1

Univariate data

1.1 Kick off with CAS
1.2 Types of data
1.3 Stem plots
1.4 Dot plots, frequency tables and histograms, and bar charts
1.5 Describing the shape of stem plots and histograms
1.6 The median, the interquartile range, the range and the mode
1.7 Boxplots
1.8 The mean of a sample
1.9 Standard deviation of a sample
1.10 Populations and simple random samples
1.11 The 68–95–99.7% rule and z-scores
1.12 Review eBookplus
1.1 Kick off with CAS

Exploring histograms with CAS

Histograms can be used to display numerical data sets. They show the shape and distribution of a data set, and can be used to gather information about the data set, such as the range of the data set and the value of the median (middle) data point.

1 The following data set details the age of all of the guests at a wedding.

<table>
<thead>
<tr>
<th>Age</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–</td>
<td>4</td>
</tr>
<tr>
<td>10–</td>
<td>6</td>
</tr>
<tr>
<td>20–</td>
<td>18</td>
</tr>
<tr>
<td>30–</td>
<td>27</td>
</tr>
<tr>
<td>40–</td>
<td>14</td>
</tr>
<tr>
<td>50–</td>
<td>12</td>
</tr>
<tr>
<td>60–</td>
<td>6</td>
</tr>
<tr>
<td>70–</td>
<td>3</td>
</tr>
<tr>
<td>80–</td>
<td>2</td>
</tr>
</tbody>
</table>

Use CAS to draw a histogram displaying this data set.

2 Comment on the shape of the histogram. What does this tell you about the distribution of this data set?

3 a Use CAS to display the mean of the data set on your histogram.
   b What is the value of the mean?

4 a Use CAS to display the median of the data set on your histogram.
   b What is the value of the median?

Please refer to the Resources tab in the Prelims section of your eBookPLUS for a comprehensive step-by-step guide on how to use your CAS technology.
1.2 Types of data

Univariate data are data that contain one variable. That is, the information deals with only one quantity that changes. Therefore, the number of cars sold by a car salesman during one week is an example of univariate data. Sets of data that contain two variables are called bivariate data and those that contain more than two variables are called multivariate data.

Data can be classified as either numerical or categorical. The methods we use to display data depend on the type of information we are dealing with.

Types of data

Numerical data is data that has been assigned a numeric value. Numerical data can be:

- discrete — data that can be counted but that can have only a particular value, for example the number of pieces of fruit in a bowl
- continuous — data that is not restricted to any particular value, for example the temperature outside, which is measured on a continuous scale.

Categorical data is data split into two or more categories. Categorical data can be:

- nominal — data that can be arranged into categories but not ordered, for example arranging shoes by colour or athletes by gender
- ordinal — data that can be arranged into categories that have an order, for example levels of education from high school to post-graduate degrees.

Discrete and continuous data

Data are said to be discrete when a variable can take only certain fixed values. For example, if we counted the number of children per household in a particular suburb, the data obtained would always be whole numbers starting from zero. A value in between, such as 2.5, would clearly not be possible. If objects can be counted, then the data are discrete.

Continuous data are obtained when a variable takes any value between two values. If the heights of students in a school were obtained, then the data could consist of any values between the smallest and largest heights. The values recorded would be restricted only by the precision of the measuring instrument. If variables can be measured, then the data are continuous.

WORKED EXAMPLE 1

Which of the following is not numerical data?

A Maths test results
B Ages
C AFL football teams
D Heights of students in a class
E Lengths of bacteria
THINK
1 To be numerical data, it is to be measurable or countable.

2 Look for the data that does not fit the criteria.

3 Answer the question.

WRITE

A: Measurable
B: Countable
C: Names, so not measurable or countable
D: Measurable
E: Measurable

3 Answer the question. The data that is not numerical is AFL football teams. The correct option is C.

WORKED EXAMPLE 1.2

Types of data

THINK
1 To be discrete data, it is to be a whole number (countable).

2 Look for the data that does not fit the criteria.

3 Answer the question.

WRITE

A: Whole number (countable)
B: Whole number (countable)
C: Whole number (countable)
D: Whole number (countable)
E: May not be a whole number (measurable)

3 Answer the question. The data that is not discrete is the height of the tallest student in a class. The correct option is E.

EXERCISE 1.2

Which of the following is not numerical data?

A Number of students in a class
B The number of supporters at an AFL match
C The amount of rainfall in a day
D Finishing positions in the Melbourne Cup
E The number of coconuts on a palm tree

Which of the following is not categorical data?

A Preferred political party
B Gender
C Hair colour
D Salaries
E Religion
3 Which of the following is not discrete data?
A Number of players in a netball team
B Number of goals scored in a football match
C The average temperature in March
D The number of Melbourne Storm members
E The number of twins in Year 12

4 Which of the following is not continuous data?
A The weight of a person
B The number of shots missed in a basketball game
C The height of a sunflower in a garden
D The length of a cricket pitch
E The time taken to run 100 m

5 Write whether each of the following represents numerical or categorical data.
   a The heights, in centimetres, of a group of children
   b The diameters, in millimetres, of a collection of ball bearings
   c The numbers of visitors to an exhibit each day
   d The modes of transport that students in Year 12 take to school
   e The 10 most-watched television programs in a week
   f The occupations of a group of 30-year-olds

6 Which of the following represent categorical data?
   a The numbers of subjects offered to VCE students at various schools
   b Life expectancies
   c Species of fish
   d Blood groups
   e Years of birth
   f Countries of birth
   g Tax brackets

7 For each set of numerical data identified in question 5 above, state whether the data are discrete or continuous.

8 An example of a numerical variable is:
A attitude to 4-yearly elections (for or against)
B year level of students
C the total attendance at Carlton football matches
D position in a queue at the pie stall
E television channel numbers shown on a dial

9 The weight of each truck-load of woodchips delivered to the wharf during a one-month period was recorded. This is an example of:
A categorical and discrete data
B discrete data
C continuous and numerical data
D continuous and categorical data
E numerical and discrete data

10 When reading the menu at the local Chinese restaurant, you notice that the dishes are divided into sections. The sections are labelled chicken, beef, duck, vegetarian and seafood. What type of data is this?

11 NASA collects data on the distance to other stars in the universe. The distance is measured in light years. What type of data is being collected?
A Discrete
B Continuous
C Nominal
D Ordinal
E Bivariate
12 The number of blue, red, yellow and purple flowers in an award winning display is counted. What type of data is being collected?

A Nominal
B Ordinal
C Discrete
D Continuous
E Bivariate

13 Students in a performing arts class watch a piece of modern dance and are then asked to rate the quality of the dance as poor, average, above average or excellent. What type of data is being collected?

14 Given the set of data: 12, 6, 21, 15, 8, 2, describe what type of numerical data this data set is.

15 If a tennis tournament seeds the players to organise the draw, what type of categorical data is this?

16 The height of the players in the basketball team is measured. The data collected is best described as:

A categorical and nominal
B categorical and numerical
C numerical and discrete
D numerical and continuous
E none of the above

1.3 Stem plots

A stem-and-leaf plot, or stem plot for short, is a way of ordering and displaying a set of data, with the advantage that all of the raw data is kept. Since all the individual values of the data are being listed, it is only suitable for smaller data sets (up to about 50 observations).

The following stem plot shows the ages of people attending an advanced computer class.

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

Key: 2|2 = 22 years old

The ages of the members of the class are 16, 22, 22, 23, 30, 32, 34, 36, 42, 43, 46, 47, 53, 57 and 61.

A stem plot is constructed by splitting the numerals of a record into two parts — the stem, which in this case is the first digit, and the leaf, which is always the last digit.

With your stem-and-leaf plot it is important to include a key so it is clear what the data values represent.
In cases where there are numerous leaves attached to one stem (meaning that the data is heavily concentrated in one area), the stem can be subdivided. Stems are commonly subdivided into halves or fifths. By splitting the stems, we get a clearer picture about the data variation.

**WORKED EXAMPLE 3**

The number of cars sold in a week at a large car dealership over a 20-week period is given as shown.

16 12 8 7 26 32 15 51 29 45
19 11 6 15 32 18 43 31 23 23

Construct a stem plot to display the number of cars sold in a week at the dealership.

**THINK**

1. In this example the observations are one- or two-digit numbers and so the stems will be the digits referring to the ‘tens’, and the leaves will be the digits referring to the units.

   Work out the lowest and highest numbers in the data in order to determine what the stems will be.

2. Before we construct an ordered stem plot, construct an unordered stem plot by listing the leaf digits in the order they appear in the data.

3. Now rearrange the leaf digits in numerical order to create an ordered stem plot. Include a key so that the data can be understood by anyone viewing the stem plot.

**WRITE**

Lowest number = 6
Highest number = 51
Use stems from 0 to 5.

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

Key: 2|3 = 23 cars

**WORKED EXAMPLE 4**

The masses (in kilograms) of the members of an Under-17 football squad are given as shown.

70.3 65.1 72.9 66.9 68.6 69.6 70.8
72.4 74.1 75.3 75.6 69.7 66.2 71.2
68.3 69.7 71.3 68.3 70.5 72.4 71.8

Display the data in a stem plot.
THINK

1 In this case the observations contain 3 digits. The last digit always becomes the leaf and so in this case the digit referring to the tenths becomes the leaf and the two preceding digits become the stem.

Work out the lowest and highest numbers in the data in order to determine what the stems will be.

2 Construct an unordered stem plot. Note that the decimal points are omitted since we are aiming to present a quick visual summary of data.

3 Construct an ordered stem plot. Provide a key.

WRITE

Lowest number = 65.1
Highest number = 75.6
Use stems from 65 to 75.

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>1</td>
</tr>
<tr>
<td>66</td>
<td>9 2</td>
</tr>
<tr>
<td>67</td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>6 3 3</td>
</tr>
<tr>
<td>69</td>
<td>6 7 7</td>
</tr>
<tr>
<td>70</td>
<td>3 8 5</td>
</tr>
<tr>
<td>71</td>
<td>2 3 8</td>
</tr>
<tr>
<td>72</td>
<td>9 4 4</td>
</tr>
<tr>
<td>73</td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>1</td>
</tr>
<tr>
<td>75</td>
<td>3 6</td>
</tr>
</tbody>
</table>

Key: 74|1 = 74.1 kg

A set of golf scores for a group of professional golfers trialling a new 18-hole golf course is shown on the following stem plot.

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

Key: 6|1 = 61

Produce another stem plot for these data by splitting the stems into:

a halves
b fifths.
THINK

a. By splitting the stem 6 into halves, any leaf digits in the range 0–4 appear next to the 6, and any leaf digits in the range 5–9 appear next to the 6*. Likewise for the stem 7.

b. Alternatively, to split the stems into fifths, each stem would appear 5 times. Any 0s or 1s are recorded next to the first 6. Any 2s or 3s are recorded next to the second 6. Any 4s or 5s are recorded next to the third 6. Any 6s or 7s are recorded next to the fourth 6 and, finally, any 8s or 9s are recorded next to the fifth 6. This process would be repeated for those observations with a stem of 7.

WRITE

a. Stem | Leaf
| 6    | 1
| 6*   | 6 6 7 8 9 9 9
| 7    | 0 1 1 2 2 3
| 7*   | 7

Key: 6|1 = 61

b. Stem | Leaf
| 6    | 1
| 6    | 6 6 7
| 6    | 6 8 9 9 9
| 7    | 0 1 1
| 7    | 2 2 3
| 7    | 7
| 7    | 7
| 7    | 7

Key: 6|1 = 61

EXERCISE 1.3 Stem plots

PRACTISE

1. The number of iPads sold in a month from a department store over 16 weeks is shown.

| 28 | 31 | 18 | 48 | 38 | 25 | 21 | 16 |
| 33 | 42 | 35 | 39 | 49 | 30 | 29 | 28 |

Construct a stem plot to display the number of iPads sold over the 16 weeks.

2. The money (correct to the nearest dollar) earned each week by a busker over an 18-week period is shown below. Construct a stem plot for the busker’s weekly earnings. What can you say about the busker’s earnings?

| 5  | 19 | 11 | 27 | 23 | 35 | 18 | 42 | 29 |
| 31 | 52 | 43 | 37 | 41 | 39 | 45 | 32 | 36 |

3. The test scores (as percentages) of a student in a Year 12 Further Maths class are shown.

| 88.0 | 86.8 | 92.1 | 89.8 | 92.6 | 90.4 | 98.3 | 94.3 | 87.7 |
| 94.9 | 98.9 | 92.0 | 90.2 | 97.0 | 90.9 | 98.5 | 92.2 | 90.8 |

Display the data in a stem plot.

4. The heights of members of a squad of basketballers are given below in metres. Construct a stem plot for these data.

| 1.96 | 1.85 | 2.03 | 2.21 | 2.17 | 1.89 | 1.99 | 1.87 |
| 1.95 | 2.03 | 2.09 | 2.05 | 2.01 | 1.96 | 1.97 | 1.91 |
5 LeBron James’s scores for his last 16 games are shown in the following stem plot.

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

Key: 2|0 = 20

Produce another stem plot for these data by splitting the stems into:

a halves
b fifths.

6 The data below give the head circumference (correct to the nearest cm) of 16 four-year-old girls.

<table>
<thead>
<tr>
<th>48</th>
<th>49</th>
<th>47</th>
<th>52</th>
<th>51</th>
<th>50</th>
<th>49</th>
<th>48</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>53</td>
<td>52</td>
<td>43</td>
<td>47</td>
<td>49</td>
<td>50</td>
</tr>
</tbody>
</table>

Construct a stem plot for head circumference, using:

a the stems 4 and 5
b the stems 4 and 5 split into halves
c the stems 4 and 5 split into fifths.

7 For each of the following, write down all the pieces of data shown on the stem plot. The key used for each stem plot is 3|2 = 32.

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

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

C Stem | Leaf
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

8 The ages of those attending an embroidery class are given below. Construct a stem plot for these data and draw a conclusion from it.

<table>
<thead>
<tr>
<th>39</th>
<th>68</th>
<th>51</th>
<th>57</th>
<th>63</th>
<th>51</th>
<th>37</th>
<th>42</th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>49</td>
<td>52</td>
<td>61</td>
<td>58</td>
<td>59</td>
<td>49</td>
<td>53</td>
</tr>
</tbody>
</table>

9 The observations shown on the stem plot at right are:

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 4 10 27 28 29 31 34 36 41</td>
<td>0 4</td>
</tr>
<tr>
<td>B 14 10 27 28 29 31 34 36 41 41</td>
<td>1</td>
</tr>
<tr>
<td>C 4 22 27 28 29 30 31 34 36 41 41</td>
<td>2 2 7 8 9 9</td>
</tr>
<tr>
<td>D 14 22 27 28 29 30 30 31 34 36 41 41</td>
<td>3 0 1 4 6</td>
</tr>
<tr>
<td>E 4 2 27 28 29 30 31 34 36 41</td>
<td>4 1 1</td>
</tr>
</tbody>
</table>

Key: 2|5 = 25
10 The ages of the mothers of a class of children attending an inner-city kindergarten are given below. Construct a stem plot for these data. Based on your display, comment on the statement ‘Parents of kindergarten children are young’ (less than 30 years old).

<table>
<thead>
<tr>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
</tr>
<tr>
<td>28</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>29</td>
</tr>
<tr>
<td>32</td>
</tr>
<tr>
<td>28</td>
</tr>
<tr>
<td>34</td>
</tr>
<tr>
<td>32</td>
</tr>
<tr>
<td>35</td>
</tr>
<tr>
<td>39</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>37</td>
</tr>
<tr>
<td>33</td>
</tr>
<tr>
<td>29</td>
</tr>
<tr>
<td>35</td>
</tr>
<tr>
<td>38</td>
</tr>
<tr>
<td>33</td>
</tr>
</tbody>
</table>

11 The number of hit outs made by each of the principal ruckmen in each of the AFL teams for Round 11 is recorded below. Construct a stem plot to display these data. Which teams had the three highest scoring ruckmen?

<table>
<thead>
<tr>
<th>Team</th>
<th>Number of hit outs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collingwood</td>
<td>19</td>
</tr>
<tr>
<td>Bulldogs</td>
<td>41</td>
</tr>
<tr>
<td>Kangaroos</td>
<td>29</td>
</tr>
<tr>
<td>Port Adelaide</td>
<td>24</td>
</tr>
<tr>
<td>Geelong</td>
<td>21</td>
</tr>
<tr>
<td>Sydney</td>
<td>31</td>
</tr>
<tr>
<td>Melbourne</td>
<td>40</td>
</tr>
<tr>
<td>Brisbane</td>
<td>25</td>
</tr>
<tr>
<td>Adelaide</td>
<td>32</td>
</tr>
<tr>
<td>St Kilda</td>
<td>34</td>
</tr>
<tr>
<td>Essendon</td>
<td>31</td>
</tr>
<tr>
<td>Carlton</td>
<td>26</td>
</tr>
<tr>
<td>West Coast</td>
<td>29</td>
</tr>
<tr>
<td>Fremantle</td>
<td>22</td>
</tr>
<tr>
<td>Hawthorn</td>
<td>33</td>
</tr>
<tr>
<td>Richmond</td>
<td>28</td>
</tr>
</tbody>
</table>

12 The 2015 weekly median rental price for a 2-bedroom unit in a number of Melbourne suburbs is given in the following table. Construct a stem plot for these data and comment on it.

