Contents:

A Describing data
B Collecting information
C Random sampling
D Presenting and interpreting data
E Grouped discrete data
F Continuous (interval) data
G Measures of centres of distributions
H Measuring the spread of data
I Comparing data
J What a sample tells us about a population
K Data based investigation
L Relative frequency
M Normal distributions
N Properties of normal distributions
O Technology and normal distributions

Review set 5
The earlier chapter Data in Context discussed how to collect, collate, and organise information. This chapter extends from that work.

Once information has been collected and collated, the data it yields can be analysed and interpreted. Mathematics provides a number of tools for that purpose. A conclusion based on a mathematical interpretation of numerical data is called a statistic. Mathematical conclusions based on numerical data are called statistics.

There are a number of steps involved when considering any statistical problem:

- identify and state the problem
- develop a method to be used in the investigation
- decide how to collect the data
- collect the data
- organise and analyse the data
- interpret the data and form a conclusion
- reconsider the underlying assumptions of the investigation.

Before we look at how to use mathematics to interpret information, we begin with how numerical information can be collected and organised.

**DESCRIPTING 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. The data 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.

There are two types of quantitative variables: discrete and 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 take the values 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 5A**

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
2a For the categorical variables in question 1, write down two or three possible categories. Discuss your answers.

2b 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.

B COLLECTING INFORMATION

Information is collected by asking questions.

Designing questions (and questionnaires) requires considerable skill because of the type of information that an answer can provide.

Wherever possible, questions should be clear and simple.

Care must be taken in the way in which a question is phrased. Leading questions can manipulate the opinions of the person answering, producing misleading results.

For example, if the Government is considering reallocating funds from education to health, a poll asking “Should the Government cut funding to schools?” is likely to produce a negative response to the reallocation, whereas a poll asking “Should the Government increase funding to hospitals?” is likely to produce a positive response.

A poll asking “Should the Government move funds from education to health?” is more likely to produce a balanced response.

CHECKING QUESTION RESPONSES

Sometimes it is a good idea to ask the same question but from a slightly different perspective so that the responses can be checked. This process can help determine if a person is consistent.

For example, suppose that student work in a statistics course is assessed by three different assessment tasks.

How students value a particular task may be reflected in the way they think it should be rated.

A first question might be:

On a scale of 0 to 100, rate the value you place on each of these tasks.
Answers to this question can tell a researcher what value a student placed on each task. To check this, a researcher may want to put a similar question later in the questionnaire. Continuing with the example above, if students value a particular task highly then it seems reasonable to think that they may expect that their work in that task should also be weighted highly in their overall assessment. A check question could be:

What percentage of your overall assessment should be given for your work in
Written tests?

This could be repeated for the other types of task.

Depending on the responses, the task of the researcher is to then refine his or her understanding of the reasons for any differences in what the person indicated.

The quality of the questions used determines the quality of the statistics that come from them. This is particularly the case for any business which relies on accurate information in order to meet the needs of its clients. Careful survey preparation gives the results credibility. For that reason a researcher will often trial a questionnaire to see if it meets its aims. This process in turn improves the value of the questionnaire.

Summary:

As you already know data comes in two types, discrete and continuous. Ideally data is collected by interview using questions given in writing to a person. In this way the researcher can avoid swaying a respondent with their voice and manner. Sometimes it is a good idea to repeat a question to check the reliability of a response.

Finally, from a mathematics point of view, questions must:

- provide data
- be framed so that only one type of data response is collected from each question
- not lead the person to a predetermined answer
- be designed to be unambiguous in their interpretation.
RANDOM SAMPLING

Information is collected from a population. A population is usually understood to be the people who live in a particular country. In statistics, however, the word population refers to all the members of a particular group being considered. That may mean all the customers or clients of a particular business, or it could be as specific as the number of ball bearings produced by a series of machines. Studying a population will provide information about all its members.

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

Data can be collected from every member of a population in what is called a census. Every 5 years the Australian government carries out a census in which it seeks basic information from the whole population.

However, it is often too expensive or impractical to obtain information from every member of the population. Often information about the whole population can be successfully gathered from a sample, or part, of the population.

A sample is a selection from the population.

For example, 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.

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 an equal chance of being selected.

We discuss four commonly used random sampling techniques.

These are:

- simple random sampling
- systematic sampling
- cluster sampling
- stratified sampling.

SIMPLE RANDOM SAMPLING

Making sure that a sample is representative of a whole population can be a difficult problem. The aim is to select the members of a sample in a random way, that is, each member of a population is equally likely to be chosen for the sample.

This random selection process aims to avoid bias.
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.

**SYSTEMATIC SAMPLING**

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.

In this process, the first member of the sample is chosen at random and every other member is chosen according to a set pattern, for example, every fourth person after that.

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

**CLUSTER SAMPLING**

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.

**STRATIFIED RANDOM SAMPLING**

Often a population is made up of diverse groups of varying size. A stratified sample aims to reflect the same proportions and particular diversity of the broader population.

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:
15% of 238 = 36 year 8’s
15% of 253 = 38 year 9’s
15% of 227 = 34 year 10’s
15% of 235 = 35 year 11’s
15% of 182 = 27 year 12’s

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 number 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 5B**

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’ opinions of its facilities that have been set up near 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 contacts 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 have 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 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.
When taking a sample it is hoped that the information gathered is **representative of the entire population**. We must take certain steps to ensure that this is so. If the sample we choose is too small, the data obtained is likely to be **less reliable** than that obtained from larger samples.

For accurate information when sampling, it is essential that:
- the individuals involved in the survey are **randomly chosen** from the population
- the number of individuals in the sample is **large enough**.

**Note:** How to choose a random sample using random numbers is covered in Year 12 Maths Applications.

For example:
Measuring a group of three fifteen-year-olds would not give a very reliable estimate of the height of fifteen-year-olds all over the world. We therefore need to choose a **random sample** that is large enough to represent the population. Note that conclusions based on a sample will never be as accurate as conclusions made from the whole population, but if we choose our sample carefully, they will be a good representation.

Care should be taken not to make a sample too large as this is costly, time consuming and often unnecessary. A balance needs to be struck so that the sample is large enough for there to be confidence in the results but not so large that it is too costly and time consuming to collect and analyse the data.

**THE SIZE OF A SAMPLE**

- For an extremely **large population** where the **population size is unknown**:
  To be very confident that a sample accurately reflects the population within \( \pm r\% \), we take a sample of size \( n \) where
  \[
  n \approx \frac{9600}{r^2}
  \]

- For a **population** size known to be \( N \):
  To be very confident that a sample accurately reflects the population within \( \pm r\% \), we take a sample of size \( n \) where
  \[
  n \approx \frac{9600N}{9600 + Nr^2}
  \]

**Example 2**
To examine the effect of a new diabetes drug, a sample of users needs to be taken. How large a sample must be taken to be very confident that the sample accurately reflects the population of users within \( \pm 2\% \) if the population size is unknown?

