

$$\hat{Y} = 2.581 + 0.879x$$

Calculate the coefficient of determination and interpret this value in context.

Solution:

$R^2 = 0.935^2 = 0.8742$.

This suggests that 87.4% of the total variation in practical scores can be explained by the linear model.

Exercise 18E

1. For each of Contexts 1 to 5, calculate the coefficient of determination and interpret it in context.

2. The age in months of 30 randomly selected primary school pupils (x) and the length of their right foot in centimetres (y) were recorded, with the following summary data obtained:

$$\Sigma x = 2964 \quad \Sigma x^2 = 294248 \quad \Sigma y = 577.3 \quad \Sigma y^2 = 11152.71 \quad \Sigma xy = 57262.7 \quad n = 30$$

   (a) Calculate the coefficient of determination.

   (b) Interpret the result from (a).

18.8 Regression Line of x on y

All of the equations of least square regression lines calculated so far have been y *on* x regression lines, calculated with a goal of minimising errors when predicting y *values*.

If there is a need to predict an x *value* based on a y value then it is instead necessary to determine the equation of an x *on* y regression line. This takes the form:

$$\hat{X} = a + by$$

Where the values of a and b for differ from those required for an y on x line, and are given instead by:

$$b = \frac{S_{xy}}{S_{yy}} \quad and \quad a = \bar{x} - b\bar{y}$$

Example

Problem: Students sitting a biology exam are given a mark for each of the two papers: theoretical (x) and practical (y). From a random sample of 12 students, the following summary statistics are obtained:

$$\Sigma x = 132 \quad \Sigma x^2 = 1742 \quad \Sigma y = 147 \quad \Sigma y^2 = 2057 \quad \Sigma xy = 1872$$

$$S_{xx} = 290 \quad S_{yy} = 256.25 \quad S_{xy} = 255 \quad n = 12 \quad r = 0.935$$

Determine the equation for the least squares regression line of x on y.

Solution:

$$b = \frac{S_{xy}}{S_{yy}} = \frac{255}{256.25} = 0.995$$

$$a = \bar{x} - b\bar{y} = \frac{132}{12} - 0.995 \times \frac{147}{12} = -1.189$$

$$\hat{X} = -1.189 + 0.995y$$

Exercise 18F

For each of Contexts 1 to 5, determine the equation of the least squares regression line for x on y, and hence calculate predicted values for each of the following.

1. From Context 1: The time a metal plate spent in the acid solution if the mass loss was 2.1%.

2. From Context 2: The calories contained in a caffeinated drink which has a carbohydrate content of 34 grams.

3. From Context 3: The systolic blood pressure of a patient whose body fat % is 11.

4. From Context 4: The maths mark of a student who gained a physics mark of 75.

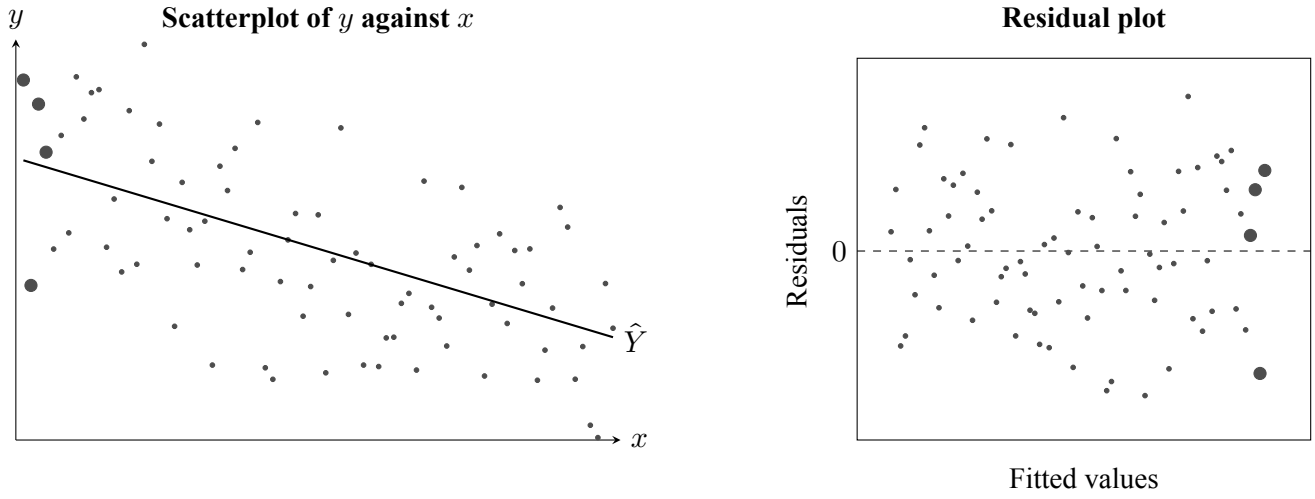
5. From Context 5: The weed killer concentration used if the area contained 8 weeds per 10m^2.

18.9 Residual Plots

When considering the use of a linear regression model for bivariate data, it is recommended to produce a plot of *residuals* (ε_i) against *fitted values* ($\hat{Y}_i$), which have been calculated using the sample regression line. The purpose of this is both to ascertain whether a linear regression model is *the most appropriate* model, as well as to judge the validity of several commonly required assumptions, including those required to obtain the estimates for α and β:

$$\mathrm{E}(\varepsilon_i) = 0 \qquad \text{and} \qquad \mathrm{V}(\varepsilon_i) = \sigma^2$$

Below is an illustration of a standard scatterplot showing a sample of bivariate data on the left, with a y on x regression line included, and a residual plot for the same data on the right. The data points highlighted with larger circles on the y against x scatterplot have their residuals similarly highlighted on the residual plot.

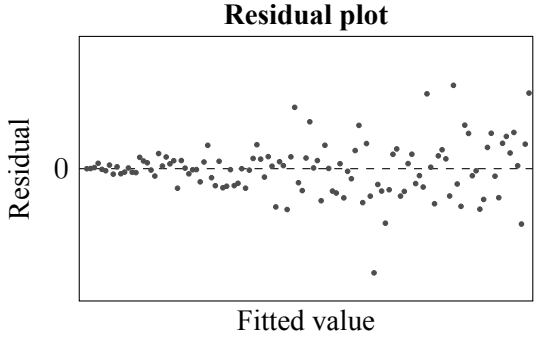


It should be observed that the residual plot broadly appears to show:

A random scatter of points...	
...centred on zero...	$\mathrm{E}(\varepsilon_i) = 0$
...with a constant variance.	$\mathrm{V}(\varepsilon_i) = \sigma^2$

This means that the residual plot would give no reason to doubt the assumptions made regarding the errors. This *ideal* residual plot would suggest that the linear regression model is **appropriate**.

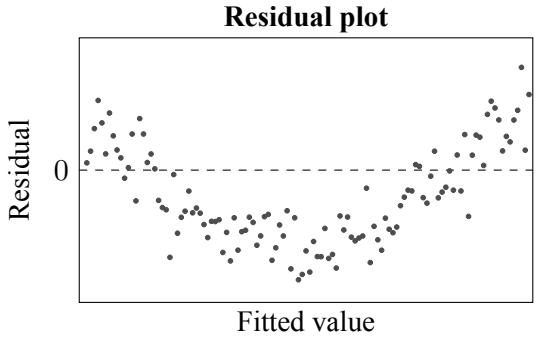
The following residual plots are accompanied with relevant comments and suitable conclusions. Note that if *any* of *"random scatter"*, *"centred on zero"* or *"constant variance"* are contradicted then the linear model is *not appropriate*.



The residual plot appears to show a 'funnel' shaped pattern of points centred on zero, with variance increasing as the fitted value increases.

$$\mathrm{E}(\varepsilon_i) = 0$$
$$\mathrm{V}(\varepsilon_i) \neq \sigma^2$$

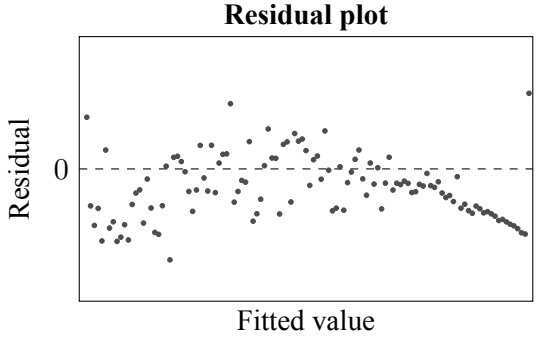
This suggests the that linear regression model proposed is **not appropriate**.



The residual plot appears to show a pattern of points following a curve, with constant variance.

$$\mathrm{E}(\varepsilon_i) \neq 0$$
$$\mathrm{V}(\varepsilon_i) = \sigma^2$$

This suggests that the linear regression model proposed is **not appropriate**.