<table>
<thead>
<tr>
<th>Suburb</th>
<th>Weekly rental ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphington</td>
<td>400</td>
</tr>
<tr>
<td>Box Hill</td>
<td>365</td>
</tr>
<tr>
<td>Brunswick</td>
<td>410</td>
</tr>
<tr>
<td>Burwood</td>
<td>390</td>
</tr>
<tr>
<td>Clayton</td>
<td>350</td>
</tr>
<tr>
<td>Essendon</td>
<td>350</td>
</tr>
<tr>
<td>Hampton</td>
<td>430</td>
</tr>
<tr>
<td>Ivanhoe</td>
<td>395</td>
</tr>
<tr>
<td>Kensington</td>
<td>406</td>
</tr>
<tr>
<td>Malvern</td>
<td>415</td>
</tr>
<tr>
<td>Moonee Ponds</td>
<td>373</td>
</tr>
<tr>
<td>Newport</td>
<td>380</td>
</tr>
<tr>
<td>North Melbourne</td>
<td>421</td>
</tr>
<tr>
<td>Northcote</td>
<td>430</td>
</tr>
<tr>
<td>Preston</td>
<td>351</td>
</tr>
<tr>
<td>St Kilda</td>
<td>450</td>
</tr>
<tr>
<td>Surrey Hills</td>
<td>380</td>
</tr>
<tr>
<td>Williamstown</td>
<td>330</td>
</tr>
<tr>
<td>Windsor</td>
<td>423</td>
</tr>
<tr>
<td>Yarraville</td>
<td>390</td>
</tr>
</tbody>
</table>
13 A random sample of 20 screws is taken and the length of each is recorded to the nearest millimetre below.

| 23 | 15 | 18 | 17 | 17 | 19 | 22 |
| 19 | 20 | 16 | 20 | 21 | 19 | 23 |
| 17 | 19 | 21 | 23 | 20 | 21 |

Construct a stem plot for screw length using:

a the stems 1 and 2
b the stems 1 and 2 split into halves
c the stems 1 and 2 split into fifths.

Use your plots to help you comment on the screw lengths.

14 The first twenty scores that came into the clubhouse in a local golf tournament are shown. Construct a stem plot for these data.

| 102 | 98 | 83 | 92 | 85 | 99 | 104 | 112 | 88 | 91 |
| 78 | 87 | 90 | 94 | 83 | 93 | 72 | 100 | 92 | 88 |

15 Golf handicaps are designed to even up golfers on their abilities. Their handicap is subtracted from their score to create a net score. Construct a stem plot on the golfer’s net scores below and comment on how well the golfers are handicapped.

76, 76, 73, 74, 69, 72, 73, 86, 73, 72, 75, 74, 77, 73, 75, 75, 71, 68, 67

16 The following data represents percentages for a recent Further Maths test.

| 63 | 71 | 70 | 89 | 88 | 69 | 76 | 83 | 93 | 80 | 73 |
| 77 | 91 | 75 | 81 | 84 | 87 | 78 | 97 | 89 | 98 | 60 |

Construct a stem plot for the test percentages, using:

a the stems 6, 7, 8 and 9
b the stems 6, 7, 8 and 9 split into halves.

17 The following data was collected from a company that compared the battery life (measured in minutes) of two different Ultrabook computers. To complete the test they ran a series of programs on the two computers and measured how long it took for the batteries to go from 100% to 0%.

| Computer 1 | 358 | 376 | 392 | 345 | 381 | 405 | 363 | 380 | 352 | 391 | 410 | 366 |
| Computer 2 | 348 | 355 | 361 | 342 | 355 | 362 | 353 | 358 | 340 | 346 | 357 | 352 |

a Draw a back-to-back stem plot (using the same stem) of the battery life of the two Ultrabook computers.
b Use the stem plot to compare and comment on the battery life of the two Ultrabook computers.
The heights of 20 Year 8 and Year 10 students (to the nearest centimetre), chosen at random, are measured. The data collected is shown in the table below.

<table>
<thead>
<tr>
<th>Year 8</th>
<th>Year 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>151</td>
<td>167</td>
</tr>
<tr>
<td>162</td>
<td>164</td>
</tr>
<tr>
<td>148</td>
<td>172</td>
</tr>
<tr>
<td>153</td>
<td>158</td>
</tr>
<tr>
<td>165</td>
<td>169</td>
</tr>
<tr>
<td>157</td>
<td>159</td>
</tr>
<tr>
<td>172</td>
<td>174</td>
</tr>
<tr>
<td>168</td>
<td>177</td>
</tr>
<tr>
<td>155</td>
<td>176</td>
</tr>
<tr>
<td>164</td>
<td>165</td>
</tr>
<tr>
<td>175</td>
<td>182</td>
</tr>
<tr>
<td>161</td>
<td>154</td>
</tr>
<tr>
<td>171</td>
<td>160</td>
</tr>
<tr>
<td>166</td>
<td>178</td>
</tr>
<tr>
<td>155</td>
<td>185</td>
</tr>
<tr>
<td>163</td>
<td>173</td>
</tr>
<tr>
<td>150</td>
<td>178</td>
</tr>
</tbody>
</table>

- Draw a back-to-back stem plot of the data.
- Comment on what the stem plot tells you about the heights of Year 8 and Year 10 students.

**Dot plots, frequency tables and histograms, and bar charts**

Dot plots, frequency histograms and bar charts display data in graphical form.

**Dot plots**

In picture graphs, a single picture represents each data value. Similarly, in dot plots, a single dot represents each data value. Dot plots are used to display discrete data where values are not spread out very much. They are also used to display categorical data.

When representing discrete data, dot plots have a scaled horizontal axis and each data value is indicated by a dot above this scale. The end result is a set of vertical ‘lines’ of evenly-spaced dots.

**worked example 6**

The number of hours per week spent on art by 18 students is given as shown.

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

Display the data as a dot plot.

**THINK**

1. Determine the lowest and highest scores and then draw a suitable scale.
2. Represent each score by a dot on the scale.

**DRAW**

Frequency tables and histograms

A histogram is a useful way of displaying large data sets (say, over 50 observations). The vertical axis on the histogram displays the frequency and the horizontal axis displays class intervals of the variable (for example, height or income).

The vertical bars in a histogram are adjacent with no gaps between them, as we generally consider the numerical data scale along the horizontal axis as continuous.

Note, however, that histograms can also represent discrete data. It is common practice to leave a small gap before the first bar of a histogram.
The data below show the distribution of masses (in kilograms) of 60 students in Year 7 at Northwood Secondary College. Construct a frequency histogram to display the data more clearly.

45.7 45.8 45.9 48.2 48.3 48.4 34.2 52.4 52.3 51.8 45.7 56.8 56.3 60.2 44.2
53.8 43.5 57.2 38.7 48.5 49.6 56.9 43.8 58.3 52.4 48.6 53.7 58.7 57.6
45.7 39.8 42.5 42.9 59.2 53.2 48.2 36.2 47.2 46.7 58.7 53.1 52.1 54.3 51.3
51.9 54.6 58.7 58.7 39.7 43.1 56.2 43.0 56.3 62.3 46.3 52.4 61.2 48.2 58.3

**WRITE/DRAW**

<table>
<thead>
<tr>
<th>Class interval</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>30–</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>35–</td>
<td>III</td>
<td>4</td>
</tr>
<tr>
<td>40–</td>
<td>II</td>
<td>7</td>
</tr>
<tr>
<td>45–</td>
<td>III</td>
<td>16</td>
</tr>
<tr>
<td>50–</td>
<td>III</td>
<td>15</td>
</tr>
<tr>
<td>55–</td>
<td>III</td>
<td>14</td>
</tr>
<tr>
<td>60–</td>
<td>III</td>
<td>3</td>
</tr>
</tbody>
</table>

**Total 60**

When constructing a histogram to represent continuous data, as in Worked example 7, the bars will sit between two values on the horizontal axis which represent the class intervals. When dealing with discrete data the bars should appear above the middle of the value they’re representing.
The marks out of 20 received by 30 students for a book-review assignment are given in the frequency table.

<table>
<thead>
<tr>
<th>Mark</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

Display these data on a histogram.

**THINK**

In this case we are dealing with integer values (discrete data). Since the horizontal axis should show a class interval, we extend the base of each of the columns on the histogram halfway either side of each score.

**DRAW**

**Bar charts**

A *bar chart* consists of bars of equal width *separated* by small, equal spaces and may be arranged either *horizontally* or *vertically*. Bar charts are often used to display categorical data.

In bar charts, the frequency is graphed against a variable as shown in both figures. The variable may or may not be numerical. However, if it is, the variable should represent discrete data because the scale is broken by the gaps between the bars.

The bar chart shown represents the data presented in Worked example 8. It could also have been drawn with horizontal bars (rows).
Segmented bar charts

A segmented bar chart is a single bar which is used to represent all the data being studied. It is divided into segments, each segment representing a particular group of the data. Generally, the information is presented as percentages and so the total bar length represents 100% of the data.

The table shown represents fatal road accidents in Australia. Construct a segmented bar chart to represent this data.

<table>
<thead>
<tr>
<th>Year</th>
<th>NSW</th>
<th>Vic.</th>
<th>Qld</th>
<th>SA</th>
<th>WA</th>
<th>Tas.</th>
<th>NT</th>
<th>ACT</th>
<th>Aust.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>376</td>
<td>278</td>
<td>293</td>
<td>87</td>
<td>189</td>
<td>38</td>
<td>67</td>
<td>12</td>
<td>1340</td>
</tr>
</tbody>
</table>


THINK

1. To draw a segmented bar chart the data needs to be converted to percentages.

<table>
<thead>
<tr>
<th>State</th>
<th>Number of accidents</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
<td>376</td>
<td>376/1340 × 100% = 28.1%</td>
</tr>
<tr>
<td>Vic.</td>
<td>278</td>
<td>278/1340 × 100% = 20.7%</td>
</tr>
<tr>
<td>Qld</td>
<td>293</td>
<td>293/1340 × 100% = 21.9%</td>
</tr>
<tr>
<td>SA</td>
<td>87</td>
<td>87/1340 × 100% = 6.5%</td>
</tr>
<tr>
<td>WA</td>
<td>189</td>
<td>189/1340 × 100% = 14.1%</td>
</tr>
<tr>
<td>Tas.</td>
<td>38</td>
<td>38/1340 × 100% = 2.8%</td>
</tr>
<tr>
<td>NT</td>
<td>67</td>
<td>67/1340 × 100% = 5.0%</td>
</tr>
<tr>
<td>ACT</td>
<td>14</td>
<td>14/1340 × 100% = 0.9%</td>
</tr>
</tbody>
</table>

WRITE/DRAW

2. To draw the segmented bar chart to scale decide on its overall length, let’s say 100 mm.

3. Therefore NSW = 28.1%, represented by 28.1 mm. Vic = 20.7%, represented by 20.7 mm and so on.

4. Draw the answer and colour code it to represent each of the states and territories.

The segmented bar chart is drawn to scale. An appropriate scale would be constructed by drawing the total bar 100 mm long, so that 1 mm represents 1%. That is, accidents in NSW would be represented by a segment of 28.1 mm, those in Victoria by a segment of 20.7 mm and so on. Each segment is then labelled directly, or a key may be used.
Using a log (base 10) scale

Sometimes a data set will contain data points that vary so much in size that plotting them using a traditional scale becomes very difficult. For example, if we are studying the population of different cities in Australia we might end up with the following data points:

<table>
<thead>
<tr>
<th>City</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adelaide</td>
<td>1304631</td>
</tr>
<tr>
<td>Ballarat</td>
<td>98543</td>
</tr>
<tr>
<td>Brisbane</td>
<td>2274460</td>
</tr>
<tr>
<td>Cairns</td>
<td>146778</td>
</tr>
<tr>
<td>Darwin</td>
<td>140400</td>
</tr>
<tr>
<td>Geelong</td>
<td>184182</td>
</tr>
<tr>
<td>Launceston</td>
<td>86393</td>
</tr>
<tr>
<td>Melbourne</td>
<td>4440328</td>
</tr>
<tr>
<td>Newcastle</td>
<td>430755</td>
</tr>
<tr>
<td>Shepparton</td>
<td>49079</td>
</tr>
<tr>
<td>Wagga Wagga</td>
<td>55364</td>
</tr>
</tbody>
</table>

A histogram splitting the data into class intervals of 100000 would then appear as follows:

A way to overcome this is to write the numbers in logarithmic (log) form. The log of a number is the power of 10 which creates this number.

\[
\log(10) = \log(10^1) = 1 \\
\log(100) = \log(10^2) = 2 \\
\log(1000) = \log(10^3) = 3 \\
\vdots \\
\log(10^n) = n 
\]

Not all logarithmic values are integers, so use the log key on CAS to determine exact logarithmic values.

For example, from our previous example showing the population of different Australian cities:

\[
\log(4440328) = 6.65 \text{ (correct to 2 decimal places)} \\
\log(184182) = 5.27 \text{ (correct to 2 decimal places)} \\
\log(49079) = 4.69 \text{ (correct to 2 decimal places)} \\
\text{and so on…}
\]
We can then group our data using class intervals based on log values (from 4 to 7) to come up with the following frequency table and histogram.

<table>
<thead>
<tr>
<th>Population</th>
<th>Log (population)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000–</td>
<td>4–5</td>
<td>4</td>
</tr>
<tr>
<td>100000–</td>
<td>5–6</td>
<td>4</td>
</tr>
<tr>
<td>1000000–</td>
<td>6–7</td>
<td>3</td>
</tr>
</tbody>
</table>

The following table shows the average weights of 10 different adult mammals.

<table>
<thead>
<tr>
<th>Mammal</th>
<th>Weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>African elephant</td>
<td>4800</td>
</tr>
<tr>
<td>Black rhinoceros</td>
<td>1100</td>
</tr>
<tr>
<td>Blue whale</td>
<td>136000</td>
</tr>
<tr>
<td>Giraffe</td>
<td>800</td>
</tr>
<tr>
<td>Gorilla</td>
<td>140</td>
</tr>
<tr>
<td>Humpback whale</td>
<td>30000</td>
</tr>
<tr>
<td>Lynx</td>
<td>23</td>
</tr>
<tr>
<td>Orang-utan</td>
<td>64</td>
</tr>
<tr>
<td>Polar bear</td>
<td>475</td>
</tr>
<tr>
<td>Tasmanian devil</td>
<td>7</td>
</tr>
</tbody>
</table>

Display the data in a histogram using a log base 10 scale.

THINK

1 Using CAS, calculate the logarithmic values of all of the weights, e.g.:
\[
\log(4800) = 3.68 \text{ (correct to 2 decimal places)}
\]

WRITE/DRAW

<table>
<thead>
<tr>
<th>Weight</th>
<th>Log (weight (kg))</th>
</tr>
</thead>
<tbody>
<tr>
<td>4800</td>
<td>3.68</td>
</tr>
<tr>
<td>1100</td>
<td>3.04</td>
</tr>
<tr>
<td>136000</td>
<td>5.13</td>
</tr>
<tr>
<td>800</td>
<td>2.90</td>
</tr>
<tr>
<td>140</td>
<td>2.15</td>
</tr>
<tr>
<td>30000</td>
<td>4.48</td>
</tr>
<tr>
<td>23</td>
<td>1.36</td>
</tr>
<tr>
<td>64</td>
<td>1.81</td>
</tr>
<tr>
<td>475</td>
<td>2.68</td>
</tr>
<tr>
<td>7</td>
<td>0.85</td>
</tr>
</tbody>
</table>

2 Group the logarithmic weights into class intervals and create a frequency table for the groupings.
Interpreting log (base 10) values

If we are given values in logarithmic form, by raising 10 to the power of the logarithmic number we can determine the conventional number.

For example, the number 3467 in log (base 10) form is 3.54, and \(10^{3.54} = 3467\).

We can use this fact to compare values in log (base 10) form, as shown in the following worked example.

**WORKED EXAMPLE 11**

The Richter Scale measures the magnitude of earthquakes using a log (base 10) scale.

How many times stronger is an earthquake of magnitude 7.4 than one of magnitude 5.2? Give your answer correct to the nearest whole number.

**THINK**

1. Calculate the difference between the magnitude of the two earthquakes.
   \[7.4 - 5.2 = 2.2\]
2. Raise 10 to the power of the difference in magnitudes.
   \[10^{2.2} \approx 158.49\] (correct to 2 decimal places)
3. Express the answer in words.
   The earthquake of magnitude 7.4 is 158 times stronger than the earthquake of magnitude 5.2.

**EXERCISE 1.4**

**Dot plots, frequency tables and histograms, and bar charts**

1. **WE6** The number of questions completed for maths homework each night by 16 students is shown below.

<table>
<thead>
<tr>
<th>5</th>
<th>6</th>
<th>5</th>
<th>9</th>
<th>10</th>
<th>10</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

   Display the data as a dot plot.

2. The data below represent the number of hours each week that 40 teenagers spent on household chores. Display these data on a bar chart and a dot plot.

   | 2 | 5 | 2 | 0 | 8 | 7 | 8 | 5 | 1 | 0 | 2 | 1 | 8 | 0 | 4 | 2 | 2 | 9 | 8 | 5 |
   | 7 | 5 | 4 | 2 | 1 | 2 | 9 | 8 | 1 | 2 | 8 | 5 | 8 | 1 | 0 | 0 | 3 | 4 | 5 | 2 | 8 |

3. **WE7** The data shows the distribution of heights (in cm) of 40 students in Year 12. Construct a frequency histogram to display the data more clearly.

   | 167 | 172 | 184 | 180 | 178 | 166 | 154 | 150 | 164 | 161 |
   | 187 | 159 | 182 | 177 | 172 | 163 | 179 | 181 | 170 | 176 |
   | 177 | 162 | 172 | 184 | 188 | 179 | 189 | 192 | 164 | 160 |
   | 166 | 169 | 163 | 185 | 178 | 183 | 190 | 170 | 168 | 159 |
4 Construct a frequency table for each of the following sets of data.
   a 4.3 4.5 4.7 4.9 5.1 5.3 5.5 5.6 5.2 3.6 2.5 4.3 2.5 3.7 4.5 6.3 1.3
   b 11 13 15 16 18 20 21 22 21 18 19 20 16 18 20 16 10 23 24 25
   27 28 30 35 28 27 26 29 30 31 24 28 29 20 30 32 33 29 30 31 33 34
   c 0.4 0.5 0.7 0.8 0.8 0.9 1.0 1.1 1.0 1.1 1.2 1.0 1.3 0.4 0.3 0.9 0.6
Using the frequency tables above, construct a histogram by hand for each set of data.

5 WEB The number of fish caught by 30 anglers in a fishing competition are given in the frequency table below.

<table>
<thead>
<tr>
<th>Fish</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Display these data on a histogram.

6 The number of fatal car accidents in Victoria each week is given in the frequency table for a year.

<table>
<thead>
<tr>
<th>Fatalities</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>7</th>
<th>9</th>
<th>10</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>12</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Display these data on a histogram.

7 WEB The following table shows how many goals each of the 18 AFL team’s leading goal kickers scored in the 2014 regular season. Construct a segmented bar chart to represent this data.