As the population size is unknown, \( n = \frac{9600}{r^2} = \frac{9600}{2^2} = 2400 \)

So, to be very confident, within \( \pm 2\% \), a sample of about 2400 needs to be taken.
A reporter for the Western Suburbs was seeking answers to the following question: ‘Do you want more money spent on roads?’

How could he investigate this statistically if there are 87,694 voters on the electoral roll and he wishes to be very confident of accuracy within ±1.5%?

It would be impractical to survey every voter on the electoral roll, so a random sample could be used. The sample size should be calculated using:

\[
n = \frac{9600N}{9600 + N\pi^2}
\]

\[
= \frac{9600 \times 87694}{9600 + (87694) \times 1.5^2}
\]

\[
\div 4069
\]

So, a sample of about 4070 should be taken.

Click on the icon to obtain a sample size calculator.

You may wish to program your graphics calculator to obtain these results.

**EXERCISE 5C**

1. A clothing manufacturer produces 450 shirts per week. Each week 25 shirts are randomly selected by the quality control staff and checked. 4 were found to be defective in one week.
   
   a. How many shirts per week form the population?
   b. Estimate the total number of shirts each week which are defective.
   c. Estimate the percentage of shirts produced each week which are satisfactory.

2. A factory produces 5000 microprocessors per week. A random sample of 400 revealed that 2 were faulty.
   
   a. What size is the population?
   b. What size is the sample?
   c. Estimate the total number of microprocessors produced in a week that are not faulty.

3. 1150 householders were selected at random from the electoral roll and asked whether they would vote for the Australian Labor Party. The survey revealed that 620 answered ‘yes’.
   
   a. If there are 12.6 million people in Australia over the age of 18, estimate how many of them would answer ‘no’.
   b. What percentage of Australians over 18 would answer ‘yes’ in your estimation?

4. Discuss how you would randomly select:
   
   a. first and second prize in a cricket club raffle
   b. three numbers from 0 to 37 on a roulette wheel.
5 “In conducting a survey to find out the percentage of people who believe the AFL grand final should always be played at the MCG (Melbourne), it would be a good idea to ask a section of the crowd at this year’s clash between Melbourne and Saint Kilda.” Discuss.

6 A newspaper conducts a survey of Australians to determine whether they believe Australia should be more restrictive in its immigration policy. How many people must be surveyed to be very confident that the survey will be accurate within 1.5% if the population size is unknown?

7 A government survey is to be held on the question of water use from our river systems. How many people should be surveyed to be very confident of accuracy within 0.8%?

8 To determine whether members of a local club would be willing to pay higher fees in order to fund the installation of new equipment, a sample of the members is surveyed. Given that there are 748 members at the club, how large a sample must be taken to be very confident that the sample accurately measures the views of all the members within 2%?

9 To find the proportion of capsicums in a crop which are suitable for sale, a sample of them will be tested. How many capsicums must be tested to be very confident that the sample accurately reflects the quality of all capsicums in the crop (within 3%), if there are 16 300 capsicums in the crop?

D PRESENTING AND INTERPRETING DATA

ORGANISING CATEGORICAL DATA

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>
DISPLAYING CATEGORICAL DATA

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

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

THE MODE

The mode of a set of categorical data is the category which occurs most frequently.

This table shows the colours of cars recorded during a busy time on a main road.

‘Red’ occurs more often than any other colour. So, the mode is ‘Red’.

<table>
<thead>
<tr>
<th>Colour</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>50</td>
</tr>
<tr>
<td>Red</td>
<td>72</td>
</tr>
<tr>
<td>Yellow</td>
<td>38</td>
</tr>
<tr>
<td>Black</td>
<td>32</td>
</tr>
<tr>
<td>Blue</td>
<td>21</td>
</tr>
<tr>
<td>Green</td>
<td>18</td>
</tr>
<tr>
<td>Silver</td>
<td>47</td>
</tr>
</tbody>
</table>
EXERCISE 5D.1

1 What is the mode of the ‘sporting injuries’ data?

2 What categories could be / are used for:
   a egg sizes          b wine bottle sizes          c eye colour?

3 Each girl at a high school has named the sport that she would most like to play. The table shows this data.
   a What percentage would like to play tennis?
   b What is the mode of the data?
   c Use technology to obtain for the data:
      i a vertical column graph         ii a pie chart.

<table>
<thead>
<tr>
<th>Sport</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tennis</td>
<td>68</td>
</tr>
<tr>
<td>Netball</td>
<td>87</td>
</tr>
<tr>
<td>Basketball</td>
<td>48</td>
</tr>
<tr>
<td>Volleyball</td>
<td>43</td>
</tr>
<tr>
<td>Badminton</td>
<td>59</td>
</tr>
<tr>
<td>Squash</td>
<td>17</td>
</tr>
</tbody>
</table>

4 There are six categories of membership, A to F, at a golf club.
   a What is the mode?
   b In a pie chart, what sector angle would be used for category: i A ii D?
   c Draw a pie chart of the data.

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:

<table>
<thead>
<tr>
<th>Current formula</th>
<th>New formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 6 8 5 8 6 4 7 6 6</td>
<td>7 3 6 7 8 6 7 7 7 7</td>
</tr>
<tr>
<td>7 5 7 9 3 7 9 6 6 6</td>
<td>8 6 7 7 7 6 6 6 4 8</td>
</tr>
<tr>
<td>8 6 7 6 8 6 7 7 7 5</td>
<td>6 7 7 4 7 5 6 6 6 6</td>
</tr>
<tr>
<td>6 7 6 9 7 5 4 6 8 7</td>
<td>7 5 8 6 5 9 7 7 8 7</td>
</tr>
<tr>
<td>6 7 4 6 8 7 6 7 6 6</td>
<td>6 7 6 8 7 7 6 6 6 7</td>
</tr>
<tr>
<td>7 8 7 7 9 7 7 8 7 7</td>
<td>9 6 6 7 6 7 5 6 8 14</td>
</tr>
</tbody>
</table>
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 ||| 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>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>60</td>
<td>1</td>
</tr>
</tbody>
</table>

A **column graph** of the frequencies or the relative frequencies can be used to display 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** 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 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.

  For example, consider the graph shown.

  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 $\frac{1}{2}$.

  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 5D.2**

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:

- 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:

- 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, 54, 42

We can organise the data into a frequency table.

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>
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>10 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>

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

**EXERCISE 5E**

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

**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.
- for the data value 116, 11 is recorded on the stem and 6 is the leaf.

The stem-and-leaf plot is:

```
Stem | Leaf       | Stem | Leaf
-----|------------|------|-------
1    | 9          | 1    | 9     
2    | 8 6 7 3 8  | 2    | 3 6 7 8 8 |
3    | 4 3 2 9 7 3 8 4 4 4 | 3    | 2 3 3 4 4 4 7 8 9 |
4    | 8 0 1 9 2 8 2 5 2 | 4    | 0 1 2 2 5 8 8 9 |
5    | 6 0 1 4      | 5    | 0 1 4 6 |
6    | 2 Note: 2 | 3 means 23 | 6    | 2    
```

The ordered stem-and-leaf plot is:

```
Stem | Leaf
-----|-------
1    | 9     
2    | 3 6 7 8 8 |
3    | 2 3 3 4 4 4 7 8 9 |
4    | 0 1 2 2 5 8 8 9 |
5    | 0 1 4 6 |
6    | 2    
```
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) has the highest frequency
- no data is lost
- the stemplot shows the spread of data for each class
- the stemplot provides a horizontal histogram
- the stemplot automatically creates a frequency table of the data.