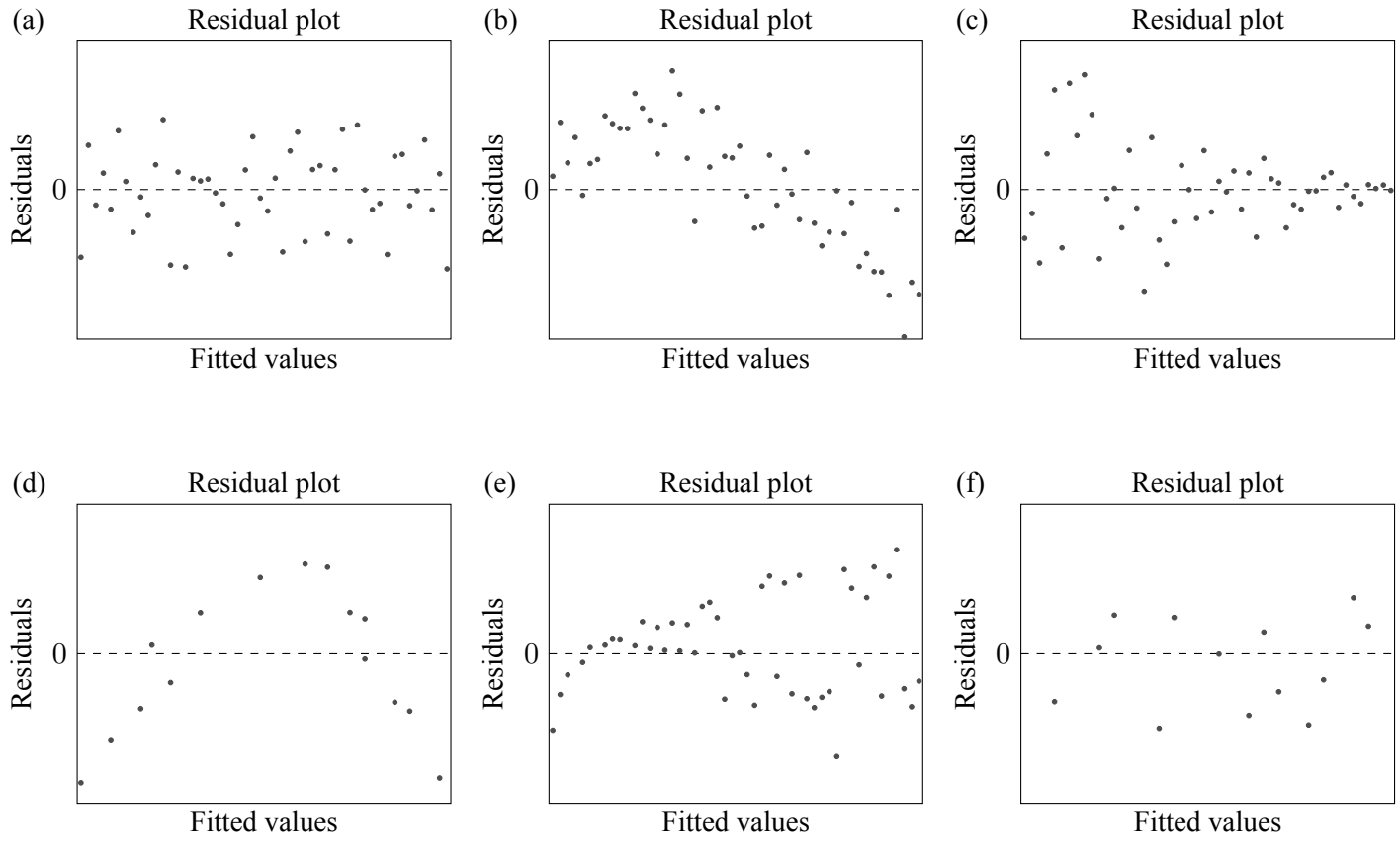
The residual plot appears to show a pattern of points following a curve, with variance decreasing as the fitted value increases.

$$\mathrm{E}(\varepsilon_i) \neq 0$$
$$\mathrm{V}(\varepsilon_i) \neq \sigma^2$$

This suggests that the linear regression model proposed is **not appropriate**.

Exercise 18G

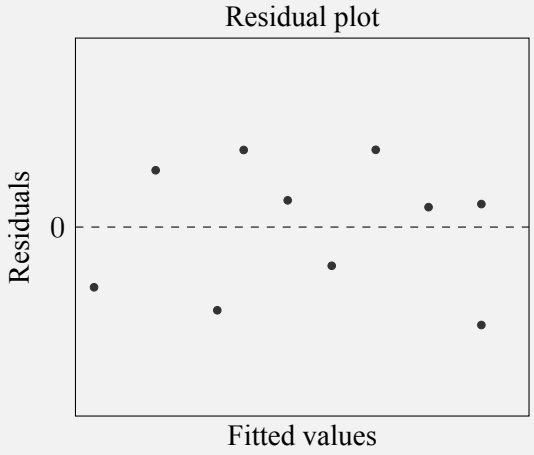
For each of the following residual plots, comment on the validity of the linear model from which they were produced.



Example 1

Problem: The mark scored in a Maths test (x) and an English test (y) is recorded for a random sample of pupils. A linear regression model between Maths mark and English mark is proposed, the equation of the y on x regression line obtained and a residual plot is constructed, shown below along with the tabulated data.

Pupil	Maths x_i	English y_i	Fit $\hat{Y}_i$	Residual $y_i - \hat{Y}_i$
A	1	3	6.98	-3.98
B	8	14	10.24	3.76
C	15	8	13.50	-5.50
D	18	20	14.90	5.09
E	23	19	17.23	1.77
F	28	17	19.56	-2.56
G	33	27	21.89	5.11
H	39	26	24.48	1.32
I	45	21	27.48	-6.48
J	45	29	27.48	1.52

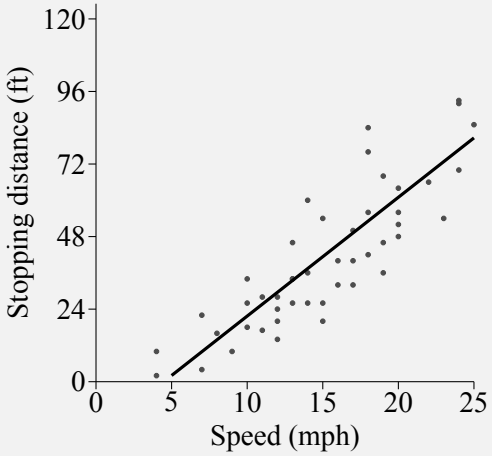


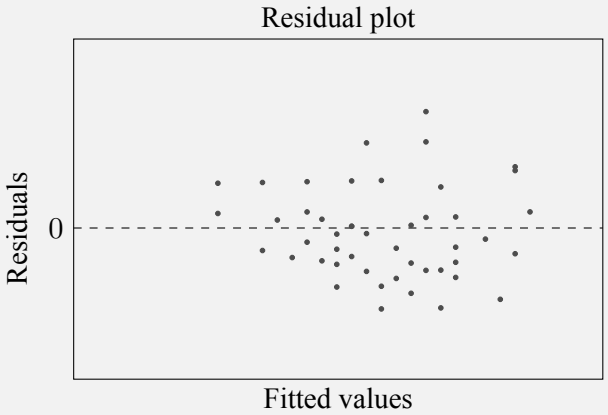
Use the residual plot to comment on the validity of the linear model.

Solution: The residual plot appears to show a random scatter of points centred on zero, and variance appears to be plausibly constant. The residual plot does not contradict any of the required assumptions for the errors, and so a linear model between English mark and Maths mark can be considered appropriate.

Example 2

Problem: A 1920s study on automobile safety involved recording the stopping distances (in feet) from selected speeds (in miles per hour) for a random sample of 47 cars. Below is a scatterplot of the results with a regression line obtained from a linear model, and a residual plot.



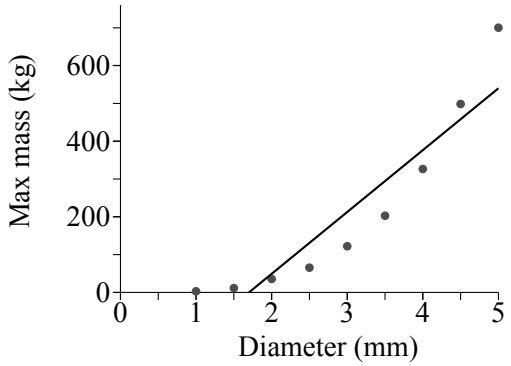


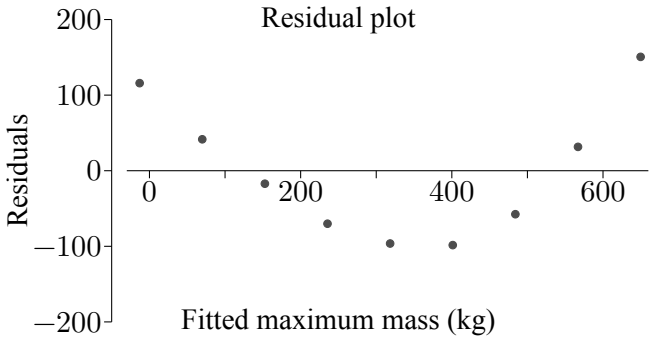
Use the residual plot to comment on the validity of the linear model.

Solution: The residual plot appears to show a random scatter of points centred on zero, with a broadly constant variance as the fitted value increases. This suggests a linear model between stopping distance and speed could be appropriate.

Exercise 18H

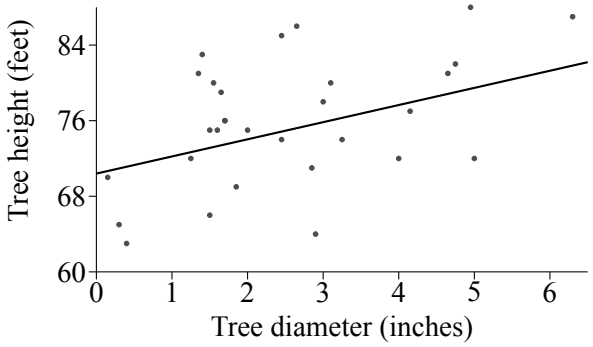
1. Metal cables are tested under laboratory conditions to determine their breaking strength. Each cable is fixed between two points, and masses attached to the middle of the cable, with the mass steadily increased until the cable breaks. The scatterplot below shows the cable diameter (x mm) against the maximum mass held before breaking (y kg).

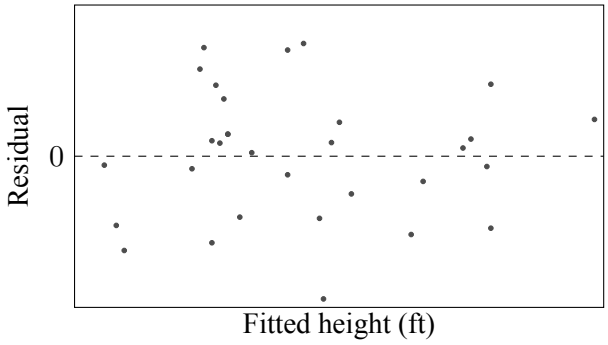




Comment on the suitability of the linear model fitted.

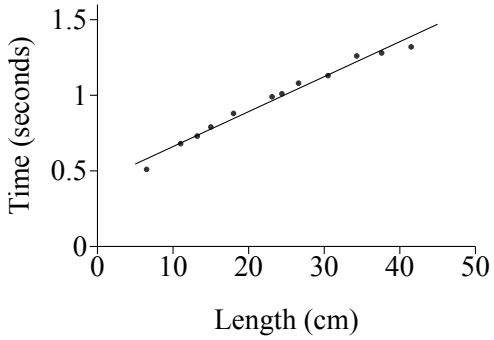
2. 31 randomly selected black cherry trees have their diameter at a set distance from the ground measured, in inches, with the height of each tree also measured, in feet. The equation of the least squares regression line for y on x is obtained, and shown on the scatterplot below alongside a residual plot.

