Contents:

A Sampling from populations
B Describing data
C Presenting and interpreting data
D Grouped discrete data
E Continuous (interval) data
F Measures of centres of distributions
G Measuring the spread of data
H Comparing data
I Sample statistics and population parameters
J Data based investigation
K Review
L Normal distributions
M The standard normal distribution
N Technology and normal distributions
O Pascal’s triangle
P Binomial distribution
Q The mean and standard deviation of a discrete variable
R Mean and standard deviation of a binomial variable
S Review
DISCUSSION

We live in an age of information overload. Data is presented to us in many ways from every possible source including newspapers, television and the internet. Some data is for our information. For example, health authorities warning about the dangers of smoking. Some is for our amusement. For example, the statistics quoted on television during sport matches. And some is to deceive us.

Generally there are six steps involved when considering any statistical problem. These are known as the 'statistical process'.

Step 1: Identify the problem.
Step 2: Formulate a method of investigation.
Step 3: Collect data.
Step 4: Analyse the data.
Step 5: Interpret the results and form a conjecture.
Step 6: Consider the underlying assumptions.

The following examples are just a few ways in which data is presented to us. When examining information, it is a good idea to see how the statistical process might have been used.

INTRODUCTION

We live in an age of information overload. Data is presented to us in many ways from every possible source including newspapers, television and the internet. Some data is for our information. For example, health authorities warning about the dangers of smoking. Some is for our amusement. For example, the statistics quoted on television during sport matches. And some is to deceive us.

Generally there are six steps involved when considering any statistical problem. These are known as the ‘statistical process’.

Step 1: Identify the problem.
Step 2: Formulate a method of investigation.
Step 3: Collect data.
Step 4: Analyse the data.
Step 5: Interpret the results and form a conjecture.
Step 6: Consider the underlying assumptions.

The following examples are just a few ways in which data is presented to us. When examining information, it is a good idea to see how the statistical process might have been used.

INTRODUCTION

We live in an age of information overload. Data is presented to us in many ways from every possible source including newspapers, television and the internet. Some data is for our information. For example, health authorities warning about the dangers of smoking. Some is for our amusement. For example, the statistics quoted on television during sport matches. And some is to deceive us.

Generally there are six steps involved when considering any statistical problem. These are known as the ‘statistical process’.

Step 1: Identify the problem.
Step 2: Formulate a method of investigation.
Step 3: Collect data.
Step 4: Analyse the data.
Step 5: Interpret the results and form a conjecture.
Step 6: Consider the underlying assumptions.

The following examples are just a few ways in which data is presented to us. When examining information, it is a good idea to see how the statistical process might have been used.

Example A

The following table shows the apparent retention rates of full-time secondary students in various states of Australia. The table is taken from the Australian Bureau of Statistics.

<table>
<thead>
<tr>
<th>All</th>
<th>NSW</th>
<th>Vic</th>
<th>Qld</th>
<th>SA</th>
<th>WA</th>
<th>Tas</th>
<th>NT</th>
<th>ACT</th>
<th>Males</th>
<th>Females</th>
<th>Persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>67.6</td>
<td>76.2</td>
<td>77.5</td>
<td>67.0</td>
<td>71.5</td>
<td>66.7</td>
<td>52.9</td>
<td>92.5</td>
<td>66.4</td>
<td>78.5</td>
<td>72.3</td>
</tr>
<tr>
<td>2000</td>
<td>67.5</td>
<td>77.2</td>
<td>77.3</td>
<td>65.4</td>
<td>71.3</td>
<td>69.5</td>
<td>49.7</td>
<td>87.1</td>
<td>66.1</td>
<td>78.7</td>
<td>72.3</td>
</tr>
<tr>
<td>2001</td>
<td>68.2</td>
<td>79.3</td>
<td>79.0</td>
<td>66.4</td>
<td>72.0</td>
<td>68.7</td>
<td>50.9</td>
<td>89.3</td>
<td>68.1</td>
<td>79.1</td>
<td>73.4</td>
</tr>
<tr>
<td>2002</td>
<td>69.9</td>
<td>80.9</td>
<td>81.3</td>
<td>66.7</td>
<td>73.7</td>
<td>72.6</td>
<td>53.0</td>
<td>88.1</td>
<td>69.8</td>
<td>80.7</td>
<td>75.1</td>
</tr>
<tr>
<td>2003</td>
<td>70.5</td>
<td>81.4</td>
<td>81.5</td>
<td>67.1</td>
<td>71.2</td>
<td>74.9</td>
<td>56.3</td>
<td>89.7</td>
<td>70.3</td>
<td>80.7</td>
<td>75.4</td>
</tr>
<tr>
<td>2004</td>
<td>71.1</td>
<td>81.1</td>
<td>81.2</td>
<td>68.0</td>
<td>72.6</td>
<td>76.4</td>
<td>59.0</td>
<td>88.5</td>
<td>70.4</td>
<td>81.4</td>
<td>75.7</td>
</tr>
<tr>
<td>2005</td>
<td>71.1</td>
<td>80.6</td>
<td>79.9</td>
<td>70.7</td>
<td>72.5</td>
<td>67.1</td>
<td>59.1</td>
<td>87.5</td>
<td>69.9</td>
<td>81.0</td>
<td>75.3</td>
</tr>
<tr>
<td>Govt.</td>
<td>65.8</td>
<td>74.0</td>
<td>73.0</td>
<td>61.7</td>
<td>65.4</td>
<td>65.5</td>
<td>70.5</td>
<td>99.6</td>
<td>63.4</td>
<td>75.7</td>
<td>69.4</td>
</tr>
<tr>
<td>Non-govt.</td>
<td>80.6</td>
<td>91.0</td>
<td>92.5</td>
<td>88.4</td>
<td>85.2</td>
<td>70.9</td>
<td>39.0</td>
<td>73.3</td>
<td>81.5</td>
<td>90.3</td>
<td>85.8</td>
</tr>
</tbody>
</table>

This graph was drawn from the table above (year 2005 data). It is one of many in the document, “Success For All”, a ministerial review of senior secondary education in South Australia.
Questions to consider:
1. Why do you think this data was collected?
2. How do you think this data was collected?
3. Why do you think the people making this report were interested in this information?
4. How does the information in the graph differ from that in the table?
5. Can we make a conjecture as to why the retention rate in the ACT is higher than that in South Australia?

Example B

An article published in a newspaper indicates that wearing bike helmets ‘reduced the number of riders, not the number of injuries’.

It states that compelling cyclists to wear helmets has not reduced head injury rates, but has discouraged people from cycling. The article claims that it has been the decline in the number of cyclists that has reduced the total number of head injuries.

The comparison of cycling and injury patterns before and after cycle helmets were made compulsory, found the laws had no effect on head injury trends, which were already falling, but cut cyclist numbers by 30%. The analysis has been criticised by some experts.

Questions to consider:
1. What was the question that was considered in this article?
2. What is the conjecture that has been made?
3. What evidence is presented in the article to support the conjecture?

Example C

A television advertisement for a dating service claims that over half a million people have visited its internet site. Amongst the claims, a few very attractive couples assure the viewer that this site has worked for them, and without this service they would not have met. The advertisement ends with the statement that half a million users cannot be wrong.

Questions to consider:
1. What is the problem the advertisers are addressing?
2. How was the information collected?
3. What are some of the underlying assumptions made by the advertisers?
Famous historical examples of errors in sampling.

Before the 1936 American presidential election the Literary Digest predicted that the republican Alf Landon would defeat the democrat Franklin D. Roosevelt by a large margin. The prediction was based on a questionnaire sent to its readers. From the ten million contacted there were over two million responses. The Literary Digest was a prestigious journal, which has correctly predicted the outcome of the last 5 elections. This time it was wrong. The error in its prediction for the 1936 election is blamed on bias in its sampling. Significantly more republican than democrat voters were readers of the Literary Digest.

During the same election George Gallup correctly predicted that Franklin D. Roosevelt would win. This prediction was based on a much smaller sample of 50,000. This prediction established Gallup’s reputation as a pollster.

Making predictions, however, is a risky business. In the 1948 presidential election, Gallup incorrectly predicted that Harry Truman would lose. Gallup put the blame on the fact that sampling had stopped 3 weeks before the election.

**RANDOM SAMPLING**

A population is the entire set about which we want to draw a conclusion.

Every 5 years the Australian government carries out a census in which it seeks basic information from the whole population.

It is often too expensive or impractical to obtain information from every member of the population. Before an election a sample of voters is asked how they will vote. With this information a prediction is made on how the population of eligible voters will vote.

A sample is a selection from the population.

In collecting samples, great care and expense is usually taken to make the selection as free from prejudice as possible, and large enough to be representative of the whole population.

A biased sample is one in which the data has been unduly influenced by the collection process and is not representative of the whole population.

To avoid bias in sampling, many different sampling procedures have been developed.

A random sample is a sample in which all members of the population have equal chance of being selected.

We discuss four commonly used random sampling techniques.

These are:  
- simple random sampling  
- systematic sampling  
- cluster sampling  
- stratified sampling.
A simple random sample of size \( n \) is a sample chosen in such a way that every set of \( n \) members of the population has the same chance of being chosen.

To select five students from your class to form a committee, the class teacher can draw five names out of a hat containing all the names of students in your class.

Suppose we wish to find the views on extended shopping hours of shoppers at a huge supermarket. As people come and go, a simple random process is not practical. In such a situation systematic sampling may be used.

To obtain a \( k\% \) systematic sample the first member is chosen at random, and from then on every \( \left( \frac{100}{k} \right) \)th member from the population.

If we need to sample 5\% of an estimated 1600 shoppers at the supermarket, i.e., 80 in all, then as \( \frac{100}{5} = 20 \), we approach every 20th shopper.

The method is to randomly select a number between 1 and 20. If this number were 13 say, we would then choose for our sample the 13th, 33rd, 53rd, 73rd, ... person entering the supermarket. This group forms our systematic sample.

Suppose we need to analyse a sample of 300 biscuits. The biscuits are in packets of 15 and form a large batch of 1000 packets. It is costly, wasteful and time consuming to take all the biscuits from their packets, mix them up and then take the sample of 300. Instead, we would randomly choose 20 packets and use their contents as our sample. This is called cluster sampling where a cluster is one packet of biscuits.

To obtain a cluster sample the population must be in smaller groups called clusters and a random sample of the clusters is taken. All members of each cluster are used.

Suppose the student leaders of a very large high school wish to survey the students to ask their opinion on library use after school hours. Asking only year 12 students their opinion is unacceptable as the requirements of the other year groups would not be addressed. Consequently, subgroups from each of the year levels need to be sampled. These subgroups are called strata.

If a school of 1135 students has 238 year 8’s, 253 year 9’s, 227 year 10’s, 235 year 11’s and 182 year 12’s and we want a sample of 15\% of the students, we must randomly choose:

\[
\begin{align*}
15\% \text{ of } 238 & = 36 \text{ year 8's} \\
15\% \text{ of } 253 & = 38 \text{ year 9's} \\
15\% \text{ of } 227 & = 34 \text{ year 10's} \\
15\% \text{ of } 235 & = 35 \text{ year 11's} \\
15\% \text{ of } 182 & = 27 \text{ year 12's}
\end{align*}
\]
To obtain a **stratified random sample**, the population is first split into appropriate groups called strata and a random sample is selected from each in proportion to the numbers in each strata.

It is not always possible to select a random sample. Dieticians may wish to test the effect fish oil has on blood platelets. To test this they need people who are prepared to go on special diets for several weeks before any changes can be observed. The usual procedure to select a sample is to advertise for volunteers. People who volunteer for such tests are usually not typical of the population. In this case they are likely to be people who are diet conscious, and have probably heard of the supposed advantage of eating fish. The dietician has no choice but to use those that volunteer.

A **convenient sample** is a sample that is easy to create.

**EXERCISE 4A**

1. In each of the following state the population, and the sample.
   a. A pollster asks 500 people if they approve of Mr John Howard as prime minister of Australia.
   b. Fisheries officers catch 200 whiting fish to measure their size.
   c. A member of a consumer group buys a basket of bread, butter and milk, meat, breakfast cereal, fruit and vegetables from a supermarket.
   d. A dietician asks 12 male volunteers over the age of 70 to come in every morning for 2 weeks to eat a muffin heavily enriched with fibre.
   e. A promoter offers every shopper in a supermarket a slice of mettwurst.

2. For each of the following describe a sample technique that could be used.
   a. Five winning tickets are to be selected in a club raffle.
   b. A sergeant in the army needs six men to carry out a dirty, tiresome task.
   c. The department of tourism in Victoria wants visitors’ opinion of its facilities set up by the Twelve Apostles along the Great Ocean Road.
   d. Cinema owners want to know what their patrons think of the latest blockbuster they have just seen.
   e. A research team wants to test a new diet to lower glucose in the blood of diabetics. To get statistically significant results they need 30 women between the ages of 65 and 75 who suffer from type II diabetes.
   f. When a legion disgraced itself in the Roman army it was decimated; that is, 10% of the soldiers in the legion were selected and killed.
   g. A council wants to know the opinions of residents about building a swimming pool in their neighbourhood.

3. In each of the following, state: (i) the intended population, (ii) the sample, (iii) any possible bias the sample might have.
   a. A recreation centre in a suburban area wants to enlarge its facilities. Nearby residents object strongly. To support its case the recreation centre asks all persons using the centre to sign a petition.
   b. Tom has to complete his statistics project by Monday morning. He is keen on sport and has chosen as part of his project ‘oxygen debt in exercise’. As a measure of
oxygen debt he has decided to measure the time it takes for the heart rate to return to normal after a 25 m sprint. Unfortunately he has not collected any data and he persuades six of his football friends to come along on Saturday afternoon to provide him with some numbers.

c A telephone survey conducted on behalf on a motor car company rings 400 households between the hours of 2 and 5 o’clock in the afternoon to ask what brand of car they drive.

d A council sends out questionnaires to all residents asking about a proposal to build a new library complex. Part of the proposal is that residents in the wards that will benefit most from the library will be asked to pay higher rates for the next two years.

4 A sales promoter decides to visit 10 houses in a street and offer special discounts on a new window treatment. The street has 100 houses numbered from 1 to 100. The sales promoter selects a random number between 1 and 10 inclusive and calls on the house with that street number. After this the promoter calls on every tenth house.

a What sampling technique is used by the sales promoter?

b Explain why every house in the street has an equal chance of being visited.

c How is this different from a simple random sample?

5 Explain why a stratified sample is a random sample.

6 How does a simple random sample differ from a cluster sample?

7 Tissue paper is made from wood pulp mixed with glue. The mixture is rolled over a huge hot roller that dries the mixture into paper. The paper is then rolled into rolls a metre or so in diameter and a few metres in width. When the roll comes off the machine a quality controller takes a sample from the end of the roll to test it.

a Explain why the samples taken by the quality controller could be biased.

b Explain why the quality controller only samples the paper at the end of the roll.

INVESTIGATION 1

In this investigation you will be exploring the web sites of a number of organisations to find out the topics and the types of data that they collect and analyse.

Note that the web addresses given here were operative at the time of writing but there is a chance that they will have changed in the meantime. If the address does not work, try using a search engine to find the site of the organisation.

What to do:

Visit the site of a world organisation such as the United Nations (www.un.org) or the World Health Organisation (www.who.int) and see the available types of data and statistics. The Australian Bureau of Statistics (www.abs.gov.au) also has a large collection of data.
DESCRIBING DATA

TYPES OF DATA

Data are individual observations of a variable. A variable is a quantity that can have a value recorded for it or to which we can assign an attribute or quality.

There are two types of variable that we commonly deal with:

CATEGORICAL VARIABLES

A categorical variable is one which describes a particular quality or characteristic. It can be divided into categories. The information collected is called categorical data.

Examples of categorical variables are:

- Getting to school: the categories could be train, bus, car and walking.
- Colour of eyes: the categories could be blue, brown, hazel, green, grey.
- Gender: male and female.

QUANTITATIVE (NUMERICAL) VARIABLES

A quantitative variable is one which has a numerical value and is often called a numerical variable. The information collected is called numerical data.

Quantitative variables can be either discrete or continuous.

A quantitative discrete variable takes exact number values and is often a result of counting.

Examples of discrete quantitative variables are:

- The number of people in a car: the variable could take the values 1, 2, 3, ....
- The score out of 50 on a test: the variable could be 0, 1, 2, 3, ...., 50.

A quantitative continuous variable takes numerical values within a certain continuous range. It is usually a result of measuring.

Examples of quantitative continuous variables are:

- The weight of newborn pups: the variable could take any value on the number line but is likely to be in the range 0.2 kg to 1.2 kg.

- The heights of Year 10 students: the variable would be measured in centimetres. A student whose height is recorded as 163 cm could have exact height between 162.5 cm and 163.5 cm.
**Example 1**

Classify these variables as categorical, quantitative discrete or quantitative continuous:

- **a** the number of heads when 4 coins are tossed
- **b** the favourite variety of fruit eaten by the students in a class
- **c** the heights of a group of 16 year old students.

**a** The values of the variables are obtained by counting the number of heads. The result can only be one of the values 0, 1, 2, 3 or 4. It is a quantitative discrete variable.

**b** The variable is the favourite variety of fruit eaten. It is a categorical variable.

**c** This is numerical data obtained by measuring. The results can take any value between certain limits determined by the degree of accuracy of the measuring device. It is a quantitative continuous variable.

---

**EXERCISE 4B**

1. For each of the following possible investigations, classify the variable as categorical, quantitative discrete or quantitative continuous:

- **a** the number of goals scored each week by a netball team
- **b** the number of children in an Australian family
- **c** the number of bread rolls bought each week by a family
- **d** the pets owned by students in a year 10 class
- **e** the number of leaves on the stems of a bottle brush species
- **f** the amount of sunshine in a day
- **g** the number of people who die from cancer each year in Australia
- **h** the amount of rainfall in each month of the year
- **i** the countries of origin of immigrants
- **j** the most popular colours of cars
- **k** the time spent doing homework
- **l** the marks scored in a class test
- **m** the items sold at the school canteen
- **n** the reasons people use taxis
- **o** the sports played by students in high schools
- **p** the stopping distances of cars doing 60 km/h
- **q** the pulse rates of a group of athletes at rest.

2. **a** For the categorical variables in question 1, write down two or three possible categories. (In all cases but one, there will be more than three categories possible.) Discuss your answers.

**b** For each of the quantitative variables (discrete and continuous) identified in question 1, discuss as a class the range of possible values you would expect.
A tally and frequency table can be used to organise categorical data.

For example, a survey was conducted on 200 randomly chosen victims of sporting injuries, to find which sport they played.

The variable ‘sport played’ is a categorical variable because the information collected can only be one of the five categories listed. The data has been counted and organised in the given frequency table:

<table>
<thead>
<tr>
<th>Sport played</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aussie rules</td>
<td>57</td>
</tr>
<tr>
<td>Netball</td>
<td>43</td>
</tr>
<tr>
<td>Rugby</td>
<td>41</td>
</tr>
<tr>
<td>Cricket</td>
<td>21</td>
</tr>
<tr>
<td>Other</td>
<td>38</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>200</strong></td>
</tr>
</tbody>
</table>

Acceptable graphs to display the ‘sporting injuries’ categorical data are:

- Segment bar graph
- Vertical column graph
- Horizontal bar graph
- Pie chart
- Segmented bar graph

For categorical data, the **mode** is the category which occurs most frequently.
ORGANISING DISCRETE NUMERICAL DATA

OPENING PROBLEM

A farmer wishes to investigate whether a new food formula increases egg production from his laying hens. To test this he feeds 60 hens with the current formula and 60 with the new one.

The hens were randomly selected from the 1486 hens on his property.

Over a period he collects and counts the eggs laid by the individual hens.

All other factors such as exercise, water, etc are kept the same for both groups.

The results of the experiment were:

Current formula
7 6 8 5 8 6 4 7 6 6
7 5 7 9 3 7 9 6 6 6
8 6 7 6 8 6 7 7 5
6 7 6 9 7 5 4 6 8 7
6 7 4 6 8 7 6 7 6 6
7 8 7 9 7 7 8 7 7

New formula
7 3 6 7 8 6 7 7 7
8 6 7 7 7 6 6 6 4 8
6 7 7 4 7 5 6 6 6 6
7 5 8 6 5 9 7 7 8 7
6 7 6 8 7 7 6 6 6 7
9 6 6 7 6 7 5 6 8 1 4

For you to consider:

• Can you state clearly the problem that the farmer wants to solve?
• How has the farmer tried to make a fair comparison?
• How could the farmer make sure that his selection is at random?
• What is the best way of organising this data?
• What are suitable methods of display?
• Are there any abnormally high or low results and how should they be treated?
• How can we best indicate the most number of eggs laid?
• How can we best indicate the spread of possible number of eggs laid?
• What is the best way to show ‘number of eggs laid’ and the spread?
• Can a satisfactory conclusion be made?

In the above problem, the **discrete quantitative variable** is: The number of eggs laid.

To organise the data a **tally/frequency table** could be used.

We count the data systematically and use a ‘|’ to indicate each data value.

Remember that $\big|\big|\big|\big|\big|$ represents 5.

The **relative frequency** of an event is the frequency of that event expressed as a fraction (or decimal equivalent) of the total frequency.
Below is the table for the new formula data:

<table>
<thead>
<tr>
<th>Number of eggs laid</th>
<th>Tally</th>
<th>Frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A column graph of the frequencies or the relative frequencies could be used to display the results.

Can you explain why the two graphs are similar?

**DESCRIBING THE DISTRIBUTION OF THE DATA SET**

It is useful to be able to recognise and classify common shapes of distributions. These shapes often become clearer if a curve is drawn through the columns of a column graph or a histogram.

Common shapes are:

- **Symmetric distributions**
  
  One half of the graph is roughly the mirror image of the other half.
  
  Heights of 18 year old women tend to be symmetric.
• Negatively skewed distributions

The left hand, or negative, side is stretched out. This is sometimes described as “having a long, negative tail”.

The time people arrive for a concert, with some people arriving very early, but the bulk close to the starting time, has this shape.

• Positively skewed distributions

The right hand, or positive, side is stretched out. This is sometimes described as “having a long, positive tail”.

The life expectancy of animals and light globes have this shape.

• Bimodal distributions

The distribution has two distinct peaks. The heights of a mixed class of students where girls are likely to be smaller than boys has this shape.

OUTLIERS

Outliers are data values that are either much larger or much smaller than the general body of data. Outliers appear separated from the body of data on a frequency graph.

For example, in the egg laying data, the farmer found one hen who laid 14 eggs which is clearly well above the rest of the data.

So, 14 is said to be an outlier.

On the column graph outliers appear well separated from the remainder of the graph.

Outliers which are genuine data values should be included in any analysis.

However, if they are a result of experimental or human error, they should be deleted and the data re-analysed.
MISLEADING PRESENTATION

Statistical data can also be presented in such a way that a **misleading impression** is given.

- A common way of doing this is by manipulating the scales on the axes of a line graph.

  The vertical scale does not start at zero. So the increase in profits looks larger than it really is. The break of scale on the vertical axis should have been indicated by \( \uparrow \).

  The graph should look like that shown alongside.

  This graph shows the true picture of the profit increases and probably should be labelled ‘A modest but steady increase in profits’.

- These two charts show the results of a survey of shoppers’ preferences for different brands of soap. Both charts begin their vertical scales at zero, but chart 1 does not use a uniform scale along the vertical axis. The scale is compressed at the lower end and enlarged at the upper end.

  This has the effect of **exaggerating the difference** between the bars on the chart. The bar for brand ‘B’, the most preferred brand, has also been darkened so that it stands out more than the other bars. Chart 2 has used a uniform scale and has treated all the bars in the same way. Chart 2 gives a more accurate picture of the survey results.

- The ‘bars’ on a bar chart (or column graph) are given a larger appearance by adding area or the appearance of volume. The height of the bar represents frequency.

  For example, consider the graph comparing sales of three different types of soft drink.

  By giving the ‘bars’ the appearance of volume the sales of ‘Kick’ drinks look to be about eight times the sales of ‘Fizz’ drinks.

  On a bar chart, frequency (sales in this case) is proportional to the height of the bar only. The graph should look like this:

  It can be seen from the bar chart that the sales of Kick are just over twice the sales of Fizz.

  There are many different ways in which data can be presented so as to give a misleading impression of the figures.
The people who use these graphs, charts, etc., need to be careful and to look closely at what they are being shown before they allow the picture to “tell a thousand words”.