<table>
<thead>
<tr>
<th>Club</th>
<th>Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adelaide Crows</td>
<td>51</td>
</tr>
<tr>
<td>Brisbane Lions</td>
<td>33</td>
</tr>
<tr>
<td>Carlton</td>
<td>29</td>
</tr>
<tr>
<td>Collingwood</td>
<td>39</td>
</tr>
<tr>
<td>Essendon</td>
<td>27</td>
</tr>
<tr>
<td>Fremantle</td>
<td>49</td>
</tr>
<tr>
<td>Geelong Cats</td>
<td>62</td>
</tr>
<tr>
<td>Gold Coast Suns</td>
<td>46</td>
</tr>
<tr>
<td>GWS Giants</td>
<td>29</td>
</tr>
<tr>
<td>Hawthorn</td>
<td>62</td>
</tr>
<tr>
<td>Melbourne</td>
<td>20</td>
</tr>
<tr>
<td>North Melbourne</td>
<td>41</td>
</tr>
<tr>
<td>Port Adelaide</td>
<td>62</td>
</tr>
<tr>
<td>Richmond</td>
<td>58</td>
</tr>
<tr>
<td>St Kilda</td>
<td>49</td>
</tr>
<tr>
<td>Sydney Swans</td>
<td>67</td>
</tr>
<tr>
<td>West Coast Eagles</td>
<td>61</td>
</tr>
<tr>
<td>Western Bulldogs</td>
<td>37</td>
</tr>
</tbody>
</table>
Information about adult participation in sport and physical activities in 2005–06 is shown in the following table. Draw a segmented bar graph to compare the participation of all persons from various age groups. Comment on the statement, ‘Only young people participate in sport and physical activities’.

### Participation in sport and physical activities — 2005–06

<table>
<thead>
<tr>
<th>Age group (years)</th>
<th>Males</th>
<th>Females</th>
<th>Persons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number (× 1000)</td>
<td>Participation rate (%)</td>
<td>Number (× 1000)</td>
</tr>
<tr>
<td>18–24</td>
<td>735.2</td>
<td>73.3</td>
<td>671.3</td>
</tr>
<tr>
<td>25–34</td>
<td>1054.5</td>
<td>76.3</td>
<td>1033.9</td>
</tr>
<tr>
<td>35–44</td>
<td>975.4</td>
<td>66.7</td>
<td>1035.9</td>
</tr>
<tr>
<td>45–54</td>
<td>871.8</td>
<td>63.5</td>
<td>923.4</td>
</tr>
<tr>
<td>55–64</td>
<td>670.1</td>
<td>60.4</td>
<td>716.3</td>
</tr>
<tr>
<td>65 and over</td>
<td>591.0</td>
<td>50.8</td>
<td>652.9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4898</strong></td>
<td><strong>64.6</strong></td>
<td><strong>5033.7</strong></td>
</tr>
</tbody>
</table>

(a) Relates to persons aged 18 years and over who participated in sport or physical activity as a player during the 12 months prior to interview.


The following table shows the average weights of 10 different adult mammals.

<table>
<thead>
<tr>
<th>Mammal</th>
<th>Weight (kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black wallaroo</td>
<td>18</td>
</tr>
<tr>
<td>Capybara</td>
<td>55</td>
</tr>
<tr>
<td>Cougar</td>
<td>63</td>
</tr>
<tr>
<td>Fin whale</td>
<td>70000</td>
</tr>
<tr>
<td>Lion</td>
<td>175</td>
</tr>
<tr>
<td>Ocelot</td>
<td>9</td>
</tr>
<tr>
<td>Pygmy rabbit</td>
<td>0.4</td>
</tr>
<tr>
<td>Red deer</td>
<td>200</td>
</tr>
<tr>
<td>Quokka</td>
<td>4</td>
</tr>
<tr>
<td>Water buffalo</td>
<td>725</td>
</tr>
</tbody>
</table>

Display the data in a histogram using a log base 10 scale, using class intervals of width 1.
10 The following graph shows the weights of animals.

If a gorilla has a weight of 207 kilograms then its weight is between that of:
A Potar monkey and jaguar.
B horse and triceratops.
C guinea pig and Potar monkey.
D jaguar and horse.
E none of the above.

11 The Richter scale measures the magnitude of earthquakes using a log (base 10) scale.
How many times stronger is an earthquake of magnitude 8.1 than one of magnitude 6.9? Give your answer correct to the nearest whole number.

12 The pH scale measures acidity using a log (base 10) scale. For each decrease in pH of 1, the acidity of a substance increases by a factor of 10.
If a liquid’s pH value decreases by 0.7, by how much has the acidity of the liquid increased?

13 Using CAS, construct a histogram for each of the sets of data given in question 4. Compare this histogram with the one drawn for question 4.

14 The following table shows a variety of top speeds.

<table>
<thead>
<tr>
<th>Speed</th>
<th>m/h</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 racing car</td>
<td>370000</td>
</tr>
<tr>
<td>V8 supercar</td>
<td>300000</td>
</tr>
<tr>
<td>Cheetah</td>
<td>64000</td>
</tr>
<tr>
<td>Space shuttle</td>
<td>2800000</td>
</tr>
<tr>
<td>Usain Bolt</td>
<td>34000</td>
</tr>
</tbody>
</table>

The correct value, to 2 decimal places, for a cheetah’s top speed using a log (base 10) scale would be:
A 5.57      B 5.47      C 7.45      D 4.81      E 11.07

15 The number of hours spent on homework for a group of 20 Year 12 students each week is shown.

<table>
<thead>
<tr>
<th>Hours</th>
<th>15</th>
<th>18</th>
<th>20</th>
<th>21</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15</td>
<td>18</td>
<td>20</td>
<td>21</td>
<td>24</td>
</tr>
</tbody>
</table>

Display the data as a dot plot.
16 The number of dogs at a RSPCA kennel each week is as shown. Construct a frequency table for these data.
7, 6, 2, 12, 7, 9, 12, 10, 5, 7, 9, 4, 5, 9, 3, 2, 10, 8, 9, 7, 9, 10, 9, 4, 3, 8, 9, 3, 7, 9

17 Using the frequency table in question 16, construct a histogram by hand.

18 Using CAS construct a histogram from the data in question 16, and compare it to the histogram in question 17.

19 A group of students was surveyed, asking how many children were in their family. The data is shown in the table.

<table>
<thead>
<tr>
<th>Number of children</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of families</td>
<td>12</td>
<td>18</td>
<td>24</td>
<td>10</td>
<td>8</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Construct a bar chart that displays the data.

20 The following graph represents the capacity of five Victorian dams.

```
<table>
<thead>
<tr>
<th>Dams</th>
<th>Capacity of Victorian dams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sugarloaf</td>
<td></td>
</tr>
<tr>
<td>O'Shaughnessy</td>
<td></td>
</tr>
<tr>
<td>Silvan</td>
<td></td>
</tr>
<tr>
<td>Yan Yean</td>
<td></td>
</tr>
<tr>
<td>Thomson</td>
<td></td>
</tr>
</tbody>
</table>
```

- **a** The capacity of Thomson Dam is closest to:
  - A 100000 ML
  - B 1000000 ML
  - C 500000 ML
  - D 5000 ML
  - E 400000 ML

- **b** The capacity of Sugarloaf Dam is closest to:
  - A 100000 ML
  - B 1000000 ML
  - C 500000 ML
  - D 5000 ML
  - E 400000 ML

- **c** The capacity of Silvan Dam is between which range?
  - A 1 and 10 ML
  - B 10 and 100 ML
  - C 1000 and 10000 ML
  - D 10000 and 100000 ML
  - E 100000 and 1000000 ML

21 The correct value, correct to 2 decimal places, to be plotted for Niagara Falls’ flow rate using a log (base 10) scale would be:

- A 3.38
- B 3.04
- C 4.06
- D 4.23
- E 5.10
22 The correct value, correct to 2 decimal places, to be plotted for Victoria Falls’ flow rate using a log (base 10) scale would be:

A 3.38  B 3.04  C 4.06  D 4.23  E 5.10

23 Using the information provided in the table below:
   a calculate the proportion of residents who travelled in 2005 to each of the countries listed
   b draw a segmented bar graph showing the major destinations of Australians travelling abroad in 2005.

<table>
<thead>
<tr>
<th></th>
<th>2004 (×1000)</th>
<th>2005 (×1000)</th>
<th>2006 (×1000)</th>
<th>2007 (×1000)</th>
<th>2008 (×1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Zealand</td>
<td>815.8</td>
<td>835.4</td>
<td>864.7</td>
<td>902.1</td>
<td>921.1</td>
</tr>
<tr>
<td>United States of America</td>
<td>376.1</td>
<td>426.3</td>
<td>440.3</td>
<td>479.1</td>
<td>492.3</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>375.1</td>
<td>404.2</td>
<td>412.8</td>
<td>428.5</td>
<td>420.3</td>
</tr>
<tr>
<td>Indonesia</td>
<td>335.1</td>
<td>319.7</td>
<td>194.9</td>
<td>282.6</td>
<td>380.7</td>
</tr>
<tr>
<td>China (excluding</td>
<td>182.0</td>
<td>235.1</td>
<td>251.0</td>
<td>284.3</td>
<td>277.3</td>
</tr>
<tr>
<td>Special Administrative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regions (SARs))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thailand</td>
<td>188.2</td>
<td>202.7</td>
<td>288.0</td>
<td>374.4</td>
<td>404.1</td>
</tr>
<tr>
<td>Fiji</td>
<td>175.4</td>
<td>196.9</td>
<td>202.4</td>
<td>200.3</td>
<td>236.2</td>
</tr>
<tr>
<td>Singapore</td>
<td>159.0</td>
<td>188.5</td>
<td>210.9</td>
<td>221.5</td>
<td>217.8</td>
</tr>
<tr>
<td>Hong Kong (SAR of China)</td>
<td>152.6</td>
<td>185.7</td>
<td>196.3</td>
<td>206.5</td>
<td>213.1</td>
</tr>
<tr>
<td>Malaysia</td>
<td>144.4</td>
<td>159.8</td>
<td>168.0</td>
<td>181.3</td>
<td>191.0</td>
</tr>
</tbody>
</table>


24 The number of people who attended the Melbourne Grand Prix are shown.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendance (000s)</td>
<td>360</td>
<td>359</td>
<td>301</td>
<td>301</td>
<td>303</td>
<td>287</td>
<td>305</td>
<td>298</td>
<td>313</td>
<td>323</td>
</tr>
</tbody>
</table>

a Construct a bar chart by hand to display the data.
b Use CAS to construct a bar chart and compare the two charts.

### 1.5 Describing the shape of stem plots and histograms

#### Symmetric distributions

The data shown in the histogram below can be described as symmetric. There is a single peak and the data trail off on both sides of this peak in roughly the same fashion.
Similarly, in the stem plot at right, the distribution of the data could be described as symmetric. The single peak for these data occur at the stem 3. On either side of the peak, the number of observations reduces in approximately matching fashion.

**Skewed distributions**

Each of the histograms shown on next page are examples of skewed distributions.

The figure on the left shows data which are **negatively skewed**. The data in this case peak to the right and trail off to the left.

The figure on the right shows **positively skewed** data. The data in this case peak to the left and trail off to the right.

---

**WORKED EXAMPLE 12**

The ages of a group of people who were taking out their first home loan is shown below.

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

Key: 1|9 = 19 years old

Describe the shape of the distribution of these data.

**THINK**

Check whether the distribution is symmetric or skewed. The peak of the data occurs at the stem 2. The data trail off as the stems increase in value. This seems reasonable since most people would take out a home loan early in life to give themselves time to pay it off.

**WRITE**

The data are positively skewed.
Describing the shape of stem plots and histograms

1. The ages of a group of people when they bought their first car are shown.

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

Key: 1|7 = 17 years old

Describe the shape of the distribution of these data.

2. The ages of women when they gave birth to their first child is shown.

Describe the shape of the distribution of the data.

3. For each of the following stem plots, describe the shape of the distribution of the data.

   a. Stem | Leaf
   0    | 1 3 |
   1    | 2 4 7 |
   2    | 3 4 4 7 8 |
   3    | 2 5 7 9 9 9 9 9 |
   4    | 1 3 6 7 |
   5    | 0 4 |
   6    | 4 7 |
   7    | 1 |

   Key: 1|2 = 12

   b. Stem | Leaf
   1    | 3 |
   2    | 6 |
   3    | 3 8 |
   4    | 2 6 8 8 9 |
   5    | 4 7 7 7 8 9 9 |
   6    | 0 2 2 4 5 |

   Key: 2|6 = 2.6

   c. Stem | Leaf
   2    | 3 5 5 6 7 8 9 9 |
   3    | 0 2 2 3 4 6 6 7 8 8 |
   4    | 2 2 4 5 6 6 6 7 9 |
   5    | 0 3 3 5 6 |
   6    | 2 4 |
   7    | 5 9 |
   8    | 2 |
   9    | 7 |
   10   | |

   Key: 10|4 = 104

   d. Stem | Leaf
   1    | 1* 5 |
   2    | 1 4 |
   2*   | 5 7 8 8 9 |
   3    | 1 2 2 3 3 3 4 4 |
   3*   | 5 5 5 6 |
   4    | 3 4 |
   4*   | |

   Key: 2|4 = 24
4 For each of the following histograms, describe the shape of the distribution of the data and comment on the existence of any outliers.

5 The distribution of the data shown in this stem plot could be described as:
   A negatively skewed
   B negatively skewed and symmetric
   C positively skewed
   D positively skewed and symmetric
   E symmetric
6 The distribution of the data shown in the histogram could be described as:
   A negatively skewed
   B negatively skewed and symmetric
   C positively skewed
   D positively skewed and symmetric
   E symmetric

7 The average number of product enquiries per day received by a group of small businesses who advertised in the Yellow Pages telephone directory is given at right. Describe the shape of the distribution of these data.

8 The number of nights per month spent interstate by a group of flight attendants is shown in the stem plot. Describe the shape of distribution of these data and explain what this tells us about the number of nights per month spent interstate by this group of flight attendants.

   Stem | Leaf
   0    | 0 0 1 1
   0    | 2 2 3 3 3 3 3 3 3
   0    | 4 4 5 5 5 5 5 5
   0    | 6 6 6 6 7
   0    | 8 8 8 9
   1    | 0 0 1
   1    | 4 4
   1    | 5 5
   1    | 7

   Key: 1|4 = 14 nights

9 The mass (correct to the nearest kilogram) of each dog at a dog obedience school is shown in the stem plot.
   a Describe the shape of the distribution of these data.
   b What does this information tell us about this group of dogs?

   Stem | Leaf
   0    | 4
   0*   | 5 7 9
   1    | 1 2 4 4
   1*   | 5 6 6 7 8 9
   2    | 1 2 2 3
   2*   | 6 7

   Key: 0|4 = 4 kg
10 The amount of pocket money (correct to the nearest 50 cents) received each week by students in a Grade 6 class is illustrated in the histogram.

a Describe the shape of the distribution of these data.

b What conclusions can you reach about the amount of pocket money received weekly by this group of students?

11 Statistics were collected over 3 AFL games on the number of goals kicked by forwards over 3 weeks. This is displayed in the histogram.

a Describe the shape of the histogram.

b Use the histogram to determine:
   i the number of players who kicked 3 or more goals over the 3 weeks
   ii the percentage of players who kicked between 2 and 6 goals inclusive over the 3 weeks.

12 The number of hours a group of students exercise each week is shown in the stem plot.

a Describe the shape of the distribution of these data.

b What does this sample data tell us about this group of students?

c How many students exercise at least 5 hours each week?

d How many students exercise between 3 and 5 hours each week?

13 The stem plot shows the age of players in two bowling teams.

a Describe the shape of the distribution of Club A and Club B.

b What comments can you make about the make-up of Club A compared to Club B?

c How many players are over the age of 70 from:
   i Club A
   ii Club B?

14 The following table shows the number of cars sold at a dealership over eight months.

<table>
<thead>
<tr>
<th>Month</th>
<th>April</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>August</th>
<th>September</th>
<th>October</th>
<th>November</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cars sold</td>
<td>9</td>
<td>14</td>
<td>27</td>
<td>21</td>
<td>12</td>
<td>14</td>
<td>10</td>
<td>18</td>
</tr>
</tbody>
</table>

a Display the data on a bar chart.

b Describe the shape of the distribution of these data.

c What does this sample data tell us about car sales over these months?

d Explain why the most cars were sold in the month of June.
The median, the interquartile range, the range and the mode

After displaying data using a histogram or stem plot, we can make even more sense of the data by calculating what are called summary statistics. Summary statistics are used because they give us an idea about:
1. where the centre of the distribution is
2. how the distribution is spread out.

We will look first at four summary statistics — the median, the interquartile range, the range and the mode — which require that the data be in ordered form before they can be calculated.

The median

The median is the midpoint of an ordered set of data. Half the data are less than or equal to the median.

Consider the set of data: 2 5 6 8 11 12 15. These data are in ordered form (that is, from lowest to highest). There are 7 observations. The median in this case is the middle or fourth score; that is, 8.

Consider the set of data: 1 3 5 6 7 8 8 9 10 12. These data are in ordered form also; however, in this case there is an even number of scores. The median of this set lies halfway between the 5th score (7) and the 6th score (8). So the median is 7.5.

Alternatively, median = \( \frac{7 + 8}{2} = 7.5 \).

When there are \( n \) records in a set of ordered data, the median can be located at the \( \left( \frac{n + 1}{2} \right) \) th position.

Checking this against our previous example, we have \( n = 10 \); that is, there were 10 observations in the set. The median was located at the \( \left( \frac{10 + 1}{2} \right) = 5.5 \) th position; that is, halfway between the 5th and the 6th terms.

A stem plot provides a quick way of locating a median since the data in a stem plot are already ordered.

WORKED EXAMPLE

Consider the stem plot below which contains 22 observations. What is the median?

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

Key: 3|4 = 34
The interquartile range

We have seen that the median divides a set of data in half. Similarly, quartiles divide a set of data in quarters. The symbols used to refer to these quartiles are $Q_1$, $Q_2$ and $Q_3$.

The middle quartile, $Q_2$, is the median.

The interquartile range $IQR = Q_3 - Q_1$.

The interquartile range gives us the range of the middle 50% of values in a set of data.

There are four steps to locating $Q_1$ and $Q_3$.

Step 1. Write down the data in ordered form from lowest to highest.

Step 2. Locate the median; that is, locate $Q_2$.