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

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>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2 3 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0 4 4 7 8 9 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0 0 1 1 2 2 3 5 5 6 8 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0 1 2 4 4 5 8 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0 3 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

f 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.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>29</td>
</tr>
<tr>
<td>29</td>
<td>39</td>
</tr>
<tr>
<td>27</td>
<td>26</td>
</tr>
<tr>
<td>29</td>
<td>36</td>
</tr>
<tr>
<td>41</td>
<td>45</td>
</tr>
<tr>
<td>29</td>
<td>25</td>
</tr>
<tr>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>33</td>
<td>34</td>
</tr>
<tr>
<td>30</td>
<td>32</td>
</tr>
<tr>
<td>25</td>
<td>36</td>
</tr>
<tr>
<td>42</td>
<td>26</td>
</tr>
<tr>
<td>36</td>
<td>25</td>
</tr>
<tr>
<td>50</td>
<td>18</td>
</tr>
<tr>
<td>33</td>
<td>34</td>
</tr>
</tbody>
</table>

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?

4 The stemplot below shows the results of a test for a group of students. The test was marked out of 35.

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

1 | 6 = 16

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

<table>
<thead>
<tr>
<th>f</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

1 | 6 = 16

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

<table>
<thead>
<tr>
<th>f</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

1 | 6 = 16
The data above was re-organised into smaller classes. Data from 10 to 14 is recorded with stem 1, and data from 15 to 19 is recorded with stem 1*, etc.

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

d Describe the distribution of the results. Compare your response with part c, above.

e What is the range of scores for the test?
f Is the data evenly spread?

F CONTINUOUS (INTERVAL) DATA

Recall that:

**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.

**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.
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 5F

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 - < 70 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 - &lt; 20</td>
<td>5</td>
</tr>
<tr>
<td>20 - &lt; 30</td>
<td>21</td>
</tr>
<tr>
<td>30 - &lt; 40</td>
<td>17</td>
</tr>
<tr>
<td>40 - &lt; 50</td>
<td>8</td>
</tr>
<tr>
<td>50 - &lt; 60</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>Freq.</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>

a How many of the seedlings are 150 mm or more?
b What percentage of the seedlings are in the 125 - < 150 mm class?
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.

G 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 the state?

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 $\bar{x}$.

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.

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

If $n = 11$, $\frac{n+1}{2} = 6$, and so the median is the 6th score.

If $n = 12$, $\frac{n+1}{2} = 6.5$, and so the median is the average of the 6th and 7th scores.
Example 5

The number of typing errors on the pages of Nigel’s assignment were: 4, 1, 3, 7, 2, 6, 5, 3, 8, 6 and 1. Find his:

a \text{ mean number of errors } 

\[
\text{mean} = \frac{4 + 1 + 3 + 7 + 2 + 6 + 5 + 3 + 8 + 6 + 1}{11} = \frac{46}{11} = 4.18
\]

b \text{ median number of errors. }

In order of size: 1, 1, 2, 3, 3, 4, 5, 6, 6, 7, 8.

\[
\text{median} = \text{6th score } \left\{ \frac{n + 1}{2} = \frac{11 + 1}{2} = 6 \right\}
\]

Example 6

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

a \text{ mean } 

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

b \text{ The ordered data set is: } 12, 13, 13, 14, 14, 15, 15, 16, 17, 18.

<table>
<thead>
<tr>
<th>Income ($'000)</th>
<th>100 - &lt; 150</th>
<th>150 - &lt; 200</th>
<th>200 - &lt; 250</th>
<th>250 - &lt; 300</th>
<th>300 - &lt; 350</th>
<th>350 - &lt; 400</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of directors</td>
<td>5</td>
<td>11</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

INTERPRETING THE MEDIAN

Regardless of its actual value, the median is the score in the middle of the data. One half (50%) of the data is below it and one half above it.

Consider this table of incomes of company directors:
Notice that this data is not evenly spread throughout the categories.
The values at the top end seem to distort the data.
The median of this data is about $190,000.

**STATISTICS USING A COMPUTER**

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

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

**STATISTICS USING A GRAPHICS CALCULATOR**

Consider the data 2, 3, 5, 4, 3, 6, 5, 7, 3, 8, 1, 7, 5, 5, 9.

For **TI-83**

Data is entered in the **STAT EDIT** menu. Press **STAT** 1 to select 1:Edit

In L1, delete all existing data. Enter the new data.

Press 2 ENTER then 3 ENTER etc., until all data is entered.

To obtain the descriptive statistics

Press **STAT** ▼ to select the **STAT CALC** menu. Press 1 to select 1:1–Var Stats

Pressing 2nd 1 (L1) ENTER gives the mean $\bar{x} = 4.87$ (to 3 sf)

Scrolling down by pressing ▼ repeatedly gives the median = 5

---

**Statistics Using a Computer**

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

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

**Statistics Using a Graphics Calculator**

Consider the data 2, 3, 5, 4, 3, 6, 5, 7, 3, 8, 1, 7, 5, 5, 9.

For **TI-83**

Data is entered in the **STAT EDIT** menu. Press **STAT** 1 to select 1:Edit

In L1, delete all existing data. Enter the new data.

Press 2 ENTER then 3 ENTER etc., until all data is entered.

To obtain the descriptive statistics

Press **STAT** ▼ to select the **STAT CALC** menu. Press 1 to select 1:1–Var Stats

Pressing 2nd 1 (L1) ENTER gives the mean $\bar{x} = 4.87$ (to 3 sf)

Scrolling down by pressing ▼ repeatedly gives the median = 5
For **Casio**

From the Main Menu, select **In delete all existing data and enter the new data. Press 2 then 3 etc until all data is entered**

**STAT.** In **List 1**, delete all existing data and enter the new data.

<table>
<thead>
<tr>
<th>List</th>
<th>List 1</th>
<th>List 2</th>
<th>List 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To obtain the descriptive statistics

Press **F6** (>) if the **GRPH** icon is not in the bottom left corner of the screen.

Press **F2** (CALC) **F1** (1VAR) which gives the mean $\bar{x} = 4.87$ (to 3 sf)

Scrolling down by pressing **Y** repeatedly gives the median = 5

**EXERCISE 5G**

In the following exercise you should use technology.

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

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. 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?

3. 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 price 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?