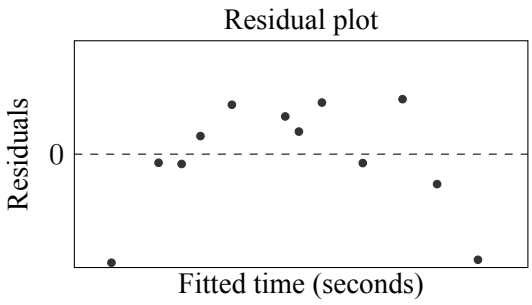




Explain why the residual plot suggests that a linear model would be suitable for the data.

3. A student investigating the relationship between the length of a pendulum and the time taken for one complete swing to occur made a simple pendulum from string and a metal disc, suspended from a clamp over the edge of a table. The time taken for one complete swing (y seconds) was recorded for 12 different string lengths (x cm).





Comment on the suitability of the linear model fitted.

18.10 Confidence and Prediction Intervals

One purpose for adopting a regression approach to examine the relationship between two variables is the wish to predict a value for the response variable (Y) given a particular value for the predictor (x).

The least squares regression line equation of $\hat{Y}_i = a + bx_i$ gives an *estimate* for the **mean** value of the response variable for any given x_i, whilst the *true* mean response value for a given x_i, which is denoted $\mathrm{E}(Y_i \mid x_i)$, is unknown.

It can be desirable to obtain a **confidence interval** for $\mathrm{E}(Y_i \mid x_i)$. It can be shown that the *point estimate*, $\hat{Y}_i$, has a sample standard error of:

$$s\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}}$$

If the errors, ε_i, are identically and independently distributed as $\mathrm{N}\left(0, \sigma^2\right)$ then the formula below, given on Page 5 of the Data Booklet, can be derived.

A confidence interval for a mean response value

A $100(1 - \alpha)\%$ confidence interval for $\mathrm{E}(Y_i \mid x_i)$ is given by:

$$\hat{Y}_i \pm t_{n-2,1-\alpha/2} \times s\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}}$$

As with confidence intervals, the level of confidence relates to the success rate of the *procedure used to generate the interval*, rather than the probability of a calculated interval containing the mean response value $\mathrm{E}(Y_i \mid x_i)$.

Example

Problem: The lengths in centimetres (x) and masses in kilograms (y) of 12 salmon were measured, with the results:

$$\Sigma x = 30 \quad \Sigma y = 540 \quad S_{xx} = 12 \quad S_{yy} = 4800 \quad S_{xy} = 150$$

Construct a 99% confidence interval for the mean length of salmon with mass 3kg.

Solution:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{150}{12} = 12.5, \quad a = \bar{y} - b\bar{x} = \frac{540}{12} - 12.5 \times \frac{30}{12} = 13.75 \quad \Longrightarrow \quad \hat{Y}_i = 13.75 + 12.5x_i$$

$$\hat{Y} = 13.75 + 12.5 \times 3 = 51.75$$

$$SSR = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = 4800 - \frac{150^2}{12} = 2925, \quad s^2 = \frac{SSR}{n-2} = \frac{2925}{10} = 292.5, \quad s = \sqrt{292.5} = 17.1$$

A 99% CI for $E(Y \mid x = 3)$ is $51.75 \pm 3.169 \times 17.1 \times \sqrt{\dfrac{1}{12} + \dfrac{(3 - 2.5)^2}{12}} = 51.75 \pm 17.49$

So a 99% confidence interval for the mean mass of a 3kg salmon is $(34.26\text{cm}\,,\ 69.24\text{cm})$.

A similar but subtly different goal to predicting a *mean* response value is predicting an **individual** response value for a given x_i, denoted as $Y_i \,|\, x_i$. In the context of the previous example, that would be predicting the length of an *individual* salmon. Since the value for a sampled individual can be expected to have more variability that the value of a sample mean, a wider interval around $\hat{Y}_i$ is needed. The required interval is called a **prediction interval**.

A prediction interval for an individual response value

A $100(1 - \alpha)\%$ prediction interval for $Y_i \,|\, x_i$ is given by:

$$\hat{Y}_i \pm t_{n-2,1-\alpha/2} \times s\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}}$$

This formula, also given on Page 5 of the Data Booklet, relies also on the errors, ε_i, being identically and independently distributed as $\text{N}\left(0, \sigma^2\right)$

Example

Problem: The lengths in centimetres (x) and masses in kilograms (y) of 12 salmon were measured, with the results:

$$\Sigma x = 30 \quad \Sigma y = 540 \quad S_{xx} = 12 \quad S_{yy} = 4800 \quad S_{xy} = 150$$

Construct a 95% prediction interval for the length of salmon with mass 4kg.

Solution:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{150}{12} = 12.5, \quad a = \bar{y} - b\bar{x} = \frac{540}{12} - 12.5 \times \frac{30}{12} = 13.75 \quad \Longrightarrow \quad \hat{Y}_i = 13.75 + 12.5x_i$$

$$\hat{Y} = 13.75 + 12.5 \times 4 = 64.25$$

$$SSR = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = 4800 - \frac{150^2}{12} = 2925, \quad s^2 = \frac{SSR}{n-2} = \frac{2925}{10} = 292.5, \quad s = \sqrt{292.5} = 17.1$$

A 95% PI for $Y \,|\, x = 3$ is $64.25 \pm 2.228 \times 17.1 \times \sqrt{1 + \frac{1}{12} + \frac{(4-2.5)^2}{12}} = 64.25 \pm 42.95$

So a 95% prediction interval for the mass of a 4kg salmon is $(21.3\text{cm}\,,\, 107.2\text{cm})$.

Exercise 18I

1. From Context 1: Construct a 95% confidence interval for the mean % mass loss of a metal plate in the acid solution for 150 hours.

2. From Context 2: Construct a 90% confidence interval for the mean carbohydrate content for a caffeinated drink with 400 calories.

3. From Context 3: Construct a 99% prediction interval for the body fat % for a patient with systolic blood pressure of 130.

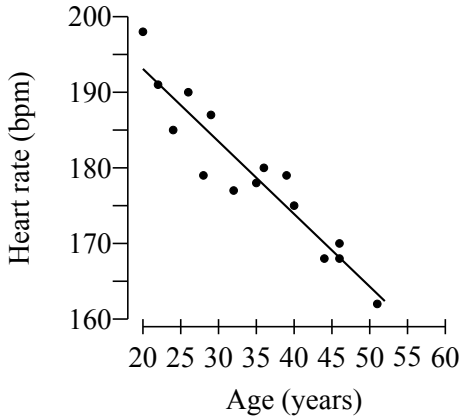
4. From Context 4: Construct a 95% prediction interval for the physics mark of a student who gained a maths mark of 74.

18.11 Transforming Data

Exercise 18J

Review Exercise

1. As part of a study on intensive exercise, a sports scientist recorded the peak heart rates (y) of a random selection of fifteen volunteers of different ages (x) who took regular exercise. The linear regression equation was calculated for the data shown in the scatter diagram and found to be $\hat{Y} = 212.3 - 0.96x$.

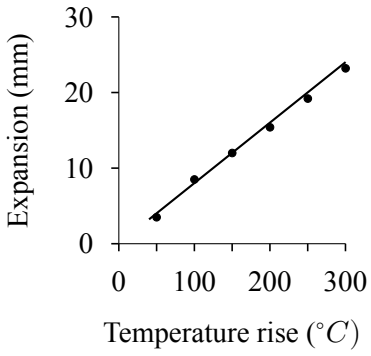


$\Sigma x = 518$ $\Sigma y = 2687$ $\Sigma xy = 91534$
$\Sigma x^2 = 19196$ $\Sigma y^2 = 482691$ $n = 15$

(a) Calculate the residual of a volunteer aged 32 with a heart rate of 177 bpm.

(b) Use a suitable hypothesis test to assess the usefulness of the linear model for prediction at the 1% level of significance.

2. When a type of metal bar is heated it expands. The amount by which it expands (e, in mm) and for a given increase in temperature (t, in °C) is recorded for a random sample of six such bars. A regression line with equation is $\hat{E} = 0.23 + 0.077t$ is obtained for the data, as well as summary statistics.

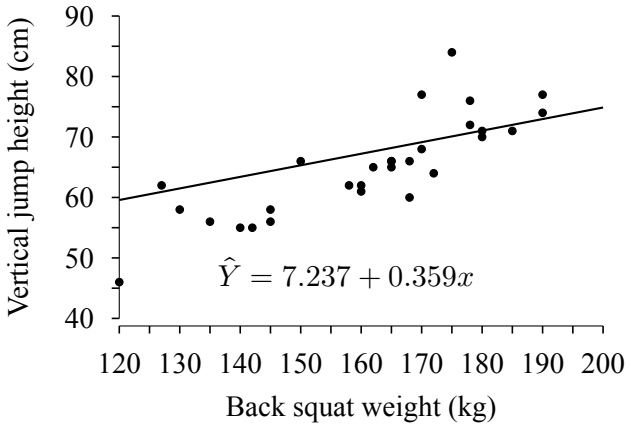


$\Sigma t = 1050$ $\Sigma e = 81.8$ $\Sigma te = 17665$
$\Sigma t^2 = 227500$ $\Sigma e^2 = 1372.54$ $n = 6$

(a) Calculate the coefficient of determination and interpret its value.

(b) Perform an appropriate hypothesis test at the 0.1% level of significance to provide evidence of a linear association between temperature rise and expansion.

3. A strength and conditioning coach wants to increase the vertical jump height performance in their trainees and considers whether back squat weight has an impact on vertical jump height performance. The summary statistics below are taken from the back squat weight (x in kg) and vertical jump height (y in cm) achieved by a random sample of 29 of the coach's trainees.