Step 3. Now consider just the lower half of the set of data. Find the middle score. This score is $Q_1$.

Step 4. Now consider just the upper half of the set of data. Find the middle score. This score is $Q_3$.

The four cases given below illustrate this method.

Case 1

Consider data containing the 6 observations: 3 6 10 12 15 21.

The data are already ordered. The median is 11.

Consider the lower half of the set, which is 3 6 10. The middle score is 6, so $Q_1 = 6$.

Consider the upper half of the set, which is 12 15 21. The middle score is 15, so $Q_3 = 15$.

Case 2

Consider a set of data containing the 7 observations: 4 9 11 13 17 23 30.

The data are already ordered. The median is 13.

Consider the lower half of the set, which is 4 9 11. The middle score is 9, so $Q_1 = 9$.

Consider the upper half of the set, which is 17 23 30. The middle score is 23, so $Q_3 = 23$. 

\[
\text{Median} = \left( \frac{n + 1}{2} \right) \text{th position} \\
= \left( \frac{22 + 1}{2} \right) \text{th position} \\
= \text{11.5th position}
\]
Case 3
Consider a set of data containing the 8 observations: 1 3 9 10 15 17 21 26.
The data are already ordered. The median is 12.5.
Consider the lower half of the set, which is 1 3 9 10. The middle score is 6,
so $Q_1 = 6$.
Consider the upper half of the set, which is 15 17 21 26. The middle score is 19,
so $Q_3 = 19$.

Case 4
Consider a set of data containing the 9 observations: 2 7 13 14 17 19 21 25 29.
The data are already ordered. The median is 17.
Consider the lower half of the set, which is 2 7 13 14. The middle score is 10,
so $Q_1 = 10$.
Consider the upper half of the set, which is 19 21 25 29. The middle score is 23,
so $Q_3 = 23$.

The ages of the patients who attended the casualty department of an inner-
suburban hospital on one particular afternoon are shown below.

<table>
<thead>
<tr>
<th>14</th>
<th>3</th>
<th>27</th>
<th>42</th>
<th>19</th>
<th>17</th>
<th>73</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>62</td>
<td>21</td>
<td>23</td>
<td>2</td>
<td>5</td>
<td>58</td>
</tr>
<tr>
<td>33</td>
<td>19</td>
<td>81</td>
<td>59</td>
<td>25</td>
<td>17</td>
<td>69</td>
</tr>
</tbody>
</table>

Find the interquartile range of these data.

**THINK**
1. Order the data.

   2 3 5 14 17 19 19 21 23
   25 27 33 42 58 59 60 62 69 73 81

2. Find the median.

   The median is 25 since ten scores lie below it and ten lie above it.

3. Find the middle score of the lower half of the data.

   For the scores 2 3 5 14 17 19 19 21 23, the middle score is 17.
   So, $Q_1 = 17$.

4. Find the middle score of the upper half of the data.

   For the scores 27 33 42 58 59 60 62 69 73 81, the middle score is 59.5.
   So, $Q_3 = 59.5$.

5. Calculate the interquartile range.

   \[
   IQR = Q_3 - Q_1 \\
   = 59.5 - 17 \\
   = 42.5
   \]

CAS can be a fast way of locating quartiles and hence finding the value of the
interquartile range.
Parents are often shocked at the amount of money their children spend. The data below give the amount spent (correct to the nearest whole dollar) by each child in a group that was taken on an excursion to the Royal Melbourne Show.

15 12 17 23 21 19 16
11 17 18 23 24 25 21
20 37 17 25 22 21 19

Calculate the interquartile range for these data.

**THINK**

1. Enter the data into CAS to generate one-variable statistics. Copy down the values of the first and third quartiles.
2. Calculate the interquartile range.

**WRITE**

\[ Q_1 = 17 \text{ and } Q_3 = 23 \]

So, IQR = \[ Q_3 - Q_1 = 23 - 17 = 6 \]

---

### The range

The *range* of a set of data is the difference between the highest and lowest values in that set.

It is usually not too difficult to locate the highest and lowest values in a set of data. Only when there is a very large number of observations might the job be made more difficult. In Worked example 15, the minimum and maximum values were 11 and 37, respectively. The range, therefore, can be calculated as follows.

\[
\text{Range} = \max_x - \min_x \\
= 37 - 11 \\
= 26
\]

While the range gives us some idea about the spread of the data, it is not very informative since it gives us no idea of how the data are distributed between the highest and lowest values.

Now let us look at another measure of the centre of a set of data: the mode.

### The mode

The *mode* is the score that occurs most often; that is, it is the score with the highest frequency. If there is more than one score with the highest frequency, then all scores with that frequency are the modes.

The mode is a weak measure of the centre of data because it may be a value that is close to the extremes of the data. If we consider the set of data in Worked...
example 13, the mode is 48 since it occurs three times and hence is the score with the highest frequency. In Worked example 14 there are two modes, 17 and 19, because they equally occur most frequently.

**EXERCISE 1.6 The median, the interquartile range, the range and the mode**

**PRACTISE**

1. **WE13** The stem plot shows 30 observations. What is the median value?

   Stem | Leaf
   ------|------
   2     | 1 1 3 4 4 4
   2*    | 5 5 7 8 9
   3     | 0 0 1 3 3 3 3
   3*    | 6 6 7 9
   4     | 0 1 1
   4*    | 6 7 9 9 9

   Key: 2|1 = 21

2. The following data represents the number of goals scored by a netball team over the course of a 16 game season. What was the team’s median number of goals for the season?

   28 36 24 46 37 21 49 32 33 41 47 29 45 52 37 24

3. **WE14** From the following data find the interquartile range.

   33 21 39 45 31 28 15 13 16 21 49 26 29 30 21 37 27 19

4. The ages of a sample of people surveyed at a concert are shown.

   21 25 24 18 19 16 19 27 32 24 15 20 31 24 29 33 27 18 19 21

   Find the interquartile range of these data.

5. **WE15** The data shows the amount of money spent (to the nearest dollar) at the school canteen by a group of students in a week.

   3 5 7 12 15 10 8 9 21 5 7 9 13 15 7 3 4 2 11 8

   Calculate the interquartile range for the data set.

6. The amount of money, in millions, changing hands through a large stocks company, in one-minute intervals, was recorded as follows.

   45.8 48.9 46.4 45.7 43.8 49.1 42.7 43.1 45.3 48.6 41.9 40.0 45.9 44.7 43.9 45.1 47.1 49.7 42.9 45.1

   Calculate the interquartile range for these data.
Write the median, the range and the mode of the sets of data shown in the following stem plots. The key for each stem plot is \(3|4 = 34\).

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

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

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

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

<table>
<thead>
<tr>
<th>e Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>2 5 8</td>
</tr>
<tr>
<td>61</td>
<td>1 3 3 6 7 8 9</td>
</tr>
<tr>
<td>62</td>
<td>0 1 2 4 6 7 8 8 9</td>
</tr>
<tr>
<td>63</td>
<td>2 2 4 5 7 8</td>
</tr>
<tr>
<td>64</td>
<td>3 6 7</td>
</tr>
<tr>
<td>65</td>
<td>4 5 8</td>
</tr>
<tr>
<td>66</td>
<td>3 5</td>
</tr>
<tr>
<td>67</td>
<td>4</td>
</tr>
</tbody>
</table>

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

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

For each of the following sets of data, write the median and the range.

| a 2 4 6 7 9 |
| b 12 15 17 19 21 |
| c 3 4 5 6 7 8 9 |
| d 3 5 7 8 12 13 15 16 |
| e 12 13 15 16 18 19 21 23 24 26 |
| f 3 8 4 2 1 6 5 |
| g 16 21 14 28 23 15 11 19 25 |
| h 7 4 3 4 9 5 10 4 2 11 |

9 a The number of cars that used the drive-in at a McBurger restaurant during each hour, from 7.00 am until 10.00 pm on a particular day, is shown below. 

14 18 8 9 12 24 25 15 18 25 24 21 25 24 14

Find the interquartile range of this set of data.
b On the same day, the number of cars stopping during each hour that the nearby Kenny’s Fried Chicken restaurant was open is shown below.

7 9 13 16 19 12 11 18 20 19 21 20 18 10 14

Find the interquartile range of these data.

c What do these values suggest about the two restaurants?

10 Write down a set of data for which \( n = 5 \), the median is 6 and the range is 7. Is this the only set of data with these parameters?

11 Give an example of a data set where:
   a the lower quartile equals the lowest score
   b the IQR is zero.

12 The quartiles for a set of data are calculated and found to be \( Q_1 = 13 \), \( Q_2 = 18 \) and \( Q_3 = 25 \). Which of the following statements is true?
   A The interquartile range of the data is 5.
   B The interquartile range of the data is 7.
   C The interquartile range of the data is 12.
   D The median is 12.
   E The median is 19.

13 For each of the following sets of data find the median, the interquartile range, the range and the mode.

a

| 16 12 8 7 26 32 15 51 29 45 |
|---|---|---|---|---|---|---|---|---|---|
| 19 11 6 15 32 18 43 31 23 23 |

b

| 22 25 27 36 31 32 39 29 20 30 |
|---|---|---|---|---|---|---|---|---|---|
| 23 25 21 19 29 28 31 27 22 29 |

14 For each set of data shown in the stem plots, find the median, the interquartile range, the range and the mode. Compare these values for both data sets.

a

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3 5 5 6 7 8 9 9</td>
</tr>
<tr>
<td>3</td>
<td>0 2 2 3 4 6 6 7 8 8</td>
</tr>
<tr>
<td>4</td>
<td>2 2 4 5 6 6 6 7 9</td>
</tr>
<tr>
<td>5</td>
<td>0 3 3 5 6</td>
</tr>
<tr>
<td>6</td>
<td>2 4</td>
</tr>
<tr>
<td>7</td>
<td>5 9</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>11 4</td>
</tr>
</tbody>
</table>

Key: 4|2 = 42

b

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

Key: 2|1 = 21
15 For the data in the stem plots shown, find the range, median, mode and interquartile range.

a Stem | Leaf
---|---
0 | 1
1 | 1 4 7
2 | 3 4 6 7
3 | 4 6 7 8 9
4 | 2 3 6 6
5 | 2 3 5
6 | 7
7 | 3

Key: 0|1 = 1

b Stem | Leaf
---|---
40 | 3 5 7
41 | 1 1 1 3 4 6 7 8 9
42 | 0 2 3 6 7 8 9
43 | 2 3 3 6 8
44 | 1 2
45 | 0

Key: 40|3 = 403

16 From the following set of data, find the median and mode.
4, 7, 9, 12, 15, 2, 3, 7, 9, 4, 7, 9, 2, 8, 13, 5, 3, 7, 5, 9, 7, 10

17 The following data represents distances (in metres) thrown during a javelin-throwing competition. Use CAS to calculate the interquartile range and median.

| 40.3 | 42.8 | 41.0 | 50.3 | 52.2 | 46.1 | 44.5 |
| 41.6 | 44.3 | 47.4 | 45.1 | 48.8 | 46.1 | 44.5 |
| 45.3 | 42.9 | 41.1 | 49.0 | 47.5 | 40.8 | 51.1 |

18 The data below shows the distribution of golf scores for one day of an amateur tournament.

a Use CAS to calculate the median, interquartile range, mode and range.

b Comment on what you suggest the average handicap of the players listed should be if par for the course is 72.

| 111 | 93 | 103 | 85 | 81 | 90 |
| 101 | 95 | 84 | 93 | 101 | 85 |
| 87 | 85 | 93 | 100 | 86 | 91 |
| 93 | 95 | 93 | 99 | 95 | 93 |
| 92 | 96 | 93 | 97 | 93 | 93 |
| 97 | 96 | 92 | 100 | 95 | 104 |

1.7 Boxplots

The five number summary statistics that we looked at in the previous section \((\min, Q_1, Q_2, Q_3, \max)\) can be illustrated very neatly in a special diagram known as a boxplot (or box-and-whisker diagram). The diagram is made up of a box with straight lines (whiskers) extending from opposite sides of the box.

A boxplot displays the minimum and maximum values of the data together with the quartiles and is drawn with a labelled scale. The length of the box is given by the interquartile range. A boxplot gives us a very clear visual display of how the data are spread out.
Boxplots can be drawn horizontally or vertically.

The boxplot at right shows the distribution of the part-time weekly earnings of a group of Year 12 students. Write down the range, the median and the interquartile range for these data.

**THINK**

1. Range = Maximum value − Minimum value. The minimum value is 20 and the maximum value is 90.

2. The median is located at the bar inside the box.

3. The ends of the box are at 40 and 80. IQR = $Q_3 - Q_1$

**WRITE**

Range = 90 − 20 = 70

Median = 50

$Q_1 = 40$ and $Q_3 = 80$

IQR = 80 − 40 = 40

Earlier, we noted three general types of shape for histograms and stem plots: symmetric, negatively skewed and positively skewed. It is useful to compare the corresponding boxplots of distributions with such shapes.
WORKED EXAMPLE 17

Explain whether or not the histogram and the boxplot shown below could represent the same data.

THINK

The histogram shows a distribution which is positively skewed.
The boxplot shows a distribution which is approximately symmetric.

WRITE

The histogram and the boxplot could not represent the same data since the histogram shows a distribution that is positively skewed and the boxplot shows a distribution that is approximately symmetric.

WORKED EXAMPLE 18

The results (out of 20) of oral tests in a Year 12 Indonesian class are:

15 12 17 8 13 18 14 16 17 13 11 12

Display these data using a boxplot and discuss the shape obtained.
Outliers

When one observation lies well away from other observations in a set, we call it an outlier. Sometimes an outlier occurs because data have been incorrectly obtained or misread. For example, here we see a histogram showing the weights of a group of 5-year-old boys.

The outlier, 33, may have occurred because a weight was incorrectly recorded as 33 rather than 23 or perhaps there was a boy in this group who, for some medical reason, weighed a lot more than his counterparts. When an outlier occurs, the reasons for its occurrence should be checked.

The lower and upper fences

We can identify outliers by calculating the values of the lower and upper fences in a data set. Values which lie either below the lower fence or above the upper fence are outliers.

An outlier is not included in the boxplot, but should instead be plotted as a point beyond the end of the whisker.
The times (in seconds) achieved by the 12 fastest runners in the 100-m sprint at a school athletics meeting are listed below.

11.2  12.3  11.5  11.0  11.6  11.4
11.9  11.2  12.7  11.3  11.2  11.3

Draw a boxplot to represent the data, describe the shape of the distribution and comment on the existence of any outliers.

**THINK**

1. Determine the five number summary statistics by first ordering the data and obtaining the interquartile range.

WRITE/DRAW

11.0  11.2  11.2  11.3  11.3
11.4  11.5  11.6  11.9  12.3  12.7

Lowest score = 11.0
Highest score = 12.7
Median = $Q_2 = 11.35$

\[ Q_1 = 11.2 \]
\[ Q_3 = 11.75 \]

IQR = $11.75 - 11.2 = 0.55$

2. Identify any outliers by calculating the values of the lower and upper fences.

\[ Q_1 - 1.5 \times IQR = 11.2 - 1.5 \times 0.55 = 10.375 \]

The lowest score lies above the lower fence of 10.375, so there is no outlier below.

\[ Q_3 + 1.5 \times IQR = 11.75 + 1.5 \times 0.55 = 12.575 \]

The score 12.7 lies above the upper fence of 12.575, so it is an outlier and 12.3 becomes the end of the upper whisker.

3. Draw the boxplot with the outlier.

4. Describe the shape of the distribution. Data peak to the left and trail off to the right with one outlier.

The data are positively skewed with 12.7 seconds being an outlier. This may be due to incorrect timing or recording but more likely the top eleven runners were significantly faster than the other competitors in the event.

**EXERCISE 1.7**

**Boxplots**

1. Write down the range, median and interquartile range for the data in the boxplot shown.

**WORKED EXAMPLE 19**

19
2 Find the median, range and interquartile range of the data displayed in the boxplot shown.

\[\text{Boxplot} \quad 6 \quad 6.5 \quad 7 \quad 7.5 \quad 8 \quad 8.5 \quad 9\]

3 **WE17** Explain whether or not the histogram and the boxplot shown could represent the same data.

4 Do the histogram and boxplot represent the same data? Explain.

5 **WE18** The results for a Physics test (out of 50) are shown.

\[
32 \quad 38 \quad 42 \quad 40 \quad 37 \quad 26 \quad 46 \quad 36 \quad 50 \quad 41 \quad 48 \quad 50 \quad 40 \quad 38 \quad 32 \quad 35 \quad 28 \quad 30
\]

Display the data using a boxplot and discuss the shape obtained.

6 The numbers of hours Year 12 students spend at their part-time job per week are shown.

\[
4 \quad 8 \quad 6 \quad 5 \quad 12 \quad 8 \quad 16 \quad 4 \quad 7 \quad 10 \quad 8 \quad 20 \quad 12 \quad 7 \quad 6 \quad 4 \quad 8
\]

Display the data using a boxplot and discuss what the boxplot shows in relation to part-time work of Year 12 students.

7 **WE19** The heights jumped (in metres) at a school high jump competition are listed:

\[
1.35 \quad 1.30 \quad 1.40 \quad 1.38 \quad 1.45 \quad 1.48 \\
1.30 \quad 1.36 \quad 1.45 \quad 1.75 \quad 1.46 \quad 1.40
\]

a Draw a boxplot to represent the data.

b Describe the shape of the distribution and comment on the existence of any outliers by finding the lower and upper fences.