EXERCISE 4C

1 State whether these quantitative (or numerical) variables are discrete or continuous:
   a the time taken to run a 1500 metre race
   b the minimum temperature reached on a July day
   c the number of tooth picks in a container
   d the weight of hand luggage taken on board an aircraft
   e the time taken for a battery to run down
   f the number of bricks needed to build a garage
   g the number of passengers on a train
   h the time spent on the internet per day.

2 50 adults were chosen at random and asked “How many children do you have?” The results were: 0 1 2 1 0 3 1 4 2 0 1 2 1 8 0 5 1 2 1 0 0 1 2 1 8
   0 1 4 1 0 9 1 2 5 0 4 1 2 3 0 0 1 2 1 3 4 9 2 3 2
   a What is the variable in this investigation?
   b Is the variable discrete or continuous? Why?
   c Construct a column graph to display the data. Use a heading for the graph, and scale and label the axes.
   d How would you describe the distribution of the data? (Is it symmetrical, positively skewed or negatively skewed? Are there any outliers?)
   e What percentage of the adults had no children?
   f What percentage of the adults had three or more children?

3 For an investigation into the number of phone calls made by teenagers, a sample of 80 sixteen-year-olds was asked the question “How many phone calls did you make yesterday?” The following column graph was constructed from the data:

   Number of phone calls made by teenagers

<table>
<thead>
<tr>
<th>number of calls</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
</tr>
</tbody>
</table>

   a What is the variable in this investigation?
   b Explain why the variable is discrete numerical.
   c What percentage of the sixteen-year-olds did not make any phone calls?
   d What percentage of the sixteen-year-olds made 3 or more phone calls?
   e Copy and complete:
      “The most frequent number of phone calls made was .........”
   f How would you describe the data value ‘12’?
   g Describe the distribution of the data.
4 The number of matches in a box is stated as 50 but the actual number of matches has been found to vary. To investigate this, the number of matches in a box has been counted for a sample of 60 boxes:

<table>
<thead>
<tr>
<th>Number of matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>53 49 51 48 51 50 49 51 50 52 51 50 49 51 50 49 51 50 50 52 51 50 49 50 48 51 49 52 50 49 49 52 51 50 50 52 50 53 48 49 50 49 52 50 49 50 49 52 51 50 50 52 50 53 48 49 50 49 52 50 49 50 49 52 51 50 50 51 50</td>
</tr>
</tbody>
</table>

a What is the variable in this investigation?
b Is the variable continuous or discrete numerical?
c Construct a frequency table for this data.
d Display the data using a bar chart.
e Describe the distribution of the data.
f What percentage of the boxes contained exactly 50 matches?

GROUPED DISCRETE DATA

A local high school is concerned about the number of vehicles passing by between 8.45 am and 9.00 am. Over 30 consecutive week days they recorded data.

The results were: 48, 34, 33, 32, 28, 39, 26, 37, 40, 27, 23, 56, 33, 50, 38, 62, 41, 49, 42, 19, 51, 48, 34, 42, 45, 34, 28, 34, 42, 45, 34

In situations like this we group the data into class intervals.

It seems sensible to use class intervals of length 10 in this case.

The tally/frequency table is:

<table>
<thead>
<tr>
<th>Number of cars</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 to 19</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>20 to 29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 to 39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40 to 49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 to 59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 to 69</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>30</td>
</tr>
</tbody>
</table>

STEM-AND-LEAF PLOTS

A stem-and-leaf plot (often called a stemplot) is a way of writing down the data in groups. It is used for small data sets.

A stemplot shows actual data values. It also shows a comparison of frequencies. For numbers with two digits, the first digit forms part of the stem and the second digit forms a leaf.

For example, for the data value 27, 2 is recorded on the stem, 7 is a leaf value.

The stem-and-leaf plot is:

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>86738</td>
</tr>
<tr>
<td>3</td>
<td>4329738444</td>
</tr>
<tr>
<td>4</td>
<td>801928252</td>
</tr>
<tr>
<td>5</td>
<td>6014</td>
</tr>
<tr>
<td>6</td>
<td>2     Note: 2</td>
</tr>
</tbody>
</table>

The ordered stem-and-leaf plot is:

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>36788</td>
</tr>
<tr>
<td>3</td>
<td>2334444789</td>
</tr>
<tr>
<td>4</td>
<td>01225889</td>
</tr>
<tr>
<td>5</td>
<td>0146</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>
The ordered stemplot arranges all data from smallest to largest.

Notice that:
- all the actual data is shown
- the minimum (smallest) data value is 19
- the maximum (largest) data value is 62
- the ‘thirties’ interval (30 to 39) occurred most often.

Note: Unless otherwise stated, stem-and-leaf plot, or stemplot, means ordered stem-and-leaf plot.

COLUMN GRAPHS

A vertical column graph can be used to display grouped discrete data.

For example, consider the local high school data.

The frequency table is:

<table>
<thead>
<tr>
<th>Number of cars</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 19</td>
<td>1</td>
</tr>
<tr>
<td>20 to 29</td>
<td>5</td>
</tr>
<tr>
<td>30 to 39</td>
<td>10</td>
</tr>
<tr>
<td>40 to 49</td>
<td>9</td>
</tr>
<tr>
<td>50 to 59</td>
<td>4</td>
</tr>
<tr>
<td>60 to 69</td>
<td>1</td>
</tr>
</tbody>
</table>

The column graph for this data is:

Note that once data has been grouped in this manner there could be a loss of useful information for future analysis.

EXERCISE 4D

1. The data set below is the test scores (out of 100) for a Science test for 42 students.

   81 56 29 78 68 69 80 89 92 58 66 56 88 51 67 64 62 55 56 75 90 92 47 59 64 89 62 39 72 80 95 68 80 64 53 43 61 71 38 44 88

   a. Construct a tally and frequency table for this data using class intervals 0 - 9, 10 - 19, 20 - 29, ......., 90 - 100.
   b. What percentage of the students scored 50 or more for the test?
   c. What percentage of students scored less than 60 for the test?
   d. Copy and complete the following:
      “More students had a test score in the interval .......... than in any other interval.”
   e. Draw a column graph of the data.

2. Following is an ordered stem-and-leaf plot of the number of goals kicked by individuals in an Aussie rules football team during a season. Find:

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2 3 7</td>
</tr>
<tr>
<td>1</td>
<td>0 4 4 7 8 9 9</td>
</tr>
<tr>
<td>2</td>
<td>0 0 1 1 2 2 3 5 5 6 8 8</td>
</tr>
<tr>
<td>3</td>
<td>0 1 2 4 4 5 8 9</td>
</tr>
<tr>
<td>4</td>
<td>0 3 7</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

   a. the minimum number kicked
   b. the maximum number kicked
   c. the number of players who kicked greater than 25 goals
   d. the number players who kicked at least 40 goals
   e. the percentage of players who kicked less than 15 goals.
How would you describe the distribution of the data?

Hint: Turn your stemplot on its side.

3 The test score, out of 50 marks, is recorded for a group of 45 Geography students.

35 29 39 27 26 29 36 41 45 29 25 50 30 33 34 22 35 48 20 32 34 39 41 46 35 35 43 45 50 30 34 36 25 42 36 25 20 18 9 40 32 33 28 33 34

a Construct an unordered stem-and-leaf plot for this data using 0, 1, 2, 3, 4 and 5 as the stems.

b Redraw the stem-and-leaf plot so that it is ordered.

c What advantage does a stem-and-leaf plot have over a frequency table?

d What is the highest lowest mark scored for the test?

e If an ‘A’ was awarded to students who scored 42 or more for the test, what percentage of students scored an ‘A’?

f What percentage of students scored less than half marks for the test?

CONTINUOUS (INTERVAL) DATA

Continuous data is numerical data which has values within a continuous range.

For example, if we consider the weights of students in a netball training squad we might find that all weights lie between 40 kg and 90 kg.

Suppose 2 students lie in the 40 kg up to but not including 50 kg, 5 students lie in the 50 kg up to but not including 60 kg, 11 students lie in the 60 kg up to but not including 70 kg, 7 students lie in the 70 kg up to but not including 80 kg, 1 student lies in the 80 kg up to but not including 90 kg.

The frequency table is shown below:

<table>
<thead>
<tr>
<th>Weight interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 - &lt; 50</td>
<td>2</td>
</tr>
<tr>
<td>50 - &lt; 60</td>
<td>5</td>
</tr>
<tr>
<td>60 - &lt; 70</td>
<td>11</td>
</tr>
<tr>
<td>70 - &lt; 80</td>
<td>7</td>
</tr>
<tr>
<td>80 - &lt; 90</td>
<td>1</td>
</tr>
</tbody>
</table>

We could use a histogram to represent the data graphically.

Weights of the students in the netball squad

HISTOGRAMS

A histogram is a vertical column graph used to represent continuous grouped data.

There are no gaps between the columns in a histogram as the data is continuous.

The bar widths must be equal and each bar height must reflect the frequency.
The time, in minutes (ignoring any seconds) for shoppers to exit a shopping centre on a given day is as follows:

17 12 5 32 7 41 37 36 27 41 24 49 38 22 62 25
19 37 21 4 26 12 32 22 39 14 52 27 29 41 21 69

a Organise this data on a frequency table. Use time intervals of 0 -, 10 -, 20 -, etc.

b Draw a histogram to represent the data.

**Example 2**

The time, in minutes (ignoring any seconds) for shoppers to exit a shopping centre on a given day is as follows:

17 12 5 32 7 41 37 36 27 41 24 49 38 22 62 25
19 37 21 4 26 12 32 22 39 14 52 27 29 41 21 69

<table>
<thead>
<tr>
<th>Time int.</th>
<th>Tally</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - &lt; 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 - &lt; 20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 - &lt; 30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 - &lt; 40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40 - &lt; 50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 - &lt; 60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 - &lt; 70</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:**
- The continuous data has been grouped into classes.
- The class with the highest frequency is called the modal class.
- The size of the class is called the class interval. In the above example it is 10.
- As the continuous data has been placed in groups it is sometimes referred to as interval data.

**INVESTIGATION 2**

**CHOOSING CLASS INTERVALS**

When dividing data values into intervals, the choice of how many intervals to use, and hence the width of each class, is important.

**What to do:**
1. Click on the icon to experiment with various data sets. You can change the number of classes. How does the number of classes alter the way we can read the data?
2. Write a brief account of your findings.

As a rule of thumb we generally use approximately \( \sqrt{n} \) classes for a data set of \( n \) individuals. For very large sets of data we use more classes rather than less.

**EXERCISE 4E**

1. The weights (kg) of players in a boy’s hockey squad were found to be:
   72 69 75 50 59 80 51 48 84 58 67 70 54 77 49 71 63 46 62 56
   61 70 60 65 52 65 68 65 77 63 71 60 63 48 75 63 66 82 72 76

   a Using classes 40 - < 50, 50 - < 60, 60 - < 70, 70 - < 80, 80 - < 90, tabulate the data using columns of weight, tally, frequency.
b How many students are in the 60 - class?
c How many students weighed less than 70 kg?
d Find the percentage of students who weighed 60 kg or more.

2 A group of young athletes was invited to participate in a hammer throwing competition.
The following results were obtained:

<table>
<thead>
<tr>
<th>Distance (metres)</th>
<th>No. of athletes</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 - 20</td>
<td>5</td>
</tr>
<tr>
<td>20 - 30</td>
<td>21</td>
</tr>
<tr>
<td>30 - 40</td>
<td>17</td>
</tr>
<tr>
<td>40 - 50</td>
<td>8</td>
</tr>
<tr>
<td>50 -</td>
<td>3</td>
</tr>
</tbody>
</table>

a How many athletes threw less than 20 metres?
b What percentage of the athletes were able to throw at least 40 metres?

3 A plant inspector takes a random sample of two week old seedlings from a nursery and measures their height to the nearest mm.
The results are shown in the table alongside.

<table>
<thead>
<tr>
<th>Height (mm)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 - &lt; 75</td>
<td>22</td>
</tr>
<tr>
<td>75 - &lt; 100</td>
<td>17</td>
</tr>
<tr>
<td>100 - &lt; 125</td>
<td>43</td>
</tr>
<tr>
<td>125 - &lt; 150</td>
<td>27</td>
</tr>
<tr>
<td>150 - &lt; 175</td>
<td>13</td>
</tr>
<tr>
<td>175 - &lt; 200</td>
<td>5</td>
</tr>
</tbody>
</table>

c The total number of seedlings in the nursery is 2079. Estimate the number of seedlings which measure:
i less than 150 mm  
ii between 149 and 175 mm.

F MEASURES OF CENTRES OF DISTRIBUTIONS

Interested to know how your performance in mathematics is going? Are you about average or above average in your class? How does that compare with the other students studying the same subject in South Australia?

To answer questions such as these you need to be able to locate the centre of a data set.

The word ‘average’ is a commonly used word that can have different meanings. Statisticians do not use the word ‘average’ without stating which average they mean. Two commonly used measures for the centre or middle of a distribution are the mean and the median.

The mean of a set of scores is their arithmetic average obtained by adding all the scores and dividing by the total number of scores. The mean is denoted, \( \mu \).
The median of a set of scores is the middle score after they have been placed in order of size from smallest to largest.

In every day language, ‘average’ usually means the ‘mean’, but when the Australian Bureau of Census and Statistics reports the ‘average weekly income’ it refers to the median income.
Example 3

In a ballet class, the ages of the students are: 17, 13, 15, 12, 15, 14, 16, 13, 14, 18.
Find a the mean age b the median age of the class members.

\[ a \text{ mean } = \frac{17 + 13 + 15 + 12 + 15 + 14 + 16 + 13 + 14 + 18}{10} \]
\[ = \frac{147}{10} \]
\[ = 14.7 \]

b The ordered data set is: 12, 13, 13, 14, 14, 15, 16, 17, 18

There are two middle scores, 14 and 15. So the median is 14.5. \{their average\}

Note: For a sample containing \( n \) scores, in order, the median is the \( \left( \frac{n+1}{2} \right) \)th score.

If \( n = 11 \), \( \frac{11+1}{2} = 6 \), and so the median is the 6th score.
If \( n = 12 \), \( \frac{12+1}{2} = 6.5 \), and so the median is the average of the 6th and 7th scores.

STATISTICS USING TECHNOLOGY

From a computer package:

Click on the icon to enter the statistics package on the CD.

Enter data set 1: 5 2 3 3 6 4 5 3 7 5 7 1 8 9 5
Enter data set 2: 9 6 2 3 5 5 7 6 7 6 3 4 4 5 8 4

Examine the side-by-side column graphs.
Click on the Box-and-Whisker spot to view the side-by-side boxplots.
Click on the Statistics spot to obtain the descriptive statistics.
Click on Print to obtain a print-out of all of these on one sheet of paper.

Notice that the package handles the following types of data:
- ungrouped discrete
- ungrouped continuous
- grouped discrete
- grouped continuous
- already grouped discrete
- already grouped continuous

From a graphics calculator

A graphics calculator can be used to find descriptive statistics and to draw some types of graphs.
(You will need to change the viewing window as appropriate.)

Consider the data set: 5 2 3 3 6 4 5 3 7 5 7 1 8 9 5

No matter what brand of calculator you use you should be able to:
• Enter the data as a list.
• Enter the statistics calculation part of the menu and obtain the descriptive statistics like these shown.

\[
\bar{x} = \frac{\sum x}{n}
\]

\(x\) is the mean

• Obtain a box-and-whisker plot such as:

(These screen dumps are from a TI-83.)

• Obtain a vertical barchart if required.

• Enter a second data set into another list and obtain a side-by-side boxplot for comparison with the first one.
Use: 9 6 2 3 5 5 7 5 6 7 6 3 4 4 5 8 4

Now you should be able to create these by yourself.

In the following exercise you should use technology to find the measures of the middle of the distribution.

You should use both forms of technology available. The real world uses computer packages.

**EXERCISE 4F**

1 Below are the points scored by two basketball teams over a 14 match series:

**Team A:** 91, 76, 104, 88, 73, 55, 121, 98, 102, 91, 114, 82, 83, 91

**Team B:** 87, 104, 112, 82, 64, 48, 99, 119, 112, 77, 89, 108, 72, 87

Which team had the higher mean score?

2 Calculate the mean and median of each data set:
   
a 44, 42, 42, 49, 47, 44, 48, 47, 49, 41, 45, 40, 49
   b 148, 144, 147, 147, 149, 148, 146, 144, 145, 143, 142, 144, 147
   c 25, 21, 20, 24, 28, 27, 25, 29, 26, 28, 22, 25

3 A survey of 40 students revealed the following number of siblings per student:

2, 0, 0, 3, 2, 0, 0, 1, 3, 3, 4, 0, 0, 5, 3, 3, 0, 1, 4, 5,
0, 1, 1, 5, 1, 0, 0, 1, 2, 2, 1, 3, 2, 1, 4, 2, 0, 0, 1, 2

   a What is the mean number of siblings per student?
   b What is the median number of siblings per student?
The selling prices of the last 10 houses sold in a certain district were as follows: $196,000, $177,000, $261,000, $242,000, $306,000, $182,000, $198,000, $179,000, $181,000, $212,000.

a Calculate the mean and median selling prices of these houses and comment on the results.

b Which measure would you use if you were:
   i a vendor wanting to sell your house
   ii looking to buy a house in the district?

Find $x$ if 7, 11, 13, 14, 15, 17, 19 and $x$ have a mean of 14.

Towards the end of season, a basketballer had played 12 matches and had an average of 18.5 points per game. In the final two matches of the season the basketballer scored 23 points and 18 points. Find the basketballer’s new average.

The mean and median of a set of 9 measurements are both 14. If 7 of the measurements are 9, 11, 13, 15, 16, 19 and 21, find the other two measurements.

Seven sample values are: 3, 8, 4, 9, 5, $a$ and $b$, where $a < b$. These have a mean of 7 and a median of 6. Find $a$ and $b$.

### MEASURES OF THE CENTRE FROM OTHER SOURCES

**Example 4**

The distribution obtained by counting the contents of 25 match boxes is shown:

Find the:
   a mean number of matches per box
   b median number of matches per box.

<table>
<thead>
<tr>
<th>Number of matches ($x$)</th>
<th>Frequency ($f$)</th>
<th>$fx$</th>
<th>Cumulative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>2</td>
<td>94</td>
<td>2</td>
</tr>
<tr>
<td>48</td>
<td>4</td>
<td>192</td>
<td>6</td>
</tr>
<tr>
<td>49</td>
<td>7</td>
<td>343</td>
<td>13</td>
</tr>
<tr>
<td>50</td>
<td>8</td>
<td>400</td>
<td>21</td>
</tr>
<tr>
<td>51</td>
<td>3</td>
<td>153</td>
<td>24</td>
</tr>
<tr>
<td>53</td>
<td>1</td>
<td>53</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>1235</td>
<td>-</td>
</tr>
</tbody>
</table>

a mean $= \frac{\sum fx}{\sum f} = \frac{1235}{25} = 49.4$

b median is the 13th score $= 49$

*Note: 6 scores are 47 or 48. 13 scores are 47, 48 or 49. \( \frac{25+1}{2} = 13 \), i.e., the 13th*
Use **technology** to answer these questions.

9 A hardware store maintains that packets contain 60 screws. To test this, a quality control inspector tested 100 packets and found the following distribution:
   a Find the mean and median number of screws per packet.
   b Comment on these results in relation to the store’s claim.
   c Which of these two measures is more reliable? Comment on your answer.

<table>
<thead>
<tr>
<th>Number of screws</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>8</td>
</tr>
<tr>
<td>57</td>
<td>11</td>
</tr>
<tr>
<td>58</td>
<td>14</td>
</tr>
<tr>
<td>59</td>
<td>18</td>
</tr>
<tr>
<td>60</td>
<td>21</td>
</tr>
<tr>
<td>61</td>
<td>8</td>
</tr>
<tr>
<td>62</td>
<td>12</td>
</tr>
<tr>
<td>63</td>
<td>8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

10 58 packets of Choc Fruits were opened and their contents counted. The following table gives the distribution of the number of Choc Fruits per packet sampled.

Find the mean and median of the distribution.

<table>
<thead>
<tr>
<th>Number in packet</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>7</td>
</tr>
<tr>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>24</td>
<td>10</td>
</tr>
<tr>
<td>25</td>
<td>14</td>
</tr>
<tr>
<td>26</td>
<td>11</td>
</tr>
<tr>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>28</td>
<td>3</td>
</tr>
</tbody>
</table>

11 The table alongside compares the mass at birth of some guinea pigs with their mass when they were two weeks old.
   a What was the mean birth mass?
   b What was the mean mass after two weeks?
   c What was the mean increase over the two weeks?

<table>
<thead>
<tr>
<th>Guinea Pig</th>
<th>Mass (g) at birth</th>
<th>Mass (g) at 2 weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>75</td>
<td>210</td>
</tr>
<tr>
<td>B</td>
<td>70</td>
<td>200</td>
</tr>
<tr>
<td>C</td>
<td>80</td>
<td>200</td>
</tr>
<tr>
<td>D</td>
<td>70</td>
<td>220</td>
</tr>
<tr>
<td>E</td>
<td>74</td>
<td>215</td>
</tr>
<tr>
<td>F</td>
<td>60</td>
<td>200</td>
</tr>
<tr>
<td>G</td>
<td>55</td>
<td>206</td>
</tr>
<tr>
<td>H</td>
<td>83</td>
<td>230</td>
</tr>
</tbody>
</table>

**Grouped class interval data:**

When data has been grouped into class intervals, it is not possible to find the measure of the centre directly from frequency tables. In these situations estimates can be made using the **midpoint** of the class to represent all scores within that interval.

The **midpoint** of a class interval is the mean of its endpoints.

For example, the midpoint for continuous data of class $40 - 50$ is $\frac{40 + 50}{2} = 45$.

The midpoint of discrete data of class $10 - 19$ is $\frac{10 + 19}{2} = 14.5$.

The **modal class** is the class with the highest frequency.
Example 5

The histogram displays the distance in metres that 28 golf balls were hit by one golfer.

a Construct the frequency table for this data and add any other columns necessary to calculate the mean and median.

b Estimate the mean for this data.

c Estimate the median for this data.

d Find the modal class for this data.

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Freq. (f)</th>
<th>Midpt. (x)</th>
<th>fx</th>
<th>Upper end point</th>
<th>Cumu frequ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>240 - &lt; 245</td>
<td>1</td>
<td>242.5</td>
<td>242.5</td>
<td>245</td>
<td>1</td>
</tr>
<tr>
<td>245 - &lt; 250</td>
<td>3</td>
<td>247.5</td>
<td>742.5</td>
<td>250</td>
<td>4</td>
</tr>
<tr>
<td>250 - &lt; 255</td>
<td>6</td>
<td>252.5</td>
<td>1515.0</td>
<td>255</td>
<td>10</td>
</tr>
<tr>
<td>255 - &lt; 260</td>
<td>2</td>
<td>257.5</td>
<td>515.0</td>
<td>260</td>
<td>12</td>
</tr>
<tr>
<td>260 - &lt; 265</td>
<td>7</td>
<td>262.5</td>
<td>1837.5</td>
<td>265</td>
<td>19</td>
</tr>
<tr>
<td>265 - &lt; 270</td>
<td>6</td>
<td>267.5</td>
<td>1605.0</td>
<td>270</td>
<td>25</td>
</tr>
<tr>
<td>270 - &lt; 275</td>
<td>2</td>
<td>272.5</td>
<td>545.0</td>
<td>275</td>
<td>27</td>
</tr>
<tr>
<td>275 - &lt; 280</td>
<td>1</td>
<td>277.5</td>
<td>277.5</td>
<td>280</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td></td>
<td>7280</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b Approximate mean \[\bar{x} = \frac{\sum fx}{\sum f} = \frac{7280}{28} = 260 \text{ metres.}\]

c There are 28 observations in this data set. The 13th to 19th distances all lie in class interval 260 - < 265, and so the median also lies in this class interval.

Now, the median is the \[\frac{28 + 1}{2} = 14.5 \text{th score}\]

\[\therefore \text{median} = 260 + \frac{2.5 \times 5}{7} \times 7 \text{ in the class}\]

\[\div 261.8\]

d There were 7 hits of distance between 260 and 265 metres, which is more than in any other. The modal class is therefore the class between 260 and 265 metres.