4. 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.
GROUPED DISCRETE DATA

**Example 7**

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</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>2</td>
</tr>
<tr>
<td>48</td>
<td>4</td>
</tr>
<tr>
<td>49</td>
<td>7</td>
</tr>
<tr>
<td>50</td>
<td>8</td>
</tr>
<tr>
<td>51</td>
<td>3</td>
</tr>
<tr>
<td>52</td>
<td>1</td>
</tr>
</tbody>
</table>

For **T1-83**

Press **STAT 1** to select **1:Edit**. Key the variable values into **L1** and the frequency values into **L2**. Press **STAT 1** to select **1:1–Var Stats** from the **STAT CALC** menu.

Enter **L1**, **L2** by pressing **2nd L1** (L1) **2nd L2** (L2) **ENTER**
- The mean is 49.4.
- Scroll down, and the median is 49.

**Note:** If you do not include **L2** you will get a screen of statistics for **L1** only.

For **Casio**

From the Main Menu, select **STAT**. Key the variable values into **List 1** and the frequency values into **List 2**.

Press **F6** (>) if the **GRPH** icon is not in the bottom left corner of the screen.

Press **F2 (CALC)** **F6 (SET)** **F2 (LIST)** **EXP** to change the frequency variable to **List 2**.

Press **EXIT F1 (1VAR)**
- The mean is 49.4.
- Scroll down, and the median is 49.
Use technology to answer these questions.

5 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>

6 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>

7 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 8

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** Find the modal class for this data.
- **c** Find the mean for this data.
- **d** Find the median for this data.

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Midpt. ($x$)</th>
<th>Freq. ($f$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>240 - &lt; 245</td>
<td>242.5</td>
<td>1</td>
</tr>
<tr>
<td>245 - &lt; 250</td>
<td>247.5</td>
<td>3</td>
</tr>
<tr>
<td>250 - &lt; 255</td>
<td>252.5</td>
<td>6</td>
</tr>
<tr>
<td>255 - &lt; 260</td>
<td>257.5</td>
<td>2</td>
</tr>
<tr>
<td>260 - &lt; 265</td>
<td>262.5</td>
<td>7</td>
</tr>
<tr>
<td>265 - &lt; 270</td>
<td>267.5</td>
<td>6</td>
</tr>
<tr>
<td>270 - &lt; 275</td>
<td>272.5</td>
<td>2</td>
</tr>
<tr>
<td>275 - &lt; 280</td>
<td>277.5</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>28</strong></td>
</tr>
</tbody>
</table>

**For TI-83**

We enter the midpoints into **L1** and the frequencies into **L2**.

We then proceed using the instructions as in **Example 7** to get

- The mean is 260 m.
- The median is 262.5 m.

**For Casio**

We enter the midpoints into **List 1** and the frequencies into **List 2**.

We then proceed using the instructions as in **Example 7** to get

- The mean is 260 m.
- The median is 262.5 m.

**Note:** The median is given here as one of the midpoints entered. Why?
8 Find the approximate mean for each of the following distributions:

\[
\begin{array}{|c|c|}
\hline
\text{Score (x)} & \text{Frequency (f)} \\
\hline 1 - 5 & 2 \\
6 - 10 & 7 \\
11 - 15 & 9 \\
16 - 20 & 8 \\
21 - 25 & 3 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|}
\hline
\text{Score (x)} & \text{Frequency (f)} \\
\hline 40 - 42 & 2 \\
43 - 45 & 1 \\
46 - 48 & 4 \\
49 - 51 & 7 \\
52 - 54 & 11 \\
55 - 57 & 3 \\
\hline
\end{array}
\]

9 30 students sit a mathematics test and the results are as follows:

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\text{Score} & 0 - 9 & 10 - 19 & 20 - 29 & 30 - 39 & 40 - 49 \\
\text{Frequency} & 1 & 4 & 8 & 14 & 3 \\
\hline
\end{array}
\]

Find the approximate value of the mean score.

10 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.

\[
\begin{array}{|c|c|}
\hline
\text{Weight (kg)} & \text{Frequency} \\
\hline 1.0 - < 1.5 & 1 \\
1.5 - < 2.0 & 2 \\
2.0 - < 2.5 & 6 \\
2.5 - < 3.0 & 17 \\
3.0 - < 3.5 & 11 \\
3.5 - < 4.0 & 8 \\
4.0 - < 4.5 & 0 \\
4.5 - < 5.0 & 1 \\
\hline
\end{array}
\]

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

\[
\begin{array}{|c|c|}
\hline
\text{Litres (L)} & \text{Frequency} \\
\hline 3000 - < 4000 & 5 \\
4000 - < 5000 & 1 \\
5000 - < 6000 & 7 \\
6000 - < 7000 & 18 \\
7000 - < 8000 & 13 \\
8000 - < 9000 & 6 \\
\hline
\end{array}
\]

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.

INVESTIGATION 3

EFFECTS OF OUTLIERS

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.

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, 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.
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 9

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, \( IQR = Q_3 - Q_1 \).

Example 10

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:

\[ \begin{align*}
\text{range} &= 5.1 - 2.3 = 2.8 \text{ kg} \\
\text{median} &= 3.4 \text{ kg} \\
IQR &= Q_3 - Q_1 \\
&= 4.15 - 2.9 \\
&= 1.25 \text{ kg}
\end{align*} \]
Example 11

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, 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 Casio we obtain:

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>min</th>
<th>Q1</th>
<th>Med</th>
<th>Q3</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>19</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>95</td>
</tr>
</tbody>
</table>

EXERCISE 5H

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, 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, 61, 59, 58, 60
c The number of beans in a pod:

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

d Stem  Leaf
1  35779
2  0135668
3  04479
4  37
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 2.2 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.
c 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 \((\text{Min}_x)\)
- the lower quartile \((Q_1)\)
- the median \((Q_2)\)
- the upper quartile \((Q_3)\)
- the maximum value \((\text{Max}_x)\)

These five numbers form the five-number summary of a data set.

In Example 11 the five-number summary and the corresponding boxplot are:

- minimum = 3
- \(Q_1 = 5\)
- median = 6
- \(Q_3 = 7\)
- maximum = 9

![Boxplot Illustration]

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 12

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 throws 7 goals or less.

From a TI-83:

So the 5-number summary is:

\[
\begin{align*}
\text{Min}_x &= 2 \\
Q_1 &= 4 \\
\text{median} &= 6 \\
Q_3 &= 7 \\
\text{Max}_x &= 9
\end{align*}
\]

a i range = \(\text{Max}_x - \text{Min}_x = 9 - 2 = 7\)

b i IQR = \(Q_3 - Q_1 = 7 - 4 = 3\)

c i range = \(\text{Max}_x - \text{Min}_x = 9 - 2 = 7\)

ii IQR = \(Q_3 - Q_1 = 7 - 4 = 3\)

d Peta throws 7 goals or less 75% of the time.
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’:

- **The upper boundary** = upper quartile + 1.5 × IQR.
  Any data larger than the upper boundary is an outlier.
- **The lower boundary** = lower quartile − 1.5 × IQR.
  Any data smaller than the lower boundary is an outlier.

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.