8 The amount of fuel (in litres) used at a petrol pump for 16 cars is listed:

\[
48.5 \quad 55.1 \quad 61.2 \quad 58.5 \quad 46.9 \quad 49.2 \quad 57.3 \quad 49.9 \\
51.6 \quad 30.3 \quad 45.9 \quad 50.2 \quad 52.6 \quad 47.0 \quad 55.5 \quad 60.3
\]

Draw a boxplot to represent the data and label any outliers.
9 For the boxplots shown, write down the range, the interquartile range and the median of the distributions which each one represents.

a

![Boxplot a](image)

b

![Boxplot b](image)

c

![Boxplot c](image)

d

![Boxplot d](image)

e

![Boxplot e](image)

10 Match each histogram below with the boxplot which could show the same distribution.

a

![Histogram a](image)

b

![Histogram b](image)

c

![Histogram c](image)

d

![Histogram d](image)

i

![Boxplot i](image)

ii

![Boxplot ii](image)

iii

![Boxplot iii](image)

iv

![Boxplot iv](image)

11 For each of the following sets of data, construct a boxplot.

a 3 5 6 8 8 9 12 14 17 18
b 3 4 4 5 5 6 7 7 8 8 9 9 10 10 12
c 4.3 4.5 4.7 4.9 5.1 5.3 5.5 5.6
d 11 13 15 15 16 18 20 21 22 21 18 19 20 16 18 20
e 0.4 0.5 0.7 0.8 0.8 0.9 1.0 1.1 1.2 1.0 1.3
12 For the distribution shown in the boxplot below, it is true to say that:
A the median is 30
B the median is 45
C the interquartile range is 10
D the interquartile range is 30
E the interquartile range is 60

13 The number of clients seen each day over a 15-day period by a tax consultant is:

\[3 \ 5 \ 2 \ 7 \ 5 \ 6 \ 4 \ 3 \ 4 \ 5 \ 6 \ 4 \ 3 \ 4\]

Represent these data on a boxplot.

14 The maximum daily temperatures (in °C) for the month of October in Melbourne are:

\[18 \ 26 \ 28 \ 23 \ 16 \ 19 \ 21 \ 27 \ 31 \ 23 \ 24 \ 26 \ 21 \ 18 \ 26 \ 27 \]
\[23 \ 21 \ 24 \ 20 \ 19 \ 25 \ 27 \ 32 \ 29 \ 21 \ 16 \ 19 \ 23 \ 25 \ 27\]

Represent these data on a boxplot.

15 The number of rides that 16 children had at the annual show are listed below.

\[8 \ 5 \ 9 \ 4 \ 9 \ 0 \ 8 \ 7 \ 9 \ 2 \ 8 \ 7 \ 9 \ 6 \ 7 \ 8\]

a Draw a boxplot to represent the data, describe the shape of the distribution and comment on the existence of any outliers.

b Use CAS to draw a boxplot for these data.

16 A concentration test was carried out on 40 students in Year 12 across Australia. The test involved the use of a computer mouse and the ability to recognise multiple images. The less time required to complete the activity, the better the student’s ability to concentrate.

The data are shown by the parallel boxplots.

a Identify two similar properties of the concentration spans for boys and girls.

b Find the interquartile range for boys and girls.

c Comment on the existence of an outlier in the boys’ data by finding the lower and upper fences.

17 The weights of 15 boxes (in kilograms) being moved from one house to another are as follows:

\[5, \ 7, \ 10, \ 15, \ 13, \ 14, \ 17, \ 20, \ 9, \ 4, \ 11, \ 12, \ 18, \ 21, \ 15\]

Draw a boxplot to display the data.
18 You work in the marketing department of a perfume company. You completed a survey of people who purchased your perfume, asking them how many times a week they used it. Analyse the data by drawing a boxplot and comment on the existence of any outliers by finding the lower and upper fences.

| 7 2 5 4 7 5 7 2 |
| 5 4 3 5 7 7 9 8 |
| 5 6 5 3 15 8 7 5 |

19 For the data set shown:

| 11 11 14 16 19 22 24 25 |
| 25 27 28 28 36 38 38 39 |

a construct a boxplot by hand  
b comment on the presence of outliers by finding the lower and upper fences  
c construct a boxplot using CAS and compare the two boxplots.

20 From the stem plot shown construct a boxplot using CAS and comment on any outliers if they exist in the data.

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1 5 8</td>
</tr>
<tr>
<td>6</td>
<td>2 5 5 7</td>
</tr>
<tr>
<td>7</td>
<td>1 2 3 3 6 7 7</td>
</tr>
<tr>
<td>8</td>
<td>1 1 3 7 7</td>
</tr>
<tr>
<td>9</td>
<td>6 7 9</td>
</tr>
<tr>
<td>10</td>
<td>1 3</td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

Key: 2|3 = 23

1.8 The mean of a sample

When dealing with sets of data, we are always working with either the population or a sample from the population. The means of a data set representing the population and a sample are calculated in the same way; however, they are represented by different symbols. The mean of a population is represented by the Greek letter $\mu$ (mu) and the mean of a sample is represented by $\bar{x}$ (x-bar), a lower case x with a bar on top. In this section we will only use $\bar{x}$ to represent the mean.

The mean of a set of data is what is referred to in everyday language as the average.

For the set of data \{4, 7, 9, 12, 18\}:

$$\bar{x} = \frac{4 + 7 + 9 + 12 + 18}{5} = 10.$$

The formal definition of the mean is:

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

where $\sum x$ represents the sum of all of the observations in the data set and $n$ represents the number of observations in the data set.
Note that the symbol, $\sum$ is the Greek letter, sigma, which represents ‘the sum of’. The mean is also referred to as a *summary statistic* and is a measure of the centre of a distribution. The mean is the point about which the distribution ‘balances’. Consider the masses of 7 potatoes, given in grams, in the photograph.

![Potatoes](image)

The mean is 145 g. The observations 130 and 160 ‘balance’ each other since they are each 15 g from the mean. Similarly, the observations 120 and 170 ‘balance’ each other since they are each 25 g from the mean, as do the observations 100 and 190. Note that the median is also 145 g. That is, for this set of data the mean and the median give the same value for the centre. This is because the distribution is symmetric.

Now consider two cases in which the distribution of data is not symmetric.

**Case 1**
Consider the masses of a different set of 7 potatoes, given in grams below.

100 105 110 115 120 160 200

The median of this distribution is 115 g and the mean is 130 g. There are 5 observations that are less than the mean and only 2 that are more. In other words, the mean does not give us a good indication of the *centre* of the distribution. However, there is still a ‘balance’ between observations below the mean and those above, in terms of the spread of all the observations from the mean. Therefore, the mean is still useful to give a measure of the central tendency of the distribution but in cases where the distribution is skewed, the median gives a better indication of the centre. For a positively skewed distribution, as in the previous case, the mean will be greater than the median. For a negatively skewed distribution the mean will be less than the median.

**Case 2**
Consider the data below, showing the weekly income (to the nearest $10) of 10 families living in a suburban street.

$600 $1340 $1360 $1380 $1400 $1420 $1420 $1440 $1460 $1500

In this case, $\bar{x} = \frac{13320}{10} = $1332, and the median is $1410.

One of the values in this set, $600, is clearly an outlier. As a result, the value of the mean is below the weekly income of the other 9 households. In such a case the mean is not very useful in establishing the centre; however, the ‘balance’ still remains for this negatively skewed distribution.
The mean is calculated by using the values of the observations and because of this it becomes a less reliable measure of the centre of the distribution when the distribution is skewed or contains an outlier. Because the median is based on the order of the observations rather than their value, it is a better measure of the centre of such distributions.

**Worked Example 20**

Calculate the mean of the set of data shown.

10, 12, 15, 16, 18, 19, 22, 25, 27, 29

**THINK**

1. Write the formula for calculating the mean, where \( \sum x \) is the sum of all scores; \( n \) is the number of scores in the set.

2. Substitute the values into the formula and evaluate.

**WRITE**

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

\[ = \frac{10 + 12 + 15 + 16 + 18 + 19 + 22 + 25 + 27 + 29}{10} \]

\[ \bar{x} = 19.3 \]

The mean, \( \bar{x} \), is 19.3.

When calculating the mean of a data set, sometimes the answer you calculate will contain a long stream of digits after the decimal point.

For example, if we are calculating the mean of the data set 44, 38, 55, 61, 48, 32, 49

Then the mean would be:

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

\[ = \frac{327}{7} \]

\[ = 46.71428571... \]

In this case it makes sense to round the answer to either a given number of decimal places, or a given number of significant figures.

**Rounding to a given number of significant figures**

When rounding to a given number of signified figures, we are rounding to the digits in a number which are regarded as ‘significant’.

To determine which digits are significant, we can observe the following rules:

- All digits greater than zero are significant
- Leading zeros can be ignored (they are placeholders and are not significant)
- Zeros included between other digits are significant
- Zeros included after decimal digits are significant
- Trailing zeros for integers are not significant (unless specified otherwise)
The following examples show how these rules work:

- 0.003561 — leading digits are ignored, so this has 4 significant figures
- 70.036 — zeros between other digits are significant, so this has 5 significant figures
- 5.320 — zeros included after decimal digits are significant, so this has 4 significant figures
- 450000 — trailing zeros are not significant, so this has 2 significant figures
- 78000.0 — the zero after the decimal point is considered significant, so the zeros between other numbers are also significant; this has 6 significant figures

As when rounding to a given number of decimal places, when rounding to a given number of significant figures consider the digit after the specified number of figures. If it is 5 or above, round the final digit up; if it is 4 or below, keep the final digit as is.

- 5067.37 — rounded to 2 significant figures is 5100
- 3199.01 — rounded to 4 significant figures is 3199
- 0.004931 — rounded to 3 significant figures is 0.00493
- 1020004 — rounded to 2 significant figures is 1000000

Calculating the mean of grouped data

When data are presented in a frequency table with class intervals and we don’t know what the raw data are, we employ another method to find the mean of these grouped data. This other method is shown in the worked example that follows and uses the midpoints of the class intervals to represent the raw data.

Recall that the Greek letter sigma, \( \sum \), represents ‘the sum of’. So, \( \sum f \) means the sum of the frequencies and is the total of all the numbers in the frequency column.

To find the mean for grouped data,

\[
\bar{x} = \frac{\sum (f \times m)}{\sum f}
\]

where \( f \) represents the frequency of the data and \( m \) represents the midpoint of the class interval of the grouped data.

**WORKED EXAMPLE 21**

The ages of a group of 30 people attending a superannuation seminar are recorded in the frequency table.

<table>
<thead>
<tr>
<th>Age (class intervals)</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20–29</td>
<td>1</td>
</tr>
<tr>
<td>30–39</td>
<td>6</td>
</tr>
<tr>
<td>40–49</td>
<td>13</td>
</tr>
<tr>
<td>50–59</td>
<td>6</td>
</tr>
<tr>
<td>60–69</td>
<td>3</td>
</tr>
<tr>
<td>70–79</td>
<td>1</td>
</tr>
</tbody>
</table>

Calculate the mean age of those attending the seminar. Give your answer correct to 3 significant figures.
THINK

1 Since we don’t have individual raw ages, but rather a class interval, we need to decide on one particular age to represent each interval. We use the midpoint, \( m \), of the class interval. Add an extra column to the table to display these.

The midpoint of the first interval is \( \frac{20 + 29}{2} = 24.5 \), the midpoint of the second interval is 34.5 and so on.

2 Multiply each of the midpoints by the frequency and display these values in another column headed \( f \times m \). For the first interval we have \( 24.5 \times 1 = 24.5 \). For the second interval we have \( 34.5 \times 6 = 207 \) and so on.

3 Sum the product of the midpoints and the frequencies in the \( f \times m \) column.

\[
\sum f = 30 \\
\sum (f \times m) = 1405
\]

4 Divide this sum by the total number of people attending the seminar (given by the sum of the frequency column).

\[
\bar{x} = \frac{1405}{30} = 46.833\ldots \approx 46.8 \text{ (correct to 3 significant figures)}.
\]

EXERCISE 1.8 The mean of a sample

PRACTISE

1 WE20 Calculate the mean of the data set shown.
9, 12, 14, 16, 18, 19, 20, 25, 29, 33, 35, 36, 39

2 Calculate the mean of the data set shown.
5.5, 6.3, 7.7, 8.3, 9.7, 6.7, 12.9, 10.5, 9.9, 5.1

3 WE21 The number of hamburgers sold at a take-away food shop is recorded in the frequency table shown.

<table>
<thead>
<tr>
<th>Hamburgers sold (class intervals)</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–4</td>
<td>6</td>
</tr>
<tr>
<td>5–9</td>
<td>8</td>
</tr>
<tr>
<td>10–14</td>
<td>2</td>
</tr>
<tr>
<td>15–19</td>
<td>3</td>
</tr>
<tr>
<td>20–24</td>
<td>5</td>
</tr>
</tbody>
</table>

Calculate the mean number of hamburgers sold each day. Give your answer correct to 4 significant figures.
4 The ages of 100 supporters who attended the grand final parade are recorded in the frequency table shown.

<table>
<thead>
<tr>
<th>Age (class intervals)</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0−9</td>
<td>10</td>
</tr>
<tr>
<td>10−19</td>
<td>21</td>
</tr>
<tr>
<td>20−29</td>
<td>30</td>
</tr>
<tr>
<td>30−39</td>
<td>18</td>
</tr>
<tr>
<td>40−49</td>
<td>17</td>
</tr>
<tr>
<td>50−59</td>
<td>4</td>
</tr>
</tbody>
</table>

Calculate the mean age of those attending the parade. Give your answer correct to 2 decimal places.

5 Find the mean of each of the following sets of data.
   a 5 6 8 8 9 (correct to 2 significant figures)
   b 3 4 4 5 5 6 7 7 7 8 8 9 9 10 10 10 12 (correct to 4 significant figures)
   c 4.3 4.5 4.7 4.9 5.1 5.3 5.5 5.6 (correct to 5 significant figures)
   d 11 13 15 15 16 18 20 21 22 (correct to 1 decimal place)
   e 0.4 0.5 0.7 0.8 0.9 1.0 1.1 1.2 1.3 (correct to 4 decimal places)

6 Calculate the mean of each of the following and explain whether or not it gives us a good indication of the centre of the data.
   a 0.7 0.8 0.85 0.9 0.92 2.3          b 14 16 16 17 17 17 19 20
   c 23 24 28 29 33 34 37 39          d 2 15 17 18 18 19 20

7 The number of people attending sculpture classes at the local TAFE college for each week during the first semester is given.

15 12 15 11 14 14 15 11 10
7 11 12 14 15 14 15 9 10 11

What is the mean number of people attending each week? (Express your answer correct to 2 significant figures.)

8 The ages of a group of junior pilots joining an international airline are indicated in the stem plot below.

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

Key: 2|1 = 21 years

The mean age of this group of pilots is:
   A 20          B 28          C 29          D 29.15          E 29.5
9 The number of people present each week at a 15-week horticultural course is given by the stem plot at right.

The mean number of people attending each week was closest to:
A 17.7  B 18  C 19.5  D 20  E 21.2

10 For each of the following, write down whether the mean or the median would provide a better indication of the centre of the distribution.

a A positively skewed distribution
b A symmetric distribution
c A distribution with an outlier
d A negatively skewed distribution

11 Find the mean of each set of data given.

<table>
<thead>
<tr>
<th>Class interval</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–2</td>
<td>1</td>
</tr>
<tr>
<td>10–2</td>
<td>3</td>
</tr>
<tr>
<td>20–2</td>
<td>6</td>
</tr>
<tr>
<td>30–2</td>
<td>5</td>
</tr>
<tr>
<td>40–2</td>
<td>12</td>
</tr>
<tr>
<td>50–2</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class interval</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–2</td>
<td>2</td>
</tr>
<tr>
<td>5–2</td>
<td>7</td>
</tr>
<tr>
<td>10–2</td>
<td>8</td>
</tr>
<tr>
<td>15–2</td>
<td>14</td>
</tr>
<tr>
<td>20–2</td>
<td>12</td>
</tr>
<tr>
<td>25–2</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class interval</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–2</td>
<td>2</td>
</tr>
<tr>
<td>5–2</td>
<td>7</td>
</tr>
<tr>
<td>10–2</td>
<td>8</td>
</tr>
<tr>
<td>15–2</td>
<td>14</td>
</tr>
<tr>
<td>20–2</td>
<td>12</td>
</tr>
<tr>
<td>25–2</td>
<td>6</td>
</tr>
</tbody>
</table>

12 The ages of people attending a beginner’s course in karate are indicated in the following frequency table.

a What is the mean age of those attending the course? (Express your answer correct to 1 decimal place.)

b Calculate the median. What does this value, compared to the mean, suggest about the shape of the distribution?

<table>
<thead>
<tr>
<th>Age</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10–14</td>
<td>5</td>
</tr>
<tr>
<td>15–19</td>
<td>5</td>
</tr>
<tr>
<td>20–24</td>
<td>7</td>
</tr>
<tr>
<td>25–29</td>
<td>4</td>
</tr>
<tr>
<td>30–34</td>
<td>3</td>
</tr>
<tr>
<td>35–39</td>
<td>2</td>
</tr>
<tr>
<td>40–44</td>
<td>2</td>
</tr>
<tr>
<td>45–49</td>
<td>1</td>
</tr>
</tbody>
</table>
13 The number of papers sold each morning from a newsagent is recorded in the frequency table shown.

<table>
<thead>
<tr>
<th>Age (class intervals)</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50–99</td>
<td>2</td>
</tr>
<tr>
<td>100–149</td>
<td>18</td>
</tr>
<tr>
<td>150–199</td>
<td>25</td>
</tr>
<tr>
<td>200–249</td>
<td>33</td>
</tr>
<tr>
<td>250–299</td>
<td>21</td>
</tr>
<tr>
<td>300–349</td>
<td>9</td>
</tr>
<tr>
<td>350–399</td>
<td>2</td>
</tr>
</tbody>
</table>

Calculate the mean number of papers sold over this period.

14 The number of fish eaten by seals at Sea Haven on a daily basis is shown. Calculate the mean number of fish eaten per day.

<table>
<thead>
<tr>
<th>Number of fish</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–9</td>
<td>2</td>
</tr>
<tr>
<td>10–19</td>
<td>4</td>
</tr>
<tr>
<td>20–29</td>
<td>5</td>
</tr>
<tr>
<td>30–39</td>
<td>18</td>
</tr>
<tr>
<td>40–49</td>
<td>19</td>
</tr>
<tr>
<td>50–59</td>
<td>24</td>
</tr>
</tbody>
</table>

15 A shipping container is filled with cargo and each piece of cargo is weighed prior to being packed on the container. Using class intervals of 10 kg, calculate the mean weight (in kg) of the pieces of cargo. Give your answer correct to four significant figures.