$\Sigma x = 4673$ $S_{xy} = 3630.45$ $n = 29$
$\Sigma x^2 = 763101$ $S_{yy} = 1899.45$

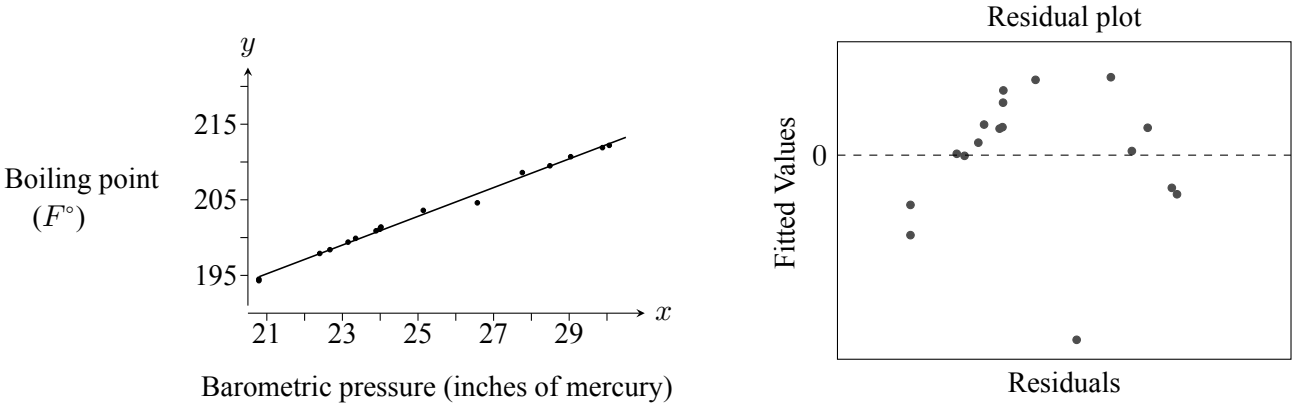
(a) Calculate an estimate for the error variance.

(b) Calculate a 95% prediction interval for vertical height for a traineee with a back squat weight of 172kg.

(c) Explain why a prediction interval is more likely to be of interest to the coach in this context than a confidence interval.

4. James Forbes was a Scottish physicist who, amongst other things, was interested in finding a way to estimate altitude from measurements of the boiling temperature of water. Some of his observations on boiling point (y, in F°) and barometric pressure (x, in inches of mercury) are summarised below.

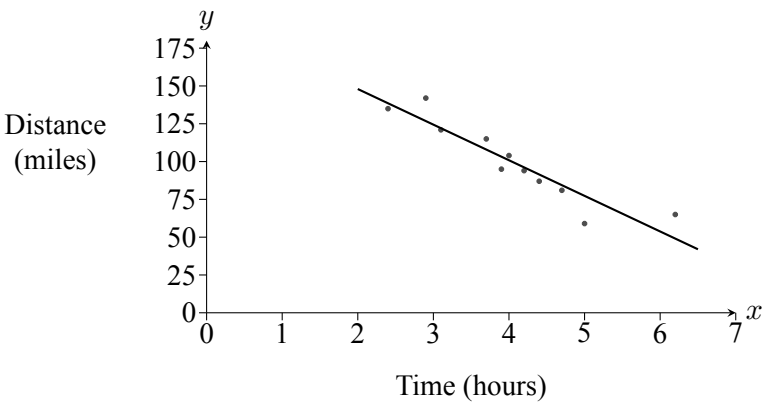
   After plotting a scatterplot for this data, a least squares regression line was obtained with the equation $\hat{Y} = 155.29 + 1.9018x$, from which a residual plot was also created.



   (a) Comment on the suitability of the linear model.

   (b) Suggest what next steps might be taken on the basis of the residual plot for the linear model.

5. A running website randomly selected eleven amateur runners who recently completed a marathon and asked them to complete a survey which included their finishing time in the marathon and information about their training routine. The scatterplot below shows, for each runner, their finishing time (x, in hours) and the total distance they ran in the final four weeks of their training (y, in miles).



Summary statistics:

$\sum x = 44.5$

$\sum x^2 = 191.21$

$\sum y = 1098$

$\sum y^2 = 116768$

$\sum xy = 4179.2$

$n = 11$

   A least squares regression line for y on x is shown on the graph, with equation $\hat{Y} = 194.8 - 23.48x$.

   (a) Calculate the correlation coefficient, and comment on its value.

   The website wishes to allow users the ability to input the number of miles in their training plan for the four weeks leading up to the rate, and receive in response an estimate for their finishing time.

   (b) Explain why the equation provided above is unsuitable for this purpose.

   (c) Obtain a suitable equation which could instead be used.

Answers

Chapter 1 Answers - Exploratory Data Analysis

Exercise 1.1

1. (a) Quantitative (b) Qualitative (c) Quantitative (d) Qualitative (e) Quantitative (f) Qualitative

2. (a) Discrete (b) Continuous (c) Discrete (d) Discrete (e) Continuous (f) Continuous

3. Categorical: *activity level* and *smoker*. Continuous: *height* and *mass*. Discrete: *pulse*.

Exercise 1.2

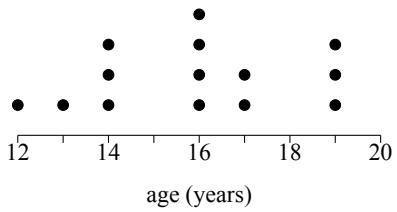
1. Population parameter (population mean) 2. Sample statistic (sample mean) 3. Sample statistic (sample mean)

Exercise 1.3

1. $\bar{x} = 35.9, s = 15.74$ 2. $\bar{f} = 5.8, s = 2.145$ 3. $\bar{x} = 29.125, s = 7.586$ 4. median $= 40.5$, IQR $= 27.5$

Exercise 1.4

1.



2.

stem-and-leaf diagram

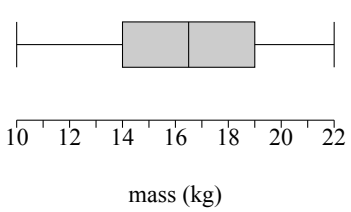
(ice cream cost)

0	90
1	40 50 50 70 80 90
2	00 10 20 50
3	60

$n = 12$

$2|50$ represents £2.50

3.



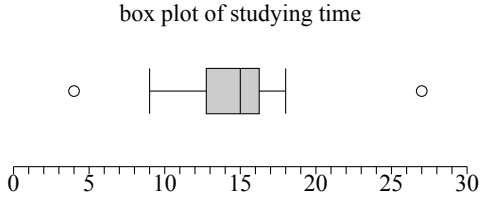
Exercise 1.5

1. $\frac{23}{94}$ or 0.2447
2. A clear key/legend linked to the plot, numbers/scale on the y axis.
3. The proportions of pupils in year group taking the bus seem to be broadly similar.
4. Approximately 15 minutes.
5. 7

Exercise 1.6

1. Lower fence $= 7.5$ and upper fence $= 51.5$. Since all values lie within this interval, there are no outliers.

(b)



2. (a) Lower fence $= 6.5$ and upper fence $= 22.5$. Since both 4 and 27 lie outwith this interval, they are outliers.

3. (a) Upper fence $= 26.5$. Since both circled points shown on the box plot are greater than 26.25, they are outliers.
   (b) If the lengths of their stays were incorrectly recorded in the hospital's records, it would be valid to remove them from the data set before conducting any further analysis. *Other valid reasons are possible.*

Exercise 1.7

1. The population of hazelnuts masses appear to follow a positively skewed distribution.
2. The population of tree heights appear to be normally distributed.
3. The population of waiting times appear to be uniformly distributed.
4. The population of number of successful passes appears to follow a bimodal distribution.

Exercise 1.8

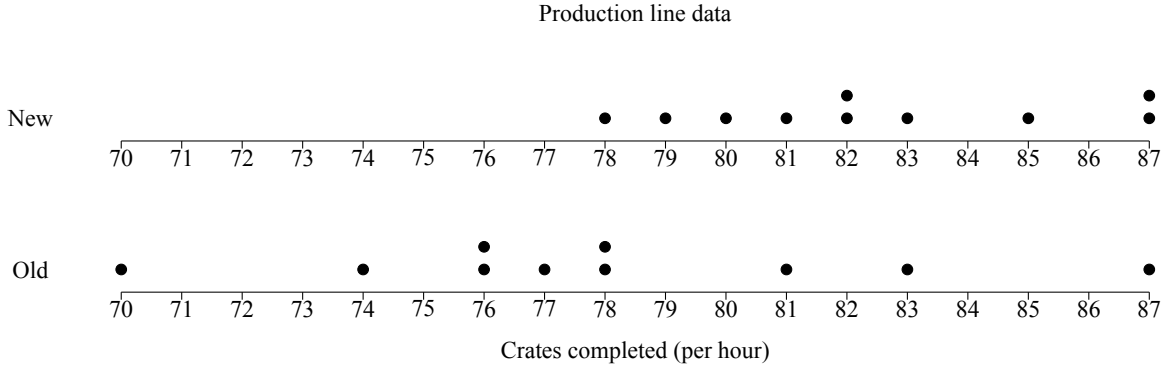
1. (a) $\bar{x}_A = 301.3, s_A = 4.57$
   $\bar{x}_B = 302.3, s_B = 12.3$

   (b) The median volume of coffee dispensed by Model B is greater.
   The amount of coffee dispensed by Model B is more varied.

2. The median score awarded by judge A was higher.
   The higher interquartile range for judge A shows that their scores were more varied.
3. (a) The daily 2pm temperature was more varied in January than in June, as shown by the slightly higher standard deviation.
   The median 2pm temperature in June was higher.

   (b) Consistent scales have not been used for the box plots, making the location and spread of each less easily comparable.
4. (a) A legend should be provided for the diagram, such as $3|2 = 3.2$ metres.

   (b) The stem-and-leaf diagram shows that trees in Woodland A were taller, on average.
5. (a)



   (b) $\bar{x}_{\text{new}} = 82.4, s_{\text{new}} = 3.13$
   $\bar{x}_{\text{old}} = 78, s_{\text{old}} = 4.76$

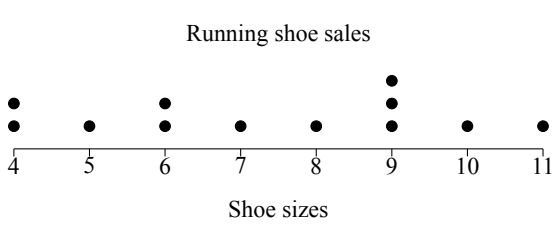
   (c) The mean number of crates completed per hours was higher for the new production line.
   The number of crates completed per hour was more consistent for the new production line.

Chapter 1 Review Exercise

1. (a) i. Track; artist. ii. Release year; no. of streams. iii. Length.

   (b) Fences could be calculated to determine whether it is an outlier.

2. (a)



   (b) mean $= 7.33$ and sd $= 2.35$

   (c) Shoe sizes are discrete data.

3. (a) The fences are 1.13 and 1.85. Since both the minimum value of 1.00 and the maximum of 1.90 lie outwith this interval, the *iris setosa* data contains at least two outliers.