Use technology to answer these questions:

12 Find the approximate mean for each of the following distributions:

a

<table>
<thead>
<tr>
<th>Score (x)</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 5</td>
<td>2</td>
</tr>
<tr>
<td>6 - 10</td>
<td>7</td>
</tr>
<tr>
<td>11 - 15</td>
<td>9</td>
</tr>
<tr>
<td>16 - 20</td>
<td>8</td>
</tr>
<tr>
<td>21 - 25</td>
<td>3</td>
</tr>
</tbody>
</table>

b

<table>
<thead>
<tr>
<th>Score (x)</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 - 42</td>
<td>2</td>
</tr>
<tr>
<td>43 - 45</td>
<td>1</td>
</tr>
<tr>
<td>46 - 48</td>
<td>4</td>
</tr>
<tr>
<td>49 - 51</td>
<td>7</td>
</tr>
<tr>
<td>52 - 54</td>
<td>11</td>
</tr>
<tr>
<td>55 - 57</td>
<td>3</td>
</tr>
</tbody>
</table>
30 students sit a mathematics test and the results are as follows:

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 9</td>
<td>1</td>
</tr>
<tr>
<td>10 - 19</td>
<td>4</td>
</tr>
<tr>
<td>20 - 29</td>
<td>8</td>
</tr>
<tr>
<td>30 - 39</td>
<td>14</td>
</tr>
<tr>
<td>40 - 49</td>
<td>3</td>
</tr>
</tbody>
</table>

Find the approximate value of the mean score.

The table shows the weight of newborn babies at a hospital over a one week period. Find the approximate mean weight of the newborn babies.

<table>
<thead>
<tr>
<th>Weight (kg)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0 - &lt; 1.5</td>
<td>1</td>
</tr>
<tr>
<td>1.5 - &lt; 2.0</td>
<td>2</td>
</tr>
<tr>
<td>2.0 - &lt; 2.5</td>
<td>6</td>
</tr>
<tr>
<td>2.5 - &lt; 3.0</td>
<td>17</td>
</tr>
<tr>
<td>3.0 - &lt; 3.5</td>
<td>11</td>
</tr>
<tr>
<td>3.5 - &lt; 4.0</td>
<td>8</td>
</tr>
<tr>
<td>4.0 - &lt; 4.5</td>
<td>0</td>
</tr>
<tr>
<td>4.5 - &lt; 5.0</td>
<td>1</td>
</tr>
</tbody>
</table>

The table shows the petrol sales in one day by a number of city service stations.

a How many service stations were involved in the survey?
b Estimate the number of litres of petrol sold for the day by the service stations.
c Find the approximate mean sales of petrol for the day.

<table>
<thead>
<tr>
<th>Litres (L)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>3000 - &lt; 4000</td>
<td>5</td>
</tr>
<tr>
<td>4000 - &lt; 5000</td>
<td>1</td>
</tr>
<tr>
<td>5000 - &lt; 6000</td>
<td>7</td>
</tr>
<tr>
<td>6000 - &lt; 7000</td>
<td>18</td>
</tr>
<tr>
<td>7000 - &lt; 8000</td>
<td>13</td>
</tr>
<tr>
<td>8000 - &lt; 9000</td>
<td>6</td>
</tr>
</tbody>
</table>

In a set of data an outlier, or extreme value, is a value which is much greater than, or much less than, the other values.

Examine the effect of an outlier on the two measures of central tendency.

Your task: Examine the effect of an outlier on the two measures of central tendency.

What to do:
1 Consider the following set of data: 1, 2, 3, 3, 4, 4, 5, 6, 7. Calculate:
   a the mean    b the median.
2 Now introduce an extreme value, say 100, to the data. Calculate:
   a the mean    b the median.
3 Comment on the effect that this extreme value has on:
   a the mean    b the median.
4 Which of the two measures of central tendency is most affected by the inclusion of an outlier?
CHOOSING THE APPROPRIATE MEASURE

The mean and median can be used to indicate the centre of a set of numbers. Which of these values is a more appropriate measure to use will depend upon the type of data under consideration.

In real estate values the median is used to measure the middle of a set of house values.

When selecting which of the two measures of central tendency to use as a representative figure for a set of data, you should keep the following advantages and disadvantages of each measure in mind.

► Mean

- The mean’s main advantage is that it is commonly used, easy to understand and easy to calculate.
- Its main disadvantage is that it is affected by extreme values within a set of data and so may give a distorted impression of the data.

For example, consider the following data: 4, 6, 7, 8, 19, 111. The total of these 6 numbers is 155, and so the mean is approximately 25.8. Is 25.8 a representative figure for the data? The extreme value (or outlier) of 111 has distorted the mean in this case.

► Median

- The median’s main advantage is that it is easily calculated and is the middle value of the data.
- Unlike the mean, it is not affected by extreme values.
- The main disadvantage is that it ignores all values outside the middle range and so its representativeness is questionable.

MEASURING THE SPREAD OF DATA

If, in addition to having measures of the middle of a data set, we also have an indication of the spread of the data, then a more accurate picture of the data set is possible.

For example:

- The mean height of 20 boys in a year 11 class was found to be 175 cm.
- A carpenter used a machine to cut 20 planks of size 175 cm long.

Even though the means of both data sets are roughly the same, there is clearly a greater variation in the heights of boys than in the lengths of planks.

Commonly used statistics that indicate the spread of a set of data are:

- the range
- the interquartile range
- the standard deviation.

The range and interquartile range are commonly used when considering the variation about a median, whereas the standard deviation is used with the mean.
THE RANGE

The range is the difference between the maximum (largest) data value and the minimum (smallest) data value.

\[ \text{range} = \text{maximum data value} - \text{minimum data value} \]

**Example 6**

Find the range of the data set: 4, 7, 5, 3, 4, 3, 6, 5, 7, 5, 3, 8, 9, 3, 6, 5, 6

Searching through the data we find: minimum value = 3 maximum value = 9

\[ \therefore \text{range} = 9 - 3 = 6 \]

**THE UPPER AND LOWER QUARTILES AND THE INTERQUARTILE RANGE**

The median divides the ordered data set into two halves and these halves are divided in half again by the quartiles.

The middle value of the lower half is called the lower quartile \((Q_1)\). One-quarter, or 25\%, of the data have a value less than or equal to the lower quartile. 75\% of the data have values greater than or equal to the lower quartile.

The middle value of the upper half is called the upper quartile \((Q_3)\). One-quarter, or 25\%, of the data have a value greater than or equal to the upper quartile. 75\% of the data have values less than or equal to the upper quartile.

\[ \text{interquartile range} = \text{upper quartile} - \text{lower quartile} \]

The interquartile range is the range of the middle half (50\%) of the data.

The data set has been divided into quarters by the lower quartile \((Q_1)\), the median \((Q_2)\) and the upper quartile \((Q_3)\).

So, the interquartile range, is \[ \text{IQR} = Q_3 - Q_1. \]

**Example 7**

Herb’s pumpkin crop this year had pumpkins which weighed (kg):
2.3, 3.1, 2.7, 4.1, 2.9, 4.0, 3.3, 3.7, 3.4, 5.1, 4.3, 2.9, 4.2

For the distribution, find the: \( a \) range \( b \) median \( c \) interquartile range

We enter the data. Using a TI we obtain:

\[ \begin{array}{l}
a \quad \text{range} = 5.1 - 2.3 = 2.8 \text{ kg} \\
b \quad \text{median} = 3.4 \text{ kg} \\
c \quad \text{IQR} = Q_3 - Q_1 \\
\quad \quad = 4.15 - 2.9 \\
\quad \quad = 1.25 \text{ kg}
\end{array} \]
Example 8

Jason is the full forward in the local Aussie rules team.
The number of goals he has kicked each match so far this season is:
6, 7, 3, 7, 9, 8, 5, 4, 6, 6, 8, 7, 6, 6, 5, 4, 5, 6
Find Jason’s:
- a mean score per match
- b median score per match
- c range of scores
- d interquartile range of scores

We enter the data. Using a TI we obtain:
- a mean $\bar{x} = 5.95$ goals
- b median = 6 goals
- c range = $9 - 3 = 6$
- d IQR = $Q_3 - Q_1$ = $7 - 5 = 2$

EXERCISE 4G

1. For each of the following sets of data, find:
   - i the range
   - ii the median
   - iii the interquartile range
   a The ages of people in a youth group:
      13, 15, 15, 17, 16, 14, 17, 16, 16, 14, 13, 14, 16, 16, 15, 14, 15, 16
   b The salaries, in thousands of dollars, of building workers:
      45, 51, 53, 58, 66, 62, 62, 61, 62, 59, 58, 60
   c The number of beans in a pod:

<table>
<thead>
<tr>
<th>Number of beans</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
<td>11</td>
<td>18</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

   d Stem | Leaf
   1     | 3 5 7 7 9
   2     | 0 1 3 5 6 6 8
   3     | 0 4 4 7 9
   4     | 3 7
   5     | 2
   Scale: 5 | 2 means 52

2. The time spent (in minutes) by 24 people at a supermarket, waiting to be served, has been recorded as follows:
   0.0  0.2  1.4  0.0  2.2  0.0  0.3  0.0
   0.6  1.8  0.4  1.9  0.0  1.2  3.8  0.7
   2.0  0.9  2.7  4.6  0.4  1.6  2.1  0.6
   a Find the median waiting time.
   b Find the range and interquartile range of the waiting time.
Copy and complete the following statements:

i  “50% of the waiting times were greater than ........ minutes.”

ii  “75% of the waiting times were less than ...... minutes.”

iii  “The minimum waiting time was ....... minutes and the maximum waiting time was ..... minutes. The waiting times were spread over ...... minutes.”

FIVE-NUMBER SUMMARY AND THE BOX-AND-WHISKER PLOT

A box-and-whisker plot (or simply a boxplot) is a visual display of some of the descriptive statistics of a data set. It shows:

- the minimum value (Min_x)
- the lower quartile (Q_1)
- the median (Q_2)
- the upper quartile (Q_3)
- the maximum value (Max_x)

These five numbers form what is known as the five-number summary of a data set.

In Example 8 the five-number summary and the corresponding boxplot are:

minimum = 3
Q_1 = 5
median = 6
Q_3 = 7
maximum = 9

Note:
- The rectangular box represents the ‘middle’ half of the data set.
- The lower whisker represents the 25% of the data with smallest values.
- The upper whisker represents the 25% of the data with greatest values.

Example 9

Peta plays netball and throws these goals in a series of matches:
5 6 7 6 2 8 9 8 4 6 7 4 5 4 3 6 6.

a Construct the five-number summary.  
b Draw a boxplot of the data.  
c Find the range.  
d Find the percentage of matches where Peta threw goals or less.

From a gcalc: So the five-number summary is:

- Min_x = 2
- Q_1 = 4
- median = 6
- Q_3 = 7
- Max_x = 9

So the 5-number summary is:

- Min_x = 2
- Q_1 = 4
- median = 6
- Q_3 = 7
- Max_x = 9
Draw a boxplot for the following data, testing for outliers and marking them, if they exist, with an asterisk on the boxplot:

| 3 | 7 | 8 | 8 | 5 | 9 | 10 | 12 | 14 | 7 | 1 | 3 | 8 | 16 | 8 | 6 | 9 | 10 | 13 | 7 |

The ordered data set is:

\[
\begin{align*}
\text{Min}_x &= 1 \\
Q_1 &= 6.5 \\
\text{median} &= 8 \\
Q_3 &= 10 \\
\text{Max}_x &= 16
\end{align*}
\]

\[\text{IQR} = Q_3 - Q_1 = 3.5\]

Test for outliers:

\[
\begin{align*}
\text{upper boundary} &= \text{upper quartile} + 1.5 \times \text{IQR} \\
&= 10 + 1.5 \times 3.5 \\
&= 15.25
\end{align*}
\]

\[
\begin{align*}
\text{lower boundary} &= \text{lower quartile} - 1.5 \times \text{IQR} \\
&= 6.5 - 1.5 \times 3.5 \\
&= 1.25
\end{align*}
\]

As 16 is above the upper boundary it is an outlier.

As 1 is below the lower boundary it is an outlier.

So, the boxplot is:

\[
\begin{array}{c}
\text{variable} \\
0 & 2 & 4 & 6 & 8 & 10 & 12 & 14 & 16 & 18 & 20 \\
\end{array}
\]

* * 

\text{BOXPLOTS AND OUTLIERS}

Outliers are extraordinary data that are usually separated from the main body of the data. Outliers are either much larger or much smaller than most of the data.

\textbf{Note:} Outliers are important data values to consider. They should not be removed before analysis unless there is a genuine reason for doing so.

There are several ‘tests’ that identify data that are outliers. A commonly used test involves the calculation of ‘boundaries’:

\begin{itemize}
\item The upper boundary = \text{upper quartile} + 1.5 \times \text{IQR}.
\item The lower boundary = \text{lower quartile} - 1.5 \times \text{IQR}.
\end{itemize}

Outliers are marked with an asterisk on a boxplot and it is possible to have more than one outlier at either end. The whiskers extend to the last value that is not an outlier.

\textbf{Example 10}

Draw a boxplot for the following data, testing for outliers and marking them, if they exist, with an asterisk on the boxplot:

\[3, 7, 8, 8, 5, 9, 10, 12, 14, 7, 1, 3, 8, 16, 8, 6, 9, 10, 13, 7\]

The ordered data set is:

\[
\begin{align*}
\text{Min}_x &= 1 \\
Q_1 &= 6.5 \\
\text{median} &= 8 \\
Q_3 &= 10 \\
\text{Max}_x &= 16
\end{align*}
\]

\[\text{IQR} = Q_3 - Q_1 = 3.5\]

Test for outliers:

\[
\begin{align*}
\text{upper boundary} &= \text{upper quartile} + 1.5 \times \text{IQR} \\
&= 10 + 1.5 \times 3.5 \\
&= 15.25
\end{align*}
\]

\[
\begin{align*}
\text{lower boundary} &= \text{lower quartile} - 1.5 \times \text{IQR} \\
&= 6.5 - 1.5 \times 3.5 \\
&= 1.25
\end{align*}
\]

As 16 is above the upper boundary it is an outlier.

As 1 is below the lower boundary it is an outlier.

So, the boxplot is:

\[
\begin{array}{c}
\text{variable} \\
0 & 2 & 4 & 6 & 8 & 10 & 12 & 14 & 16 & 18 & 20 \\
\end{array}
\]

* * 

\text{Note:} Outliers are important data values to consider. They should not be removed before analysis unless there is a genuine reason for doing so.

There are several ‘tests’ that identify data that are outliers. A commonly used test involves the calculation of ‘boundaries’:

\begin{itemize}
\item The upper boundary = \text{upper quartile} + 1.5 \times \text{IQR}.
\item The lower boundary = \text{lower quartile} - 1.5 \times \text{IQR}.
\end{itemize}

Outliers are marked with an asterisk on a boxplot and it is possible to have more than one outlier at either end. The whiskers extend to the last value that is not an outlier.
3 A boxplot has been drawn to show the distribution of marks (out of 120) in a test for a particular class.

- What was the **highest mark** scored?
- What was the **lowest mark** scored?
- What was the **median test score** for this class?
- What was the **range of marks** scored for this test?
- What percentage of students scored 35 or more for the test?
- What was the **interquartile range** for this test?
- The top 25% of students scored a mark between ... and ....
- If you scored 63 for this test, would you be in the top 50% of students in this class?
- Comment on the symmetry of the distribution of marks.

4 A set of data has a lower quartile of 28, median of 36 and an upper quartile of 43.

- Calculate the **interquartile range** for this data set.
- Calculate the boundaries that identify outliers.
- Which of the data 20, 11, 52, 61 would be outliers?

5 Hati examines a new variety of pea and does a count on the number of peas in 33 pods. Her results were:

4, 7, 9, 3, 1, 11, 5, 4, 6, 6, 4, 4, 12, 8, 2, 3, 3, 6, 7, 8, 4, 3, 2, 5, 5, 5, 8, 7, 6, 5

- Find the **median, lower quartile and upper quartile** of the data set.
- Find the **interquartile range** of the data set.
- What are the lower and upper boundaries for outliers?
- Are there any outliers according to **c**?
- Draw a boxplot of the data set.

6 Sam counts the number of nails in several boxes and tabulates the data as shown below:

<table>
<thead>
<tr>
<th>Number of nails</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>14</td>
</tr>
<tr>
<td>24</td>
<td>17</td>
</tr>
<tr>
<td>25</td>
<td>11</td>
</tr>
<tr>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
</tr>
</tbody>
</table>

- Find the five-number summary for this data set.
- Find the **range** and **IQR** for the data set.
- Are there any outliers? Test for them.
- Construct a boxplot for the data set.
THE STANDARD DEVIATION

The standard deviation measures the average deviation of data values from the mean and may reveal more about the variation of the data set than the IQR.

For technical reasons beyond the level of this book, the formula for a standard deviation used for a sample is slightly different from the one used for the population.

For a sample of size \( n \), the standard deviation is defined as:

\[
\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_n - \mu)^2}{n - 1}} = \sqrt{\frac{\sum(x - \mu)^2}{n - 1}}
\]

Note: In this formula:
- \( x_i \) is a value of the sample and \( \mu \) is the sample mean.
- \((x_i - \mu)^2\) measures how far \( x_i \) is from the mean \( \mu \). The square ensures that all of the quantities are positive.
- If the sum of all of the \((x_i - \mu)^2\) is small, it indicates that most of the data values are close to the mean. Dividing this sum by \((n - 1)\) gives an indication of how far, on average, the data is from the mean.
- The square root is used to obtain the correct units. For example, if \( x_i \) is the weight of a student in kg, \( s^2 \) would be in kg^2.
- If there are only two sample points \( x_1 \) and \( x_2 \), the formula for standard deviation is the Pythagorean distance of \((x_1, x_2)\) from \((\mu, \mu)\). This is no accident. It is one reason the standard deviation is the most widely used measure for the spread of data.

For a population, if \( N \) is the population size, and \( \mu \) is the population mean, then the standard deviation is

\[
\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}
\]

In this formula:
- the Greek letter \( \mu \) (mu) is used for the population mean
- the Greek letter \( \sigma \) (sigma) is used for the population standard deviation.

Example 11

A greengrocer chain is to purchase oranges from two different wholesalers. They take five random samples of 40 oranges to examine them for skin blemishes. The counts for the number of blemished oranges are:

<table>
<thead>
<tr>
<th>Wholesaler</th>
<th>Healthy Eating</th>
<th>Freshfruit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>4 16 14 8 8</td>
<td>9 12 11 10 13</td>
</tr>
</tbody>
</table>

Find the mean and standard deviation for each data set, and hence compare the wholesale suppliers.
Wholesaler Healthy Eating:

<table>
<thead>
<tr>
<th>x</th>
<th>x - ( \bar{x} )</th>
<th>(x - ( \bar{x} ))^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>-6</td>
<td>36</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>0</td>
</tr>
</tbody>
</table>

The mean \( \bar{x} = \frac{50}{5} = 10 \)

The standard deviation \( s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \)

\( = \sqrt{\frac{96}{5-1}} \)

\( = 4.90 \)

Wholesaler Freshfruit:

<table>
<thead>
<tr>
<th>x</th>
<th>x - ( \bar{x} )</th>
<th>(x - ( \bar{x} ))^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>55</td>
<td>0</td>
</tr>
</tbody>
</table>

The mean \( \bar{x} = \frac{55}{5} = 11 \)

The standard deviation \( s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \)

\( = \sqrt{\frac{10}{5-1}} \)

\( = 1.58 \)

Wholesaler Freshfruit supplied oranges with one more blemish, on average, but with less variability (smaller \( s \) value) than those supplied by Healthy Eating.

7 The column graphs show two distributions.

Sample A

Sample B

a By looking at the graphs, which distribution appears to have wider spread?

b Find the mean of each sample.

c For each sample, use the table method of Example 11 to find the standard deviation.

Use technology to answer the following questions.

8 The following table shows the change in cholesterol levels in 6 volunteers after a two week trial of special diet and exercise.

<table>
<thead>
<tr>
<th>Volunteer</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change in cholesterol</td>
<td>0.8</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.4</td>
<td>2.8</td>
</tr>
</tbody>
</table>

a Find the standard deviation of the data.

b Recalculate the standard deviation with the outlier removed.

c What is the effect on the standard deviation of an extreme value?
9 The number of points scored by Andrew and Brad in the last 8 basketball matches are tabulated below.

| Points by Andrew | 23 | 17 | 31 | 25 | 25 | 19 | 28 | 32 |
| Points by Brad   | 9  | 29 | 41 | 26 | 14 | 44 | 38 | 43 |

a Find the mean and standard deviation for the number of points scored by each player.
b Which of the two players is more consistent?

10 Two samples of 20 have these symmetric distributions.

![Sample A graph]

![Sample B graph]

a By looking at the graphs determine which sample has the wider spread.
b For each distribution, find the i median ii range iii IQR
c Find the standard deviation of each distribution.
d What measures of spread are useful here?

11 Two samples of 22 have these symmetric distributions.

![Sample A graph]

![Sample B graph]

a By looking at the graphs determine which one has the wider spread.
b Find for each distribution the i median ii range iii IQR
c Find the standard deviation of each distribution.
d What measures of spread are useful here?

GROUPED DATA

For grouped data the sample standard deviation $s$, is

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

where $f$ is the frequency of the score $x$.

Note that the sample size is $n = \sum f$. 

SA_11FSC
Example 12

Find the standard deviation of the distribution of the number of children in a family:

<table>
<thead>
<tr>
<th>Children</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Families</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\bar{x} = \frac{\sum fx}{\sum f} = \frac{30}{10} = 3
\]

and

\[
s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n-1}} = \sqrt{\frac{12}{9}} = 1.15
\]

12 Without using technology, find the mean and standard deviation of the following maths test results.

<table>
<thead>
<tr>
<th>Test score ((x))</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency ((f))</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

13 The number of chocolates in 58 boxes is displayed in the given frequency table.

<table>
<thead>
<tr>
<th>Number of chocolates</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>31</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>13</td>
<td>12</td>
<td>12</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

Find the mean and standard deviation of this distribution.

Example 13

The weights (in kg) of 25 calves are displayed in the table shown.

Find the approximate value of the standard deviation by using class midpoints.

<table>
<thead>
<tr>
<th>Weight class (kg)</th>
<th>Centre of class ((x))</th>
<th>Frequency</th>
<th>(f)</th>
<th>(f(x - \bar{x})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 - 59</td>
<td>55</td>
<td>10</td>
<td>55</td>
<td>676</td>
</tr>
<tr>
<td>60 - 69</td>
<td>65</td>
<td>9</td>
<td>195</td>
<td>768</td>
</tr>
<tr>
<td>70 - 79</td>
<td>75</td>
<td>13</td>
<td>675</td>
<td>324</td>
</tr>
<tr>
<td>80 - 89</td>
<td>85</td>
<td>6</td>
<td>510</td>
<td>96</td>
</tr>
<tr>
<td>90 - 99</td>
<td>95</td>
<td>4</td>
<td>380</td>
<td>784</td>
</tr>
<tr>
<td>100 - 109</td>
<td>105</td>
<td>2</td>
<td>210</td>
<td>1152</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td></td>
<td><strong>25</strong></td>
<td><strong>2025</strong></td>
<td><strong>3800</strong></td>
</tr>
</tbody>
</table>

\[
\bar{x} = \frac{\sum fx}{\sum f} = \frac{2025}{25} = 81
\]

and

\[
s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n-1}} = \sqrt{\frac{3800}{25-1}} = 12.6
\]
14 The lengths of 30 trout are displayed in the frequency table. Use the approach in Example 13 to find the best estimate of the mean length and the standard deviation of the lengths.

<table>
<thead>
<tr>
<th>Length (cm)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 - &lt; 32</td>
<td>1</td>
</tr>
<tr>
<td>32 - &lt; 34</td>
<td>1</td>
</tr>
<tr>
<td>34 - &lt; 36</td>
<td>3</td>
</tr>
<tr>
<td>36 - &lt; 38</td>
<td>7</td>
</tr>
<tr>
<td>38 - &lt; 40</td>
<td>11</td>
</tr>
<tr>
<td>40 - &lt; 42</td>
<td>5</td>
</tr>
<tr>
<td>42 - &lt; 44</td>
<td>2</td>
</tr>
</tbody>
</table>

15 The weekly wages (in dollars) of 90 steel yard workers are displayed in the given frequency table.

Use technology to find the approximate mean wage and the standard deviation of the wages.