<table>
<thead>
<tr>
<th>Weight (kg)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>78.3</td>
<td></td>
</tr>
<tr>
<td>67.3</td>
<td></td>
</tr>
<tr>
<td>82.8</td>
<td></td>
</tr>
<tr>
<td>44.3</td>
<td></td>
</tr>
<tr>
<td>90.5</td>
<td></td>
</tr>
<tr>
<td>57.3</td>
<td></td>
</tr>
<tr>
<td>55.9</td>
<td></td>
</tr>
<tr>
<td>42.3</td>
<td></td>
</tr>
<tr>
<td>48.8</td>
<td></td>
</tr>
<tr>
<td>63.4</td>
<td></td>
</tr>
<tr>
<td>69.7</td>
<td></td>
</tr>
<tr>
<td>70.4</td>
<td></td>
</tr>
<tr>
<td>77.1</td>
<td></td>
</tr>
<tr>
<td>79.4</td>
<td></td>
</tr>
<tr>
<td>47.6</td>
<td></td>
</tr>
<tr>
<td>52.9</td>
<td></td>
</tr>
<tr>
<td>45.4</td>
<td></td>
</tr>
<tr>
<td>60.1</td>
<td></td>
</tr>
<tr>
<td>73.4</td>
<td></td>
</tr>
<tr>
<td>88.6</td>
<td></td>
</tr>
<tr>
<td>41.9</td>
<td></td>
</tr>
<tr>
<td>63.7</td>
<td></td>
</tr>
</tbody>
</table>

16 The number of cups of coffee drunk by 176 Year 12 students in the two weeks leading to their exams is shown.

<table>
<thead>
<tr>
<th>Number of cups</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–9</td>
<td>3</td>
</tr>
<tr>
<td>10–19</td>
<td>5</td>
</tr>
<tr>
<td>20–29</td>
<td>7</td>
</tr>
<tr>
<td>30–39</td>
<td>24</td>
</tr>
<tr>
<td>40–49</td>
<td>29</td>
</tr>
<tr>
<td>50–59</td>
<td>41</td>
</tr>
<tr>
<td>60–69</td>
<td>32</td>
</tr>
<tr>
<td>70–79</td>
<td>30</td>
</tr>
<tr>
<td>80–89</td>
<td>3</td>
</tr>
<tr>
<td>90–99</td>
<td>2</td>
</tr>
</tbody>
</table>

a What is the mean number of cups of coffee drunk in the two-week period?
b Calculate the median.
c What does the median value when compared to the mean value suggest about the shape of the distribution?
1.9

Standard deviation of a sample

The standard deviation gives us a measure of how data are spread around the mean. For the set of data {8, 10, 11, 12, 12, 13}, the mean, $\bar{x} = 11$.

The amount that each observation ‘deviates’ (that is, differs) from the mean is calculated and shown in the table at right.

The deviations from the mean are either positive or negative depending on whether the particular observation is lower or higher in value than the mean. If we were to add all the deviations from the mean we would obtain zero.

If we square the deviations from the mean we will overcome the problem of positive and negative deviations cancelling each other out. With this in mind, a quantity known as sample variance ($s^2$) is defined:

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

Technically, this formula for variance is used when the data set is a sub-set of a larger population; that is, a sample of the population.

Variance gives the average of the squared deviations and is also a measure of spread. A far more useful measure of spread, however, is the standard deviation, which is the square root of variance ($s$). One reason for it being more useful is that it takes the same unit as the observations (for example, cm or number of people). Variance would square the units, for example, cm$^2$ or number of people squared, which is not very practical.

Other advantages of the standard deviation will be dealt with later in the topic.

**In summary,**

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

where $s$ represents sample standard deviation

$\sum$ represents ‘the sum of’

$x$ represents an observation

$\bar{x}$ represents the mean

$n$ represents the number of observations.

**Note:** This is the formula for the standard deviation of a sample. The standard deviation for a population is given by $\sigma$ (sigma) and is calculated using a slightly different formula, which is outside the scope of this course.

While some of the theory or formulas associated with standard deviation may look complex, the calculation of this measure of spread is straightforward using CAS. Manual computation of standard deviation is therefore rarely necessary, however application of the formula is required knowledge.
The price (in cents) per litre of petrol at a service station was recorded each Friday over a 15-week period. The data are given below.

152.4  160.2  159.6  168.6  161.4  156.6  164.8  162.6
161.0  156.4  159.0  160.2  162.6  168.4  166.8

Calculate the standard deviation for this set of data, correct to 2 decimal places.

**THINK**
1. Enter the data into CAS to determine the sample statistics.
2. Round the value correct to 2 decimal places.

**WRITE**

1. Enter the data into CAS to determine the sample statistics. 
   \[ s = 4.51592 \]
2. Round the value correct to 2 decimal places. 
   \[ = 4.52 \text{ cents/L} \]

The number of students attending SRC meetings during the term is given in the stem plot shown. Calculate the standard deviation for this set of data, correct to 4 significant figures, by using the formula \( s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \).

**THINK**
1. Calculate the value of the mean (\( \bar{x} \)).
2. Set up a table to calculate the values of \((x - \bar{x})^2\).

**WRITE**

1. Calculate the value of the mean (\( \bar{x} \)). 
   \[ \bar{x} = \frac{\sum x}{n} = \frac{4 + 8 + 8 + 11 + 13 + 14 + 15 + 18 + 23 + 25}{10} = 13.9 \]
2. Set up a table to calculate the values of \((x - \bar{x})^2\).

\[
\begin{array}{|c|c|c|}
\hline
x & x - \bar{x} & (x - \bar{x})^2 \\
\hline
4 & -9.9 & 98.01 \\
8 & -5.9 & 34.81 \\
8 & -5.9 & 34.81 \\
11 & -2.9 & 8.41 \\
13 & -0.9 & 0.81 \\
14 & 0.1 & 0.01 \\
15 & 1.1 & 1.21 \\
18 & 4.1 & 16.81 \\
23 & 9.1 & 82.81 \\
25 & 11.1 & 123.21 \\
\hline
\end{array}
\]

\[ \sum (x - \bar{x})^2 = 400.9 \]
3 Enter the values into the formula to calculate the standard deviation \( s \).

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

\[
= \sqrt{\frac{400.9}{9}}
\]

\[
= 6.6741\ldots
\]

4 Round the value correct to 4 significant figures.

\[
= 6.674 \text{ (correct to 4 significant figures)}
\]

The standard deviation is a measure of the spread of data from the mean. Consider the two sets of data shown.

Each set of data has a mean of 10. The top set of data has a standard deviation of 1 and the bottom set of data has a standard deviation of 3.

As we can see, the larger the standard deviation, the more spread are the data from the mean.

**EXERCISE 1.9**

**Standard deviation of a sample**

1. The Australian dollar is often compared to the US dollar. The value of the Australian dollar compared to the US dollar each week over a six-month period is shown.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
</tr>
</tbody>
</table>

Calculate the standard deviation for this set of data, correct to 2 decimal places.

2. The test results for a maths class are shown.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>2</td>
</tr>
<tr>
<td>98</td>
<td>1</td>
</tr>
<tr>
<td>75</td>
<td>1</td>
</tr>
<tr>
<td>81</td>
<td>1</td>
</tr>
<tr>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>55</td>
<td>1</td>
</tr>
<tr>
<td>52</td>
<td>1</td>
</tr>
<tr>
<td>78</td>
<td>1</td>
</tr>
<tr>
<td>90</td>
<td>1</td>
</tr>
<tr>
<td>76</td>
<td>1</td>
</tr>
</tbody>
</table>

Calculate the standard deviation for the set of data, correct to 2 decimal places.
3. Calculate the standard deviation of the data set shown, correct to 4 significant places, by using the formula

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

4. The number of students attending the service learning meetings in preparation for the year’s fundraising activities is shown by the stem plot. Calculate the standard deviation for this set of data, correct to 3 decimal places, by using the formula

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

5. For each of the following sets of data, calculate the standard deviation correct to 2 decimal places.

   a. 3 4 4.7 5.1 6 6.2
   b. 7 9 10 10 11 13 14
   c. 12.9 17.2 17.9 20.2 26.4 28.9
   d. 41 43 44 45 46 47 49
   e. 0.30 0.32 0.37 0.39 0.41 0.43 0.45

6. First-quarter profit increases for 8 leading companies are given below as percentages.

   2.3 0.8 1.6 2.1 1.7 1.3 1.4 1.9

   Calculate the standard deviation for this set of data and express your answer correct to 2 decimal places.

7. The heights in metres of a group of army recruits are given below.

   1.8 1.95 1.87 1.77 1.75 1.79 1.81 1.83 1.76 1.80 1.92 1.87 1.85 1.83

   Calculate the standard deviation for this set of data and express your answer correct to 2 decimal places.

8. Times (correct to the nearest tenth of a second) for the heats in the 100 m sprint at the school sports carnival are given at right.

   11 0
   11 2 3
   11 4 4 5
   11 6 6
   11 8 8 9
   12 0 1
   12 2 2 3
   12 4 4
   12 6
   12 9

   Key: 11|0 = 11.0 s
9 The number of outgoing phone calls from an office each day over a 4-week period is shown in the stem plot below.

| Stem | Leaf | Key: 2|1 = 21 calls |
|------|------|------|
| 0    | 8 9  |      |
| 1    | 3 4 7 9 |     |
| 2    | 0 1 3 7 7 |    |
| 3    | 3 4  |      |
| 4    | 1 5 6 7 8 |   |
| 5    | 3 8  |      |

Calculate the standard deviation for this set of data and express your answer correct to 4 significant figures.

10 A new legal aid service has been operational for only 5 weeks. The number of people who have made use of the service each day during this period is set out at right. The standard deviation (to 2 decimal places) of these data is:

A 6.00
B 6.34
C 6.47
D 15.44
E 16.00

11 The speed of 20 cars (in km/h) is monitored along a stretch of road that is a designated 80 km/h zone. Calculate the standard deviation of the data, correct to 2 decimal places.

80, 82, 77, 75, 80, 80, 81, 78, 79, 78, 80, 80, 85, 70, 79, 81, 81, 80, 80, 80

12 Thirty pens are randomly selected off the conveyor belt at the factory and are tested to see how long they will last, in hours. Calculate the standard deviation of the data shown, correct to 3 decimal places.

20, 32, 38, 22, 25, 34, 47, 31, 26, 29, 30, 36, 28, 40, 31, 26, 37, 38, 32, 36, 35, 25, 29, 30, 40, 35, 38, 39, 37, 30

13 Calculate the standard deviation of the data shown, correct to 2 decimal places, representing the temperature of the soil around 25 germinating seedlings:

<table>
<thead>
<tr>
<th>Temperature (°C)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>28.9</td>
<td></td>
</tr>
<tr>
<td>27.4</td>
<td></td>
</tr>
<tr>
<td>23.6</td>
<td></td>
</tr>
<tr>
<td>25.6</td>
<td></td>
</tr>
<tr>
<td>21.1</td>
<td></td>
</tr>
<tr>
<td>22.9</td>
<td></td>
</tr>
<tr>
<td>29.6</td>
<td></td>
</tr>
<tr>
<td>25.7</td>
<td></td>
</tr>
<tr>
<td>27.4</td>
<td></td>
</tr>
<tr>
<td>23.6</td>
<td></td>
</tr>
<tr>
<td>22.4</td>
<td></td>
</tr>
<tr>
<td>24.6</td>
<td></td>
</tr>
<tr>
<td>21.8</td>
<td></td>
</tr>
<tr>
<td>26.4</td>
<td></td>
</tr>
<tr>
<td>24.9</td>
<td></td>
</tr>
<tr>
<td>25.0</td>
<td></td>
</tr>
<tr>
<td>23.5</td>
<td></td>
</tr>
<tr>
<td>26.1</td>
<td></td>
</tr>
</tbody>
</table>

14 Aptitude tests are often used by companies to help decide who to employ. An employer gave 30 potential employees an aptitude test with a total of 90 marks. The scores achieved are shown.

67, 67, 68, 68, 68, 69, 69, 72, 72, 73, 73, 74, 74, 74, 75, 75, 77, 78, 78, 79, 79, 79, 79, 81, 81, 81, 82, 83, 83, 83, 86

Calculate the mean and standard deviation of the data, correct to 1 decimal place.
15 The number of players attending basketball try-out sessions is shown by the stem plot. Calculate the standard deviation for this set of data, correct to 4 significant figures.

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

Key: 2|1 = 21

16 The scores obtained out of 20 at a dancing competition are shown. Calculate the standard deviation of the scores, correct to 3 decimal places.

<table>
<thead>
<tr>
<th>18.5</th>
<th>16.5</th>
<th>18.0</th>
<th>12.5</th>
<th>13.0</th>
<th>18.0</th>
<th>15.5</th>
<th>17.5</th>
<th>18.5</th>
<th>19.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.0</td>
<td>12.5</td>
<td>16.5</td>
<td>13.5</td>
<td>19.0</td>
<td>20.0</td>
<td>17.5</td>
<td>19.5</td>
<td>16.0</td>
<td>15.5</td>
</tr>
</tbody>
</table>

1.10 Populations and simple random samples

Populations

A group of Year 12 students decide to base their statistical investigation for a maths project on what their contemporaries — that is, other Year 12 students — spend per year on Christmas and birthday presents for their family members. One of their early decisions is to decide what the population is going to be for their investigation. That is, are they looking at Year 12 students in Australia or in Victoria or in metropolitan Melbourne or in their suburb or just in their school? In practice, it is difficult to look at a large population unless, of course, you have a lot of resources available to you! The students decide that their population will be the Year 12 students at their school. This means that any conclusions they draw as a result of their investigation can be generalised to Year 12 students at their school but not beyond that.

Samples

Given that there are 95 students in Year 12 at the school, it would be too time-consuming to interview all of them. A smaller group known as a sample is therefore taken from the population. The way in which this smaller group is chosen is of paramount importance. For the investigation to have credibility, the sample should be a random selection from the population and every member of that population should have an equal chance of being chosen in the sample. Also, the selection of one person from the population should not affect whether or not another person is chosen; that is, the selections should be independent. A simple random sample provides such a sample.

The students conducting the investigation decide to choose a sample of 12 fellow students. While it would be simplest to choose 12 of their mates as the sample, this would introduce bias since they would not be representative of the population as a whole.

The students obtain a list of names of the 95 students in Year 12. They then write next to the name of each student a number from 1 to 95. Using a calculator, the students generate 12 random numbers between 1 and 95. Alternatively, the students could have used a table of random numbers. Any point on the table can be taken as the starting point. The
students decide which direction to move through the table; for example, across the table to the right or to the left or down. Once a direction is chosen, they must stay with that movement and write down the 2-digit numbers as they go along.

The numbers chosen by the students are then matched to the numbers on the name list and the students in their sample can be identified.

These 12 students are then asked what they spent in the last year on family presents.

The students conducting the investigation can then record the data.

Random numbers can also be generated with the aid of CAS.

The mean of a data set which represents a population is \( \mu \).

The mean of a data set which represents a sample is \( \bar{x} \).

The standard deviation of a data set which represents a population is \( \sigma \).

The standard deviation of a data set which represents a sample is \( s \).

**Displaying the data**

The raw data for a sample of 12 students from our population of 95 are given below in dollars.

25 30 35 38 34 22 30 40 35 25 32 40

Since there are not many responses, a stem plot is an appropriate way of displaying the data.

To summarise and comment further on the sample, it is useful to use some of the summary statistics covered earlier in this topic. The most efficient way to calculate these is to use CAS. Using the steps outlined in the previous sections, we obtain a list of summary statistics for these data.

\[
\begin{align*}
\bar{x} &= 32.2 \\
\text{median} &= 33 \\
Q_1 &= 27.5 \\
Q_3 &= 36.5
\end{align*}
\]

To measure the centre of the distribution, the median and the mean are used. Since there are no outliers and the distribution is approximately symmetric, the mean is quite a good measure of the centre of the distribution. Also, the mean and the median are quite close in value.

---

**WORKED EXAMPLE 24**

Generate 5 random numbers (integers) between 1 and 50.

**THINK**

1. Find the appropriate menu in CAS to generate random integers.
2. Generate 5 random numbers between 1 and 50.
3. An example of a set of numbers is displayed.

**WRITE**

\[\{48, 46, 8, 26, 21\}\].

---
To measure the spread of the distribution, the standard deviation and the interquartile range are used. Since \( s = 6 \), and since the distribution is approximately bell-shaped, we would expect that approximately 95\% of the data lie between 32.2 + 12 = 44.2 and 32.2 - 12 = 20.2 (as shown in section 1.11). It is perhaps a little surprising to think that 95\% of students spend between $20.20 and $44.20 on family presents. One might have expected there to be greater variation on what students spend. The data, in that sense, are quite bunched.

The interquartile range is equal to 36.5 - 27.5 = 9. This means that 50\% of those in the sample spent within $9 of each other on family presents. Again, one might have expected a greater variation in what students spent. It would be interesting to know whether students confer about what they spend and therefore whether they tended to allocate about the same amount of money to spend.

At another school, the same investigation was undertaken and the results are shown in the following stem plot.

The summary statistics for these data are as follows:

\[ \bar{x} = 47.5, \ s = 16.3, \ Q_1 = 35, \ \text{median} = 50, \ Q_3 = 60. \]

The distribution is approximately symmetric, albeit very spread out. The mean and the median are therefore reasonably close and give us an indication of the centre of the distribution. The mean value for this set of data is higher than for the data obtained at the other school. This indicates that students at this school in this year level, in general, spend more than their counterparts at the other school. Reasons for this might be that this school is in a higher socio-economic area and students receive greater allowances, or perhaps at this school there is a higher proportion of students from cultures where spending more money on family presents is usual.

The range of money spent on family presents at this school and at this particular year level is $55. This is certainly much higher than at the other school. The interquartile range at this school is $25. That is, the middle 50\% of students spend within $25 of each other which is greater than the students at the other school.

**EXERCISE 1.10**

**Populations and simple random samples**

1. **PRACTISE** Generate 5 random numbers (integers) between 1 and 100.
2. Generate 10 random numbers (integers) between 1 and 250.
3 Students are selecting a sample of students at their school to complete an investigation. Which of the following are examples of choosing this sample randomly?