   (b) The median length of *iris versicolor* petals appears to be greater than the length of *iris setosa* petals. The length of *iris versicolor* petals appears to be more varied than those of the *iris setosa* species, as seen by their interquartile ranges of 0.18 and 0.60 centimetres respectively.

Chapter 2 Answers - An Introduction to Probability Theory

Exercise 2.1

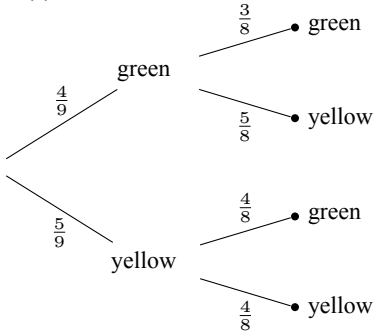
1. (a) P(vowel) $= \dfrac{5}{26}$
   (b) P(consonant) $= \dfrac{21}{26}$
   (c) P(statistics) $= \dfrac{5}{26}$
2. (a) P(blue) $= \dfrac{4}{15}$
   (b) P(green or red) $= \dfrac{11}{15}$
   (c) P(not red) $= \dfrac{3}{5}$
3. (a) P(orange) $= \dfrac{1}{3}$
   (b) P(orange and strawberry) $= 0$
   (c) P(not caramel) $= \dfrac{7}{12}$
4. (a) P(Scotland) $= \dfrac{2}{19}$
   (b) P(Wales or NI) $= \dfrac{13}{76}$
   (c) P(not England) $= \dfrac{21}{76}$
5. (a) P(blue and 5) $= \dfrac{1}{18}$
   (b) P(5) $= \dfrac{1}{9}$
   (c) P(blue or 5) $= \dfrac{7}{18}$
6. (a) P(G) $= \dfrac{9}{10}$
   (b) P(G and A) $= \dfrac{27}{80}$
   (c) P(not A) $= \dfrac{5}{8}$
7. (a) P(nurse) $= \dfrac{25}{64}$
   (b) P(not ward 2) $= \dfrac{39}{64}$
   (c) P(ward 2 and not doctor) $= \dfrac{21}{64}$

Exercise 2.2

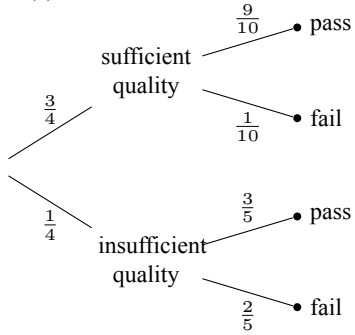
1. (a) {H1,H2,H3,H4,H5,H6
   T1,T2,T3,T4,T5,T6}
   (b) P(even and tails) $= \dfrac{1}{4}$
2. (a) P(tea and toast) $= \dfrac{1}{12}$
   (b) P(museli and not coffee) $= \dfrac{1}{6}$
   (c) P(not porridge and not tea) $= \dfrac{1}{2}$
3. (a) P(tails twice in a row) $= \dfrac{3}{8}$
   (b) P(tails at least twice) $= \dfrac{1}{2}$
4. P(sum $= 3$) $= \dfrac{1}{20}$
5. P(product < 5) $= \dfrac{2}{9}$
6. P(at least one number $\leqslant 2$) $= \dfrac{2}{3}$
7. (a) P(20p first) $= \dfrac{1}{5}$
   (b) P(20p selected) $= \dfrac{2}{5}$
   (c) P(total $\geqslant 15$p) $= \dfrac{1}{2}$
   (d) P(total < 12p) $= \dfrac{2}{5}$
8. (a) 56
   (b) i. P(A:whiteboard and C:pens) $= \dfrac{1}{56}$
   ii. P(neither chosen) $= \dfrac{15}{28}$
   iii. P(at least one chosen) $= \dfrac{13}{28}$

Exercise 2.3

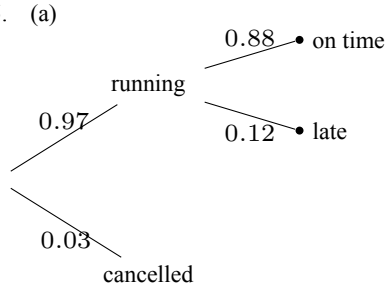
1. (a)



   (b) P(both green) $= \frac{1}{6}$
   (c) P(at least one green) $= \frac{13}{18}$
2. P(both same colour) $= \frac{9}{19}$
3. P(neither O_+) $= 0.4281$
4. (a)

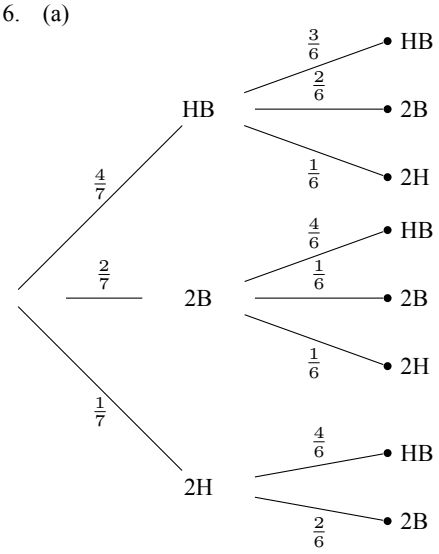


   (b) P(both pass) $= 0.825$
5. (a)



   (b) P(on time) $= 0.8536$

6. (a)



   (b) P(both same) $= \frac{1}{3}$
   (c) P(different) $= \frac{2}{3}$
   (d) P(neither HB) $= \frac{1}{7}$

7. (a) P(no stops) $= 0.224$
   (b) P(at least one stop) $= 0.776$
   (c) P(exactly two stops) $= 0.252$

Exercise 2.4

1. (a) P$(A) = \frac{5}{9}$
   (b) P$(B) = \frac{4}{9}$
   (c) P$(\bar{A}) = \frac{4}{9}$
   (d) P$(\bar{B}) = \frac{5}{9}$
   (e) P$(A \cap B) = \frac{1}{3}$
   (f) P$(A \cup B) = \frac{2}{3}$
   (g) P$(\bar{A} \cap B) = \frac{1}{9}$
   (h) P$(A \cap \bar{B}) = \frac{2}{9}$
2. (a) $x = 0.35$
   (b) i. P$(N) = 0.55$
   ii. P$(\bar{M}) = 0.5$
   iii. P$(M \cup N) = 0.7$
3. (a) P(prime) $= \frac{3}{8}$
   (b) P(odd) $= \frac{1}{2}$
   (c) P($\overline{\text{prime}}$) $= \frac{5}{8}$
   (d) P($\overline{\text{odd}}$) $= \frac{1}{2}$
   (e) P(prime $\cap$ odd) $= \frac{5}{16}$
   (f) P(prime $\cup$ odd) $= \frac{9}{16}$
   (g) P($\overline{\text{prime}} \cap$ odd) $= \frac{3}{16}$
   (h) P(prime $\cap \overline{\text{odd}}$) $= \frac{1}{16}$
4. (a) P$(C \cap D) = 0.34$
   (b) P$(C \cup D) = 0.91$
   (c) P$(\bar{C} \cap \bar{D}) = 0.09$
   (d) P$(\bar{C} \cup \bar{D}) = 0.66$
   (e) P$(C \cap \bar{D}) = 0.45$
   (f) P$(\bar{C} \cap D) = 0.12$
   (g) P$(C \cup \bar{D}) = 0.88$
   (h) P$(\bar{C} \cup D) = 0.55$
5. P(biology) $= \frac{13}{24}$
6. P(PM $\cap$ NO_2) $= \frac{1}{10}$

Exercise 2.5

1. P$(X \cup Y) = 0.6$
2. P$(A \cup B) = \frac{29}{30}$
3. P$(E \cup F) = 0.61$
4. P$(C \cup D) = \frac{11}{14}$
5. P$(A \cap B) = 0.163$
6. P$(X \cap Y) = \frac{9}{80}$
7. P$(V \cap W) = 0.2$
8. P$(\bar{J}) = \frac{7}{72}$
9. P$(X \cup Y) = 0.9$
10. P$(\bar{B}) = \frac{1}{6}$
11. P(fish) $= 0.85$
12. P$(\bar{N}) = \frac{3}{8}$
13. P$(\bar{R}) = 0.19$
14. (a) P$(X \cup Y) = 0.9 \neq 1$
    $\therefore X, Y$ not exhaustive
    (b) P$(R \cup T) = 1$
    $\therefore R, T$ exhaustive
    (c) P$(E \cup F) = 0.85 \neq 1$
    $\therefore E, F$ not exhaustive
15. (a) P$(W \cup V) = 0.3 \neq 0$
    $\therefore W, V$ not mutually exclusive
    (b) P$(X \cap Y) = 0$
    $\therefore X, Y$ mutually exclusive
    (c) P$(A \cap B) = 0$
    $\therefore A, B$ mutually exclusive

Chapter 2 Review Exercise

1. (a) P($\overline{\text{yellow}} \cup 1$) $= \frac{2}{3}$
   (b) P(neither yellow) $= \frac{5}{18}$
2. (a) P(English $\cap$ N5) $= \frac{21}{64}$
   (b) P(History $\cup$ Higher) $= \frac{43}{64}$
3. P$(F) = 0.9$
4. P(takes payment) $= 0.8926$
5. (a) P(not a tie) $= \frac{5}{6}$
   (b) P(attacker wins) $= \frac{7}{12}$
   (c) P(bonus point) $= \frac{1}{8}$
6. P$(N \cup \bar{A}) = \frac{7}{25}$
7. P$(B \cap C) = 0.2$

Chapter 3 Answers - Sampling Theory

Note: Whilst the answers included here are intended to include an appropriate level of detail, variations in the amount and type of information given in exam questions will require a flexible and common-sense approach to be used in any exam. Past exam questions and marking schemes should be used as a guide, as well as course reports and relevant CPD events.

Exercise 3.1

1. (a) i. People who use his local supermarket.
   ii. All shoppers who use the self-checkout tills on that Sunday morning.
   (b) The sampling frame excludes people who prefer to shop at other times, such as during the evening or midweek. It also only includes those using self-checkout tills, likely excluding those who have a large trolly of shopping. Both of these issues are likely to make the sample unrepresentative of the target population, and may introduce bias.
   (c) The sampling methods is non-random and so will likely introduce bias to the data, making it unsuitable.
2. (a) There may be so many flowers that it is too time-consuming and impractical to make a sampling frame comprised of individual flowers.
   (b) A better choice of sampling unit for the sampling frame may be square metres of field, forming a numbered grid.
   (c) A random number generator will help avoid possible bias in the selection of square metres to include in the sample, by ensuring it is random and not influenced by the botanist. It should help produce a sample that is more representative of the population of wildflowers.