<table>
<thead>
<tr>
<th>Wages ($)</th>
<th>Number of workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>600 - &lt; 700</td>
<td>5</td>
</tr>
<tr>
<td>700 - &lt; 800</td>
<td>16</td>
</tr>
<tr>
<td>800 - &lt; 900</td>
<td>27</td>
</tr>
<tr>
<td>900 - &lt; 1000</td>
<td>16</td>
</tr>
<tr>
<td>1000 - &lt; 1100</td>
<td>12</td>
</tr>
<tr>
<td>1100 - &lt; 1200</td>
<td>8</td>
</tr>
<tr>
<td>1200 - &lt; 1300</td>
<td>4</td>
</tr>
<tr>
<td>1300 - &lt; 1400</td>
<td>2</td>
</tr>
</tbody>
</table>

**STATISTICS IN ACTION**

The Senior Secondary Assessment Board of South Australia (SSABSA) is responsible for awarding grades to thousands of students in their final years in school. The grades awarded to students undertaking Publicly Examined Subjects are a combination of the examination results and grades submitted by teachers.

To ensure that the final mark is fair, marks submitted by schools are compared with the examination results. Should, for example, a school’s average result be higher than the average result attained in the examination, the school’s marks would be decreased. But what happens if a normally good student performs very badly in the exam?

Consider an extreme case. Suppose the school has given this student 100%, but in the exam the student gets 0%. If there were only 10 students doing this subject in a school, the school assessed mark could be adjusted down by 10% for each student in the school! It might seem that a poor result by a single student could affect the marks of every student in the class.

SSABSA is, of course, aware of this type of problem. Extreme cases, that is, outliers, like this are removed and are not taken into account in calculating by how much school assessed marks should be adjusted.

**COMPARING DATA**

A great deal of research is concerned with improving performance. A basketball coach wants to know whether a new coaching technique is better than an old one. Medical researchers want to know if a new medicine is better than existing ones.
The statistical process breaks posing and solving problems into several stages.
- Identify the problem.
- Formulate a method of investigation.
- Collect data.
- Analyse the data.
- Interpret the result and form a conjecture.
- Consider the underlying assumptions.

CASE STUDY 1

COMPARING NUMERICAL DATA

Imagine the following situation.

In a certain school all students are encouraged to learn to play a musical instrument and become a member of one of the many school bands. The senior music teacher in a school is concerned about wind players breathing at the wrong places when playing music. The teacher knows that some of the players who do not have this difficulty have taken lessons in breath control.

The problem

Will taking precious time away from practice to give instructions on breathing to all the wind players be worthwhile?

Formulating a method of investigation

The music teacher decides to compare students who have taken lessons in breath control with those who have not. A sample of 23 students who have taken lessons in breath control is matched as closely as possible for age and the time they have been playing an instrument with a group of 23 students who have not had lessons in breath control. The players will be asked to play up and down a simple scale, and the length of time they play until they take their fifth breath will be measured.

Collecting the data

It is estimated that it will take about 5 minutes to measure the time for each student, so that a total time of about 4 hours is required to collect the data. Eight teachers are asked to help collect the data and a timetable is drawn up for the next two weeks.

The time to the nearest tenth of a second for each player is shown in numerical order.

Data from only 21 of the trained players was recorded:

23.8  48.2  51.3  51.6  53.5  57.9  58.1  60.2  60.6  61.1  62.4
66.6  67.3  68.7  68.8  72.1  72.2  72.7  77.6  83.5  92.8

Data from only 19 of the untrained players was recorded:

24.2  28.3  35.8  35.8  38.3  38.3  38.4  41.3  45.9  47.9
49.5  51.1  62.3  66.8  67.9  72.6  73.4  75.3  82.9

Analysing the data

A back-to-back stemplot of times to the nearest second and parallel boxplots show the difference between the two sets of data.
There is one outlier amongst the trained players. The median playing time for the trained players is higher than that for the untrained players. The interquartile range and, if we ignore the outlier, the range for the trained players is smaller than that for the untrained players, indicating that the trained players are more consistent. From the stemplot, the shape of the distribution for the trained players appears to be symmetric whereas that for the untrained players appears to be bimodal.

From this it could be conjectured that training players in breath control increases the ability of players to play for longer without the need to take extra breaths.

**Considering the underlying assumptions**

It is assumed that the main difference between the two groups of students is one of training. It is also possible that students who are more enthusiastic about playing music are more likely to make extra efforts such as learning about breathing. The bimodal nature of the distribution for the untrained players indicates there could be two different types of players; the better ones seem no different from the trained players.

Before taking any action, the enthusiasm of players should be further explored by, for example, finding out how long players spend practising on their instruments each week.

In *Case Study 1* we considered two variables:

- the categorical variable ‘level of training’ with two values, either some or none
- the continuous variable ‘length of playing’.

The level of training was used to explain the length of playing. The level of training is known as the **explanatory variable** or an **independent variable**. It is called ‘independent’ because it can be changed, for example, by providing training.

The length of playing without taking a breath is known as a **response variable** or **dependent variable**. The outcome depends on the level of training.

Dependent variables vary according to the changes in the independent variable. In mathematics, independent variables are usually plotted along the horizontal axis and dependent variables along the vertical axis.

In *Case Study 1*, the explanatory variable, level of training, was a categorical variable, whereas the response variable was a continuous variable.
Fifty volunteers participated in a double blind trial to test a drug to lower cholesterol levels. In a double blind trial neither the volunteers nor the experimenters knew who was given what treatment. Twenty five volunteers were given the medicine and the other twenty five were given a placebo which was a sugar pill that looked the same as the drug but would have no effect at all. The results are summarised below:

**Cholesterol levels of the 25 participants who took the drug:**
4.8  5.6  4.7  4.2  4.8  4.6  4.8  5.2  4.8  5.0  4.7  5.1  4.4  
4.7  4.9  6.2  4.7  4.7  4.4  5.6  3.2  4.4  4.6  5.2  4.7

**Cholesterol levels of the 25 participants who took the placebo:**
7.0  8.4  8.8  6.1  6.6  7.6  6.5  7.9  6.2  6.8  7.5  6.0  8.2  
5.7  8.3  7.9  6.7  7.3  6.1  7.4  8.4  6.6  6.5  7.6  6.1

**Example 14**

Fifty volunteers participated in a double blind trial to test a drug to lower cholesterol levels. In a double blind trial neither the volunteers nor the experimenters knew who was given what treatment. Twenty five volunteers were given the medicine and the other twenty five were given a placebo which was a sugar pill that looked the same as the drug but would have no effect at all. The results are summarised below:

**Cholesterol levels of the 25 participants who took the drug:**
4.8  5.6  4.7  4.2  4.8  4.6  4.8  5.2  4.8  5.0  4.7  5.1  4.4  
4.7  4.9  6.2  4.7  4.7  4.4  5.6  3.2  4.4  4.6  5.2  4.7

**Cholesterol levels of the 25 participants who took the placebo:**
7.0  8.4  8.8  6.1  6.6  7.6  6.5  7.9  6.2  6.8  7.5  6.0  8.2  
5.7  8.3  7.9  6.7  7.3  6.1  7.4  8.4  6.6  6.5  7.6  6.1

a What are the variables that are considered in this study? 
b Draw a back-to-back stemplot for the two sets of data. 
c Draw parallel boxplots for the data. 
d Interpret the result, and make a conjecture. 
e Consider the underlying assumptions.

**a** The independent variable is the categorical variable ‘treatment’ with two possible values, placebo and drug. The dependent variable is the continuous variable ‘cholesterol level’.

**b** Since the data is close together we split the stem to construct the stemplot. For example, the stem of 3 is recorded for values from 3.0 to < 3.5, and 3" for values from 3.5 to < 4.0.

**Placebo**

<table>
<thead>
<tr>
<th>3</th>
<th>3&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4&quot;</td>
</tr>
<tr>
<td>5</td>
<td>5&quot;</td>
</tr>
<tr>
<td>7</td>
<td>7&quot;</td>
</tr>
<tr>
<td>8 1 1 0</td>
<td>6</td>
</tr>
<tr>
<td>8 7 6 6 5 5</td>
<td>6&quot;</td>
</tr>
<tr>
<td>4 3 0</td>
<td>7</td>
</tr>
<tr>
<td>9 9 6 6 5</td>
<td>7&quot;</td>
</tr>
<tr>
<td>4 4 3 2</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>8&quot;</td>
</tr>
</tbody>
</table>

**Drug**

<table>
<thead>
<tr>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
</tr>
<tr>
<td>4 4 4</td>
</tr>
<tr>
<td>6 6 7 7 7 7 7 8 8 8 8 9</td>
</tr>
<tr>
<td>0 1 2 2</td>
</tr>
<tr>
<td>6 6</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

**Leaf unit:** 0.1 units of cholesterol

**c**

---

**Example 14**

Fifty volunteers participated in a double blind trial to test a drug to lower cholesterol levels. In a double blind trial neither the volunteers nor the experimenters knew who was given what treatment. Twenty five volunteers were given the medicine and the other twenty five were given a placebo which was a sugar pill that looked the same as the drug but would have no effect at all. The results are summarised below:

**Cholesterol levels of the 25 participants who took the drug:**
4.8  5.6  4.7  4.2  4.8  4.6  4.8  5.2  4.8  5.0  4.7  5.1  4.4  
4.7  4.9  6.2  4.7  4.7  4.4  5.6  3.2  4.4  4.6  5.2  4.7

**Cholesterol levels of the 25 participants who took the placebo:**
7.0  8.4  8.8  6.1  6.6  7.6  6.5  7.9  6.2  6.8  7.5  6.0  8.2  
5.7  8.3  7.9  6.7  7.3  6.1  7.4  8.4  6.6  6.5  7.6  6.1

a What are the variables that are considered in this study? 
b Draw a back-to-back stemplot for the two sets of data. 
c Draw parallel boxplots for the data. 
d Interpret the result, and make a conjecture. 
e Consider the underlying assumptions.

**a** The independent variable is the categorical variable ‘treatment’ with two possible values, placebo and drug. The dependent variable is the continuous variable ‘cholesterol level’.

**b** Since the data is close together we split the stem to construct the stemplot. For example, the stem of 3 is recorded for values from 3.0 to < 3.5, and 3" for values from 3.5 to < 4.0.

**Placebo**

<table>
<thead>
<tr>
<th>3</th>
<th>3&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4&quot;</td>
</tr>
<tr>
<td>5</td>
<td>5&quot;</td>
</tr>
<tr>
<td>7</td>
<td>7&quot;</td>
</tr>
<tr>
<td>8 1 1 0</td>
<td>6</td>
</tr>
<tr>
<td>8 7 6 6 5 5</td>
<td>6&quot;</td>
</tr>
<tr>
<td>4 3 0</td>
<td>7</td>
</tr>
<tr>
<td>9 9 6 6 5</td>
<td>7&quot;</td>
</tr>
<tr>
<td>4 4 3 2</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>8&quot;</td>
</tr>
</tbody>
</table>

**Drug**

<table>
<thead>
<tr>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
</tr>
<tr>
<td>4 4 4</td>
</tr>
<tr>
<td>6 6 7 7 7 7 7 8 8 8 8 9</td>
</tr>
<tr>
<td>0 1 2 2</td>
</tr>
<tr>
<td>6 6</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

**Leaf unit:** 0.1 units of cholesterol

**c**
There are two outliers among the volunteers taking the drug, but as there is little overlap between the two groups, they do not affect the conclusion. Those taking the drug had a lower median cholesterol level and a smaller IQR. From comparing the two medians we conjecture that the drug lowered the cholesterol level, and from the IQR we conjecture that the drug made the cholesterol level more uniform.

It is assumed that the group of volunteers had similar cholesterol levels before the trial, and that the lowering of cholesterol was due to the drug and not to bias in the sample. The randomness of the selection procedure should have made such bias very unlikely, but the cholesterol level of all volunteers should also have been recorded before the experiment.

**EXERCISE 4H**

1. Fancy chocolate frogs are enclosed in fancy paper and sell at 50 cents each. Plain chocolate frogs are sold in bulk and sell for $2.50 for 100 grams. To test which of the two varieties is more economical to buy, Lucy bought 20 individual fancy frogs and 20 plain frogs. The weights in grams of the frogs are given below.

   **Plain frogs:**
   - 11.8
   - 11.9
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0
   - 12.0

   **Fancy frogs:**
   - 15.2
   - 15.2
   - 15.1
   - 14.7
   - 14.7
   - 14.7
   - 14.7
   - 14.6
   - 14.6
   - 14.6
   - 14.6
   - 14.6
   - 14.6
   - 14.6
   - 14.6
   - 14.6
   - 14.6
   - 14.6
   - 14.6
   - 14.6
   - 14.6

   a. What are the variables that are being considered in this study?
   b. Find the five-number summary for the two samples, and draw two parallel boxplots.
   c. From your analysis in part b, form a conjecture about the weights of the frogs.
   d. Which of the two types of frogs would be more economical to buy?

2. Two taxi drivers, John and Peter, argued about who was more successful. To settle the argument they agreed that they would randomly select 25 days on which they had worked over the past two months and record the amount of money they had collected on each day.

   The amount collected to the nearest dollar is shown below.

   **Peter:**
   - 194
   - 199
   - 188
   - 195
   - 168
   - 205
   - 196
   - 183
   - 93
   - 154
   - 147
   - 270
   - 116
   - 132
   - 253
   - 205
   - 191
   - 182
   - 118
   - 140
   - 155
   - 190
   - 223
   - 233
   - 208

   **John:**
   - 276
   - 152
   - 127
   - 163
   - 180
   - 161
   - 110
   - 153
   - 110
   - 147
   - 152
   - 223
   - 139
   - 139
   - 142
   - 141
   - 97
   - 116
   - 129
   - 215
   - 241
   - 159
   - 174
   - 158
   - 160

   a. What are the variables that are being considered?
   b. Construct a back-to-back stemplot for this data.
   c. Which of the two drivers do you conjecture was more successful?
3 A new cancer drug was being developed. It was claimed that it helped lengthen the survival time of patients once they were diagnosed with a certain form of cancer. The drug was first tested on rats to see if it was effective on them. Forty rats were infected with the type of cancer cells that the drug was supposed to fight. Then, using a random allocation process, two groups of twenty rats were formed. One group was given the drug and one group was not. The experiment was to run for a maximum of 192 days. The survival time in days of each rat in the experiment was recorded.

Survival time of rats that were given the drug:
64  78  106  106  106  127  127  134  148  186
192* 192* 192* 192* 192* 64  78  106  106

Survival time of rats that were not given the drug:
37  38  42  43  43  43  43  48  49
51  51  55  57  59  62  66  69  86  37

* denotes that the rat was still alive at the end of the experiment.

a What are the variables that are being considered?
b Construct a back-to-back stemplot for this data.
c Make a conjecture based on the analysis in part b.

4 Bill decided to compare the effect of two fertilisers; one organic, the other inorganic. Bill prepared three identical plots named A, B and C. In each plot he planted 40 radish seeds. After planting, each plot was treated in an identical manner, except that plot A received no fertiliser, plot B received the organic fertiliser, and plot C received the inorganic fertiliser.

The data supplied below is the length to the nearest cm of foliage of the individual plants that survived up to the end of the experiment.

Data from plot A:
27  29  9  10  8  36  36  42  32  32  32  30  38
32  30  34  22

Data from plot B:
51  54  56  41  50  47  47  46  48  52  34  20  28
47  58  56  63  66  54  48  48  53  47  29  46  33
45  58  34

Data from plot C:
55  76  65  61  67  69  68  64  76  59  56  79  70
69  70  76  43  70  62  60  58  79  65  75  60  39
68  68  63  54  61  72  58  77  66  65  47  50

a What are the variables that are being considered?
b Construct a five-number summary for each of the three samples of data.
c Construct parallel boxplots for the three sets of data.
d Make conjecture based on your analysis in c. Does the fact that there are more plants in plot C make a difference to your conjecture?
An educational researcher believes that girls are not as good at science as boys. To study this claim, 40 thirteen year old girls and 40 thirteen year old boys were selected to answer twenty basic science questions. The results of the test are shown below.

**Boys**
14 13 18 17 12 15 12 14 15 14 16 14 14 14 15 17 17 19 14 9 15 12 15 14 15 10 12

**Girls**
15 16 15 13 16 18 11 12 16 15 10 14 14 14 13 17 14 11 12 15 17 16 14 18 14 16 15 14 14 19 13 15 12 12 11 13 13 17 14 10

a. What are the variables that are considered in this study?

b. Plot a column graph for each set of data.

c. Find five-number summaries for each data set.

d. Plot parallel boxplots for the two sets of data.

e. Calculate the median, IQR, the mean and standard deviations for each data set.

f. What conclusions can you draw from the analysis?

---

**Example 15**

The independent variable is the categorical variable ‘gender’ with two values, either girl or boy. The dependent variable is the discrete variable ‘number of correct answers’ with possible integer values of 0 to 20.

Technology was used to construct the two column graphs.

The five-number summaries are recorded in the following table.

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Q₁</th>
<th>Median</th>
<th>Q₃</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys</td>
<td>9</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>Girls</td>
<td>10</td>
<td>13</td>
<td>14</td>
<td>16</td>
<td>19</td>
</tr>
</tbody>
</table>

---

**Boys**

**Girls**
The information is summarised in the following table.

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>IQR</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys</td>
<td>14</td>
<td>2</td>
<td>14.1</td>
<td>2.06</td>
</tr>
<tr>
<td>Girls</td>
<td>14</td>
<td>3</td>
<td>14.2</td>
<td>2.20</td>
</tr>
</tbody>
</table>

The analysis, particularly the summary statistics found in e, does not show any clear differences between the two data sets. From this we conjecture that there is no difference between girls and boys in answering test questions on science.

An athletics coach wanted to test a new diet on a group of 20 runners. To measure any difference he recorded their time to run 50 metres before the diet started and four weeks after the diet had started. The results in seconds are recorded below.

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys</td>
<td>5.84</td>
<td>5.34</td>
</tr>
<tr>
<td></td>
<td>5.46</td>
<td>5.38</td>
</tr>
<tr>
<td></td>
<td>5.55</td>
<td>5.51</td>
</tr>
<tr>
<td></td>
<td>5.60</td>
<td>5.62</td>
</tr>
<tr>
<td></td>
<td>5.67</td>
<td>5.71</td>
</tr>
<tr>
<td></td>
<td>5.89</td>
<td>5.84</td>
</tr>
<tr>
<td></td>
<td>6.53</td>
<td>6.71</td>
</tr>
<tr>
<td></td>
<td>6.70</td>
<td>6.97</td>
</tr>
<tr>
<td></td>
<td>6.96</td>
<td>7.32</td>
</tr>
<tr>
<td></td>
<td>7.27</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.4</td>
<td>7.36</td>
</tr>
<tr>
<td></td>
<td>7.41</td>
<td>7.65</td>
</tr>
<tr>
<td></td>
<td>7.58</td>
<td>7.81</td>
</tr>
<tr>
<td></td>
<td>7.70</td>
<td>8.06</td>
</tr>
<tr>
<td></td>
<td>8.17</td>
<td>8.06</td>
</tr>
<tr>
<td></td>
<td>8.68</td>
<td>8.42</td>
</tr>
<tr>
<td></td>
<td>8.93</td>
<td>8.63</td>
</tr>
<tr>
<td></td>
<td>9.66</td>
<td>8.91</td>
</tr>
<tr>
<td></td>
<td>10.0</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>10.01</td>
<td>10.92</td>
</tr>
</tbody>
</table>

a What are the variables under consideration?
b Find the median, IQR, mean and standard deviation for each data set.
c Form a conjecture about the effectiveness of the diet.

Note: In this section we have compared data sets which are either similar or very different.

Consider the two sample sets:

**Sample 1:**
- 9.39
- 9.95
- 7.97
- 9.95
- 10.22
- 10.6
- 13.12
- 9.71
- 9.81
- 9.32

**Sample 2:**
- 11.19
- 12.76
- 12.8
- 10.24
- 10.01
- 9.19
- 12.47
- 10.76
- 12.7
- 10.43

This table shows the basic statistics for the two sets.

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>IQR</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>10.1</td>
<td>1.45</td>
<td>10.3</td>
<td>1.29</td>
</tr>
<tr>
<td>Sample 2</td>
<td>10.6</td>
<td>2.18</td>
<td>10.7</td>
<td>1.68</td>
</tr>
</tbody>
</table>

There is some evidence that the data in sample 2 is a little larger than that of sample 1, but the difference could be due to the random nature in which the samples were collected.

To decide whether the difference is unlikely to be due to chance alone for small apparent differences requires knowledge of hypothesis testing which is not covered in this book.
CASE STUDY 2

A court in a recreation centre can be used either for tennis or for basketball. The manager wants to make the best use of this court.

The problem
When should the court be used for tennis and when should it be used for basketball?

Formulating the method of investigation
The manager identifies three different types of customers using the court:
- those coming during weekends
- those coming Monday to Friday during day time
- those coming Monday to Friday in the evening.

The manager decided to ask the customers using the centre when they used the court and what sport they preferred to play.

Collecting data
A sheet of paper is left by the front office for customers to fill in their preferences. The form with a few entries is shown.

<table>
<thead>
<tr>
<th>Time</th>
<th>Preferred sport</th>
</tr>
</thead>
<tbody>
<tr>
<td>weekends, day time, evening</td>
<td>Tennis, basketball</td>
</tr>
<tr>
<td>1 weekend</td>
<td>tennis</td>
</tr>
<tr>
<td>2 evening</td>
<td>basketball</td>
</tr>
<tr>
<td>3 daytime</td>
<td>tennis</td>
</tr>
<tr>
<td>4 evening</td>
<td>basketball</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Analysing the data
The data that was collected is summarised in this table of counts.

<table>
<thead>
<tr>
<th>Time</th>
<th>Tennis</th>
<th>Basketball</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weekends</td>
<td>19</td>
<td>78</td>
<td>97</td>
</tr>
<tr>
<td>Day time</td>
<td>23</td>
<td>9</td>
<td>32</td>
</tr>
<tr>
<td>Evening</td>
<td>17</td>
<td>53</td>
<td>70</td>
</tr>
</tbody>
</table>

There are two related variables:
- time with 3 possible values: weekends, day time, evening
- sport played with 2 possible values: tennis, basketball.

To compare the numbers in the table we must first convert them to proportions or percentages.

Complete the following table of percentages.

<table>
<thead>
<tr>
<th>Time</th>
<th>Tennis</th>
<th>Basketball</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weekends</td>
<td>\frac{19}{97} \times 100 = 19.6</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Day time</td>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Evening</td>
<td></td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>
Interpreting the results and forming a conjecture

About 75% to 80% of weekend and evening customers prefer to play basketball, whereas only 30% of the day time players prefer this sport.

The manager decides to use the court in proportion to the interest shown. On this basis the manager timetables the court for tennis on Sunday morning, for the 4 days times Monday, Tuesday, Thursday and Friday, and for Wednesday evening. For the other times the court will be used for basketball.

The manager conjectures that this division is an accurate reflection of the interests of its customers.

Consider the underlying assumptions

One assumption is that the number of people who show an interest in the use of the court reflects the amount of time the court should be used. Basketball is a team sport and a game of basketball involves more people than a game of tennis and may also be more difficult to schedule. This should be examined further.

In Case study 2 the two variables were:

- time, with possible values: weekends, day and evening
- sport played, with two values: tennis and basketball.

It is not always easy to decide which is the dependent or independent variable. In this case ‘time’ was taken as the independent variable and ‘sport played’ as the dependent variable.

The changes in household size in the town of Calcakoo since 1950 are to be investigated. The table given shows the number of private households of different size in 1950, 1975 and 2000. This data was gathered from a census.

- Name the dependent and independent variables involved in this investigation.
- Calculate a table of appropriate percentages, to the nearest percent, that will help the investigation.
- Use technology to produce a side-by-side column graph.
- Using the analysis in b and c, form a conjecture about change in household sizes in Calcakoo.
7 A market research company is contracted to investigate the ages of the people who listen to the only radio stations in the area. The radio stations are KPH, EWJ and MFB.