**Example 13**

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, 6, 8, 9, 10, 13, 7

The ordered data set is:

\[
\begin{aligned}
&1 3 3 5 6 7 7 7 8 8 8 9 9 10 10 12 13 14 16 \\
&\text{Min}_x = 1 \quad \text{Q}_1 = 6.5 \quad \text{median} = 8 \quad Q_3 = 10 \quad \text{Max}_x = 16 \\
\end{aligned}
\]

IQR = Q₃ − Q₁ = 3.5

**Test for outliers:**

- upper boundary = upper quartile + 1.5 × IQR
  = 10 + 1.5 × 3.5
  = 15.25

- lower boundary = lower quartile − 1.5 × IQR
  = 6.5 − 1.5 × 3.5
  = 1.25

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:

```
0 2 4 6 8 10 12 14 16 18 20
```

variable

Notice that the whisker is drawn 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.
   a What was the highest mark scored?
   b What was the lowest mark scored?
   c What was the range of marks scored for this test?
   d What percentage of students scored 35 or more for the test?
   e What was the interquartile range for this test?
   f The top 25% of students scored a mark between and ....
   g If you scored 63 for this test, would you be in the top 50% of students in this class?
   h 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.
   a Calculate the interquartile range for this data set.
   b Calculate the boundaries that identify outliers.
   c Which of the data 20, 11, 52, 67 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 are:
   4, 7, 9, 3, 1, 11, 5, 4, 6, 6, 4, 4, 12, 8, 2, 3, 3, 6, 7, 8, 4, 4, 3, 2, 5, 5, 5, 8, 7, 6, 5
   a Find the median, lower quartile and upper quartile of the data set.
   b Find the interquartile range of the data set.
   c What are the lower and upper boundaries for outliers?
   d Are there any outliers according to c?
   e 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>

   a Find the five-number summary for this data set.
   b Find the range and IQR for the data set.
   c Are there any outliers? Test for them.
   d 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:

\[
s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}
\]

Note: In this formula:
- \( x_i \) is a data value of the sample and \( \bar{x} \) is the sample mean.
- \((x_i - \bar{x})^2\) measures how far \( x_i \) is from the mean \( \bar{x} \). The square ensures that all of the quantities are positive.
- If the sum of all of the \((x_i - \bar{x})^2\) 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\(^2\).

On a TI-83, the sample standard deviation is obtained in 1-Var Stats as \( S_X \).
On a Casio the sample standard deviation is obtained in 1-Var Stats as \( x\sigma_{n-1} \).

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

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

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 14**

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>4</th>
<th>16</th>
<th>14</th>
<th>8</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshfruit</td>
<td>9</td>
<td>12</td>
<td>11</td>
<td>10</td>
<td>13</td>
<td></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>96</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>10</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.

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 14 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.

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.

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

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

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

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

We do not need to use this formula as it is easier to use technology.
12 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</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.

14 The lengths of 30 trout are displayed in the frequency table. 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>

I 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.
Interpreting the results

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 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 the 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

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.

**Example 15**

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

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.

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

*Leaf unit: 0.1 units of cholesterol*

---

c

<table>
<thead>
<tr>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

---
d 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.

e 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 5I

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 for a pack of 5. To test which of the two varieties is more economical to buy in terms of weight, 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
12.1 12.1 12.1 12.1 12.1 12.2 12.2 12.3 12.3

Fancy frogs:
15.2 15.2 15.1 14.8 14.7 14.7 14.6 14.6 14.6
14.6 14.6 14.6 14.5 14.5 14.5 14.5 14.5 14.4

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?
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.

<table>
<thead>
<tr>
<th>Survival time of rats that were given the drug</th>
<th>Survival time of rats that were not given the drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>64    78    106   106   106   127  127  134  148  186</td>
<td></td>
</tr>
<tr>
<td>192*  192*  192*  192*  192*  64   78   106  106</td>
<td></td>
</tr>
<tr>
<td>37    38    42    43    43    43   43   48   49</td>
<td></td>
</tr>
<tr>
<td>51    51    55    59    62    66   69   86   37</td>
<td></td>
</tr>
</tbody>
</table>

* 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.

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.

<table>
<thead>
<tr>
<th>Data from plot A</th>
<th>Data from plot B</th>
<th>Data from plot C</th>
</tr>
</thead>
<tbody>
<tr>
<td>27    29    9    10    8    36    36    42    32    32    32    30    38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39    38    50    34    41    39    40    12    14    35    35    42    25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32    30    34    22    47    58    56    63    66    54    48    48    53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47    58    56    63    66    54    48    48    53    47    29    46    33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45    58    34    55    76    65    61    67    69    68    64    76    59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59    79    70    69    70    43    70    62    60    58    79    65    75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60    39    68    63    54    61    72    58    77    66    65    47    50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 14 15 12 14 14 14 14
15 13 14 14 16 16 14 13 11 12 12 15 15 13
17 17 19 14 9 15 12 12 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 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?

a 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.
b Technology was used to construct the two column graphs.
c 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>

d
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.

**Before**
- 5.84 5.46 5.55 5.60 5.67 5.89 6.53 6.70 6.96 7.27
- 7.4 7.41 7.58 7.70 8.17 8.68 8.93 9.66 10.0 10.01

**After**
- 5.34 5.38 5.51 5.62 5.71 5.8 5.84 6.71 6.97 7.32
- 7.36 7.65 7.81 8.06 8.06 8.42 8.63 8.91 9.1 10.92

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

Consider the two sample sets:

**Sample 1:**
- 9.31 7.91 10.69 11.99 10.49 9.23 12.87 10.33 11.85 11.21

**Sample 2:**
- 11.19 12.76 12.8 10.24 10.01 9.19 12.47 10.76 12.7 10.43
- 10.95 7.89 11.6 11.99 10.27 8.74 9.69 12.71 6.83 10.46

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.
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>weekend</td>
<td>tennis, basketball</td>
</tr>
<tr>
<td>evening</td>
<td>tennis</td>
</tr>
<tr>
<td>daytime</td>
<td>tennis, basketball</td>
</tr>
<tr>
<td>evening</td>
<td>basketball</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Tennis</th>
<th>Basketball</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weekends</td>
<td>19</td>
<td>78</td>
</tr>
<tr>
<td>Day time</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>Evening</td>
<td>17</td>
<td>53</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.
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 day 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.
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 17

Write a report for analysis carried out in Example 15.

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>2</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 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>2 1 1 1 0 6</td>
<td>2</td>
</tr>
<tr>
<td>8 7 6 6 5 5 6*</td>
<td>4 3 0 7</td>
</tr>
<tr>
<td>9 9 6 6 5 7*</td>
<td>4 4 3 2 8</td>
</tr>
<tr>
<td>8</td>
<td>8*</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, Exercise 51.
   B The new cancer drug in question 3, Exercise 51.
   C Your analysis of the diet in question 5, Exercise 51.
The object of selecting a sample is to gather from it statistics that reflect the characteristics of the population. For example, in order to estimate a population mean we usually calculate and use a sample mean.