A Choosing students queuing at the tuckshop
B Assigning numbers to a list of student names and using a random number table to select random numbers
C Calling for volunteers
D Choosing the girls in an all-girls science class
E Choosing students in a bus on the way home

4 Generate 10 random numbers (integers) between 1 and 100.
5 Generate 20 random numbers (integers) between 1 and 500.

6 Which is larger: A population or a sample? Explain why.

7 When selecting students for a simple random sample of a year level, the students selected should be:

A of similar age
B a group of mates
C independent
D female
E the tallest students

8 The students selected for a simple random sample of a year level should be selected by:

A a group of mates
B a group who all dance
C a selection of males
D the students with the best test results
E using random numbers

9 Would the mean be a good measure of the centre of the distribution shown at right? Explain.

10 The mean is a good measure of the centre of a distribution if the data is:

A skewed left
B symmetric
C skewed right
D has outliers
E bimodal

11 The interquartile range is 12, since \( Q_1 = 24 \) and \( Q_3 = 36 \). The percentage of data that fit between 24 and 36 is:

A 12%
B 30%
C 50%
D 68%
E 95%

12 Conduct an investigation into how much money students in your year level earn per week (this might be an allowance or a wage). Write a report on your findings, ensuring you include:
a an explanation of the population for your investigation
b the manner in which your sample was selected
c the number in your sample
d your results as raw data
e your results in a stem plot or histogram
f the summary statistics for your data.
Comment on your results based on the summary statistics.

13 Repeat question 12, but this time investigate the following for students in your year level:
   a the number of hours spent on homework each week
   b the number of hours spent working in part-time jobs.

14 Conduct a similar investigation to that which you completed in questions 12 and 13; however, this time sample students in another year group. Compare these data with those obtained for your year level.

The 68–95–99.7% rule and z-scores

The 68–95–99.7% rule

The heights of a large number of students (a population) at a graduation ceremony were recorded and are shown in the histogram at right.

This set of data is approximately symmetric and has what is termed a bell shape. Many sets of data fall into this category and are often referred to as normal distributions. Examples are birth weights and people’s heights. Data which are normally distributed have their symmetrical, bell-shaped distribution centred on the mean value, \( \mu \) (the mean of the population).

A feature of this type of distribution is that we can predict what percentage of the data lie 1, 2 or 3 standard deviations (\( \sigma \), the standard deviation of the population) either side of the mean using what is termed the 68–95–99.7% rule.

The 68–95–99.7% rule for a bell-shaped curve states that approximately:
1. 68% of data lie within 1 standard deviation either side of the mean
2. 95% of data lie within 2 standard deviations either side of the mean
3. 99.7% of data lie within 3 standard deviations either side of the mean.
In Figure 1, 68% of the data shown lie between the value which is 1 standard deviation below the mean, that is $\mu - \sigma$, and the value which is 1 standard deviation above the mean, that is $\mu + \sigma$.

In Figure 2, 95% of the data shown lie between the value which is 2 standard deviations below the mean, that is $\mu - 2\sigma$, and the value which is 2 standard deviations above the mean, that is $\mu + 2\sigma$.

In Figure 3, 99.7% of the data shown lie between the value which is 3 standard deviations below the mean, that is $\mu - 3\sigma$, and the value which is 3 standard deviations above the mean, that is $\mu + 3\sigma$.

The wrist circumferences of a large group of people were recorded and the results are shown in the histogram below. The mean of the set of data is 17.7 and the standard deviation is 0.9. Write down the wrist circumferences between which we would expect approximately:

a 68% of the group to lie
b 95% of the group to lie
c 99.7% of the group to lie.

**THINK**

a The distribution can be described as approximately bell-shaped and therefore the 68–95–99.7% rule can be applied. Approximately 68% of the people have a wrist circumference between $\mu - \sigma$ and $\mu + \sigma$ (or one standard deviation either side of the mean).

b Similarly, approximately 95% of the people have a wrist size between $\mu - 2\sigma$ and $\mu + 2\sigma$.

c Similarly, approximately 99.7% of the people have a wrist size between $\mu - 3\sigma$ and $\mu + 3\sigma$.

**WRITE**

a $\mu - \sigma = 17.7 - 0.9 = 16.8$
$\mu + \sigma = 17.7 + 0.9 = 18.6$
So approximately 68% of the people have a wrist size between 16.8 and 18.6 cm.

b $\mu - 2\sigma = 17.7 - 1.8 = 15.9$
$\mu + 2\sigma = 17.7 + 1.8 = 19.5$
Approximately 95% of people have a wrist size between 15.9 cm and 19.5 cm.

c $\mu - 3\sigma = 17.7 - 2.7 = 15.0$
$\mu + 3\sigma = 17.7 + 2.7 = 20.4$
Approximately 99.7% of people have a wrist size between 15.0 cm and 20.4 cm.

Using the 68–95–99.7% rule, we can work out the various percentages of the distribution which lie between the mean and 1 standard deviation from the
mean and between the mean and 2 standard deviations from the mean and so on. The diagram at right summarises this.

Note that 50% of the data lie below the mean and 50% lie above the mean due to the symmetry of the distribution about the mean.

The distribution of the masses of a large number of packets of ‘Fibre-fill’ breakfast cereal is known to be bell-shaped with a mean of 250 g and a standard deviation of 5 g. Find the percentage of Fibre-fill packets with a mass which is:

**a** less than 260 g  
**b** less than 245 g  
**c** more than 240 g  
**d** between 240 g and 255 g.

**THINK**

1. Draw the bell-shaped curve. Label the axis. $\mu = 250$, $\mu + \sigma = 255$, $\mu + 2\sigma = 260$ etc.

**WRITE/DRAW**

**a** 260 g is 2 standard deviations above the mean. Using the summary diagram, we can find the percentage of data which is less than 260 g.

**b** 245 g is 1 standard deviation below the mean.

**c** 240 g is 2 standard deviations below the mean.

**d** Now, 240 g is 2 standard deviations below the mean while 255 g is 1 standard deviation above the mean.

**WORKED EXAMPLE 26**

The distribution of the masses of a large number of packets of ‘Fibre-fill’ breakfast cereal is known to be bell-shaped with a mean of 250 g and a standard deviation of 5 g. Find the percentage of Fibre-fill packets with a mass which is:

**a** less than 260 g  
**b** less than 245 g  
**c** more than 240 g  
**d** between 240 g and 255 g.

**THINK**

1. Draw the bell-shaped curve. Label the axis. $\mu = 250$, $\mu + \sigma = 255$, $\mu + 2\sigma = 260$ etc.

**WRITE/DRAW**

**a** Mass of 260 g is 2 standard deviations above the mean. Percentage of distribution less than 260 g is $13.5\% + 34\% + 34\% + 13.5\% + 2.35\% + 0.15\% = 97.5\%$
   
   or $13.5\% + 34\% + 50\% = 97.5\%$

**b** Mass of 245 g is 1 standard deviation below the mean. Percentage of distribution less than 245 g is $13.5\% + 2.35\% + 0.15\%$
   
   = 16\%  
   
   or $50\% - 34\%$
   
   = 16\%

**c** Mass of 240 g is 2 standard deviations below the mean. Percentage of distribution more than 240 g is $13.5\% + 34\% + 34\% + 13.5\% + 2.35\% + 0.15\%$
   
   = 97.5\%  
   
   or $13.5\% + 34\% + 50\%$
   
   = 97.5\%

**d** Mass of 240 g is 2 standard deviations below the mean. Mass of 255 g is 1 standard deviation above the mean. Percentage of distribution between 240 g and 255 g is $13.5\% + 34\% + 34\% = 81.5\%$
The number of matches in a box is not always the same. When a sample of boxes was studied it was found that the number of matches in a box approximated a normal (bell-shaped) distribution with a mean number of matches of 50 and a standard deviation of 2. In a sample of 200 boxes, how many would be expected to have more than 48 matches?

THINK

1 Find the percentage of boxes with more than 48 matches. Since 48 = 50 − 2, the score of 48 is 1 standard deviation below the mean.

2 Find 84% of the total sample.

WRITE

48 matches is 1 standard deviation below the mean. Percentage of boxes with more than 48 matches

\[ \text{Percentage} = \frac{34\% + 50\%}{2} = 84\% \]

Number of boxes = 84% of 200

\[ = 168 \text{ boxes} \]

Standard z-scores

To find a comparison between scores in a particular distribution or in different distributions, we use the z-score. The z-score (also called the standardised score) indicates the position of a certain score in relation to the mean.

A z-score of 0 indicates that the score obtained is equal to the mean, a negative z-score indicates that the score is below the mean and a positive z-score indicates a score above the mean.

The z-score measures the distance from the mean in terms of the standard deviation. A score that is exactly one standard deviation above the mean has a z-score of 1. A score that is exactly one standard deviation below the mean has a z-score of −1.

To calculate a z-score we use the formula:

\[ z = \frac{x - \mu}{\sigma} \]

where \( x \) = the score, \( \mu \) = the mean of the population and \( \sigma \) = the standard deviation of the population.

WORKED EXAMPLE 27

In an IQ test, the mean IQ is 100 and the standard deviation is 15. Dale’s test results give an IQ of 130. Calculate this as a z-score.

THINK

1 Write the formula.

2 Substitute for \( x \), \( \mu \) and \( \sigma \).

3 Calculate the z-score.

WRITE

\[ z = \frac{x - \mu}{\sigma} = \frac{130 - 100}{15} = \frac{30}{15} = 2 \]

Dale’s z-score is 2, meaning that his IQ is exactly two standard deviations above the mean.

Not all z-scores will be whole numbers; in fact most will not be. A whole number indicates only that the score is an exact number of standard deviations above or below the mean.
Using Worked example 28, an IQ of 88 would be represented by a z-score of \(-0.8\), as shown below.

\[
z = \frac{x - \mu}{\sigma}
\]
\[
= \frac{88 - 100}{15}
\]
\[
= -0.8
\]

The negative value indicates that the IQ of 88 is below the mean but by less than one standard deviation.

### Comparing data

An important use of z-scores is to compare scores from different data sets. Suppose that in your maths exam your result was 74 and in English your result was 63. In which subject did you achieve the better result?

At first glance, it may appear that the maths result is better, but this does not take into account the difficulty of the test. A mark of 63 on a difficult English test may in fact be a better result than 74 if it was an easy maths test.

The only way that we can fairly compare the results is by comparing each result with its mean and standard deviation. This is done by converting each result to a z-score.

If, for maths, \(\mu = 60\) and \(\sigma = 12\), then

\[
z = \frac{x - \mu}{\sigma}
\]
\[
= \frac{74 - 60}{12}
\]
\[
= 1.17
\]

And if, for English, \(\mu = 50\) and \(\sigma = 8\), then

\[
z = \frac{x - \mu}{\sigma}
\]
\[
= \frac{63 - 50}{8}
\]
\[
= 1.625
\]

The English result is better because the higher z-score shows that the 63 is higher in comparison to the mean of each subject.

---

**WORKED EXAMPLE 29**

Janine scored 82 in her physics exam and 78 in her chemistry exam. In physics, \(\mu = 62\) and \(\sigma = 10\), while in chemistry, \(\mu = 66\) and \(\sigma = 5\).

**a** Write both results as a standardised score.

**b** Which is the better result? Explain your answer.

**THINK**

**a** 1 Write the formula for each subject.

2 Substitute for \(x\), \(\mu\) and \(\sigma\).

3 Calculate each z-score.

**b** Explain that the subject with the highest z-score is the better result.

**WRITE**

**a** Physics: \(z = \frac{x - \mu}{\sigma}\)

\[
= \frac{82 - 62}{10}
\]
\[
= 2
\]

Chemistry: \(z = \frac{x - \mu}{\sigma}\)

\[
= \frac{78 - 66}{5}
\]
\[
= 2.4
\]

**b** The chemistry result is better because of the higher z-score.
In each example the circumstances must be analysed carefully to see whether a higher or lower $z$-score is better. For example, if we were comparing times for runners over different distances, the lower $z$-score would be the better one.

**EXERCISE 1.11  The 68–95–99.7% rule and $z$-scores**

1. **WE25** The concentration ability of a large group of adults is tested during a short task which they are asked to complete. The length of the concentration span of those involved during the task is shown.

The mean, $\mu$, is 49 seconds and the standard deviation, $\sigma$, is 14 seconds.

Write down the values between which we would expect approximately:

- **a** 68% of the group’s concentration spans to fall
- **b** 95% of the group’s concentration spans to fall
- **c** 99.7% of the group’s concentration spans to fall.

2. The monthly rainfall in Mathmania Island was found to follow a bell-shaped curve with a mean of 45 mm and a standard deviation of 1.7 mm. Write down the rainfall range which we would expect approximately:

- **a** 68% of the group to lie
- **b** 95% of the group to lie
- **c** 99.7% of the group to lie.

3. **WE26** The distribution of masses of potato chips is known to follow a bell-shaped curve with a mean of 200 g and a standard deviation of 7 g. Find the percentage of the potato chips with a mass which is:

- **a** more than 214 g
- **b** more than 200 g
- **c** less than 193 g
- **d** between 193 g and 214 g.

4. The distribution of heights of a group of Melbourne-based employees who work for a large international company is bell-shaped. The data have a mean of 160 cm and a standard deviation of 10 cm.

Find the percentage of this group of employees who are:

- **a** less than 170 cm tall
- **b** less than 140 cm tall
- **c** greater than 150 cm tall
- **d** between 130 cm and 180 cm in height.

5. **WE27** The number of marbles in a bag is not always the same. From a sample of boxes it was found that the number of marbles in a bag approximated a normal distribution with a mean number of 20 and a standard deviation of 1. In a sample of 500 bags, how many would be expected to have more than 19 marbles?

6. The volume of fruit juice in a certain type of container is not always the same. When a sample of these containers was studied it was found that the volume of juice they contained approximated a normal distribution with a mean of 250 mL.
and a standard deviation of 5 mL. In a sample of 400 containers, how many would be expected to have a volume of:

a more than 245 mL
b less than 240 mL
c between 240 and 260 mL?

7 In a Physics test on electric power, the mean result for the class was 76% and a standard deviation of 9%. Drew’s result was 97%. Calculate his mark as a z-score.

8 In a maths exam, the mean score is 60 and the standard deviation is 12. Chifune’s mark is 96. Calculate her mark as a z-score.

9 Bella’s Specialist Maths mark was 83 and her English mark was 88. In Specialist Maths the mean was 67 with a standard deviation of 9, while in English the mean was 58 with a standard deviation of 14.
   a Convert the marks in each subject to a z-score.
   b Which subject is the better result for Bella? Explain.

10 Ken’s English mark was 75 and his maths mark was 72. In English, the mean was 65 with a standard deviation of 8, while in maths the mean mark was 56 with a standard deviation of 12.
   a Convert the mark in each subject to a z-score.
   b In which subject did Ken perform better? Explain your answer.

11 In each of the following, decide whether or not the distribution is approximately bell-shaped.

a

b

c
d

e

f
12 Copy and complete the entries on the horizontal scale of the following distributions, given that $\mu = 10$ and $\sigma = 2$.

13 Copy and complete the entries on the horizontal scale of the following distributions, given that $\mu = 5$ and $\sigma = 1.3$.

14 A research scientist measured the rate of hair growth in a group of hamsters. The findings are shown in the histogram.
The mean growth per week was 1.9 mm and the standard deviation was 0.6 mm. Write down the hair growth rates between which approximately:
   a 68% of the values fall
   b 95% of the values fall
   c 99.7% of the values fall.

15 The heights of the seedlings sold in a nursery have a bell-shaped distribution.
The mean height is 7 cm and the standard deviation is 2.
Write down the values between which approximately:
   a 68% of seedling heights will lie
   b 95% of seedling heights will lie
   c 99.7% of seedling heights will lie.

16 A distribution of scores is bell-shaped and the mean score is 26. It is known that 95% of scores lie between 21 and 31.
It is true to say that:
   A 68% of the scores lie between 23 and 28
   B 97.5% of the scores lie between 23.5 and 28.5
C the standard deviation is 2.5  
D 99.7% of the scores lie between 16 and 36  
E the standard deviation is 5

17 The number of days taken off in a year by employees of a large company has a distribution which is approximately bell-shaped. The mean and standard deviation of this data are shown below.  
Mean = 9 days  
Standard deviation = 2 days

Find the percentage of employees of this company who, in a year, take off:

a more than 15 days  
b fewer than 5 days  
c more than 7 days  
d between 3 and 11 days  
e between 7 and 13 days.

18 A particular bolt is manufactured such that the length is not always the same. The distribution of the lengths of the bolts is approximately bell-shaped with a mean length of 2.5 cm and a standard deviation of 1 mm.

a In a sample of 2000 bolts, how many would be expected to have a length:
   i between 2.4 cm and 2.6 cm  
   ii less than 2.7 cm  
   iii between 2.6 cm and 2.8 cm?

b The manufacturer rejects bolts which have a length of less than 2.3 cm or a length of greater than 2.7 cm. In a sample of 2000 bolts, how many would the manufacturer expect to reject?

19 In a major exam, every subject has a mean score of 60 and a standard deviation of 12.5. Clarissa obtains the following marks on her exams. Express each as a z-score.

   a English 54  
   b Maths 78  
   c Biology 61  
   d Geography 32  
   e Art 95

20 In a normal distribution the mean is 58. A score of 70 corresponds to a standardised score of 1.5. The standard deviation of the distribution is:

A 6  
B 8  
C 10  
D 12  
E 9

21 In the first maths test of the year, the mean mark was 60 and the standard deviation was 12. In the second test, the mean was 55 and the standard deviation was 15. Barbara scored 54 in the first test and 50 in the second test. In which test did Barbara do better? Explain your answer.

22 The table below shows the average number of eggs laid per week by a random sample of chickens with 3 different types of living conditions.

<table>
<thead>
<tr>
<th>Number of eggs per week</th>
<th>Cage chickens</th>
<th>Barn chickens</th>
<th>Free range chickens</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>4.8</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>4.9</td>
<td>4.6</td>
<td>3.8</td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>4.3</td>
<td>4.1</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
Number of eggs per week

<table>
<thead>
<tr>
<th></th>
<th>Cage chickens</th>
<th>Barn chickens</th>
<th>Free range chickens</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4</td>
<td>4.7</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>4.2</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>5.8</td>
<td>3.9</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>5.6</td>
<td>4.9</td>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>4.1</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>4.0</td>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td>4.9</td>
<td>4.4</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td>4.5</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>4.6</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>4.1</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>4.2</td>
<td>4.1</td>
<td></td>
</tr>
</tbody>
</table>

a Copy and complete the following table by calculating the mean and standard deviation of barn chickens and free range chickens correct to 1 decimal place.