The research company surveyed a random sample of 1000 people from the area and asked them which radio station they mainly listened to. The results are summarised in the table above.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>KPH</th>
<th>EWJ</th>
<th>MFB</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 30</td>
<td>57</td>
<td>61</td>
<td>196</td>
</tr>
<tr>
<td>30 - 60</td>
<td>51</td>
<td>103</td>
<td>57</td>
</tr>
<tr>
<td>&gt; 60</td>
<td>224</td>
<td>91</td>
<td>160</td>
</tr>
</tbody>
</table>

a State the two variables that are under investigation.
b Classify each of these variables as either dependent or independent.
c Produce an appropriate table of percentages so that analysis can be carried out.
d Use technology to produce a side-by-side column graph.
e Form a conjecture about what radio station various age groups listen to.

WRITING REPORTS

In writing a report it helps to plan it on the statistical process. Other points to bear in mind:

- Keep your language simple. Short sentences are easier to read than long sentences.
- Stick to the information and do not overstate your case.
- Do not use personal pronouns. Instead of “I collected the data.” use “The data was collected.”

The tradition in science is that any idea can be challenged, but the person holding that idea may not be. Keeping an article impersonal means that any criticism of the article is not (supposed to be) directed at the person who wrote it.

Example 16

Write a report for analysis carried out in Example 14.

1 State the problem
   This study examined the effectiveness of a cholesterol lowering drug.

2 Formulate method of investigation
   This drug was tried on human subjects in a double blind trial.

3 Collect data
   The actual raw data is usually not displayed in a report, but may be added as an appendix.
   50 volunteers were split randomly into two groups of 25 each. One group was given the drug, the other was given a placebo with neither the researcher nor the subjects knowing which treatment was applied. Blood samples were tested for cholesterol at the end of the study.
4 Analyse the data

Select the best display for the information. In this case the stemplot has been chosen. The boxplot essentially shows the same information and could also be used. Only display different graphs if they contain different information.

The results of the study are displayed in the back-to-back stemplot.

<table>
<thead>
<tr>
<th>Placebo</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3*</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2 4 4 4</td>
</tr>
<tr>
<td>4*</td>
<td>6 6 7 7 7 7 8 8 8 9</td>
</tr>
<tr>
<td>5</td>
<td>0 1 2 2</td>
</tr>
<tr>
<td>7</td>
<td>6 6</td>
</tr>
<tr>
<td>5*</td>
<td>6 6</td>
</tr>
<tr>
<td>2 1 1 1 0</td>
<td>6 2</td>
</tr>
<tr>
<td>8 7 6 6 5 5</td>
<td>6* 7</td>
</tr>
<tr>
<td>4 3 0</td>
<td></td>
</tr>
<tr>
<td>9 9 6 6 5</td>
<td>7* 8</td>
</tr>
<tr>
<td>4 4 3 2 8</td>
<td></td>
</tr>
<tr>
<td>8 8*</td>
<td></td>
</tr>
</tbody>
</table>

Leaf unit: 0.1 units of cholesterol

5 Interpret results and form a conjecture

From the stemplot it can be clearly seen that those taking the drug had lower cholesterol levels at the end of the study. The median cholesterol level of the subjects taking the drug was 4.7 units compared with 7.0 for those receiving the placebo. There appears to be an outlier of 3.2 units in the subjects taking the drug. It is conjectured that the drug lowers cholesterol levels.

6 Consider underlying assumptions

It is assumed that the volunteers had similar cholesterol levels before the trial, and that the lower cholesterol levels of those taking the drug were due to the drug and not to bias in the sample. The randomness of the selection procedure would have made such bias unlikely, but the cholesterol level of all volunteers should have been checked at the beginning of the experiment.

7 Round off the report with a conclusion

From this experiment it can be seen that the drug lowered the cholesterol level. It should however be noted that the level of one subject dropped as low as 3.2 units and this could be dangerously low. It is recommended that if this drug is to be used, patients are monitored for possible harmful side effects.

8 Write a report of your analysis of one of the following:

A The chocolate frogs in question 1.
B The new cancer drug in question 3.
C Your analysis of the diet in question 5.
This section examines how well a sample statistic reflects a population parameter if the only errors are statistical errors.

### INVESTIGATION 4

Following is a table in random order of birth weights to the nearest 0.01 kg of 216 babies born without complications. It is arranged into 6 blocks of 36 each.

We shall examine how well a sample of 15 reflects the population of 216 babies.

<table>
<thead>
<tr>
<th>Block</th>
<th>Weights of Babies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.23 3.10 3.06 3.14 3.33 3.04</td>
</tr>
<tr>
<td></td>
<td>3.05 3.32 3.10 3.25 3.44 2.79</td>
</tr>
<tr>
<td></td>
<td>2.90 2.86 3.39 3.20 4.19 3.23</td>
</tr>
<tr>
<td></td>
<td>2.69 2.81 3.64 3.33 3.78 3.56</td>
</tr>
<tr>
<td></td>
<td>3.49 3.31 3.42 3.14 3.04 3.24</td>
</tr>
<tr>
<td></td>
<td>3.01 3.14 3.05 4.15 3.68 3.18</td>
</tr>
<tr>
<td>2</td>
<td>3.62 3.03 3.58 3.70 2.92 3.68</td>
</tr>
<tr>
<td></td>
<td>3.33 3.23 3.51 3.26 3.25 3.55</td>
</tr>
<tr>
<td></td>
<td>3.19 3.35 3.02 3.89 3.36 2.97</td>
</tr>
<tr>
<td></td>
<td>3.23 2.93 2.38 3.18 3.44 3.16</td>
</tr>
<tr>
<td></td>
<td>3.47 2.82 3.16 3.89 2.89 3.27</td>
</tr>
<tr>
<td></td>
<td>3.31 3.17 2.8 3.32 3.60 2.95</td>
</tr>
<tr>
<td>3</td>
<td>3.48 3.40 3.08 2.94 3.38 3.83</td>
</tr>
<tr>
<td></td>
<td>3.04 3.18 3.13 3.27 3.25 3.50</td>
</tr>
<tr>
<td></td>
<td>3.41 3.42 3.81 3.03 3.49 3.08</td>
</tr>
<tr>
<td></td>
<td>3.62 2.91 3.66 3.50 2.84 3.34</td>
</tr>
<tr>
<td></td>
<td>2.78 2.84 3.24 2.95 3.24 2.81</td>
</tr>
<tr>
<td></td>
<td>3.53 3.24 3.27 3.27 3.17 3.29</td>
</tr>
<tr>
<td>4</td>
<td>3.19 3.26 3.46 3.23 3.41 3.34</td>
</tr>
<tr>
<td></td>
<td>3.74 2.93 3.17 2.98 3.16 3.03</td>
</tr>
<tr>
<td></td>
<td>3.70 3.49 3.26 3.34 2.89 2.98</td>
</tr>
<tr>
<td></td>
<td>3.08 3.24 2.68 3.66 3.24 2.83</td>
</tr>
<tr>
<td></td>
<td>3.12 3.19 3.43 3.53 3.51 3.45</td>
</tr>
<tr>
<td></td>
<td>3.45 3.83 3.21 2.83 3.07 3.32</td>
</tr>
<tr>
<td>5</td>
<td>3.24 3.33 3.02 3.10 3.25 3.57</td>
</tr>
<tr>
<td></td>
<td>3.76 3.23 3.75 3.66 3.55 3.73</td>
</tr>
<tr>
<td></td>
<td>3.38 2.66 3.23 3.26 3.25 2.91</td>
</tr>
<tr>
<td></td>
<td>3.36 3.88 3.26 3.83 3.52 3.81</td>
</tr>
<tr>
<td></td>
<td>3.28 3.11 3.57 3.39 3.12 3.17</td>
</tr>
<tr>
<td></td>
<td>2.69 2.94 3.98 2.79 3.69 3.68</td>
</tr>
<tr>
<td>6</td>
<td>3.18 3.25 3.12 3.24 3.09 3.61</td>
</tr>
<tr>
<td></td>
<td>3.24 3.17 3.65 3.50 3.47 3.53</td>
</tr>
<tr>
<td></td>
<td>3.23 3.58 3.28 3.25 3.13 2.96</td>
</tr>
<tr>
<td></td>
<td>2.91 3.23 3.67 3.27 3.08 3.74</td>
</tr>
<tr>
<td></td>
<td>3.32 3.62 3.86 2.97 3.09 3.19</td>
</tr>
<tr>
<td></td>
<td>2.59 3.51 3.42 3.16 2.95 4.02</td>
</tr>
</tbody>
</table>
What to do:

1 a Select a sample of 15 babies from this population by:
   • rolling a die to select one of the 6 blocks
   • rolling the die again to select a column in the block
   • rolling the die again to select a baby’s weight by row number in the column
   • select 15 entries reading from left to right across the page from the weight you found.

   For example, if the 3 rolls of the die produced {3, 5, 2}, the sample weights would be:
   3.25, 3.50, 3.74, 2.93, 3.17, 2.98, 3.16, 3.03, 3.41, 3.42, 3.81, 3.03, 3.49, 3.70, 3.49

b Comment on whether the sample found in a is a random sample (i.e., does every baby have the same chance of being selected).

2 For your sample:
   a find the five-number summary
   b find the mean and standard deviation.

3 Repeat this process for another four samples.

   Compare your results of 2 and 3 with other students in your class.

4 The table of babies’ birth weights is also contained in column A2 to A217 of the spreadsheet “Babies’ Weights”. Treating the data in the spreadsheet as the population:
   a find the five-number summary of the population
   b find the mean and standard deviation of the population.

   Compare the results with those of your sample.

You should have discovered that the sample mean, median, IQR and the standard deviation are close to the population mean, median, IQR and standard deviation, but that the sample range is not always a good indicator of the population range.

In the following investigation you will explore how well you can predict the mode of a population from the mode of a sample.

A lot of money and time is spent by pollsters trying to predict the outcome of an upcoming election. Usually a sample of the voting population is selected and asked how they will vote. We can simulate this process and get accurate results, particularly after we know the outcome of the election!

INVESTIGATION 5 PREDICTING ELECTION RESULTS

In the 2006 South Australian election five parties contested the district of Frome. The order in which they appeared on the ballot paper was: DEM, LIB, GRN, ALP, FFP. There were 20 713 voters in Frome.

In this investigation you are asked to predict what the outcome would be from a sample taken from the population of voters from Frome.
What to do:

1. Use your calculator to generate a list of 10 random integers between 1 and 20713. The people corresponding to the random integers make up your sample of 10.

2. To decide how the person corresponding to the random integer \( n \) is going to vote, use the rules summarised in the table. In this table we have also added UND for those persons interviewed who are still undecided how they will vote.

<table>
<thead>
<tr>
<th>Random integer ( n )</th>
<th>Party</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n \leq 350 )</td>
<td>DEM</td>
</tr>
<tr>
<td>( 350 &lt; n \leq 9000 )</td>
<td>LIB</td>
</tr>
<tr>
<td>( 9000 &lt; n \leq 17000 )</td>
<td>GRN</td>
</tr>
<tr>
<td>( 17000 &lt; n \leq 18000 )</td>
<td>ALP</td>
</tr>
<tr>
<td>( n &gt; 18000 )</td>
<td>FFP</td>
</tr>
<tr>
<td></td>
<td>UND</td>
</tr>
</tbody>
</table>

For example, for the random integer \( n = 6598 \), the person corresponding to that number would vote LIB.

Hint: To make it easy to see how your sample will vote, sort the random numbers in numerical order first.

3. Copy and fill in the following table of counts for your sample.

<table>
<thead>
<tr>
<th>Party</th>
<th>Count</th>
<th>Total number of votes</th>
<th>Percentage of votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FFP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UND</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. From your sample, predict the percentage of votes each party can expect to receive in the election.

5. Repeat this four more times. Does each of your five samples predict that the same party will win the election?

6. A sample of 10 is far too small. Repeat the above procedure for samples of size 50.

7. From the rule of allocating parties to the random integers, decide the actual results of this election. Compare this with the conclusions from your samples.

ESTIMATING POPULATION MEANS AND STANDARD DEVIATIONS

A great deal of statistical analysis is concerned with how much difference can be expected between a sample and a population.

The sample mean is one of the most important and best understood statistics.

To see what is involved, consider a machine that produces a population of chocolates with mean weight 15 grams and standard deviation 3 grams.

The histogram shows the individual weights of 200 such chocolates where the mean and standard deviation are very close to 15 g and 3 g respectively.

Chocolates are packed into boxes with 10 in each. This gives us a new population of boxes of chocolates. The following table displays the contents of 16 boxes. The individual weights of the chocolates in each box are given to the nearest gram. The mean of the weights in each box is also recorded in the last row of the table.
Note that:

- The mean weights of the boxes are clustered around 15.
- The spread of the mean weights is much less than the spread of the individual chocolates.

For this sample of 16 boxes, the mean is

\[ \frac{16.2 + 16.3 + 15.1 + \ldots + 15.2}{16} = 15.3 \]

This mean is about the same as the mean of the population of individual chocolates.

The standard deviation, which measures how far the means are spread out, is 0.745. This is roughly \( \frac{1}{4} \) of that of the standard deviation for the individual chocolates.

Because we are talking about two closely related distributions, the language and notation can become confusing.

**NOTATION**

1. We have a population with variable \( X \) that has mean \( \mu \) and standard deviation \( \sigma \). In the above example, \( X \) is the weight of a chocolate. The mean \( \mu = 15 \) grams and the standard deviation \( \sigma = 3 \) grams.

2. We take a sample of size \( n \) from this population and use the mean \( \overline{x} \) to estimate the mean \( \mu \) of the original population. In the above example, \( n = 10 \) (10 chocolates in one box), and \( \overline{x} \) is the mean weight of chocolates in a box.

3. All possible sample means make up a new population with a variable usually denoted by \( \overline{X} \). The possible values of \( \overline{X} \) are the sample means \( \overline{x} \). In the above example, the possible values of \( \overline{X} \) are the mean weights of 10 chocolates in a box. These are: 16.2, 16.3, 15.1, etc.
4 The mean of the variable $\bar{X}$ is usually denoted by $\mu(\bar{X})$ (or $\mu_{\bar{X}}$).

The standard deviation of $\bar{X}$ is usually denoted by $\sigma(\bar{X})$ (or $\sigma_{\bar{X}}$).

The problem is to find how $\mu(\bar{X})$ and $\sigma(\bar{X})$ are related to $\mu$ and $\sigma$.

5 When we are taking samples from the population of sample means, the standard deviation of such samples is usually denoted by $s_{\bar{X}}$. $s_{\bar{X}}$ is used as an estimate of $\sigma(\bar{X})$.

In the above example, $s_{\bar{X}} = 0.745$.

**INVESTIGATION 6  A COMPUTER BASED RANDOM SAMPLER**

In this investigation we explore the relationship between:
- the population mean $\mu$ and the mean of sample means $\mu_{\bar{X}}$
- the population standard deviation $\sigma$ and the standard deviation of sample means $\sigma_{\bar{X}}$

We examine samples taken from symmetric distributions as well as one that is skewed.

We start by sampling from a population which has a symmetrical distribution. The population is normally distributed with a mean of 50 and standard deviation of 15.

**What to do:**

1 Click on the icon given alongside. This opens a worksheet named Samples with a number of buttons.

2 Select a sample size from the drop down box. Start with a sample size of 10.

3 Click on the “find samples” button to create 200 different samples of size 10 from the population described above.

4 Click on the “find sample means” button to find the means of each of the 200 different samples.

5 Click on the “analyse” button. This lists the two hundred sample means, finds the standard deviation $s_{\bar{X}}$ of the 200 means, draws a histogram of the 200 sample means and superimposes a smooth curve on the histogram.

To see this output, open the worksheet named Analysis.

Note that the first graph on this worksheet is the graph of the distribution for the population. Note that the axes differ from that of the other graphs.

6 Make a copy of the table and enter the value of $(s_{\bar{X}})^2$ in the first column next to $n = 10$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$s_{\bar{X}}^2$</th>
<th>$s_{\bar{X}}^2$</th>
<th>$s_{\bar{X}}^2$</th>
<th>$s_{\bar{X}}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>160</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7 Go back to the worksheet named Samples and change the sample size to 20. Repeat steps 3, 4 and 5. Enter the value of $(s_{\bar{X}})^2$ next to $n = 20$ in the table.

8 Repeat for samples of size 40, 80 and 160.
We are trying to see how \((s_{\overline{X}})^2\) is related to the standard deviation of the population. However, \((s_{\overline{X}})^2\) can vary quite a lot.

To spot the pattern more clearly, repeat the experiment another 3 times.

From your experiment, determine a relationship between the square of the sample standard deviation \((s_{\overline{X}})^2\) and the square of the population standard deviation.

Now click on the icon to sample data from a population with a uniform distribution. These distributions are very commonly used in computer games where, for example, cards have to be selected at random. Complete an analysis of this data by repeating the above procedure and recording all results.

Now click on the icon to sample data from a population with an exponential distribution. These distributions are notoriously skew. They are commonly used in modelling lifetimes, such as the lifetime of light globes. Complete an analysis of this data by repeating the above procedure and recording all results.

From the investigation you should have discovered that:

- The mean of the samples is approximately the mean of the population.
- As the sample size \(n\) increases there is less variability amongst the sample means, i.e., as the sample size increases, \(s_{\overline{X}}\) decreases.
  More precisely, if \(\sigma\) is the population standard deviation, the standard deviation of the means \(s_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\).
- The histogram of the sample means is symmetric, even if the population is not symmetric.

\[ s_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \]  

is called the **sampling error**.

The sampling error is a measure of the variability of the sample means.

The sampling error gives some idea of how much a sample mean can be expected to differ from the population mean. As \(n\) becomes large, the sampling error \(s_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\) becomes small. This means that the sample means are close together, and the larger the value of \(n\), the closer the sample means will crowd together around the population mean \(\mu\).

This important result is one version of the **law of large numbers**, which says that for a large sample size a particular sample mean \(\overline{X}\) is likely to be close to the population mean \(\mu\).

**EXERCISE 4I**

1. A population has standard deviation \(\sigma = 10\).
   a. Calculate the sampling error \(s_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\) for each of the following sample sizes:
      i. \(n = 4\)    ii. \(n = 9\)    iii. \(n = 16\)    iv. \(n = 25\)    v. \(n = 100\)
b Plot a graph of the sampling error against the sample size \( n \).

c What happens to the sampling error for large values of \( n \)?

2 The table shows the test results of 4 classes of 11 students each.

- a Calculate the mean and standard deviation of the 44 marks.
- b Calculate the mean mark for each of the 4 classes.
- c Calculate the mean and the standard deviation of the four means you found in b.
- d Compare the results you obtained in a with those you obtained in c.

<table>
<thead>
<tr>
<th>Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>7</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

3 Two histograms both of size 50 are shown below. One is of a population with mean \( \mu = 15 \) and standard deviation \( \sigma = 6 \). The other is of the sample means of size \( n = 36 \). Note that the vertical scales of the two histograms are not the same.

- a Which of the two is the histogram for the sample means?
- b What is the standard deviation for the sample means?

4 A machine produces nails with a mean length of 5 cm and standard deviation of 0.2 cm. These nails are packaged into lots of 50.

- a What would you expect the mean length of a nail in a package to be?
- b If you considered many such packages, what would you expect the standard deviation of the mean lengths of the packages to be?

5 A farmer wants to find the expected mean weight of a flock of lambs by weighing a random sample. He guesses that the standard deviation of the weights is roughly 6 kg.

- a If the farmer wants the sampling error \( \frac{\sigma}{\sqrt{n}} \) to be about 1 kg, how many lambs should he weigh?
- b If the farmer wants the standard error to be \( \frac{1}{2} \) kg (half the standard error in a), how many more lambs should he weigh?
A POSSIBLE PROJECT

The work you have covered in this chapter so far should give you sufficient knowledge to carry out your own statistical investigation. Begin by choosing a problem or topic that interests you. Outline your view of the problem question and the data you need to answer it. Discuss your problem and proposed analysis with your teacher. If you need, refine both the problem and your proposed method of solving it.

Collect, in sufficient quantity, the data needed. Aim to ensure your data is randomly selected. Use the software available to produce any graphs and statistical calculations.

Prepare a report of your work. You may choose how you present your work. You may present it as:

- A newspaper or magazine article
- A powerpoint slide presentation
- A video
- A wordprocessed document.

In your report include:

- A description of the problem or issue that you are investigating.
- A simple account of the method you have employed to carry out the investigation.
- Your conclusion.
- A discussion of any weaknesses in your method that may cause your conclusion to be suspect.

Click on this icon to obtain suggestions for projects involving samples and surveys.

REVIEW SET 4A

1. For each of the following variables give:
   i. a list of possible levels (categories) if there are only a few, or state that there are many or infinite levels
   ii. measurement units (where appropriate)
   iii. the variable type (categorical, quantitative discrete or quantitative continuous)

   a. country of residence
   b. breathing rate
   c. height

2. The data supplied is the diameter (in cm) of a number of bacteria colonies as measured by a microbiologist 12 hours after seeding.

   a. Produce a stemplot for this data.
   b. Find the median and range of the data.
   c. Comment on the skewness of the data.
3 The back-to-back stemplot alongside represents the times for the 100 metre freestyle recorded by members of a swimming squad.

The scale used is: leaf unit: 0.1 seconds

<table>
<thead>
<tr>
<th>Girls</th>
<th>Boys</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>0 2 7</td>
</tr>
<tr>
<td>7 6 3</td>
<td>34</td>
</tr>
<tr>
<td>5 8 7 9 9</td>
<td></td>
</tr>
<tr>
<td>8 4 3 0</td>
<td>35</td>
</tr>
<tr>
<td>8 3 3</td>
<td>36</td>
</tr>
<tr>
<td>7 8 8</td>
<td></td>
</tr>
<tr>
<td>7 6 6 6</td>
<td>37</td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
</tr>
</tbody>
</table>

a Copy and complete the following table:

<table>
<thead>
<tr>
<th>Girls</th>
<th>Boys</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>outliers</td>
<td></td>
</tr>
<tr>
<td>shape</td>
<td></td>
</tr>
<tr>
<td>centre (median)</td>
<td></td>
</tr>
<tr>
<td>spread (range)</td>
<td>1 41</td>
</tr>
</tbody>
</table>

b Write an argument that supports the conclusion you have drawn about the girls’ and boys’ swimming times.

4 A community club wants to survey its 500 members about a new membership package, by choosing a sample of 30. The club has an alphabetical list of members.

a How would you use a random number generator to find a simple random sample? Explain how you would handle the situation of the same two numbers appearing in your selection.

b It is decided to select one person at random and select the 30 members in alphabetical order starting from the one they selected at random. Is this a random sample?

5 This data shows the distance, in metres, Glen McGraw was able to throw a cricket ball.

<table>
<thead>
<tr>
<th>Distance (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>71.2</td>
</tr>
<tr>
<td>65.1</td>
</tr>
<tr>
<td>68.0</td>
</tr>
<tr>
<td>71.1</td>
</tr>
<tr>
<td>74.6</td>
</tr>
<tr>
<td>68.8</td>
</tr>
<tr>
<td>83.2</td>
</tr>
<tr>
<td>85.0</td>
</tr>
<tr>
<td>74.5</td>
</tr>
<tr>
<td>87.4</td>
</tr>
<tr>
<td>84.3</td>
</tr>
<tr>
<td>77.0</td>
</tr>
<tr>
<td>82.8</td>
</tr>
<tr>
<td>84.4</td>
</tr>
<tr>
<td>80.6</td>
</tr>
<tr>
<td>75.9</td>
</tr>
<tr>
<td>89.7</td>
</tr>
<tr>
<td>83.2</td>
</tr>
<tr>
<td>97.5</td>
</tr>
<tr>
<td>82.9</td>
</tr>
<tr>
<td>90.5</td>
</tr>
<tr>
<td>85.5</td>
</tr>
<tr>
<td>90.7</td>
</tr>
<tr>
<td>92.9</td>
</tr>
<tr>
<td>95.6</td>
</tr>
<tr>
<td>85.5</td>
</tr>
<tr>
<td>64.6</td>
</tr>
<tr>
<td>73.9</td>
</tr>
<tr>
<td>80.0</td>
</tr>
<tr>
<td>86.5</td>
</tr>
</tbody>
</table>

a Determine the highest and lowest value for the data set.

b Produce between 6 and 12 groups in which to place all the data values.

c Prepare a frequency distribution table.

d For this data, draw a: i frequency histogram ii relative frequency histogram.

e Determine: i the mean ii the median.

6 A grower has picked a crop of oranges. The mean weight of the oranges is 800 grams with a standard deviation of 90 grams. The farmer bags the oranges in bags of 20. Let \( \overline{Y} \) be the mean weight of oranges in a bag.

a What is the mean of \( \overline{Y} \)?

b What is the standard deviation of \( \overline{Y} \)?