It is unlikely that the sample statistic will be exactly the same as the population parameter. There are many reasons for that difference, including:

- **Statistical errors** (or **random errors**): These are caused by natural variability of any random process and are unavoidable in statistical analysis.

- **Bias** or **systematic errors**: These are caused by faults in the sampling process and great care and expense is usually taken to avoid such problems.

- **Errors in measurement and recording data**.

The two investigations below give you an opportunity to consider the impact of the size and quality of a random sample.

### 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 1</th>
<th>Block 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.23</td>
<td>3.62</td>
</tr>
<tr>
<td>3.10</td>
<td>3.03</td>
</tr>
<tr>
<td>3.06</td>
<td>3.58</td>
</tr>
<tr>
<td>3.14</td>
<td>3.70</td>
</tr>
<tr>
<td>3.33</td>
<td>2.92</td>
</tr>
<tr>
<td>3.04</td>
<td>3.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Block 3</th>
<th>Block 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.48</td>
<td>3.19</td>
</tr>
<tr>
<td>3.40</td>
<td>3.26</td>
</tr>
<tr>
<td>3.08</td>
<td>3.46</td>
</tr>
<tr>
<td>2.94</td>
<td>3.23</td>
</tr>
<tr>
<td>3.38</td>
<td>3.41</td>
</tr>
<tr>
<td>3.33</td>
<td>3.42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Block 5</th>
<th>Block 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.24</td>
<td>3.18</td>
</tr>
<tr>
<td>3.33</td>
<td>3.25</td>
</tr>
<tr>
<td>3.02</td>
<td>3.12</td>
</tr>
<tr>
<td>3.10</td>
<td>3.24</td>
</tr>
<tr>
<td>3.25</td>
<td>3.09</td>
</tr>
<tr>
<td>3.57</td>
<td>3.61</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**

In 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 20713 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 20,713. 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 9500 )</td>
<td>GRN</td>
</tr>
<tr>
<td>( 9500 &lt; n \leq 17000 )</td>
<td>ALP</td>
</tr>
<tr>
<td>( 17000 &lt; n \leq 18000 )</td>
<td>FFP</td>
</tr>
<tr>
<td>( n &gt; 18000 )</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.

DATA BASED INVESTIGATION

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:
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.
- The analysis you carried out. This includes a copy of your data, any graphics and summary statistics you produced and the argument that you wrote to support your conclusion.
- 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.

### RELATIVE FREQUENCY

The given frequency table shows the results of 30 students on a test marked out of 10.

To this table we can add a relative frequency column.

The relative frequency of an item is its frequency as a fraction of the total number of items in the sample, i.e., \( \text{relative frequency} = \frac{\text{item frequency}}{\text{total frequency}} \)

So, the relative frequency for a score of 4 is \( \frac{2}{30} = 0.067 \)

We now have:

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
<th>Relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
<td>0.067</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.167</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>0.133</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>0.200</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>0.167</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>0.200</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.067</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>1</td>
</tr>
</tbody>
</table>

From a relative frequency we can easily convert to a percentage by multiplying by 100%.

For example, \( 0.067 = 0.067 \times 100\% = 6.7\% \)

we can say that 6.7% of the results were a ‘4’.

Notice also, that the probability of a randomly selected student from the group getting a ‘4’ is 0.067 or 6.7%.
The following data is of the number of mint drops in a packet.

28 32 29 32 33 29 32 27 28 27 30 26 31 27
28 32 33 28 29 31 32 28 31 30 29 30 27 32 29
32 31 29 32 31 27 28 29 27 31

a Tabulate the data including columns of number, tally, frequency and relative frequency.

b What percentage of the packets contained:
   i 30 mints   ii at least 30 mints?

c If a packet is selected from this group, what is the probability that it contains:
   i 31 mints   ii at most 28 mints?

d In a batch of 3600 packets, estimate the number containing 30 mints.

Comment on the result.

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Number} & \text{Tally} & \text{Frequency} & \text{Relative frequency} \\
\hline
26 & | & 1 & 0.025 \\
27 & || | & 6 & 0.150 \\
28 & | | | & 6 & 0.150 \\
29 & | | | | & 7 & 0.175 \\
30 & | | | | | & 3 & 0.075 \\
31 & | | | | | | & 7 & 0.175 \\
32 & | | | | | | | & 8 & 0.200 \\
33 & | | | | | | | | & 2 & 0.050 \\
\hline
\text{Total} & & 40 & 1.000 \\
\hline
\end{array}
\]

b i Relative frequency for 30 mints = 0.075
\[\therefore \text{percentage of packets containing 30 mints} = 7.5\%\]

ii Relative frequency for at least 30 mints = 0.075 + 0.175 + 0.200 + 0.050
\[\therefore \text{percentage of packets containing at least 30 mints} = 50\%\]

c i \[\Pr(31 \text{ mints}) = 0.175\]
ii \[\Pr(\text{at most 28 mints}) = 0.025 + 0.150 + 0.150 = 0.325\]

d Estimated number containing 30 mints = 0.075 \times 3600
\[= 270 \text{ packets.}\]

Comment: This estimate could be unreliable due to the small sample from which the estimate is made.
EXERCISE 5L

1 Allheat produces heating elements expected to last 100 hours. To test the life of the elements 50 of them were selected at random and tested until they failed. The life times were:

106 97 94 107 105 93 104 99 110 102 108 104 96
104 100 98 109 106 101 99 101 104 107 106 95 94
100 102 103 94 93 101 105 107 96 100 99 97 97
95 98 102 105 107 105 94 93 105 102 101

- a Tabulate the data and include a relative frequency column.
- b Draw a relative frequency histogram of the data.
- c What percentage of the results were:
  - i more than 100 hours
  - ii between 98 hours and 104 hours?
- d Determine the probability that a randomly selected element lasts for:
  - i no more than 99 hours
  - ii more than 103 hours
- e For a batch of 1500 elements, estimate the number lasting more than 100 hours.

2 Footballers were invited to participate in a long kicking competition.

The results were:

<table>
<thead>
<tr>
<th>Distance (m)</th>
<th>20 - &lt; 30</th>
<th>30 - &lt; 40</th>
<th>40 - &lt; 50</th>
<th>50 - &lt; 60</th>
<th>60 - &lt; 70</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of footballers</td>
<td>3</td>
<td>26</td>
<td>31</td>
<td>16</td>
<td>4</td>
</tr>
</tbody>
</table>

- a Tabulate the data and include a relative frequency column.
- b Draw a relative frequency histogram of the data.
- c How many footballers took part in the competition?
- d How many kicked:
  - i more than 40 m
  - ii at most 50 m?
- e What percentage kicked between 40 m and 60 m?
- f What is the probability that a randomly selected footballer kicked the ball:
  - i less than 40 m
  - ii between 30 m and 50 m?