Exercise 3.2

1. (a) One-stage cluster sampling.
   (b) Stratified sampling.
   (c) Systematic sampling.
   (d) Simple random sampling.
   (e) Quota sampling.
   (f) Convenience Sampling.
2. (a) Each of the different types of room may have their own distinct characteristics in terms of ageing or the quality of maintenance. Stratified sampling will ensure that a proportionate number from each type of room will be included in the sample.
   (b) 30 out of 360 cabins means 1 in every 12 cabins should be sampled, so the sample should contain: 2 Deluxe cabins; 6 Premium cabins and 22 Standard cabins. To sample the Deluxe cabins, a numbered list of those cabins should be obtained as a suitable sampling frame, and then a random number generator used to produce 2 different random numbers. Those cabins corresponding to the numbers selected should be included in the sample. This process should then be repeated for the Premium cabins and the Standard cabins, using a numbered sampling frame for each and randomly selecting the appropriate number of cabins of each type. Combining all of the selected cabins will make a stratified sample.
3. (a) Convenience sampling.
   (b) He only samples coral reefs close to his town, ignoring that those near other towns or villages, or rural locations, may have a different level of health. This means the sample is unlikely to be representative of the the population of coral reefs along the West Coast of Scotland. (Alternatively, the non-random selection method may be discussed.)
4. (a) Using cluster sampling means less physical travelling to a larger number of petrol stations, which would cost more and take more time.
   (b) To sample 40 pumps, 5 petrol stations are needed if one-stage cluster sampling is used. A sampling frame of all the petrol stations numbered from 1 to 126 should be drawn up. A random number generator can then be used to select 5 numbers, and the petrol stations corresponding to those numbers should be included in the sample. From each of those petrol stations, all of the pumps should be sampled.
5. (a) Systematic sampling.
   (b) She didn't use a random starting point, so it is not a true random sampling process that is being used.
6. (a) Quota sampling.
   (b) This is a non-random sampling method which is liable to produce a sample not representative of the target population, and may introduce bias into the data gathered.

Chapter 3 Review Exercise

1. (a) i. Convenience sampling.
   ii. This is a non-random sampling method which may produce a sample that is not representative of the target population, and may introduce bias to the data.
   (b) i. Simple random sampling.
   ii. With the 150 employees in an ordered list, the random cell selection tool could be used to select the first employee to be included in the sample. From there, every 5th employee on the list could be selected, looping back around to the start of the list if needed until the sample of size 30 has been obtained.
   (c) i. A numbered list of the restaurants in the chain could be obtained. A random number generator could be used to pick several numbers, and those restaurants corresponding to the numbers could be selected to be sampled. For a one-stage cluster sample, all employees at those restaurants should be sampled.
   ii. One advantage of this method is that it will require less restaurants to be visted for an equivalent sample size, reducing travel costs. One disadvantage is that visiting less restaurants may make it more likely than any restaurants in the chain experiencing problems have none of their employees included in the sample.
2. For an overall sample size of 150 out of 5000 (3%), the sample needs to include 3 people from Management, 12 from Sales and 135 from the Shop Floor. For each strata, an ordered sampling frame should be obtained, and numbers assigned. For the Management sample, a random number generator should be used to choose 3 numbers, and those staff corresponding to those numbers should be selected to be in the sample. The same process should be repeated to choose 12 from the Sales samling frame, and 135 from the Shop Floor sampling frame. Put together, this will make a sample of size 150 with proportions from each strata matching those of the target population.
3. The pizzas sampled will only be those from takeaway restaurants, and may not be representative of pizzas sold in the UK as a whole. Additionally, only those takeaways with a presence on social media will be included.
4. (a) It is unlikely to be possible for the researcher to obtain a list of all pupils in Scotland studying Higher Mathematics.
   (b) From an ordered list of all centres in Scotland which presents pupils for Higher Maths, simple random sampling could be used to select a number of schools. For each of those schools selected, an ordered list of all pupils taking Higher Maths should be obtained, and used to selected a random number of pupils from each selected school. Each of those pupils selected should be included in the overall sample.
5. A random letter selector and a random number generator could be used together to select a letter and a number, to correspond to grids, such as ’C’ and ’5’ to select grid C5. This process could be repeated to create a list of grids to for the botanist to visit and study the health of flowers in those grids.

Chapter 4 Answers - Further Probability Theory

Exercise 4.1

1. (a) $\mathrm{P}(X) = \frac{5}{9}$
   (b) $\mathrm{P}(X|Y) = \frac{3}{4}$
   (c) $\mathrm{P}(X \cap Y) = \frac{1}{3}$
   (d) $\mathrm{P}(X \cup Y) = \frac{2}{3}$
   (e) $\mathrm{P}(Y) = \frac{4}{9}$
   (f) $\mathrm{P}(Y|X) = \frac{3}{5}$
   (g) $\mathrm{P}(Y \cap X) = \frac{1}{3}$
   (h) $\mathrm{P}(Y \cup X) = \frac{2}{3}$
2. (a) $\mathrm{P}(\text{scored}) = \frac{73}{135}$
   (b) $\mathrm{P}(\text{scored} \mid \text{3-pointer}) = \frac{15}{41}$
   (c) $\mathrm{P}(\text{3-pointer}) = \frac{41}{135}$
   (d) $\mathrm{P}(\text{3-pointer} \mid \text{scored}) = \frac{15}{73}$
3. (a) P(rainy)$= 0.3$
   (b) P(not rainy)$= 0.7$
   (c) P(leaves | rainy)$= 0.4$
   (d) P(does not leave | rainy)$= 0.6$
   (e) P(leaves | not rainy)$= 0.9$
   (f) P(does not leave | not rainy)$= 0.1$
4. (a) $\mathrm{P}(A) = \frac{2}{5}$
   (b) $\mathrm{P}(B) = \frac{2}{5}$
   (c) $\mathrm{P}(C) = \frac{1}{3}$
   (d) $\mathrm{P}(A \cap B) = \frac{2}{15}$
   (e) $\mathrm{P}(B \cup C) = \frac{3}{5}$
   (f) $\mathrm{P}(A \cap B \cap C) = \frac{1}{15}$
   (g) $\mathrm{P}(A|C) = \frac{3}{5}$
   (h) $\mathrm{P}(C|B) = \frac{1}{3}$
   (i) $\mathrm{P}(\overline{B}|C) = \frac{3}{5}$
   (j) $\mathrm{P}(C|\overline{A}) = \frac{2}{9}$
   (k) $\mathrm{P}((A \cap B)|C) = \frac{1}{5}$

Exercise 4.2

1. (a) $\mathrm{P}(A|B) = 0.2$
   (b) $\mathrm{P}(\text{red} \mid \text{blue}) = 0.4849$
   (c) $\mathrm{P}(X|Y) = \frac{2}{3}$
   (d) $\mathrm{P}(M|N) = 0.3233$
   (e) i. $\mathrm{P}(\overline{A}|B) = 0.8$
   ii. $\mathrm{P}(\overline{\text{red}} \mid \text{blue}) = 0.5151$
   iii. $\mathrm{P}(\overline{X}|Y) = \frac{1}{3}$
   iv. $\mathrm{P}(\overline{M}|N) = 0.6767$
2. $\mathrm{P}(\overline{V}|W) = 0.3$
3. $\mathrm{P}(5 \mid \text{yellow}) = 0.2$
4. $\mathrm{P}(\text{lose} \mid \text{concede} \geq 2) = \frac{5}{8}$
5. $\mathrm{P}(M|N) = 0.25$
6. $\mathrm{P}(V \cup W) = 0.725$

Exercise 4.3

1. (a) Independent.
   (b) Not independent.
2. (a) Not independent.
   (b) Independent.
   (c) Independent.
3. (a) Show e.g. $\mathrm{P}(V) = \mathrm{P}(V|W)$
   (b) Show e.g. $\mathrm{P}(T) \neq \mathrm{P}(T|W)$

Exercise 4.4

1. (a) $\mathrm{P}(X \cap Y) = 0.072$
   (b) $\mathrm{P}(\text{blue} \cap \text{yellow}) = \frac{3}{8}$
   (c) $\mathrm{P}(E \cap F) = 0.246$
2. (a) $\mathrm{P}(A \cap B) = 0.392$
   (b) $\mathrm{P}(A \cup B) = 0.708$
3. $\mathrm{P}(X \cup Y) = 0.88$
4. $\mathrm{P}(\text{working} \cap \text{Calculo}) = 0.12$
5. $\mathrm{P}(\text{out-of-date} \cap \text{weak}) = 0.096$
6. $\mathrm{P}(\text{blue} \cap \text{pink}) = 0.38$

Exercise 4.5

1. (a) $\mathrm{P}(X|Y) = \frac{7}{12}$
   (b) $\mathrm{P}(V|W) = 0.625$
2. $\mathrm{P}(1 \mid \text{red}) = 0.6$
3. $\mathrm{P}(\text{gene} \mid \text{positive}) = 0.2186$
4. (a) $\mathrm{P}(M|N) = \frac{8}{15}$
   (b) $\mathrm{P}(\overline{M}|N) = \frac{7}{15}$
5. $\mathrm{P}(\text{pass} \mid \text{car}) = 0.8446$
6. $\mathrm{P}(\text{Pacific 525} \mid \text{music}) = 0.7467$
7. $\mathrm{P}(\overline{E}|F) = 0.1946$

Exercise 4.6

1. P(UK) $= 0.8708$
2. P(survives) $= 0.6745$
3. P(win) $= 0.695$
4. P(likes) $= 0.54$
5. P(not make second part) $= 0.28$
6. P(do not survive) $= 0.268$
7. P(win) $= 0.3882$

Exercise 4.7

1. (a) P(damaged) $= 0.0655$
   (b) P(express | damaged) $= 0.1069$
2. (a) P(flagged) $= 0.0308$
   (b) P(malicious | flagged) $= 0.3584$
3. P(guessed | wrong) $= 0.7273$
4. (a) P(visits) $= 0.033$
   (b) P(social media | visits) $= 0.8409$
5. P(lost | rest day) $= 0.0559$
6. P(flaw | doesn't leave) $= 0.8048$
7. (a) P(has condition | positive) $= 0.4580$
   (b) P(has condition | positive test) $= 0.4975$