7 A market research company surveyed a random sample of 800 people from Sun City and asked them which of the three mayoral candidates they preferred. The results are summarised in the table alongside:

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 25</td>
<td>A</td>
</tr>
<tr>
<td>25 - 55</td>
<td>33</td>
</tr>
<tr>
<td>&gt; 55</td>
<td>182</td>
</tr>
</tbody>
</table>

a State the two variables which are under investigation.

b Classify each of these variables as either dependent or independent.

c Produce an appropriate table of percentages so that analysis can be carried out.

d Why is a table of percentages required in this situation?
REVIEW SET 4B

1 The 100 students in years 11 and 12 of a high school were asked whether (y) or not (n) they owned a mobile phone. The replies, as they were received, were

nyyny nyny nyny nyny nyny nyny nyny nyny nyny nyny nyny nyny nyny nyny nyny

a Calculate the proportion of all students who said they owned a mobile phone.
b What proportion of the first i 5 ii 10 students said they owned a mobile phone? Are these samples representative of all Year 11 and 12 students?
c Use your calculator to generate a set of random numbers to select a simple random sample of i 5 ii 10 iii 20 students from the 100 students above. Calculate the proportion of each sample who said they owned a mobile phone.

2 Find the five-number summary and the interquartile range for each of the following data sets that have already been placed in rank order. Then draw a boxplot for each data set:

a 4.0, 10.1, 13.4, 14.2, 15.0, 16.5, 22.2, 22.4, 23.1, 30.0
b 11, 15, 17, 21, 23, 25, 25, 27, 47, 49, 49

3 Find, using your calculator, the mean and standard deviation of the following sets of data values:

a 117, 129, 105, 124, 123, 128, 131, 124, 123, 125, 108
b 6.1, 5.6, 7.2, 8.3, 6.6, 8.4, 7.7, 6.2

4 The given parallel boxplots represent the 100-metre sprint times for the members of two athletics squads.

a Determine the 5-number summaries for both A and B.
b Determine the i range ii interquartile range for each group.
c Copy and complete: i The members of squad ...... generally ran faster times.
ii The times in squad ...... were more varied.

5 The batting averages for the Australian and Indian teams for the 2001 test series in India were as follows:

Australia 109.8, 48.6, 47.0, 33.2, 32.2, 29.8, 24.8, 20.0, 10.8, 10.0, 6.0, 3.4, 1.0
India 83.83, 56.33, 50.67, 28.83, 27.00, 26.00, 21.00, 20.00, 17.67, 11.33, 10.00, 6.00, 4.00, 4.00, 1.00, 0.00

a Record the 5-number summary for each country.
b Construct parallel boxplots for the data.
c Compare and comment on the centres and spread of the data sets.
d Should any outliers be discarded and the data reanalysed?

Imagine that you are the researcher. Write a report to the person who contracted you to carry out the research.
A machine fills cartons with orange juice with mean volume of 255 mL and standard deviation of 1.5 mL. These cartons are then wrapped together in slabs of 12. Let \( \overline{V} \) be the mean volume of the cartons in a slab.

\[ a \quad \text{What is the mean of } \overline{V}? \quad b \quad \text{Calculate the standard deviation of } \overline{V}. \]

A manufacturer of light globes claims that the newly invented type has a life 20% longer than the current globe type. Forty of each globe type are randomly selected and tested. Here are the results to the nearest hour.

**Old type:**
103 96 113 111 126 100 122 110 84 117 111 87 90 121 99 114 105 121 93 109 87 127 117 131 115 116 82 130 113 95 103 113 104 104 87 118 75 111 108 112

**New type:**
146 131 132 160 128 119 133 117 139 123 191 117 132 107 141 136 146 142 123 144 133 124 153 129 118 130 134 151 145 131 109 129 109 131 145 125 164 125 133 135

\[ a \quad \text{Determine the 5-number summary and interquartile range for each of the data sets.} \quad b \quad \text{Produce side-by-side boxplots.} \quad c \quad \text{Discuss the manufacturer’s claim.} \]

REVIEW SET 4C

1. Briefly state which sampling technique you would use to select a random sample for each of the following.
   \[ a \quad \text{A chocolate manufacturer wants to test the quality of the chocolates at the end of a production line when the chocolates have already been boxed.} \quad b \quad \text{South Australia has 11 electorates. A pollster wants to predict how South Australians will vote in the next election.} \quad c \quad \text{A club wants to select 5 winners in a raffle.} \]

2. Jenny’s golf scores for her last 20 rounds were:

\[
90, \ 106, \ 84, \ 103, \ 112, \ 100, \ 105, \ 81, \ 104, \ 98, \ 107, \ 95, \ 104, \ 108, \ 99, \ 101, \ 106, \ 102, \ 98, \ 101
\]

   \[ a \quad \text{Find the } i \text{ median } \quad ii \quad \text{lower quartile } \quad iii \quad \text{upper quartile} \quad b \quad \text{Find the interquartile range of the data set.} \quad c \quad \text{Find the mean and standard deviation of her scores.} \]

3. Two taxi drivers, Peter and John, decided to measure who was more successful by comparing the amount of money they collected per hour. They randomly selected 25 hours to make the comparison.

**Peter** (dollars per hour)

<table>
<thead>
<tr>
<th>17.27</th>
<th>11.31</th>
<th>15.72</th>
<th>18.92</th>
<th>9.55</th>
<th>12.98</th>
<th>19.12</th>
<th>18.26</th>
<th>22.79</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.09</td>
<td>11.68</td>
<td>15.84</td>
<td>12.81</td>
<td>24.03</td>
<td>15.03</td>
<td>12.95</td>
<td>12.25</td>
<td></td>
</tr>
<tr>
<td>20.09</td>
<td>18.64</td>
<td>18.94</td>
<td>13.92</td>
<td>11.69</td>
<td>15.52</td>
<td>15.21</td>
<td>18.59</td>
<td></td>
</tr>
</tbody>
</table>
John ($ per hour)
23.70 13.30 12.18 14.20 15.74 14.01 10.05 13.34 14.18
10.05 12.20 13.50 18.64 13.29 12.65 13.54 13.44
8.83 11.09 12.29 18.94 20.08 13.84 14.57 13.63

a. Produce parallel boxplots for this data.
b. Is there any evidence one driver is more successful than the other?

4 Explain why each of the following sampling techniques might be biased.
   a. A researcher uses the members of the under fourteen football team in a town to test the claim that boys in Australia are overweight.
   b. A manager of a shop wants to know what customers think of the services provided by the shop. The manager questions the first 10 customers that enter the shop Monday morning.
   c. A promoter of Dogoon washing powder approaches a random sample of households and offers them a prize if they say they use Dogoon washing powder.

5 A variable $X$ has mean $\mu = 23$ and standard deviation $\sigma = 3$.
   Let $\overline{X}$ be the mean weights of samples of size $n = 16$ taken from $X$.
   Find the mean and standard deviation of $\overline{X}$.

6 The histogram shows the weights in kg of a sample of turkeys on a farm.
   a. What is the sample size?
   b. Construct a frequency and relative frequency table for this data.
   c. Use the table to estimate the mean weight and standard deviation.
   d. What proportion of the turkeys weigh more than 10 kg?

7 The number of peanuts in a jar varies slightly from jar to jar. A sample of 30 jars for two brands $X$ and $Y$ was taken and the number of peanuts in each jar was recorded.

<table>
<thead>
<tr>
<th>Brand X</th>
<th>Brand Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>871</td>
<td>909</td>
</tr>
<tr>
<td>885</td>
<td>906</td>
</tr>
<tr>
<td>878</td>
<td>913</td>
</tr>
<tr>
<td>882</td>
<td>891</td>
</tr>
<tr>
<td>889</td>
<td>898</td>
</tr>
<tr>
<td>885</td>
<td>901</td>
</tr>
<tr>
<td>916</td>
<td>904</td>
</tr>
<tr>
<td>913</td>
<td>907</td>
</tr>
<tr>
<td>886</td>
<td>900</td>
</tr>
<tr>
<td>905</td>
<td>901</td>
</tr>
<tr>
<td>907</td>
<td>900</td>
</tr>
<tr>
<td>898</td>
<td>900</td>
</tr>
<tr>
<td>892</td>
<td>907</td>
</tr>
<tr>
<td>913</td>
<td>907</td>
</tr>
<tr>
<td>927</td>
<td>906</td>
</tr>
<tr>
<td>907</td>
<td>901</td>
</tr>
<tr>
<td>904</td>
<td>904</td>
</tr>
<tr>
<td>904</td>
<td>903</td>
</tr>
<tr>
<td>903</td>
<td>903</td>
</tr>
<tr>
<td>903</td>
<td>900</td>
</tr>
<tr>
<td>888</td>
<td>901</td>
</tr>
</tbody>
</table>

a. Produce a back-to-back stemplot for the data for each brand.
b. Complete this table:

<table>
<thead>
<tr>
<th></th>
<th>Brand X</th>
<th>Brand Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>outliers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>shape</td>
<td></td>
<td></td>
</tr>
<tr>
<td>centre (median)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spread (range)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

c. Use the above information to compare Brand X and Brand Y.
The value of many quantities is the combined effect of a number of random factors. For example, the birth weight of a baby is the combined result of many unpredictable factors such as genes, mother’s food intake and the environment she lives in. The weight of a packet of corn flakes is a combination of the weights of each of the individual flakes.

**EXERCISE 4L**

1. List at least three factors that affect each of the following:
   a. the height of 17 year old girls
   b. the weight of potatoes grown in one field
   c. the time to travel to school
   d. the mark achieved in an examination.

2. Think of four quantities that are a combined result of at least 3 factors.

The next investigation explores the distribution of a quantity that is the combined result of different factors.

**INVESTIGATION 7**

The botanist Robert Brown (1773 - 1858) observed that when very fine pollen grains were suspended in water they were seen to vibrate erratically. Brown thought that this indicated some life force within the pollen, but later it was seen that finely ground glass behaved in the same manner. This effect is now known as “Brownian motion”. When the atomic theory of matter was developed, it was soon realised that the particles were bombarded by water molecules. There were likely to be more hits from one side than another, making a very small particle jump.

This was one of the first direct actions of molecules that was observed, and scientists, particularly Einstein, were able to use Brownian motion to draw important conclusions about the nature of molecules.

In this investigation we shall simulate Brownian motion in one dimension.

It is difficult to show the erratic movement of the particles, but it is possible to measure the distance a particle has moved away from its initial position after a certain time.

In one dimension, Brownian motion has also been given the name of “Drunken-person walk.”

Imagine a drunk deciding it is time to go home. Not quite knowing in which direction to move, the drunk staggers a few steps in one direction, then pauses before again staggering off in a random direction.

This process is easy to simulate.

**What to do:**

1. Suppose the drunk is 10 m from home.
   Toss a coin. If the coin turns up heads move the drunk one metre towards home. If the coin turns up tails move the drunk one metre away from home.

2. After 10 trials see how far the drunk is from home.

Repeat this several times and draw a histogram of your results.
In **Investigation 8** you will build a spreadsheet that can generate large quantities of data. This can also be done on a graphics calculator but with a limited amount of data.

The following investigation is very important. Although the list of instructions may look long, most of it is concerned with the details of setting up a spreadsheet.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial position</td>
<td>step1</td>
<td>step2</td>
<td>step3</td>
<td>step4</td>
<td>step5</td>
<td>step18</td>
<td>step19</td>
<td>step20</td>
<td>Final position</td>
<td>Mean</td>
</tr>
</tbody>
</table>

2 Enter the number 10 in cell A2, and copy it down to fill column A2 to A401 with the number 10.

3 In cell B2, under the cell with heading ‘Step 1’, enter the formula \( =2\times\text{RAND}() - 1 \)
   Note that \( =\text{RAND}(()) \) finds a random number between 0 and 1.
   What does the formula \( =2\times\text{RAND}() - 1 \) calculate?

4 Copy the formula from cell B2 to all cells from B2 to K2.
   This should give you the first 10 random steps. The other steps will be used in the next part of the investigation.

5 In cell V2, under the heading ‘Final position’, enter the formula \( =\text{SUM}(A2:U2) \)
   What does the formula in cell V2 calculate?

6 Copy the formulae in cells B2 to V2 down to fill in the block B2 to V401.
   - The number in W2 under the heading ‘Mean’, is the mean, and the number in X2 under the heading ‘Standard deviation’ is the standard deviation of the data in column V2 to V401.
   - The number in Y2 indicates the number of observations that fall within 1 standard deviation of the mean. For example, if the mean \( \mu = 10.03 \) and the standard deviation \( \sigma = 1.85 \), Y2 indicates the number of observations between \( 10.03 - 1.85 = 8.18 \) and \( 10.03 + 1.85 = 11.88 \). Similarly, the cells Z2 and AA2 indicate the number of observations that lie within 2 and 3 standard deviations of the mean respectively.
   - The graph that appears to the right of column X is the histogram of the data in the column V2 to V401.

If you are having difficulties setting up this spreadsheet, click on the tag ‘Random walk 2’ at the bottom of the spreadsheet. This will open a completed version.

- Describe the shape of the histogram.
- What is the mean and what is the standard deviation?

7 Copy and fill in the following table for 5 different trials. The entries of the first line may not agree with your values. Pressing F9 will calculate a different set of random numbers.
THE NORMAL DISTRIBUTION

From Investigation 8 you should have discovered that changing the number and values of factors may change the mean and standard deviation, but leaves the following unchanged.

- The shape of the histogram is symmetric about the mean.
- Approximately 68% of the data lies between one standard deviation below the mean and one standard deviation above the mean.
- Approximately 95% of the data lies between two standard deviations below the mean to two standard deviations above the mean.
- Approximately 99.7% of the data lies between three standard deviations below the mean to three standard deviations above the mean.

Data lying outside the range of this last result is a rare event. With a sample of 400, you would only expect about 1 or 2 cases. If you want to measure this more accurately, you will have to adjust the spreadsheet to get much larger samples.

A smooth curve drawn through the midpoint of each column of the histogram would ideally look like the graph displayed.

This is the graph of a normal distribution.

A graph of a normal distribution of a population with mean \( \mu \) and standard deviation \( \sigma \) has the following features:
It is symmetric about the mean $\mu$.

- It is concave (or concave down) from one standard deviation to the left of the mean to one standard deviation to the right of the mean, i.e., between $\mu - \sigma$ and $\mu + \sigma$. For the other values it is convex (or concave up).

- The area below the curve and the horizontal axis between $\mu - \sigma$ and $\mu + \sigma$ is approximately 68% of the total area.

- The area below the curve and the horizontal axis between $\mu - 2\sigma$ and $\mu + 2\sigma$ is approximately 95% of the total area.

- The area below the curve and the horizontal axis between $\mu - 3\sigma$ and $\mu + 3\sigma$ is approximately 99.7% of the total area.

The information is displayed in the graph alongside.

Curves with this shape are known as **normal curves**. Because of their characteristic shape, they are also called **bell shaped curves**.

Variables which are the combined result of many random factors are often approximately normal.

A normal variable $X$ with mean $\mu$ and standard deviation $\sigma$ is denoted by $X \sim N(\mu, \sigma^2)$.

**Example 17**

A population has a normal distribution with mean $\mu$, and standard deviation $\sigma$. Find the proportion of the population that is less than $\mu + \sigma$.

The proportion less than $\mu$ is 50%.

| The proportion between $\mu$ and $\mu + \sigma$ is 34%.  |
| So the proportion less than $\mu + \sigma$ is 50% + 34% = 84%.  |

3 Explain why it is feasible that the distribution of each of the following variables is normal:

- a the maximum temperature in your city on January 12th over a period of years
- b the diameters of bolts immediately after manufacture
- c weights of oranges picked from the same tree
- d the time it takes to complete a motor car on a production line.
4 A population has a normal distribution with mean \( \mu \) and standard deviation \( \sigma \).
Calculate the percentage of the distribution \( \leq a \) shaded on the graph for the values of \( a \) shown in the table.

<table>
<thead>
<tr>
<th>( a )</th>
<th>Approx. percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu + 3\sigma )</td>
<td>84%</td>
</tr>
<tr>
<td>( \mu + 2\sigma )</td>
<td></td>
</tr>
<tr>
<td>( \mu + \sigma )</td>
<td></td>
</tr>
<tr>
<td>( \mu )</td>
<td></td>
</tr>
<tr>
<td>( \mu - \sigma )</td>
<td></td>
</tr>
<tr>
<td>( \mu - 2\sigma )</td>
<td></td>
</tr>
<tr>
<td>( \mu - 3\sigma )</td>
<td></td>
</tr>
</tbody>
</table>

5 Five hundred Year 11 students sat for a Mathematics examination. Their marks were normally distributed with a mean of 75 and standard deviation of 8.

a Copy and complete this bell-shaped curve and assign scores to the markings on the horizontal axis.

b If a pass mark is 51% and a credit is 83%, will the proportion of students who fail be greater than or less than those who gain a credit?

c How many students would you expect to have scored marks:
   i between 59 and 91
   ii more than 83
   iii less than 59
   iv between 67 and 91?

Example 18

The mean of a variable \( X \) that is normally distributed is 40 and the standard deviation is 5. What percentage of the values:

a are less than 45

b lie between 30 and 45?

a 45 is \( \mu + \sigma \) and approximately 84% is less than \( \mu + \sigma \).

b 30 is \( \mu - 2\sigma \) and 45 is \( \mu + \sigma \).

\[ \therefore \text{ approximate percentage is} \]

\[ 84\% - 2.5\% = 81.5\% \]
Example 19

Two rival soft drink companies sell the same flavoured cola in cans that are stamped 375 mL. The information given is known about the volume of the population of cans from each distributor:

<table>
<thead>
<tr>
<th>Company</th>
<th>form</th>
<th>mean (mL)</th>
<th>standard deviation (mL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>normal</td>
<td>378</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>normal</td>
<td>378</td>
<td>3</td>
</tr>
</tbody>
</table>

a For each company determine the proportion of cans that would have volumes:

i less than or equal to the stamped volume of 375 mL

ii greater than or equal to 381 mL.

b Determine the proportion of cans that company B will produce that will lie between or equal to the limits 372 mL and 381 mL.

c Determine the proportion of cans that company A will produce that will lie between or equal to the limits 380 mL and 381 mL.

a i \( \mu_A = 378 \) \( \sigma_A = 1 \) 375 is \( \mu_A - 3\sigma_A \)
\[ \therefore \text{ approximately } 0.15\% \]
\( \mu_B = 378 \) \( \sigma_B = 3 \) 375 is \( \mu_B - \sigma_B \)
\[ \therefore \text{ approximately } 16\% \]

ii For A, \( 381 = \mu_A + 3\sigma_A \)
\[ \therefore 100\% - 99.85\% = 0.15\% \text{ (approximately)} \]

For B, \( 381 = \mu_B + \sigma_B \)
\[ \therefore 100\% - 84\% = 16\% \text{ (approximately)} \]

b For B, \( 372 = \mu_B - 2\sigma_B \) and \( 381 = \mu_B + \sigma_B \)
\[ \therefore \text{ proportion is } 84\% - 2.5\% = 81.5\% \]

c For A, \( 380 = \mu_A + 2\sigma_A \) and \( 381 = \mu_A + 3\sigma_A \)
\[ \therefore \text{ proportion is } 99.85\% - 97.5\% = 2.35\% \]
6 It is known that when a specific type of radish is grown in a certain manner without fertiliser the weights of the radishes produced are normally distributed with a mean of 40 g and a standard deviation of 10 g. When the same type of radish is grown in the same way except for the inclusion of fertiliser, it is known that the weights of the radishes produced are normally distributed with a mean of 140 g and a standard deviation of 40 g.

a Determine the proportion of radishes grown without fertiliser with weights less than 50 grams.

b Determine the proportion of radishes grown with fertiliser with weights less than 60 grams.

c Determine the proportion of radishes grown with and without fertiliser with weights equal to or between 20 and 60 grams.

d Determine the proportion of radishes grown with and without fertiliser that will have weights greater than or equal to 60 grams.

7 A clock manufacturer investigated the accuracy of its clocks after 6 months of continuous use. They found that the mean error was 0 minutes with a standard deviation of 2 minutes. If a buyer purchases 800 of these clocks, find the expected number of them that will be:

a on time or up to 4 minutes fast after 6 months of continuous use

b on time or up to 6 minutes slow after 6 months of continuous use

c between 4 minutes slow and 6 minutes fast after 6 months of continuous use.

8 A bottle filling machine fills, on average, 20 000 bottles a day with a standard deviation of 2000. If we assume that production is normally distributed and the year comprises 260 working days, calculate the approximate number of working days that:

a under 18 000 bottles are filled

b over 16 000 bottles are filled

c between 18 000 and 24 000 bottles (inclusive) are filled.

INVESTIGATION 9

THE SHAPE OF THE NORMAL CURVE

The purpose of this investigation is to explore how the graph of the normal distribution changes with various values of the mean $\mu$ and the standard deviation $\sigma$. You can use your graphics calculator to plot these values or use the graphics package by clicking on the icon.

What to do: $f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}$ is the normal function.

1 Fix the value of the standard deviation $\sigma$ to be 1, and draw the graph for different values of the mean: $\mu = 0 \quad b \quad \mu = 1 \quad c \quad \mu = 2$.

Comment on how the graph changes with changes in the value of the mean $\mu$.

2 Fix the value of the mean $\mu$ to be 0, and draw the graph for different values of the standard deviation: $\sigma = 1 \quad b \quad \sigma = 2 \quad c \quad \sigma = 0.5$.

3 Experiment with a few other values of $\mu$ and $\sigma$. 
From Investigation 9 you should have discovered that:

- A change in the mean $\mu$ does not change the shape of the graph, but shifts it horizontally. The graph is centred about the mean $\mu$.
- A change in the standard deviation $\sigma$ changes the spread of the graph. An increase in the standard deviation increases the spread, and a decrease in the standard deviation decreases the spread. The graph changes its concavity one standard deviation from the mean.

9 Sketch the graph of each of the following normal distributions by hand on one set of axes.

a mean $\mu = 25$ and standard deviation $\sigma = 5$

b mean $\mu = 30$ and standard deviation $\sigma = 2$

c mean $\mu = 21$ and standard deviation $\sigma = 10$

10 Three histograms of sample size 200 are shown below. The samples were all selected from a normal distribution; one with mean $\mu = -1$ and standard deviation 3, one with mean $\mu = 1$ and standard deviation 1, and one with mean $\mu = 2$ and standard deviation 0.5. Note that the scales are not the same for the three histograms.

a Identify each of the three histograms with its distribution.

b For each of the three histograms estimate the proportion of outcomes between the mean and one standard deviation to the right of the mean.

THE STANDARD NORMAL DISTRIBUTION

Suppose that a physics test is marked out of 100 and that the marks are normally distributed with a mean mark of 70% and standard deviation of 10%. For the same class a mathematics test is marked out of 20 and the marks are also normally distributed with mean 12 and standard deviation of 1.5.

If Barbara had a mark of 15 for her mathematics and 80% for her physics, which of the two is her better score?

A mark of 15/20 for mathematics is equivalent to 75%, which is less than 80% for physics. It might seem that Barbara is doing better in physics than in mathematics.

If, however, we compare the marks on a normal curve, a different picture emerges.
The formula for calculating z-scores is:

\[ z\text{-score} = \frac{\text{raw score} - \mu}{\sigma}. \]
Suppose the distribution of the weights in grams of chocolate frogs is \( N(11, 0.75^2) \), i.e., the distribution is normal with \( \mu = 11 \) and \( \sigma = 0.75 \).

**a** Sketch a graph that displays both the actual weight as well as the \( z \)-score along the horizontal axis.

**Exercise 4M**

Suppose the birth weight of babies born without complications is normally distributed with mean \( \mu = 3.29 \) kg and standard deviation \( \sigma = 0.20 \) kg.

**a** A baby has a birth weight of 3.69 kg. What is the \( z \)-score for this baby?

**b** The \( z \)-score for a baby’s birth weight was –2. What was the baby’s birth weight?

**c** If a baby has a birth weight \( z \)-score between –3 and 3 it is thought to be ‘normal’. Between what birth weights is this?