3 A horticulturalist took random samples of two week old seedlings and measured their height to the nearest millimetre. The results were:

<table>
<thead>
<tr>
<th>Height (mm)</th>
<th>300-&lt;325</th>
<th>325-&lt;350</th>
<th>350-&lt;375</th>
<th>375-&lt;400</th>
<th>400-&lt;425</th>
<th>425-&lt;450</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>12</td>
<td>20</td>
<td>45</td>
<td>28</td>
<td>14</td>
<td>6</td>
</tr>
</tbody>
</table>

- a Construct a table of columns for this data and include one for relative frequency.
- b How many seedlings were in the sample?
- c Draw a relative frequency histogram for the data.
- d How many seedlings were less than 375 mm tall?
- e What percentage of the seedlings were between 350 mm and 425 mm tall?
- f Estimate the probability that a randomly chosen seedling was:
  - i between 350 and 400 mm high
  - ii less than 400 mm high
- g In a sample of 2000 seedlings, estimate the number that are less than 375 mm tall.
In Section L, we observed the shape of some typical relative frequency histograms. Some of these histograms seem to be symmetric and bell-shaped. These are called normal distributions. Many naturally occurring phenomena have distributions which are normal, or approximately normal. For example, the maximum shell width of Goolwa cockles is normally distributed. Click on the icon to see how a distribution is built up. So, a typical sample from a normal distribution looks like the one given below, although not perfectly symmetrical like this one.

HOW THE NORMAL DISTRIBUTION ARISES

Consider a typical variable like ‘the birth weight of babies’. The birth weight is the combined effect of a number of random factors. These could include:

- genetic make up of both parents
- mother’s food intake and type of food eaten
- mother’s exercise during pregnancy
- environmental factors

Another example could be:

Consider the apples stripped from an apple tree. They do not all have the same weight. This variation may be due to several factors which could include:

- different genetic factors
- different times when the flowers were fertilised
- different amounts of sunlight reaching the leaves and fruit
- different weather conditions (some may be affected by the prevailing winds more than others), etc.

The result is that much of the fruit could have weights centred about, for example, a mean weight of 214 grams, and there are far fewer apples that are much heavier or lighter. Invariably, a bell-shaped distribution of weights would be observed and the normal distribution model fits the data fairly closely.
THE NORMAL DISTRIBUTION CURVE

Below we have the graph of the normal distribution of scores. Notice its symmetry.

The normal distribution is a theoretical, or idealised model of many real life distributions. In a normal distribution, data is equally distributed about the mean. The mean also coincides with the median of the data.

The normal distribution lies at the heart of statistics. Many naturally occurring phenomena have a distribution that is normal, or approximately normal.

Some examples are:
- the chest sizes of Australian males
- the distribution of errors in many manufacturing processes
- the lengths of adult female tiger sharks
- the length of cilia on a cell
- scores on tests taken by a large population
- repeated measurements of the same quantity
- yields of corn, wheat, etc.

EXERCISE 5M

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.

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 the heights of 16 year old girls
   d the time it takes to complete a motor car on a production line.
INVESTIGATION 6

THE NORMAL CURVES PROPERTIES

What to do:

1. Find for \( n = 300 \), the sample’s mean (\( \bar{x} \)), median and standard deviation (\( s \)).
2. Find the proportion of the sample values which lie in the intervals \( \bar{x} \pm s \), \( \bar{x} \pm 2s \), \( \bar{x} \pm 3s \).
3. Select another random sample for \( n = 200 \) and repeat 2.
4. Repeat again, each time recording your results.
5. Increase \( n \) to 1000 and obtain more data for proportions in the intervals described in 2. Repeat several times.
6. Write a brief report of your findings.

From Investigation 6 you should have discovered that changing the number and values of data in a sample 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 and 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.

**Example 19**

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?

First we draw a normal curve for $\mu = 40, \sigma = 5$

**EXERCISE 5N**

1. The chest measurements of 17 year old footballers is normally distributed with mean 88 cm and standard deviation 7 cm.

Find the percentage of footballers with chest measurements:

- $a$ between 81 cm and 95 cm
- $b$ between 88 cm and 102 cm
- $c$ between 74 cm and 95 cm
- $d$ between 67 cm and 109 cm
- $e$ more than 81 cm
- $f$ less than 102 cm
2 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.

   ![Bell-shaped curve](image)

   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?

3 A company sells toasters with a mean life of 5 years and a standard deviation of 1 year. The company will replace a toaster if it is faulty within 2 years of sale. If they sell 5000 toasters, how many can they expect to replace if life expectancy is normally distributed?

4 A restauranteur found that the average time spent by diners was 2 hours, with a standard deviation of 30 minutes. Assuming that the time spent by diners is normally distributed, and that there are 200 diners each week, calculate:
   a the number of diners who stay between 2 and 3 hours
   b the number who stay longer than 3 hours
   c the number who stay less than 1\frac{1}{2} hours.

5 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.
6. 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. Assuming the error of the clocks is normally distributed, 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.

7. 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.

8. 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 will 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.

---

**TECHNOLOGY AND NORMAL DISTRIBUTIONS**

In Section N questions were based around the standard 68% : 95% : 99.7% proportions. We can find other probabilities for the normal distribution using a graphics calculator.

Suppose \( X \) is normally distributed with mean 10 and standard deviation 2.

How do we find \( \Pr(8 \leq X \leq 11) \)?
The length of salmon caught in South Australian waters is normally distributed with mean 41.4 cm and standard deviation 4.2 cm. What percentage of a large catch would be expected to have a length of:

a  between 38 and 44 cm   
b  more than 43 cm   
c  less than 35 cm?

Let $X$ cm = length of a salmon. Using a TI-83:

a  $Pr(38 < X < 44)$
  $= \text{normalcdf}(38, 44, 41.4, 4.2)$
  $= 0.523$
  $= 52.3\%$

b  $Pr(X > 43)$
  $= \text{normalcdf}(43, 99, 41.4, 4.2)$
  $= 0.352$
  $= 35.2\%$

c  $Pr(X < 35)$
  $= \text{normalcdf}(-99, 35, 41.4, 4.2)$
  $= 0.0638$
  $= 6.38\%$
REVIEW SET 5A

EXERCISE 5O

1 The weight of new born alpacas is normally distributed with mean 7.2 kg and a standard deviation of 0.34 kg. Determine the percentage of new born alpacas with a weight:
   a between 6.5 kg and 7.5 kg  b more than 7 kg  c less than 7.7 kg

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, how many would we expect to be less than 30 cm in length?

5 The army recruiting department knows that the height 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?

REVIEW SET 5A

1 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 i median ii range of the data.
   c Comment on the skewness of the data.

2 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 randomly choose your sample?
   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?
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>7 6 3</td>
<td>34</td>
</tr>
<tr>
<td>8 7 4 3 0</td>
<td>8 8 3 3</td>
</tr>
<tr>
<td>8 7 6 6</td>
<td>8 7 8</td>
</tr>
<tr>
<td>6</td>
<td>38</td>
</tr>
<tr>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>1</td>
<td>40</td>
</tr>
</tbody>
</table>

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

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

e Determine: i the mean ii the median.