Chapter 4 Review Exercise

1. P$(E|F) = \dfrac{7}{9}$
2. (a) i. P$(W \cup A) = \dfrac{17}{20}$
   ii. P$(\overline{W} \cap B) = \dfrac{1}{8}$
   (b) P(power) $= 0.825$
   (c) P(Mac | $\overline{\text{power}}$) $= \dfrac{1}{7}$
3. (a) P(wins) $= 0.3917$
   (b) P(1st | lost) $= 0.1370$
4. P$(M|B) = \dfrac{1}{5} =$ P(M)
   hence P, M are independent
5. (a) P(incorrect) $= 0.2875$
   (b) P(cat | incorrect) $= 0.2348$

1. Hi

1a) $0.9 \implies$ not exhaustive

1b) $1 \implies$ exhaustive

1c) $0.15 \implies$ not exhaustive

1d) $\frac{5}{6} \implies$ not exhaustive

2a) $0.3 \implies$ not mutually exclusive

2b) $\frac{3}{5} \implies$ not mutually exclusive

2c) $0.6 \implies$ not mutually exclusive

2d) $0 \implies$ mutually exclusive

3a) 0.64

3b) 0.82

3c) 0.56

3d) 0.9

4) $\frac{47}{72}$

5) 0.9

6) $\frac{1}{6}$

7) 0.85

8) $\frac{5}{8}$

9) 1

10a) 0.89

10b) 0.05

10c) 0.19

11) $P(E \cup F) = \frac{71}{60} > 1$ Probabilities cannot be greater than 1.

1b $\frac{1}{6}$

1c $\frac{13}{18}$

2 $\frac{9}{19}$

3 0.4281

4b 0.825

5b 0.8536

6b $\frac{1}{3}$

6c $\frac{2}{3}$

6d $\frac{1}{7}$

7a 0.224

7b 0.776

7c 0.252

1a) $0.9 \implies$ not exhaustive

1b) $1 \implies$ exhaustive

1c) $0.15 \implies$ not exhaustive

1d) $\frac{5}{6} \implies$ not exhaustive

2a) $0.3 \implies$ not mutually exclusive

2b) $\frac{3}{5} \implies$ not mutually exclusive

2c) $0.6 \implies$ not mutually exclusive

2d) $0 \implies$ mutually exclusive

3a) 0.64

3b) 0.82

3c) 0.56

3d) 0.9

4) $\frac{47}{72}$

5) 0.9

6) $\frac{1}{6}$

7) 0.85

8) $\frac{5}{8}$

9) 1

10a) 0.89

10b) 0.05

10c) 0.19

11) $P(E \cup F) = \frac{71}{60} > 1$ Probabilities cannot be greater than 1.

1a)i) $\frac{5}{9}$

1a)ii) $\frac{4}{9}$

1a)iii) $\frac{3}{4}$

1a)iv) $\frac{3}{5}$

1a)v) $\frac{1}{3}$

1b) $\frac{3}{4}$

2a) 0.3616

2b) 0.2044

3a) 0.7

3b) 0.3

4) 0.4

5) 0.25

6a)i) $\frac{2}{15}$

6a)ii) $\frac{4}{5}$

6a)iii) $\frac{4}{15}$

6a)iv) $\frac{2}{3}$

6a)v) $\frac{1}{5}$

6b)i) $\frac{2}{7}$

6b)ii) $\frac{4}{7}$

6b)iii) $\frac{4}{7}$

6b)iv) $\frac{3}{7}$

6b)v) $\frac{3}{8}$

7) 0.725

8) $P(\text{express}) = 0.35$
$P(\text{regular}) = 0.65$
$P(\text{undamaged}|\text{express}) = 0.98$
$P(\text{undamaged}|\text{express}) = 0.91$

Chapter 5 - An Introduction to Linear Regression

1. Hi

Chapter 6 - Random Variables

1. Hi

Chapter 7 - Discrete Distributions

Exercise 6A

1. $E(X) = 3$ and $V(X) = 2$
2. $E(X) = 8$ and $V(X) = \frac{56}{3}$
3. (a) $E(X) = 4.5$ and $V(X) = 5.25$
   (b) $E(2X + 5) = 14$ and $V(2X + 5) = 21$
4. (a) $X \sim U(20)$
   (b) $E(X) = 10.5$ and $V(X) = 33.25$
   (c) $E(1 - 3X) = -30.5$ and $V(1 - 3X) = 299.25$
5. (a) $E(X - Y) = -1$
   (b) $V(X - Y) = \frac{8}{3}$
6. (a) $k = 10$
   (b) $P(2 \leqslant X < 6) = 0.4$
   (c) $SD(X) = \frac{\sqrt{33}}{2}$
7. (a) $k = 8$
   (b) $E(1 - 2Y) = -8$
8. $X \sim U(11)$ and $Y \sim U(7)$

Exercise 6B

1. 0.2503
2. 0.2787
3. 0.0579
4. 0.0988
5. (a) $X \sim B(4, \frac{1}{7})$
   (b) 0.3599
6. 0.0214

Exercise 6C

1. (a) 0.9894
   (b) 0.0367
   (c) 0.9527
   (d) 0.0106
   (e) 0.0473
2. (a) 0.9527
   (b) 0.2235
   (c) 0.5756
3. (a) 0.7414
   (b) 0.0003
   (c) 0.4829
4. (a) 0.1468
   (b) 0.5033
   (c) 0.1927

Exercise 6D

1. (a) $\mu = 0.8, \sigma^2 = 0.64$
   (b) $\mu = 2, \sigma^2 = \frac{4}{3}$
   (c) $\mu = 0.4, \sigma^2 = 0.36$
   (d) $\mu = 12, \sigma^2 = 7.2$
2. (a) $S \sim B(10, \frac{1}{4})$
   (b) $E(S) = 2.5$ and $V(S) = 1.875$
   (c) 0.0197
3. (a) $p = 0.2$
   (b) $SD(X) = 0.96$
4. (a) $p = \frac{2}{3}$
   (b) $E(Y) = 8$
5. (a) $n = 20$ and $p = \frac{4}{5}$
   (b) $P(5 < X < 10) = 0.0006$

Exercise 6E

1. (a) 0.1336
   (b) 0.0821
   (c) 0.2873
   (d) 0.7127
2. (a) $X \sim Po(4)$
   (b) 0.0183

Exercise 6F

1. (a) 0.01512
   (b) 0.5543
   (c) 0.0620
   (d) 0.3937
2. (a) 0.02231
   (b) 0.0611
   (c) 0.4412
   (d) 0.1912
3. (a) $X \sim Po(1)$
   (b) 0.0190

Exercise 6G

1. 0.9473
2. 0.9389
3. 0.2661
4. 0.1353
5. 0.0144

Review Exercise

1. (a) $\frac{5}{8}$
   (b) $E(X) = 4.5$ and $V(X) = 5.25$
2. (a) 0.1954
   (b) 0.2071
3. (a) 0.1294
   (b) 0.4769
4. 0.4335
5. (a) 0.8298
   (b) $\mu = 17$ and $\sigma = 1.597$
   (c) The germination of the seeds is independent.
6. (a) 0.6703
   (b) 0.1954
7. (a) $P(M|V) = 0.2$
   (b) $X \sim B(8, 0.3)$
   (c) 0.0580
8. (a) 0.2
   (b) 0.1528
   (c) 0.0267

Chapter 8 Answers - Continuous Distributions

Exercise 8.1

1. (a) $\frac{2}{3}$
   (b) 0.7667
   (c) 0.1
2. (a) $\frac{1}{3}$
   (b) 0
   (c) 0.6583
3. (a) $\frac{1}{6}$
   (b) $\frac{5}{6}$
   (c) $\frac{5}{12}$
4. (a) $E(X) = 1, SD(X) = 2.89$
   (b) $E(3X - 1) = 2$, $V(5 - 2X) = \frac{100}{3}$
   (c) 0.64
   (d) 0.4
5. (a) $\frac{5}{6}$
   (b) 7.5 mins
   (c) 4.33 mins
6. $\frac{4}{7}$
7. (a) 0.4
   (b) $X \sim B(20, 0.4)$,0.5841
8. (a) 8
   (b) 4.04
   (c) 0.5
   (d) 0.5771
9. (a) $X \sim U(2, 14)$
   (b) $\sqrt{12}$
   (c) $\frac{5}{12}$
10. (a) $Y \sim U(16, 22)$
    (b) 0.4082

Exercise 8.2

1. (a) $52, 60, 68$ marked
   (b) $237, 243, 249$ marked
   (c) $5.4, 8, 10.6$ marked
   (d) $12.6, 15, 17.4$ marked
2. $3, 7, 11$ marked
3. $5, 8, 11$ marked
4. $26, 32, 38$ marked

Exercise 8.3

1. (a) 0
   (b) 0.9573
   (c) 0.9573
   (d) 0.5
   (e) 0.6368
   (f) 0.9788
   (g) 0.0212
   (h) 0.1170
   (i) 0.1170
   (j) 0.7123
   (k) 0.9251
   (l) 0.2843
2. (a) 0.0740
   (b) 0.2888
   (c) 0.0339
   (d) 0.1618
   (e) 0.6612
   (f) 0.6826
   (g) 0.9544
   (h) 0.9500
   (i) 0.8990

Exercise 8.4

1. (a) 0.7475
   (b) 0.1587
   (c) 0.6306
   (d) 0.4719
2. (a) 0.1587
   (b) 0.4305
   (c) 0.9660
   (d) 0.3891
3. (a) $X \sim N(67.4, 6^2)$
   (b) i. 0.6676
   ii. 0.6554
   iii. 0.2647
4. (a) i. 0.8997
   ii. 0.9500
   (b) i. 143
   ii. 595
   iii. 536

Exercise 8.5

1. (a) 1.28
   (b) 1.64
   (c) 3.09
   (d) 2.33
   (e) 2.58
   (f) -1.96
   (g) -2.58
   (h) -1.28
   (i) -1.96
   (j) 3.09
   (k) 0.81
   (l) 2.75
2. (a) $X \sim N(2.4, 0.2^2)$
   (b) 0.0228
   (c) 2.866 litres
3. (a) 8.692 mins
   (b) 7.612 to 8.788 mins
4. $\sigma = 9.9$
5. $\mu = 64.1, \sigma = 14.8$