**d** What proportion of babies born would be considered ‘normal’ in weight?

The graph shows the normal curve with different scales along the horizontal axis.

\[
\begin{array}{cccccccc}
\text{\( \mu = 3.29 \)} & 2.69 & 2.89 & 3.09 & 3.29 & 3.49 & 3.69 & 3.89 \\
\text{\( \sigma = 0.20 \)} & -3 & -2 & -1 & 0 & 1 & 2 & 3 \\
\text{weights} & \mu - 3\sigma & \mu - 2\sigma & \mu - \sigma & \mu & \mu + \sigma & \mu + 2\sigma & \mu + 3\sigma \\
\end{array}
\]

**Example 20**

Suppose the birth weight of babies born without complications is normally distributed with mean \( \mu = 3.29 \) kg and standard deviation \( \sigma = 0.20 \) kg.

**a** From the graph you can see that the baby’s weight is 2 standard deviations above the mean, \( \therefore \) the \( z \)-score of the baby’s weight is 2

or using the formula, \( z \)-score = \( \frac{\text{raw score} - \mu}{\sigma} \) = \( \frac{3.69 - 3.29}{0.2} \) = 2.0

(This means that this baby’s weight is 2 standard deviations above the mean.)

**b** From the graph we can see that the baby’s weight is 2.89 kg

or using the formula, \( -2 = \frac{\text{raw score} - 3.29}{0.2} \)

\( \therefore -0.4 = \text{raw score} - 3.29 \)

\( \therefore \text{raw score} = 2.89 \) kg

**c** A \( z \)-score of –3 corresponds to a weight of 3 standard devs. below the mean.

A \( z \)-score of 3 corresponds to a weight of 3 standard devs. above the mean.

From the graph this corresponds to a weight between 2.69 and 3.89 kg.

**d** 99.7% of all outcomes from the normal distribution lie within 3 standard deviations from the mean.

So, about 99.7% of all babies have ‘normal’ birth weight.

**EXERCISE 4M**

1 Suppose the distribution of the weights in grams of chocolate frogs is \( N(11, 0.75^2) \), i.e., the distribution is normal with \( \mu = 11 \) and \( \sigma = 0.75 \).

**a** Sketch a graph that displays both the actual weight as well as the \( z \)-score along the horizontal axis.
Find the \( z \)-score for each of the following observations:

i. \( 12.5 \) g

ii. \( 9.5 \) g

iii. \( 13.25 \) g

b Chocolate frogs are considered to be underweight if the \( z \)-score is less than \(-2\). Which chocolate frogs would be considered underweight?

c A supplier buys a box with 1000 chocolate frogs.

How many of the frogs could be expected to be underweight?

2 The volume of soft drink filled in bottles by a machine is normally distributed with mean 378 mL and standard deviation 1.5 mL.

a What is the \( z \)-score of a bottle with 375 mL?

b What is the volume in a bottle with a \( z \)-score of 3?

c If all bottles with a \( z \)-score of less than \(-2\) are rejected, what proportion of the bottles would be rejected?

3 Find the proportion of observations with \( z \)-scores:

a between \(-1\) and \(+1\)

b between \(-2\) and \(+2\)

c between \(-1\) and 2

d less than 0

e between \(-2\) and 0

f larger than 0

g less than 1

h larger than 2

Example 21

Suppose that Hua’s time to sprint one hundred metres is normally distributed with mean 11.7 seconds and standard deviation 0.6 seconds.

a What would be the \( z \)-score if Hua ran the distance in 10 seconds?

b If Hua’s \( z \)-score was 2.3, how long did he take to run the distance?

a \( Hua’ \text{ s } z \)-score \( = \frac{10 - 11.7}{0.6} \) = \(-2.83\)

b If Hua’s \( z \)-score was 2.3, \( \frac{\text{raw score} - 11.7}{0.6} = 2.3 \)

\( \therefore \text{raw score} = 2.3 \times 0.6 + 11.7 \)

\( \therefore \text{raw score} = 13.08 \)

So, Hua took 13.1 seconds to finish the sprint.

4 The table shows Sergio’s midyear exam results.

The exam results for each subject are normally distributed with the mean \( \mu \) and standard deviation \( \sigma \) shown in the table.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Sergio’s score</th>
<th>Subject ( \mu )</th>
<th>Subject ( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics</td>
<td>83%</td>
<td>73%</td>
<td>10.8%</td>
</tr>
<tr>
<td>Chemistry</td>
<td>77%</td>
<td>50%</td>
<td>11.6%</td>
</tr>
<tr>
<td>Mathematics</td>
<td>84%</td>
<td>74%</td>
<td>10.1%</td>
</tr>
<tr>
<td>German</td>
<td>91%</td>
<td>86%</td>
<td>9.6%</td>
</tr>
<tr>
<td>Biology</td>
<td>72%</td>
<td>62%</td>
<td>12.2%</td>
</tr>
</tbody>
</table>

a Find the \( z \)-score for each of Sergio’s subjects.

b Arrange Sergio’s performance in each subject from ‘best’ to ‘worst’ in terms of how many standard deviations they are from the mean.
THE STANDARD NORMAL DISTRIBUTION

If $x$ is an observation from a normal distribution with mean $\mu$ and standard deviation $\sigma$, then, the $z$-score of $x$ is $z = \frac{x - \mu}{\sigma}$.

Recall that a variable is a quantity that can have different values for different individuals of the population. In statistics variables are commonly denoted by capital letters.

For example, birth weight of babies could be denoted by $X$.

If we select one individual from the population we observe one possible outcome of the variable. Observations are commonly denoted by lower case letters.

For example, if a specific baby is found to have a birth weight of 3.63 kg we could record this as $x = 3.63$ kg.

Note: Distributions will be denoted by capital letters. Observations, or outcomes, from distributions will be denoted by lower case letters.

If a variable $X$ has a normal distribution $N(\mu, \sigma^2)$, then the standard normal distribution $Z$, where $Z = \frac{X - \mu}{\sigma}$ is the number of standard deviations $X$ is above the mean.

$Z$ is a normal distribution with mean 0 and standard deviation 1, i.e., $Z$ is $N(0, 1)$.

The probability a variable lies within an interval is the proportion of values the variable has within that interval.

**Example 22**

Find the probability that the standard normal distribution $Z$ lies between $-1$ and 2.

The graph of the $Z$ distribution is shown:

The probability $Z$ lies between $-1$ and 0 is about 0.34.

The probability $Z$ lies between 0 and 2 is about 0.475.

So, the probability $Z$ lies between $-1$ and 2 is $0.34 + 0.475 = 0.815$

The probability $Z$ lies between $-1$ and 2 is written $\Pr(-1 \leq Z \leq 2)$ or $\Pr(-1 < Z < 2)$.

Since $Z$ is a continuous (interval) variable, the probability $Z$ is exactly equal to $-1$ or 2 is zero, and $\Pr(-1 \leq Z \leq 2) = \Pr(-1 < Z < 2)$.

5 Calculate each of the following probabilities. In each case sketch a graph of the distribution showing the region of interest.

- $a$ $\Pr(-2 < Z < 2)$
- $b$ $\Pr(-3 < Z < 1)$
- $c$ $\Pr(-1 < Z < 2)$
- $d$ $\Pr(Z < 2)$
- $e$ $\Pr(Z < -2)$
- $f$ $\Pr(Z \geq 2)$
- $g$ $\Pr(Z \geq -1)$
- $h$ $\Pr(Z \geq 1)$
6 The volume in mL, $X$, of a can of soft drink is known to have the normal distribution $N(355, 2^2)$.
   a A can is found to have a volume $x = 353$ mL. Find the $z$-score of $x$.
   b Find $Pr(Z > -1)$.
   c What proportion of the cans will have a volume greater than 353 mL?
   d Ann buys 60 cans of this soft drink. How many cans can Ann expect to have volume less than 353 mL?

7 The height of players in cm, $X$, in a basketball competition is known to have a normal distribution $N(181, 4^2)$.
   a Find $Pr(Z > 2)$.
   b Find $Pr(X > 189)$.
   c What proportion of the players are between 169 and 189 cm tall?
   d A player, Des, has height $x = 185$ cm. What proportion of the players are taller than Des?

8 The time, $X$ minutes, it takes Lisa to travel to school has normal distribution $N(20, 3^2)$.
   a Find $Pr(Z < 2)$.
   What does this mean in terms of the time it takes Lisa to go to school?
   b Find $Pr(Z < 3)$.
   c If Lisa leaves home at 8:31 a.m. every morning, what proportion of the time will she get to school before 9:00 a.m.?

9 The length of fish in cm, $L$, has a normal distribution $N(36, 4^2)$.
   a If a fish has a $z$-score of $-2$, what is the length, $l$, of the fish?
   b A fish farmer wants to sell all fish which are longer than 32 cm. What proportion of the fish will be left?

TECHNOLOGY AND NORMAL DISTRIBUTIONS

So far we have only used integer $z$-scores to calculate probabilities. By refining the methods used in Investigation 8, you could make a table of probabilities for other $z$-scores. This information is, of course, available on your calculator.

**Note:** $Pr(a < X < b)$ can be calculated on a **TI83** using $\text{normalcdf}(a, b, \mu, \sigma)$.

Click on the icon for the instructions for your calculator.

**Example 23**

The weights of cherries, $X$ grams, is normally distributed with mean 18.3 g and standard deviation 2.5 g. Use technology to find the probability that a cherry weighs:
   a between 17 g and 22 g
   b less than 18 g
   c more than 20 g

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>$Pr(17 &lt; X &lt; 22) = \text{normalcdf}(17, 22, 18.3, 2.5) \approx 0.629$</td>
</tr>
<tr>
<td>b</td>
<td>$Pr(X &lt; 18) = \text{normalcdf}(-\infty, 18, 18.3, 2.5) \approx 0.452$</td>
</tr>
<tr>
<td>c</td>
<td>$Pr(X &gt; 20) = \text{normalcdf}(20, \infty, 18.3, 2.5) \approx 0.248$</td>
</tr>
</tbody>
</table>
Note: −E99 is the largest negative number which can be entered. It is 10−99.

EXERCISE 4N

1 If $X \sim N(23, 3^2)$, use technology to calculate the following probabilities:
   a $\Pr(19 \leq X \leq 25)$  
   b $\Pr(X \geq 22)$  
   c $\Pr(X \leq 25)$

2 The weights of three month old Siamese kittens are normally distributed with mean 629 grams and standard deviation 86 grams. Find:
   a the probability of a randomly selected three month old kitten having a weight of:
      i more than 500 g  
      ii less than 700 g  
      iii between 600 g and 750 g
   b the proportion of three month old kittens weighing
      i more than 550 g  
      ii between 500 g and 700 g.

3 The lengths of bolts in a batch are normally distributed with mean 100.3 mm and a standard deviation of 1.2 mm. All bolts are supposed to be at least 100 mm long. What proportion of the bolts have an unacceptable length?

4 The length of salmon caught off Kangaroo Island are found to be normally distributed with mean 341 mm and standard deviation 37 mm.
   a What proportion of the salmon are:
      i greater than 40 cm long  
      ii between 30 and 40 cm long?
   b In a catch of 120 salmon, how many would we expect to be less than 30 cm in length?

5 The army recruiting department knows that the heights of young adults suitable to join is normally distributed with mean 182 cm and standard deviation 8.1 cm. 2186 young adults apply to join, but the policy is to admit only those whose heights are between 175 cm and 190 cm. How many of them would you expect to be admitted?

FINDING QUANTILES (k-VALUES)

Consider the normal distribution alongside. It has mean 32 and standard deviation 3.1.

Suppose we want to find $k$ such that $\Pr(X \leq k) = 0.9$.

$k$ can be found using $\text{invNorm}(0.9, 32, 3.1)$ on a TI83.

So, $\Pr(X \leq k) = 0.9$

means that $k = \text{invNorm}(0.9, 32, 3.1)$

$\therefore k \approx 35.97$

{ $\text{2nd}$ $\text{DISTR}$ $3$ brings up $\text{invNorm}(\ )$}
Example 24

The Geography examination results are normally distributed with mean 64.2 and standard deviation 9.21. It is decided that 85% of the students will pass the examination. What is the lowest score needed to pass?

Let $X$ denote a result in the Geography exam. Then, $X \sim N(64.2, 9.21^2)$. We need to find $k$ such that $\Pr(X \leq k) = 0.15$.

Thus, $k = \text{invNorm}(0.15, 64.2, 9.21) = 54.65$.

So, the lowest pass mark would be 55.

6 For the example above, what would be the lowest score needed to pass if it is decided that:
   a 10% will fail
   b 80% will pass?

7 A machine cuts off lengths of steel rod. The lengths are normally distributed with mean 125 cm and standard deviation 0.82 cm.

It is known that 7.5% of the rods are rejected because they are not long enough for the job they were intended for.

What is the shortest acceptable length of rod?

8 The heights of army recruits is normally distributed with mean 182.5 cm and standard deviation of 8.1 cm.

It is known that over the last 25 years 10.7% of applicants to the army were rejected as they were too tall and 15.2% were rejected as being too short.

What is the acceptable height range (to the nearest cm) the army used?

PASCAL’S TRIANGLE

In many situations there are just two outcomes:
- you either pass or fail an exam
- a baby is either a boy or a girl
- medical treatment either improves a patient or it does not
- you either support Port Power or you do not.

A repetition of a number of independent trials in which there are two possible results is called a **binomial experiment**.

Remember that, two trials, or events, are **independent** if the outcome in one does not affect the result of the other.

A student guessing every answer in a multiple choice test in which every question had 4 possible answers, would be carrying out a binomial experiment.
If a variable can only have two possible outcomes, they are often called success or failure. Since tossing a coin can have two outcomes, heads or tails, a coin is often used to simulate a binomial experiment. In this case we could count a head as a success and a tail as a failure.

**SAMPLING SPACE AND PASCAL’S TRIANGLE**

In this section we explore the sampling space, i.e., all possible outcomes, of tossing a coin \( n \) times.

To examine the structure of the sampling space, we shall construct a tree diagram. In this particular diagram it is not unusual to turn the tree diagram through \( 90^\circ \).

**INVESTIGATION 10**

**ALGEBRA AND TREE DIAGRAMS**

In this activity we wish to investigate any link between the algebraic expansion of \( (H+T)^n \) for \( n = 1, 2, 3, 4 \) and \( 5 \) and tree diagrams with up to 5 levels of branches.

**What to do:**

1. Find in simplest form the expansion of \( (H+T)^2 \).
   - What is the connection to the tree diagram alongside?

2. Find in simplest form the expansion of \( (H+T)^3 \).
   - (You may choose to use the binomial expansion from Year 10 or expand \( (H+T)^2 (H+T) = (H^2 + 2HT + T^2)(H+T) \), etc)
   - Extend the tree diagram of 1 to another level of branches. What do you notice?

3. Repeat 2, but for \( (H+T)^4 \). What do you notice?

4. Repeat 2, but for \( (H+T)^5 \). What do you notice?

From the above investigation, you should have re-discovered Pascal’s triangle.

Notice that:

\[
(H+T)^1 = 1H + 1T \\
(H+T)^2 = 1H^2 + 2HT + 1T^2 \\
(H+T)^3 = 1H^3 + 3H^2T + 3HT^2 + 1T^3 \\
(H+T)^4 = 1H^4 + 4H^3T + 6H^2T^2 + 4HT^3 + 1T^4 \\
\]

This triangle of numbers is called Pascal’s triangle.
Notice for the case $n = 3$, $3$ heads occurs $1$ time
2 heads occurs $3$ times
1 head occurs $3$ times
0 heads occurs $1$ time

and for the case $n = 4$, $4$ heads occurs $1$ time
3 heads occurs $4$ times
2 heads occurs $6$ times
1 head occurs $4$ times
0 heads occurs $1$ time

**DISCUSSION**

- How do we find the next few rows of Pascal’s triangle with minimal effort?
- How can Pascal’s triangle be used to write down the expansions of: $(H + T)^6$, $(H + T)^7$, $(H + T)^8$, etc.

**$C_n^r$ NOTATION**

We can identify any member of Pascal’s triangle by using the form $C_n^r$ where $n$ is the row of Pascal’s triangle and $r$ is the $(r + 1)$th member of the row $(r = 0, 1, 2, 3, 4, .... n)$

**Note:** In the $r$th row of Pascal’s triangle there are $n + 1$ numbers.

The following diagram shows Pascal’s triangle, using $C_n^r$ notation.

\[
\begin{array}{cccc}
1 & 1 & & \\
1 & 2 & 1 & \\
1 & 3 & 3 & 1 & \\
1 & 4 & 6 & 4 & 1 & \\
1 & 5 & 10 & 10 & 5 & 1 & \\
1 & 6 & 15 & 20 & 15 & 6 & 1 & \\
\end{array}
\]

The numbers $C_n^r$ are also known as **binomial coefficients**.

In some books $C_n^r$ appears as $\binom{n}{r}$ or $nC_r$ or $^nC_r$.

The $C$ stands for **combinations** and $C_n^r$ is the combined number (or total) of $r$ successes in $n$ trials.

**EXERCISE 4O**

1. Use the 6th row of the Pascal’s triangle to generate the 7th row.
2. Use Pascal’s triangle to find: $a \ C_2^4 \quad b \ C_3^5 \quad c \ C_4^6$
3. Write down the 5th row of Pascal’s triangle and use it to find the number of possible ways of getting:
   - $a \ 5$ heads
   - $b \ 4$ heads
   - $c \ 3$ heads
   - $d \ 2$ heads
   - $e \ 1$ head
   - $f \ 0$ heads
   - $g \ at \ most \ 3$ heads
A bag contains 3 red counters and 2 green counters. A person selects a counter from the bag, notes its colour and replaces the counter. This is repeated 6 times.

a How many possible ways are there of selecting:
   i 5 red counters
   ii 5 or more red counters?

b How, if at all, does your answer change if ‘red’ is interchanged with ‘green’?

To find $C_{13}^{10}$ from Pascal’s triangle would be a time consuming task. Use your calculator to find these binomial coefficients:

a $C_{3}^{10}$  

b $C_{5}^{11}$  

c $C_{4}^{14}$  

d $C_{6}^{16}$  

e $C_{13}^{20}$

Use Pascal’s triangle to find:

a $C_{3}^{10}$  

b $C_{4}^{11}$  

c $C_{5}^{14}$  

d $C_{6}^{16}$  

e $C_{13}^{20}$

Use the above results to conjecture the simplified value of $C_{0}^{n} + C_{1}^{n} + C_{2}^{n} + C_{3}^{n} + \ldots + C_{n-1}^{n} + C_{n}^{n}$.

## BINOMIAL DISTRIBUTION

Suppose a spinner has three blue edges and one white edge. Then, on each occasion it is spun, we will get a blue or a white.

The chance of finishing on blue is $\frac{3}{4}$ and on white is $\frac{1}{4}$.

If $p$ is the probability of getting a blue, and $q$ is the probability of getting a white then $p = \frac{3}{4}$ and $q = \frac{1}{4}$ (the chance of failing to get a blue).

Consider twirling the spinner three times. Let the variable $X$ be the number of blue results that could occur. The possible outcomes of $X$ are $x = 0, 1, 2, or 3$.

The possible outcomes, with their probabilities, are displayed on the tree diagram.
Thus, if there are \( n \) independent trials of a binomial experiment, where \( p \) is the probability of a success in any one trial, then

\[
\Pr(\text{\( x \) successes and \( n-x \) failures}) = C_x^n p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.
\]

**The General Case**

In the case of \( n \) trials (instead of 3), there are \( C_n^x \) ways of selecting \( x \) blues and \( n-x \) non-blues.

If \( p \) is the probability of selecting a blue from a single trial, then for \( n \) trials,

\[
\Pr(\text{\( x \) blues}) = C_x^n p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.
\]

Thus,

**Example 25**

Which of the following are binomial experiments?

- **a** 50 identical coins are tossed at once. The outcome is the number of heads resulting.
- **b** Three marbles are drawn without replacement from a bag that contains 3 blue and 5 green marbles.
- **c** Two chess players are to compete against each other for a series of 24 games.

- **a** As each coin has the same probability of falling heads, this would be the same as tossing one coin 50 times. So, we have a binomial experiment.
- **b** Since the marbles are not replaced, the three events of drawing marbles are not independent. This is not a binomial experiment.
- **c** In a game of chess there are three possible outcomes, win, lose or draw. This is not a binomial experiment where each event can only have two outcomes.
Which of the following is a binomial experiment? Justify your answer.

**a** A student randomly selects answers in a multiple choice question. There are 4 choices for each question, and there are 20 questions.

**b** A plastic cup is tossed 100 times to see how often it falls on its side, top or bottom.

**c** An archer shoots at a target 30 times. Either the arrows hit the target or they miss the target.

**d** A quality controller selects 4 discs from a sample of 20 and tests them for quality. The discs are either satisfactory or they are not.

**e** A six sided die is rolled 100 times to see if it is fair.

**f** A six sided die is rolled 100 times to find the proportion of the time a 4 appears.

---

The probability a dart player hits the bull’s eye is 0.25. If the player has 5 attempts at hitting the bull’s eye, find the probability that he hits the bull’s eye:

**a** exactly 2 times

**b** at most 2 times

**c** at least 3 times.

---

The local train service is not very reliable. It is known that the 7:37 train will run late on average 2 days out of every five week days. For any week of the year taken at random, find the probability of the 7:37 train being on time:

**a** all 5 week days

**b** only on Monday

**c** on at most 2 week days

**d** on at least 2 week days.

---

**Example 26**

The probability John beats Charles at tennis is 0.7. If John and Charles play 5 matches, find the probability that John beats Charles:

**a** exactly 3 times

**b** at most twice

**c** three or more times.

Let $X$ be the number of games John wins. $p = 0.7$ and $1 - p = 0.3$.

**a** The probability John beats Charles exactly three times is

$$Pr(X = 3) = C_5^3 (0.7)^3 (0.3)^2 = 10 (0.7)^3 (0.3)^2 \div 0.309$$

**b** The probability John beats Charles at most twice is

$$Pr(X \leq 2) = Pr(X = 0) + Pr(X = 1) + Pr(X = 2)$$

$$= C_5^0 (0.7)^0 (0.3)^5 + C_5^1 (0.7)^1 (0.3)^4 + C_5^2 (0.7)^2 (0.3)^3$$

$$\div 0.163$$

**c** The probability John beats Charles three or more times is:

$$Pr(X \geq 3) = 1 - Pr(X \leq 2)$$

$$= 1 - 0.163 \{ \text{from b} \}$$

$$\div 0.837$$

---

**EXERCISE 4P**

1 Which of the following is a binomial experiment? Justify your answer.

**a** A student randomly selects answers in a multiple choice question. There are 4 choices for each question, and there are 20 questions.

**b** A plastic cup is tossed 100 times to see how often it falls on its side, top or bottom.

**c** An archer shoots at a target 30 times. Either the arrows hit the target or they miss the target.

**d** A quality controller selects 4 discs from a sample of 20 and tests them for quality. The discs are either satisfactory or they are not.

**e** A six sided die is rolled 100 times to see if it is fair.

**f** A six sided die is rolled 100 times to find the proportion of the time a 4 appears.

2 The probability a dart player hits the bull’s eye is 0.25. If the player has 5 attempts at hitting the bull’s eye, find the probability that he hits the bull’s eye:

**a** exactly 2 times

**b** no more than 2 times

**c** at least 3 times.

3 The local train service is not very reliable. It is known that the 7:37 train will run late on average 2 days out of every five week days. For any week of the year taken at random, find the probability of the 7:37 train being on time:

**a** all 5 week days

**b** only on Monday

**c** on at most 2 week days

**d** on at least 2 week days.
Quality controllers usually select items without replacing them. This means that quality control is not a binomial experiment. Since, however, quality controllers usually make a selection from very large numbers, not replacing items makes very little difference, and the binomial distribution is often a good approximation.

4 Suppose that there are 50 defective CD’s in a batch of 1000.
   a What is the probability that a randomly selected disc is defective?
   b A quality controller selects 6 discs at random from the batch. Use the binomial distribution to estimate the probability that:
      i exactly 2 discs are defective
      ii at most two discs are defective
      iii at least 3 discs are defective.

5 Records show that 6% of items assembled on a production line are faulty. A random sample of 5 items is selected. Find the probability that:
   a none will be faulty   b at most 1 will be faulty   c up to 4 will be faulty

Your calculator is able to calculate binomial probabilities. The functions provided are:

- **Binomial probability (density) function**
  This calculates the binomial probability of one event.
  The TI command \( \text{binompdf}(100, 0.5, 45) \) calculates the probability of 45 heads appearing when a coin with probability of 0.5 of heads appearing is tossed 100 times.