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

71.2 65.1 68.0 71.1 74.6 68.8 83.2 85.0 74.5 87.4
84.3 77.0 82.8 84.4 80.6 75.9 89.7 83.2 97.5 82.9
90.5 85.5 90.7 92.9 95.6 85.5 64.6 73.9 80.0 86.5

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.

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 20</td>
</tr>
<tr>
<td>&gt; 25</td>
<td>B 45</td>
</tr>
<tr>
<td>&gt; 25</td>
<td>C 195</td>
</tr>
<tr>
<td>25 - 55</td>
<td>A 33</td>
</tr>
<tr>
<td>&gt; 55</td>
<td>B 90</td>
</tr>
<tr>
<td>&gt; 55</td>
<td>C 77</td>
</tr>
<tr>
<td>&gt; 55</td>
<td>A 182</td>
</tr>
<tr>
<td>&gt; 55</td>
<td>B 42</td>
</tr>
<tr>
<td>&gt; 55</td>
<td>C 116</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?
e Imagine that you are the researcher. Write a report to the person who contracted you to carry out the research.

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

a Calculate the proportion of all students who said they owned a mobile phone.
b What proportion of the first 5 and 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, 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?

6 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: 146 131 132 160 128 119 133 117 139 123 109 129 109 131
            191 117 132 107 141 136 146 142 123 144 145 125 164 125
             133 124 153 129 118 130 134 131 145 131 133 135
   a Determine the 5-number summary and interquartile range for each of the data sets.
   b Produce side-by-side boxplots.
   c Discuss the manufacturer’s claim.
REVIEW SET 5C

1 Briefly state which sampling technique you would use to select a random sample for each of the following.
   a 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.
   b South Australia has 11 electorates. A pollster wants to predict how South Australians will vote in the next election.
   c 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, 100, 105, 81, 104, 98, 107, 95, 104, 108, 99, 101, 106, 102, 98, 101
   a Find the median ii lower quartile iii upper quartile
   b Find the interquartile range of the data set.
   c 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 ($ per hour)
   17.27 11.31 15.72 18.92 9.55 12.98 19.12 18.26 22.79
   16.69 11.68 15.84 12.81 24.03 15.03 12.95 12.25
   20.09 18.64 18.94 13.92 11.69 15.52 15.21 18.59

   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 a parallel boxplot 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 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?

6 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>901</td>
</tr>
<tr>
<td>889</td>
<td>898</td>
</tr>
<tr>
<td>885</td>
<td>901</td>
</tr>
<tr>
<td>894</td>
<td>900</td>
</tr>
<tr>
<td>904</td>
<td>907</td>
</tr>
<tr>
<td>897</td>
<td>901</td>
</tr>
<tr>
<td>899</td>
<td>900</td>
</tr>
<tr>
<td>908</td>
<td>895</td>
</tr>
<tr>
<td>901</td>
<td>903</td>
</tr>
<tr>
<td>898</td>
<td>896</td>
</tr>
<tr>
<td>904</td>
<td>903</td>
</tr>
<tr>
<td>894</td>
<td>901</td>
</tr>
<tr>
<td>893</td>
<td>901</td>
</tr>
<tr>
<td>895</td>
<td>904</td>
</tr>
<tr>
<td>909</td>
<td>895</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.

REVIEW SET 5D

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 May is normally distributed with mean 20.7°C and standard deviation of 3°C. How many days in May can be expected to have temperatures between 17.7°C and 26.7°C?

3 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.

4 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?
5 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.

6 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>

7 The length of a shell is normally distributed with mean 24 mm and standard deviation 3.2 mm.

a What percentage of shells have width:
   i between 20 mm and 30 mm
   ii less than 25 mm
   iii more than 18 mm?

b If 20 000 shells were selected at random, how many (to the nearest 100) would you expect to have a width between 20 mm and 30 mm?

REVIEW SET 5E

1 The diameter of apples picked from an orchard is normally distributed with mean 70 mm and standard deviation 4.1 mm.

a What proportion of the apples have diameter:
   i less than 67 mm
   ii between 72 mm and 77 mm?

b Only apples with diameter of at least 63 mm are considered suitable for sale. From a batch of 500 apples, how many will be suitable for sale?

2 A class of 32 students were asked how many cousins they have, and the results were:

<table>
<thead>
<tr>
<th>3</th>
<th>6</th>
<th>4</th>
<th>7</th>
<th>5</th>
<th>5</th>
<th>3</th>
<th>2</th>
<th>6</th>
<th>6</th>
<th>7</th>
<th>3</th>
<th>4</th>
<th>4</th>
<th>6</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

a Tabulate the data and include a relative frequency column.

b Draw a relative frequency histogram of the data.

c What percentage of the class have:
   i less than 5 cousins
   ii at least 6 cousins?

d What is the probability that a randomly selected student from the class will have:
   i between 3 and 6 cousins
   ii at most 5 cousins?
3 A survey of workers in an office building shows that the average time spent at work per day by the workers is normally distributed with mean 8.5 hours and standard deviation 24 minutes.
   a What percentage of the workers spend less than 8 hours per day at work?
   b If the building contains 300 workers, how many of them work between 8 and 9 hours per day?

4 The table alongside shows the favourite weekend activities for a class of students.
   a How many students are in the class?
   b What percentage of students prefer to:
      i go to the movies  
      ii play sport?
   c What is the mode of the data?
   d Draw a pie chart of the data.

4 The table alongside shows the favourite weekend activities for a class of students.
   a How many students are in the class?
   b What percentage of students prefer to:
      i go to the movies  
      ii play sport?
   c What is the mode of the data?
   d Draw a pie chart of the data.

5 Over a period of 80 days, a restaurant owner kept a record of the number of customers at his restaurant each day. The results were:

<table>
<thead>
<tr>
<th>No. of customers</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>80 - &lt; 100</td>
<td>10</td>
</tr>
<tr>
<td>100 - &lt; 120</td>
<td>14</td>
</tr>
<tr>
<td>120 - &lt; 140</td>
<td>22</td>
</tr>
<tr>
<td>140 - &lt; 160</td>
<td>16</td>
</tr>
<tr>
<td>160 - &lt; 180</td>
<td>18</td>
</tr>
</tbody>
</table>

   a Construct a table of columns for this data, including a relative frequency column.
   b Draw a relative frequency histogram of the data.
   c On what percentage of the days were there at least 140 customers?
   d Estimate the probability that, on a randomly chosen day, the restaurant has:
      i between 120 and 160 customers  
      ii at most 100 customers.
   e If the restaurant is open for 360 days per year, how many days in a year would you expect the restaurant to have at least 120 customers?

6 Bottles of spring water are marked to contain 600 mL of water. Analysis of these bottles reveals that the volume of water in the bottles is normally distributed with mean 604 mL and standard deviation 3 mL.
   a What proportion of bottles contain at least 600 mL of water?
   b The bottles are only acceptable if they contain between 597 mL and 610 mL of water. In a batch of 800 bottles, how many will be rejected?