Exercise 8.6

1. (a) $N(34, 13.09)$
   (b) 0.6103
2. 0.1867
3. 0.1379
4. 0.0192

Exercise 8.7

1. (a) $np = 16, nq = 24\ np, nq > 5$
   (b) i. 0.9265
   ii. 0.0031
   iii. ≈ 0
2. (a) $np = 54, nq = 27\ np, nq > 5$
   (b) i. 0.3632
   ii. 0.0338
   iii. 0.6971
3. $X \sim B(600, 0.85)\ 0.1151$
4. 0.2578
5. 0.0044

Exercise 8.8

1. (a) 0.6217
   (b) 0.0661
   (c) 0.5398
2. (a) 0.6736
   (b) 0.5640
   (c) 0.1357
3. (a) 0.2611
   (b) 0.2349
   (c) 0.0550
4. 0.8907
5. (a) 0.7384
   (b) 3.8
6. (a) 0.4562
   (b) 0.1718
   (c) 0.8315
7. (a) 0.2389
   (b) 0.4641
8. 0.8529

Chapter 8 Review Exercise

1. (a) 4
   (b) 3
   (c) $\frac{1}{3}$
2. (a) 0.7977
   (b) 0.3413
3. (a) $H \sim N(18.1, 0.7^2)$
   (b) 0.0033
   (c) 19.248 feet
4. (a) 0.1353
   (b) 0.0733
5. 22.99
6. (a) 0.3
   (b) 0.6172
   (c) 0.4562
7. (a) 0.0951

Chapter 9 - Distribution of the Sample Mean

Exercise 9.1

1. (a) i. 0.3821
   ii. 0.4404
   (b) i. $N(24, \frac{4^2}{3})$
   ii. 0.3015
   iii. 0.3974
2. 0.1922
3. (a) 0.3085
   (b) 0.1170
4. 0.0228
5. (a) 0.0082
   (b) 0.3974
6. 0.9992

Exercise 9.2

1. (a) $X \approx N(2.3, \frac{5^2}{40})$
   (b) $Y \sim N(83, \frac{9}{25})$
   (c) $D \sim N(126, \frac{10}{7})$
   (d) Comment.
2. 0.9857
3. 0.0071
4. 0.1038
5. 0.3758
6. (a) 0.0526

Review Exercise

1. (a) $\bar{X} \sim N(15, \frac{3^2}{8})$
   (b) 0.9706
2. 0.9981
3. (a) Comment.
   (b) $\bar{X} \approx N(600, \frac{70^2}{40})$
   (c) 0.1489
4. (a) i. $X \sim N(60, 2^2)$
   ii. 0.3085
   (b) i. $\bar{X} \sim N(60, \frac{2^2}{8})$
   ii. 0.1271
   (c) Comment.
5. (a) $X \sim Po(4.5)$
   (b) 0.1736
   (c) $X \approx N(4.5, \frac{4.5}{30})$
   (d) 0.0985

Chapter 10 - An Introduction to Hypothesis Testing

Exercise 10.1

1. $H_0 : \mu = 8.6$
   $H_1 : \mu > 8.6$
2. $H_0 : \mu = 12$
   $H_1 : \mu > 12$
3. $H_0 : \mu = 4.2$
   $H_1 : \mu < 4.2$
4. $H_0 : \mu = 2.5$
   $H_1 : \mu \neq 2.5$
5. $H_0 : \mu = 65$
   $H_1 : \mu > 65$
6. $H_0 : \mu = 15.7$
   $H_1 : \mu \neq 15.7$

Exercise 10.2

1. $0.0162 < 0.05 \implies$ *reject* H_0
2. $0.1056 > 0.01 \implies$ *do not reject* H_0
3. $0.0485 < 0.1 \implies$ *reject* H_0
4. $0.0764 > 0.05 \implies$ *do not reject* H_0

Exercise 10.3

1. $0.0286 < 0.1 \implies$ *reject* H_0
2. $0.0702 > 0.001 \implies$ *do not reject* H_0
3. $0.0485 < 0.05 \implies$ *reject* H_0
4. $0.0220 < 0.1 \implies$ *reject* H_0
5. $0.0294 < 0.05 \implies$ *reject* H_0
6. $0.0488 > 0.01 \implies$ *do not reject* H_0
7. $0.0023 < 0.05 \implies$ *reject* H_0
8. $0.0250 > 0.01 \implies$ *do not reject* H_0

Exercise 10.4

1. $1.67 < 1.96 \implies$ *do not reject* H_0
2. $-2.5 < -1.64 \implies$ *reject* H_0
3. $2.24 < 2.33 \implies$ *do not reject* H_0
4. $-0.77 > -2.33 \implies$ *do not reject* H_0
5. $2.00 > 1.96 \implies$ *reject* H_0

Exercise 10.5

1. $-2.985 < -1.895 \implies$ *reject* H_0
2. $0.636 < 1.476 \implies$ *do not reject* H_0
3. $2.317 < 2.447 \implies$ *do not reject* H_0
4. $-3.531 < -2.821 \implies$ *reject* H_0

Exercise 10.6

1. $1.52 < 1.64 \implies$ *do not reject* H_0
2. $2.42 > 2.33 \implies$ *reject* H_0
3. $-2.78 < -1.64 \implies$ *reject* H_0
4. $1.17 < 1.96 \implies$ *do not reject* H_0
5. $-2.04 > -2.58 \implies$ *do not reject* H_0
6. $-2.2 < -1.64 \implies$ *reject* H_0

Review Exercise

1. $-2.54 < -2.33 \implies$ *reject* H_0
2. $2.158 < 2.365 \implies$ *do not reject* H_0
3. $-1.94 < -1.64 \implies$ *reject* H_0
4. $-3.00 < -2.33 \implies$ *reject* H_0
5. $-2.170 < -1.895 \implies$ *reject* H_0

Chapter 11 - Confidence Intervals

Exercise 10.1

1. $(55.18, 58.82)$
2. $(12.76, 14.84)$
3. $(14.53, 15.87)$
4. $(37.38, 40.62)$
5. $(49.03, 51.29)$
6. $(5.41, 5.99)$
7. (a) $(37.64, 38.76)$
   (b) $(37.46, 38.94)$
   (c) 31
8. 10.0

Exercise 10.2

1. $(499.1, 510.9) \implies$ within, comment
2. $(419.5, 426.5) \implies$ within, comment
3. $(10.11, 11.69) \implies$ outwith, comment
4. $(2.439, 2.961) \implies$ within, comment
5. $(19.95, 20.85) \implies$ outwith, comment
6. $(28580, 30020 \implies$ within, comment$)$

Exercise 10.3

1. $(6.88, 7.87)$
2. $(4048.5, 4714.4) \implies$ within, comment
3. $(960.0, 1024.0)$
4. $(3.636, 4.464) \implies$ outwith, comment
5. $(2.760, 4.740) \implies$ outwith, comment

Exercise 10.4

1. $(0.598, 0.736)$
2. $(0.133, 0.427)$
3. $(0.327, 0.387)$
4. $(0.086, 0.149) \implies$ within, comment
5. $(0.021, 0.199) \implies$ outwith, comment
6. $(0.026, 0.083) \implies$ within, comment
7. $n = 147$
8. (a) Systematic sampling
   (b) $(0.264, 0.446) \implies$ within, comment

Review Exercise

1. $(63.98, 82.77)$
2. $(0.116, 0.444)$
3. (a) $(125.8, 126.8)$
   (b) Outwith, comment
4. (a) 12.5
   (b) $s = 25.77$
5. $(0.679, 0.888) \implies$ outwith, comment

Chapter 12 - Control Charts

Exercise 12E

2. $3.324 < 6.635 \implies$ *do not reject* H_0

3. $2.326 < 4.605 \implies$ *do not reject* H_0

Chapter 12 - Chi-Squared Goodness-of-Fit Test

Exercise 12E

2. $3.324 < 6.635 \implies$ *do not reject* H_0

3. $2.326 < 4.605 \implies$ *do not reject* H_0

Chapter 14 - Chi-Squared Test for Association

1. Hi

Chapter 15 - Two-Sample Parametric Tests

1. Hi

Chapter 16 - Wilcoxon Signed Rank Test

For any questions for which a level of significance is not specified, a 5% level of significance has been used. Full answers should include, of course, conclusions in context

Exercise 16A

1. $9 > 3 \implies$ do not reject H_0
2. $5.5 < 3 \implies$ reject H_0
3. $8 > 5 \implies$ do not reject H_0
4. $7.5 > 5 \implies$ do not reject H_0
5. $10 \leqslant 10 \implies$ reject H_0
6. $9 < 10 \implies$ reject H_0
7. $6.5 > 5 \implies$ do not reject H_0

Exercise 16B

1. $8.5 < 13 \implies$ reject H_0
2. $11 > 5 \implies$ do not reject H_0
3. $46 > 21 \implies$ do not reject H_0
4. $1.5 < 2 \implies$ reject H_0
5. $4 < 8 \implies$ reject H_0

Exercise 16C

1. $-0.34 > -1.96 \implies$ do not reject H_0
2. $-1.67 < -1.64 \implies$ reject H_0
3. $-0.27 > -1.64 \implies$ do not reject H_0
4. $W = 80.5$
5. $W = 55$

Review Exercise

1. $5 \leqslant 5 \implies$ reject H_0
2. $14 > 8 \implies$ do not reject H_0
3. $-2.40 > -1.96 \implies$ reject H_0

Chapter 17 - Mann Whitney Rank Sum Test

1. Hi

Chapter 18 - Further Linear Regression

1. Hi

18.11.1 Spare

1a) Independent

1b) Not independent

2a) Independent

2b) Not independent

2c) Not independent

Exercise 1G

1) 0.8708

2) 0.6745

3) 0.695

4) 0.268

5) 0.3882

Exercise 1H

1a) 0.4524

1b) 0.2836

1c) 0.6735

1d) 0.3333

1e) 0.0787

2) 0.8409

3) 0.7273

4) 0.0559

5) 0.3043

6) 0.3667

7) 0.8048

Exercise 1I

1) $\frac{7}{12}$

2a) 0.625

2b) 0.375

4) 0.3584

Review Exercise

1) $\frac{7}{9}$

2a) $\frac{17}{20}$

2b)i) $\frac{37}{40}$

2b)ii) $\frac{1}{8}$

3a) 0.3917

3b) 0.1370

4) $P(M|B) = P(M) \therefore$ independent

5a) 0.6525

5b) 0.3669

\end{multicols}