- **Binomial cumulative (density) function**
  This calculates the cumulative probability of events less than or equal to an event happening. The TI command \( \text{binomcdf}(100, 0.5, 32) \) calculates the probability of a fair coin turning up heads less than or equal to 32 times.

### Example 27

A biased coin has probability of 0.3 of turning up heads.
If the coin is tossed 100 times, find the probability that:
   a 42 heads appear   b at most 23 heads appear   c at least 25 heads appear

Using TI instructions:

\[
\begin{align*}
\text{a } & \quad P(X = 42) = \text{binompdf}(100, 0.3, 42) = 0.00321 \\
\text{b } & \quad P(X \leq 23) = \text{binomcdf}(100, 0.3, 23) = 0.0755 \\
\text{c } & \quad P(X \geq 25) = 1 - P(X \leq 24) = 1 - \text{binomcdf}(100, 0.3, 24) = 0.886
\end{align*}
\]
6 A fair coin is tossed 100 times. Use your calculator to find the probability that:
   a exactly 53 heads appear  b at most 48 heads appear  c at least 53 heads appear

7 A six sided is rolled 50 times. Use your calculator to find the probability that a 6 appears:
   a exactly 8 times  b at most 7 times  c at least 7 times

Example 28

A spinner has probability of 0.4 of coming up on a green colour.
If the spinner is spun 60 times, find the probability the spinner comes up on a green colour between, and including, 23 and 26 times.

Let \( X \) be the number of times the spinner comes up green.

\[
P(23 \leq X \leq 26) = P(X \leq 26) - P(X \leq 22)
= \text{binompdf}(60, 0.4, 26) - \text{binompdf}(60, 0.4, 22)
\approx 0.397
\]

8 Calculate the probabilities for each of the following variables.
   a \( X \sim \text{Bin}(40, 0.4) \), find \( P(20 \leq X \leq 30) \).
   b \( X \sim \text{Bin}(60, 0.5) \), find \( P(30 \leq X \leq 40) \).
   c \( W \sim \text{Bin}(100, 0.5) \), find \( P(59 \leq W \leq 61) \).
   d \( U \sim \text{Bin}(120, 0.7) \), find \( P(81 \leq U \leq 87) \).

9 Let \( X \) be the number of heads if a fair coin is tossed 100 times.
   Find the number \( a \) so that \( P(50 - a \leq X \leq 50 + a) \approx 95\% \).

Q THE MEAN AND STANDARD DEVIATION OF A DISCRETE VARIABLE

Consider tossing one coin 100 times where \( X \) is the number of heads resulting from one toss.

One such sample is given alongside:

For this sample the mean \( \mu = 0.51 \), and the standard deviation \( s \approx 0.5024 \). Check this!

We could find many more samples of 100 and discover that the \( \mu \)-values hover around 0.5 and likewise the \( s \)-values are also around 0.5 .

This conclusion comes from our random sampling.

But, what are the theoretical values of the population mean \( \mu \) and the population standard deviation \( \sigma \), and how can we calculate them easily?

To do this for the one coin problem above, observe this theoretical probability table:
As \( \bar{x} = \frac{\sum f_i x_i}{\sum f_i} \), then in our example, \( \bar{x} = \frac{f_1 x_1 + f_2 x_2}{f_1 + f_2} = \frac{f_1 x_1 + f_2 x_2}{N} \).

\[
\therefore \bar{x} = x_1 \left( \frac{f_1}{N} \right) + x_2 \left( \frac{f_2}{N} \right)
\]

However, for large samples of size \( N \), the sample mean \( \bar{x} \) will be approximately equal to the population mean \( \mu \), and \( \frac{f_i}{N} \) will be approximately the probability \( p_i \) of \( x_i \) occurring.

This shows that the population mean is \( \mu = x_1 p_1 + x_2 p_2 \) and in the above example \( \mu = 0 \times 0.5 + 1 \times 0.5 = 0.5 \).

Likewise, \( s = \sqrt{\frac{f_1 (x_1 - \bar{x})^2 + f_2 (x_2 - \bar{x})^2}{N}} = \sqrt{\frac{f_1}{N} (x_1 - \bar{x})^2 + \frac{f_2}{N} (x_2 - \bar{x})^2} \)

And so, by the same reasoning, for the population, \( \sigma = \sqrt{(x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2} \)

In our one coin example this gives \( \sigma = \sqrt{(0 - 0.5)^2 \times 0.5 + (1 - 0.5)^2 \times 0.5} = 0.5 \)

We say that the discrete random variable \( X \) has mean \( \mu \) and standard deviation \( \sigma \).

In general,

- the **population mean** is: \( \mu = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n \)
- the **population standard deviation** is:

\[
\sigma = \sqrt{(x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \ldots + (x_n - \mu)^2 p_n}
\]

**Example 29**

A unbiased coin is tossed twice.

Let \( X \) be the number of heads resulting.

The probability table for \( X \) is:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( p_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( \frac{1}{4} )</td>
</tr>
<tr>
<td>1</td>
<td>( \frac{1}{2} )</td>
</tr>
<tr>
<td>2</td>
<td>( \frac{1}{4} )</td>
</tr>
</tbody>
</table>

Find the mean and standard deviation of \( X \).

**Solution**

\[
\mu = 0 \left( \frac{1}{4} \right) + 1 \left( \frac{1}{2} \right) + 2 \left( \frac{1}{4} \right) = 1
\]

\[
\sigma = \sqrt{(0 - 1)^2 \left( \frac{1}{4} \right) + (1 - 1)^2 \left( \frac{1}{2} \right) + (2 - 1)^2 \left( \frac{1}{4} \right)}
\]

\[
= \sqrt{\frac{1}{4} + 0 + \frac{1}{4}}
\]

\[
= 0.707
\]

A probability histogram for \( X \) in Example 29 is:
EXERCISE 4Q

1 Suppose we are tossing a single coin and \( X \) is the number of heads resulting.
   a Construct a probability distribution table for \( X \).
   b Draw a probability histogram for \( X \).
   c Find the population mean and standard deviation of \( X \).

2 Suppose we are rolling a single die and \( X \) is the number of sixes resulting.
   a Construct a probability distribution table for \( X \).
   b Draw a probability histogram for \( X \).
   c Find the population mean and standard deviation of \( X \).

3 Suppose we are tossing three coins and \( X \) is the number of heads resulting.
   a Construct a probability distribution table for \( X \).
   b Draw a probability histogram for \( X \).
   c Find the population mean and standard deviation of \( X \).

4 A pair of dice is rolled and \( S \) is the sum of the numbers on the uppermost faces.
   a Draw a grid which shows the sample space.
   b Hence, copy and complete the following table for \( S \):

<table>
<thead>
<tr>
<th>Sum</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>( \frac{1}{36} )</td>
</tr>
<tr>
<td>3</td>
<td>( \frac{2}{36} )</td>
</tr>
<tr>
<td>4</td>
<td>( \frac{3}{36} )</td>
</tr>
<tr>
<td>5</td>
<td>( \frac{4}{36} )</td>
</tr>
<tr>
<td>6</td>
<td>( \frac{5}{36} )</td>
</tr>
<tr>
<td>7</td>
<td>( \frac{6}{36} )</td>
</tr>
<tr>
<td>8</td>
<td>( \frac{5}{36} )</td>
</tr>
<tr>
<td>9</td>
<td>( \frac{4}{36} )</td>
</tr>
<tr>
<td>10</td>
<td>( \frac{3}{36} )</td>
</tr>
<tr>
<td>11</td>
<td>( \frac{2}{36} )</td>
</tr>
<tr>
<td>12</td>
<td>( \frac{1}{36} )</td>
</tr>
</tbody>
</table>

c Draw a probability histogram for \( S \).

d Enter the data from \( S \) in L1 and the data from \( P \) in L2. Treat the data entered as if it were grouped data and calculate the mean and standard deviation of \( S \). What do you notice?

R

MEAN AND STANDARD DEVIATION OF A BINOMIAL VARIABLE

If we toss an ordinary unbiased coin 20 times we expect it to fall heads half the time. That is, we expect \( np = 20 \times \frac{1}{2} = 10 \) times.

Similarly, if we roll a die with probability \( \frac{1}{6} \) of a four turning up, then in 30 rolls we would expect \( np = 30 \times \frac{1}{6} = 5 \) fours to turn up.

This suggests that

\[
\text{The mean of a binomial distribution is } \mu = np.
\]

Note: The standard deviation can be found using: \( \sigma = \sqrt{np(1-p)} \).

Let us examine Example 29 again using these formulae.
An unbiased coin is tossed twice. Let \( X \) be the number of heads resulting. Find the mean and standard deviation of \( X \).

This is clearly a binomial experiment. In this example, \( n = 2 \) and \( p = \frac{1}{2} \).

\[
\mu = np = 2 \times \frac{1}{2} = 1 \quad \text{and} \quad \sigma = \sqrt{np(1-p)} = \sqrt{2 \left( \frac{1}{2} \right) \left( 1 - \frac{1}{2} \right)} = \sqrt{\frac{1}{2}} \approx 0.707
\]

**EXERCISE 4R**

Repeat the first three questions of the last exercise, and check the formulae above.

1. Suppose we are tossing a single coin and \( X \) is the number of heads resulting.
   - a. What are the values of \( n \) and \( p \)?
   - b. Find the population mean and standard deviation of \( X \).

2. Suppose we are rolling a single die and \( X \) is the number of sixes resulting.
   - a. What are the values of \( n \) and \( p \)?
   - b. Find the population mean and standard deviation of \( X \).

3. Suppose we are tossing three coins and \( X \) is the number of heads resulting.
   - a. What are the values of \( n \) and \( p \)?
   - b. Find the population mean and standard deviation of \( X \).

The formula for calculating \( \sigma \) is not obvious to see. However, in the following investigation we will see how closely experimental means and standard deviations compare with these formulae.

**INVESTIGATION 11**

**THE MEAN AND STANDARD DEVIATION OF THE BINOMIAL VARIABLE**

1. Click on the icon to open up the powerful binomial sorting simulator shown alongside.

The simulator drops balls onto 4 bars. When it hits the first bar the ball moves to the right with probability \( p \) and to the left with probability \( 1 - p \).

The ball then moves on to hit the next three bars until it comes to the bottom in one of 5 positions.

You can also think of this as flipping a coin 4 times with probability \( p \) of heads turning up. Turning to the right at a bar means 1 head, and turning to the left means one tail. The coin is tossed again at the next bar. The final 5 possible outcomes can be thought of as 0, 1, 2, 3 or 4 heads in 4 tosses of the coin.
2 Click the “Start” button to repeat the experiment 1000 times.
In the example shown in the diagram, this data can be summarised in the frequency table:

<table>
<thead>
<tr>
<th>Number of heads</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>66</td>
</tr>
<tr>
<td>1</td>
<td>246</td>
</tr>
<tr>
<td>2</td>
<td>374</td>
</tr>
<tr>
<td>3</td>
<td>244</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
</tr>
</tbody>
</table>

3 You can now find the mean and standard deviation of the data in the frequency table in the usual way. An easier way to calculate these quantities is to open the Statistics package by clicking on the icon.

This shows the open Statistical package.

You can fill in the information from the frequency table in the two left hand columns.
The package automatically calculates a number of statistics, including the mean of 2.006 and the standard deviation of 1.017 for the illustrated data.
Both of these agree closely with formula calculated values

\[
\mu = np \quad \text{and} \quad \sigma = \sqrt{np(1-p)}
\]

\[
= 4 \times \frac{1}{2} = 2 \quad \quad = \sqrt{4\left(\frac{1}{2}\right)^2} = 1
\]

4 Obtain the experimental binomial distribution results for 1000 repetitions with
- \(a\) \(n = 4\) and \(p = 0.5\)
- \(b\) \(n = 5\) and \(p = 0.6\)
- \(c\) \(n = 6\) and \(p = 0.75\)
- \(d\) For each of the distributions obtained in \(a, b,\) and \(c,\) find the mean \(\mu\) and standard deviation \(\sigma\) from the statistics pack.

5 Compare your experimental values with \(\mu = np\) and \(\sigma = \sqrt{np(1-p)}\).
Suppose $X$ is Binomial $(50, p)$, i.e., $X$ is Binomial $(50, \frac{1}{6})$.

Find the mean and standard deviation of $X$ for each of the following values of $p$:

a) $0.1$

b) $0.2$

c) $0.5$

d) $0.8$

e) $0.9$

f) $1$

Let $X$ be Binomial $(100, p)$.

a) Find the mean and standard deviation of $X$.

b) What is the probability an outcome is less than 1 standard deviation below the mean?

Example 31

A fair die is rolled 42 times. $X$ is the number of fours that could result.

Find the mean and standard deviation of the $X$ distribution.

This is a binomial distribution with $n = 42$, and $p = \frac{1}{6}$, i.e., $X$ is Binomial $(42, \frac{1}{6})$.

$\mu = np = 42 \times \frac{1}{6} = 7$ and $\sigma = \sqrt{np(1-p)} = \sqrt{42\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)} \approx 2.42$

Example 32

Suppose $X$ is Binomial $(100, 0.25)$.

a) Find the mean and standard deviation of $X$.

b) What is the probability an outcome is less than 1 standard deviation below the mean?

a) The mean $\mu = 100 \times 0.25 = 25$

The standard deviation $\sigma = \sqrt{100 \times 0.25 \times 0.75} = 4.33$

b) If an outcome is less than 1 standard deviation below the mean, it is less than $25 - 4.33 = 20.67$.

Since the outcomes are all integers this is the probability that the outcomes are less than or equal to 20.

$P(X \leq 20) = 0.149$  
{using technology}

4 Suppose $X$ is Binomial $(50, p)$. Find the mean and standard deviation of $X$ for each of the following values of $p$:

a) $0.1$  
b) $0.2$  
c) $0.5$  
d) $0.8$  
e) $0.9$  
f) $1$

5 Let $X$ be Binomial $(100, p)$.

a) Sketch the graph of $\sigma^2$ against $p$.  
b) For what value of $p$ is $\sigma$ a maximum?

6 Valerie tossed a coin 100 times and only 43 heads appeared.

a) Assuming the coin was fair, how many standard deviations is this less than the mean?

b) Assuming the coin was fair, what is the probability of getting less than 43 heads in 100 tosses of the coin?

7 A biased coin has probability of 0.4 of turning up heads. Let $X$ be the number of heads in 200 tosses of the coin.

a) Calculate the mean $\mu$, and the standard deviation $\sigma$.

b) Find the probabilities that $X$ lies within:

i) 1 standard deviation of the mean

ii) 2 standard deviations of the mean

iii) 3 standard deviations of the mean

c) Compare your answer with those from a normal distribution.
Here are some graphs of binomial distributions:

\( n = 30, \ p = 0.7 \)  
\( n = 35, \ p = 0.45 \)

Some distributions look very much like normal distributions with symmetry about the mean. The following questions should be done using technology.

Click on the demo icon or the icon for your calculator for help if necessary.

Remember to use \( \mu = np \) and \( \sigma = \sqrt{np(1-p)} \).

\[8\] Assume Shaq’s free throw percentage over his basketball career is 50% and imagine that in one game he has 20 free throw attempts. Let \( X \) be the number of successful free throw attempts.

- **a** Use technology to obtain a histogram or scatterplot of the binomial distribution for \( X \).
- **b** Calculate the mean and standard deviation of the random variable \( X \).
- **c** Hence, plot a continuous normal distribution over the histogram or scatterplot.
- **d** Calculate the probability of Shaq making between 8 and 15 (inclusive) free throws.
- **e** Could the probability in **d** be calculated using the normal distribution as an approximation?

\[9\] Repeat **8** assuming Michael’s free throw percentage over his career is 90% and that he has 20 attempts.

\[10\] Repeat **8** assuming Dennis’s free throw percentage over his career is 25% and that he has 20 attempts.
**EXTENSION**

One trial \((n = 1)\)

In the case of \(n = 1\) where \(p\) is the probability of success and \(1 - p\) is the probability of failure, the number of successes \(x\) could be 0 or 1.

The table of probabilities is

<table>
<thead>
<tr>
<th>(x)</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p)</td>
<td>1 - (p)</td>
<td>(p)</td>
</tr>
</tbody>
</table>

Now \(\mu = x_1p_1 + p_2x_2\)

\[\begin{align*}
\mu &= 0(1 - p) + 1(p) \\
&= p
\end{align*}\]

\[
\sigma^2 = (x_1 - \mu)^2p_1 + (x_2 - \mu)^2p_2
\]

\[\begin{align*}
&= (0 - p)^2(1 - p) + (1 - p)^2p \\
&= p^2(1 - p) + (1 - p)^2p \\
&= p(1 - p)[p + 1 - p] \\
&= p(1 - p)
\end{align*}\]

\[\therefore \sigma = \sqrt{p(1 - p)}\]

Two trials \((n = 2)\)

In the case where \(n = 2\),

\[
\begin{align*}
\Pr(0) &= C^2_0 p^0(1 - p)^2 = (1 - p)^2 \\
\Pr(1) &= C^2_1 p^1(1 - p)^1 = 2p(1 - p) \\
\Pr(2) &= C^2_2 p^2(1 - p)^0 = p^2
\end{align*}\]

as \(x = 0, 1\) or 2

So, the table of probabilities is:

<table>
<thead>
<tr>
<th>(x)</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p)</td>
<td>((1 - p)^2)</td>
<td>(2p(1 - p))</td>
<td>(p^2)</td>
</tr>
</tbody>
</table>

Now \(\mu = x_1p_1 + x_2p_2 + x_3p_3\)

\[\begin{align*}
\mu &= 0(1 - p)^2 + 2p(1 - p) + 2p^2 \\
&= 0 + 2p - 2p^2 + 2p^2 \\
&= 2p
\end{align*}\]

\[
\sigma^2 = (x_1 - \mu)^2p_1 + (x_2 - \mu)^2p_2 + (x_3 - \mu)^2p_3
\]

\[\begin{align*}
&= (-2p)^2(1 - p)^2 + (1 - 2p)^2(1 - p) + (2 - 2p)^2p^2 \\
&= 4p^2(1 - p)^2 + (1 - 2p)^22p(1 - p) + 4(1 - p)^2p^2 \\
&= 2p(1 - p)[2p(1 - p) + (1 - 2p)^2 + 2p(1 - p)] \\
&= 2p(1 - p)[2p - 2p^2 + 1 - 4p + 4p^2 + 2p - 2p^2] \\
&= 2p(1 - p)(1)
\end{align*}\]

\[\therefore \sigma = \sqrt{2p(1 - p)}\]

**Challenge:**

11 If \(X\) is Bin(3, \(p\)), show that:

- \(a\) the mean of the distribution is 3\(p\)
- \(b\) the standard deviation is \(\sigma = \sqrt{3p(1 - p)}\)

12 If \(X\) is Bin(4, \(p\)), show that:

- \(a\) the mean of the distribution is 4\(p\)
- \(b\) the standard deviation is \(\sigma = \sqrt{4p(1 - p)}\)
REVIEW SET 4D

1. **a** List at least 3 factors that could affect the weight of Granny Smith apples grown in an orchard.
   **b** Suggest why the diameters of snails in your garden could be normally distributed.

2. Suppose the maximum temperature in April is normally distributed with mean 20.7°C and standard deviation of 3°C. How many days in April can be expected to have temperatures between 17.7°C and 26.7°C?

3. Make a free-hand sketch for the following distributions on the same set of axes.
   **a** $N(0, 1^2)$  
   **b** $N(3, 3^2)$  
   **c** $N(-2, (\frac{1}{2})^2)$

4. An examination result is normally distributed with $\mu = 62$ and $\sigma = 9$.
   **a** What is the $z$-score of a mark of 53?
   **b** If a credit is to be awarded to any student with a $z$-score larger than 2, what is the minimum mark to get a credit?
   **c** If 30 students sat for this examination, how many were awarded a credit?

5. In a school competition, the time to run 100 metres was normally distributed with mean 13.7 seconds and standard deviation of 1.1 seconds. The 500 metre race was normally distributed with mean of 148 seconds and standard deviation of 22 seconds. If Henry ran the 100 metres in 12 seconds, and Peter ran the 500 metres in 115 seconds, who was the better performer?

6. The seventh row of the Pascal’s triangle is: 1 7 21 35 35 21 7 1
   **a** Write down row eight of Pascal’s triangle.
   **b** What is $C_4^8$?

7. A multiple choice exam has 20 questions, and each question has 3 possible choices. If the pass mark is 10, what is the probability a student passes this exam by randomly guessing each answer?

8. $X$ is Bin(20, 0.8).
   **a** What is the mean and standard deviation of $X$?
   **b** What number is 2 standard deviations below the mean?
1 The following are the graphs of the normal distributions \(N(0, 2^2), \ N(3, 1^2)\) and \(N(0, 0.2^2)\). Explain which graph belongs to which distribution.

2 A ticket printing machine prints on average 50,000 tickets a day with standard deviation of 2,500. Assuming that production is normally distributed and a year comprises 260 working days, calculate the approximate number of working days that:
   a) under 47,500 tickets are printed
   b) over 45,000 tickets are printed
   c) between 47,500 and 55,000 tickets (inclusive) are printed.

3 The results of a Physics exam are normally distributed with mean 65 and standard deviation 7. The results of a Mathematics exam are normally distributed with a mean of 75 and standard deviation 15.
   Allan scored a mark of 65 for Physics and a mark of 90 for Mathematics, giving him a total score of 155.
   Anne scored a mark of 72 for Physics and a mark of 80 for Mathematics giving her a total score of 152.
   On the basis of these scores, instead of Anne, Allan was given an academic prize. Was this fair?

4 The weight of apples from a farmer’s apple crop is normally distributed with mean of 62 g and standard deviation 5.7 g.
   a) Find the z-score of an apple weighing 47 g.
   b) If the farmer can only sell apples with weights between 57 and 80 grams, what proportion of the crop can the farmer sell?

5 If \(Z\) is the standard normal distribution, calculate each of the following:
   a) \(P(-1 \leq Z \leq 2)\)
   b) \(P(Z > 1.5)\)

6 Use your calculator to calculate \(C_{10}^0\), \(C_{10}^5\), and \(C_{10}^{10}\).
   Use these values to calculate \(C_{11}^7\) and \(C_{8}^{11}\).

7 Suppose \(X\) is Bin(40, 0.4)
   a) What are the mean and standard deviation of \(X\)?
   b) What proportion of the outcomes of \(X\) would you expect to lie within 2 standard
How does this compare with the normal distribution?

c A coin has probability of 0.4 of coming up heads. If the coin is tossed 50 times, what is the probability that more than 23 heads appear?

## REVIEW SET 4F

1. The mean of a variable $X$ that is normally distributed is 68 and the standard deviation is 14. What percentage of the values:
   - a are less than 82
   - b lie between 40 and 82?

2. Draw each of the following distributions accurately on one set of axes.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Form</th>
<th>mean (cm)</th>
<th>standard deviation (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>normal</td>
<td>45</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>normal</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>normal</td>
<td>41</td>
<td>8</td>
</tr>
</tbody>
</table>

3. Two dairy produce companies sell the same type of margarine in containers that are stamped 275 g. The information given about the weight of the population of containers from each distributor is known:

<table>
<thead>
<tr>
<th>Company</th>
<th>Form</th>
<th>mean (g)</th>
<th>standard deviation (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>normal</td>
<td>279</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>normal</td>
<td>279</td>
<td>4</td>
</tr>
</tbody>
</table>

   - a For each company, determine the proportion of containers which have weights:
     - i less than or equal to the stamped weight of 275 g
     - ii greater than or equal to 283 g
   - b Determine the proportion of containers that company B will produce that will lie between or are equal to 271 g and 283 g.
   - c Determine the proportion of containers that company A will produce that will lie between or are equal to 275 g and 281 g.

4. Explain why it is feasible that the distribution of each of the following variables is normal.
   - a The length of nails immediately after manufacture.
   - b The time it takes the 7:30 Belair train to travel to Adelaide.

5. If $X$ is $N(30, 4^2)$ and $Y$ is $N(150, 30^2)$, which of the following outcomes is more extreme, $x = 36$ or $y = 190$?

6. A medical treatment has 70% chance of being successful. Suppose a doctor is to apply this treatment to three patients.
   - a Draw a tree diagram to show all possible outcomes.
   - b From the tree diagram find the probability that the treatment is successful for 2 out of the three patients.

7. Find the probability that if a fair coin is tossed 100 times, exactly 54 heads will appear.