<table>
<thead>
<tr>
<th>Living conditions</th>
<th>Cage</th>
<th>Barn</th>
<th>Free range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b A particular free range chicken lays an average of 4.3 eggs per week. Calculate the z-score relative to this sample correct to 3 significant figures.

The number of eggs laid by free range chickens is normally distributed. A free range chicken has a z-score of 1.

c Approximately what percentage of chickens lay fewer eggs than this chicken?

d Referring to the table showing the number of eggs per week, prepare five-number summaries for each set of data.

i State the median of each set of data.

ii What could be concluded about the egg-producing capabilities of chickens in different living conditions?
The Maths Quest Review is available in a customisable format for you to demonstrate your knowledge of this topic.

The Review contains:
- **Multiple-choice** questions — providing you with the opportunity to practise answering questions using CAS technology
- **Short-answer** questions — providing you with the opportunity to demonstrate the skills you have developed to efficiently answer questions using the most appropriate methods
- **Extended-response** questions — providing you with the opportunity to practise exam-style questions.

A summary of the key points covered in this topic is also available as a digital document.

**REVIEW QUESTIONS**

Download the Review questions document from the links found in the Resources section of your eBookPLUS.

**studyON** is an interactive and highly visual online tool that helps you to clearly identify strengths and weaknesses prior to your exams. You can then confidently target areas of greatest need, enabling you to achieve your best results.

To access eBookPLUS activities, log on to [www.jacplus.com.au](http://www.jacplus.com.au)

**Interactives**

A comprehensive set of relevant interactives to bring difficult mathematical concepts to life can be found in the Resources section of your eBookPLUS.
1 Answers

EXERCISE 1.2
1 D  2 D
3 C  4 B
5 Numerical: a, b, c
6 Categorical: c, d, e, f, g
7 Discrete: c
Continuous: a, b
8 C
9 C
10 Categorical
11 B
12 A
13 Categorical and ordinal
14 Discrete
15 Ordinal
16 D

EXERCISE 1.3
1 Stem | Leaf
1 168
2 15889
3 013589
4 289
Key: 1|6 = 16

2 Stem | Leaf
0 5
1 189
2 379
3 125679
4 1235
5 2
Key: 0|5 = $5

The busker’s earnings are inconsistent.

3 Stem | Leaf
86 8
87 7
88 0
89 8
90 2489
91 0126
92 93
94 39
95
96
97 0
98 359
Key: 86|8 = 86.8%

4 Stem | Leaf
18 579
19 156679
20 13359
21 7
22 1
Key: 18|5 = 1.85 cm

5 a Stem | Leaf
2 022
2* 5688
3 333
3* 778999
Key: 20 = 20 points

b Stem | Leaf
2 0
2 22
2 5
2 6
2 88
3 333
3 77
3 8999
Key: 20 = 20 cm

6 a Stem | Leaf
4 37788999
5 00001223
Key: 43 = 43 cm

b Stem | Leaf
4 3
4* 7788999
5 00001223
5* 3
Key: 43 = 43 cm

c Stem | Leaf
4
4 3
4
4 77
4 88999
5 00001
5 223
5
5
Key: 43 = 43 cm

7 a 1 2 5 8 12 13 16
16 17 21 23 24 25 26 27 30 32
b 10 11 23 23 30 35 39 41
42 47 55 62
c 101 102 115 118 122 123 123 136 136 137 141 143 144 155 155 156 157
d 50 51 53 53 54 55 55 56 56 57 59
e 1 4 5 8 10 12 16 19 19 21 21 25 29

8 Stem | Leaf
3  7 9
4  2 9 9
5  1 1 2 3 7 8 9
6  1 3 3 8

Key: 3|7 = 37 years
    It seems to be an activity for older people.

9 C

10 Stem | Leaf
1  9
2
2  5 8 8 9 9 9
3  0 0 2 2 2 3 3 4
3  5 5 7 8 9

Key: 2|5 = 25 years
    More than half of the parents are 30 or older with a considerable spread of ages, so this statement is not very accurate.

11 Stem | Leaf
1  9
2  1 2 4
2  5 6 8 9 9
3  1 1 2 3 4
3  0 1

Key: 2|1 = 21 hit outs

Bulldogs, Melbourne, St Kilda

12 Stem | Leaf
3  0
3  4
3  5 0 0 1
3  6 5
3  7 3
3  8 0 0
3  9 0 0 5
4  0 0 6
4  1 0 5
4  2 1 3
4  3 0 0
4  4
4  5 0

Key: 33|0 = $330
    The stem plot shows a fairly even spread of rental prices with no obvious outliers.

13 a Stem | Leaf
1  5 6 7 7 8 9 9 9 9
2  0 0 0 1 1 2 3 3 3

Key: 1|5 = 15 mm

b Stem | Leaf
1  5 6 7 7 8 9 9 9
2  0 0 0 1 1 2 3 3 3
2  2 2

Key: 1|5 = 15 mm

Values are bunched together; they vary little.

14 Stem | Leaf
7  2 8 8 3 3 5 7 8 8
9  0 1 2 2 3 4 8 9
10 0 2 4
11 2

Key: 7|2 = 72 shots

15 Stem | Leaf
6  8 9
7  1 1 2 2 3 3 3 4 4
7  5 5 5 6 6 7
8
8  6

Key: 7|1 = 71 net score

The handicapper has done a good job as most of the net scores are around the same scores; that is, in the 70s.

16 a Stem | Leaf
6  0 3 9
7  0 1 3 5 6 7 8
8  0 1 3 4 7 8 9 9
9  1 3 7 8

Key: 6|0 = 60%

b Stem | Leaf
6  0 3
6  9
7  0 1 3
7  5 6 7 8
8  0 1 3 4
8  7 8 9 9
9  1 3
9  7 8

Key: 6|0 = 60%
17 a Computer 1
5
8 2
6 3
6
1 0
2 1
5
0

Computer 2
34
35
36
37
38
39
40
41

Key: 34|0 = 340 minutes

b Computer 1 lasts longer but is not as consistent. Computer 2 is more consistent but doesn’t last as long.

18 a

Year 8
9 8
7 5 5 5 3 1 0
8 6 5 4 3 2 1 0
5 2 1
14
15
16
17
18

Year 10
2 4 6 8 9
0 4 5 7 7 9
2 3 4 6 7 8 8
2 5

Key: 14|8 = 148 cm

b As you would expect the Year 10 students are generally taller than the Year 8 students; however, there is a large overlap in the heights.

EXERCISE 1.4

1 a

Number of students

0 1 2 3 4 5 6 7 8 9 10

Number of questions completed

2 a

Number of students

0 1 2 3 4 5 6 7 8 9 10

Number of hours

2 c

Score

0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3

Frequency
1
2
1
1
2
2
2
2
1
1
1
Participation in activities

- 18–24 years: 14.2%
- 25–34 years: 21.1%
- 35–44 years: 20.3%
- 45–54 years: 18.1%
- 55–64 years: 14%
- 65 and over: 12.5%

The statement seems untrue as there are similar participation rates for all ages. However, the data don’t indicate types of activities.

Check your histograms against those shown in the answer to question 4.

Check your histogram against that shown in the answer to question 17.
23 a

<table>
<thead>
<tr>
<th>Country</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ</td>
<td>26.5%</td>
</tr>
<tr>
<td>US</td>
<td>13.5%</td>
</tr>
<tr>
<td>UK</td>
<td>12.8%</td>
</tr>
<tr>
<td>India</td>
<td>10.1%</td>
</tr>
<tr>
<td>China</td>
<td>7.5%</td>
</tr>
<tr>
<td>Thailand</td>
<td>6.4%</td>
</tr>
<tr>
<td>Fiji</td>
<td>6.2%</td>
</tr>
<tr>
<td>Singapore</td>
<td>6.0%</td>
</tr>
<tr>
<td>HK</td>
<td>5.9%</td>
</tr>
<tr>
<td>Malaysia</td>
<td>5.1%</td>
</tr>
</tbody>
</table>

24 a

![Bar chart showing attendance by year](chart.png)

b Check your bar chart against that shown in the answer to part a.

EXERCISE 1.5

1 Positively skewed
2 Negatively skewed
3 a Symmetric
   b Negatively skewed
   c Positively skewed
   e Symmetric
4 a Symmetric, no outliers
   b Symmetric, no outliers
   c Symmetric, no outliers
   d Negatively skewed, no outliers
   e Negatively skewed, no outliers
   f Positively skewed, no outliers
5 E
6 C
7 Negatively skewed
8 Positively skewed. This tells us that most of the flight attendants in this group spend a similar number of nights (between 2 and 5) interstate per month. A few stay away more than this and a very few stay away a lot more.

9 a Symmetric
   b This tells us that there are few low-weight dogs and few heavy dogs but most dogs have a weight in the range of 10 to 19 kg.
10 a Symmetric
   b Most students receive about $8 (give or take $2).
11 a Positively skewed
   b i 15
      ii 85%
12 a Positively skewed
   b Since most of the data is linked to the lower stems, this suggests that some students do little exercise, but those students who exercise, do quite a bit each week. This could represent the students in teams or in training squads.
13 a Club A: negatively skewed
   Club B: positively skewed
   b Since Club A has more members of its bowling team at the higher stems as compared to Club B; you could say Club A has the older team as compared to Club B.
   c i Club A: 11 members over 70 years of age
      ii Club B: 4 members over 70 years of age.
14 a

![Bar chart showing attendance by month](chart2.png)

b Positively skewed
   c June, July and November represent the months with the highest number of sales.
   d This is when the end of financial year sales occur.

EXERCISE 1.6

1 Median = 33
2 Median = 36.5 goals
3 IQR = 14
4 IQR = 8
5 IQR = 6.5
6 IQR = 3.3
7

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>Range</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>37</td>
<td>56</td>
<td>38, 49</td>
</tr>
<tr>
<td>b</td>
<td>5</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>c</td>
<td>11</td>
<td>18</td>
<td>8, 11</td>
</tr>
<tr>
<td>d</td>
<td>42.5</td>
<td>18</td>
<td>43</td>
</tr>
<tr>
<td>e</td>
<td>628</td>
<td>72</td>
<td>613, 628, 632</td>
</tr>
</tbody>
</table>
8

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>b</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>c</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>d</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>e</td>
<td>18.5</td>
<td>14</td>
</tr>
<tr>
<td>f</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>g</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>h</td>
<td>4.5</td>
<td>9</td>
</tr>
<tr>
<td>i</td>
<td>23</td>
<td>21</td>
</tr>
</tbody>
</table>

9 a 10
   b 8
   c The IQRs (middle 50%) are similar for the two restaurants, but they don’t give any indication about the number of cars in each data set.

10 An example is 2 3 6 8 9. There are many others.

11 a The lowest score occurs several times. An example is 2 2 2 3 5 6.
   b There are several data points that have the median value. An example is 3 5 5 5 5 7.

12 C

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>Interquartile range</th>
<th>Range</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>21</td>
<td>18</td>
<td>45</td>
<td>15, 23, 32</td>
</tr>
<tr>
<td>b</td>
<td>27.5</td>
<td>8</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>c</td>
<td>3.7</td>
<td>3</td>
<td>5.9</td>
<td>3.7</td>
</tr>
</tbody>
</table>

13

<table>
<thead>
<tr>
<th></th>
<th>Median</th>
<th>Interquartile range</th>
<th>Range</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>42</td>
<td>21</td>
<td>91</td>
<td>46</td>
</tr>
<tr>
<td>b</td>
<td>32</td>
<td>7</td>
<td>30</td>
<td>34</td>
</tr>
</tbody>
</table>

The data in set a have a greater spread than in set b, although the medians are similar. The spread of the middle 50% (IQR) of data for set a is bigger than for set b but the difference is not as great as the spread for all the data (range).

15 a Range = 72, Median = 37.5, Mode = 46, IQR = 22
   b Range = 47
       Median = 422
       Mode = 411
       IQR = 20

16 Median = 7, Mode = 7

17 $Q_1 = 42.2, Q_2 = 48.15, IQR = 5.95, Median = 45.1$

18 a Median = 93, $Q_1 = 91.5, Q_3 = 97, IQR = 5.5, Range = 30, Mode = 93$
   b The average handicap of the golfer’s should be around 21.

19 a $\text{Mode} = 46, \text{Median} = 23, \text{Range} = 16$
   b $\text{Mode} = 32, \text{Median} = 29, \text{Range} = 17$

20 a $\text{Mode} = 6, \text{Median} = 41, \text{Range} = 30$
   b $\text{Mode} = 29, \text{Median} = 27, \text{Range} = 31$

EXERCISE 1.7

1 Range = 39
   Median = 25
   IQR = 19

2 Range = 3
   Median = 7.5
   IQR = 1.4

3 They could represent the same data.
4 They could represent the same data.
5

6 Negatively skewed; 50% of results are between 32 and 42.

7 a The data is symmetrical and 1.75 is an outlier.
   b The data is symmetrical and 1.75 is an outlier.

8 30.3 is an outlier.

9

<table>
<thead>
<tr>
<th></th>
<th>Range</th>
<th>Interquartile range</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>12</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>b</td>
<td>7</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>c</td>
<td>350</td>
<td>100</td>
<td>250</td>
</tr>
<tr>
<td>d</td>
<td>100</td>
<td>30</td>
<td>65</td>
</tr>
<tr>
<td>e</td>
<td>20</td>
<td>10</td>
<td>25</td>
</tr>
</tbody>
</table>

10 a iii b iv c i d ii

11 The boxplots should show the following:

<table>
<thead>
<tr>
<th>Minimum value</th>
<th>$Q_1$</th>
<th>Median</th>
<th>$Q_3$</th>
<th>Maximum value</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3</td>
<td>6</td>
<td>8.5</td>
<td>14</td>
</tr>
<tr>
<td>b</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>c</td>
<td>4.3</td>
<td>4.6</td>
<td>5</td>
<td>5.4</td>
</tr>
<tr>
<td>d</td>
<td>11</td>
<td>15.5</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>e</td>
<td>0.4</td>
<td>0.7</td>
<td>0.9</td>
<td>1.1</td>
</tr>
</tbody>
</table>

12 D
13 Number of clients seen in a day

14 See boxplot at foot of the page*

15 a

The data are negatively skewed with an outlier on the lower end. The reason for the outlier may be that the person wasn’t at the show for long or possibly didn’t like the rides.

16 a Two similar properties: both sets of data have the same minimum value and similar IQR value.

b Boys IQR = 16
Girls IQR = 16.5

c The reason for an outlier in the boys’ data may be that the student did not understand how to do the test, or he stopped during the test rather than working continuously.

17 Median = 13, \( Q_1 = 9, Q_3 = 17, \) Min = 4, Max = 21

18 Median = 5
\( Q_1 = 4.5 \)
\( Q_3 = 7 \)
Min = 2, Max = 15
IQR = 2.5
15 is an outlier

19 a Median = 25
\( Q_1 = 17.5 \)
\( Q_3 = 32 \)
Min = 11, Max = 39
IQR = 14.5

b No outliers

c Check your boxplot against that shown in the answer to part a.

20 Median = 86
\( Q_1 = 75 \)
\( Q_3 = 97 \)
Min = 23, Max = 113
IQR = 22

23 is an outlier

EXERCISE 1.8
1 23.46
2 8.26
3 10.54
4 26.80

5 a 7.2
b 7.125
c 4.9875
d 16.7
e 0.8818

6 a 1.0783 No, because of the outlier.
b 17 Yes
c 30.875 Yes
d 15.57 No, because of the outlier.

7 12

8 D

9 A

10 a Median b Mean c Median d Median

11 a 36.09
b 16.63
c 168.25
d 18.55

12 a 24.4

b Median = 22
The distribution is positively skewed — confirmed by the table and the boxplot.

13 214.5 papers

14 Approximately 41 fish

15 63.14 kg

16 a Approximately 53 cups

b The median is 54.5, approximately 55 cups.
c The data is negatively skewed.

EXERCISE 1.9
1 3.54 cents
2 14.27%
3 9.489
4 7.306
5 a 1.21
b 2.36
6.01
d 2.45
e 0.06
6 0.48%
7 0.06 m
8 0.51 seconds
9 15.49
C
11 2.96 km/h
12 6.067 pens
13 2.39 °C
x = 75.7, s = 5.6
15 3.786 players
16 2.331

EXERCISE 1.10
1 Answers will vary.
2 Answers will vary.
3 B
4 Answers will vary.
5 Answers will vary.
6 Population is larger, since a sample is taken from the population.
7 C
8 E
9 Yes, because the distribution is reasonably symmetric with no outliers
10 B
11 C
12 Answers will vary.
13 Answers will vary.
14 Answers will vary.

EXERCISE 1.11
1 a 68% of group’s concentration span falls between 35 secs and 63 secs
b 95% of group’s concentration span falls between 21 secs and 77 secs
c 99.7% of group’s concentration span falls between 7 secs and 91 secs
2 a 68% of the group to lie between 43.3 mm and 46.7 mm
b 95% of the group to lie between 41.6 mm and 48.4 mm
c 99.7% of the group to lie between 39.9 mm and 50.1 mm
3 a 2.50%  b 50%  c 16%  d 81.5%
4 a 84%  b 2.5%  c 84%  d 97.35%
5 420 bags

6 a 336 containers   b 10 containers
   c 380 containers.
7 2.33
8 3
9 a Specialist: \( \mu = 67, \sigma = 9 \)
   English: \( \mu = 58, \sigma = 14 \)
   \( z_s = 1.78, z_e = 2.14 \)
b English has the higher result as it has the higher \( z \)-score.
10 a English 1.25, Maths 1.33
   b Maths mark is better as it has a higher \( z \)-score.
11 a Yes  b Yes  c No
d No  e No  f Yes
12 a 8 and 12  b 6 and 14  c 4 and 16
13 a 3.7 and 6.3  b 2.4 and 7.6  c 1.1 and 8.9
14 a 1.3 mm and 2.5 mm  
   b 0.7 mm and 3.1 mm  
   c 0.1 mm and 3.7 mm
15 a 5 and 9  b 3 and 11  c 1 and 13
16 C
17 a 0.15%  b 2.5%  c 84%
d 83.85%  e 81.5%
18 a i 1360  ii 1950  iii 317
   b 100
19 a −0.48  b 1.44  c 0.08
d −2.24  e 2.8
20 B
21 Second test, Barbara’s \( z \)-score was −0.33 compared to −0.5 in the first test.
22 a Barn: \( \mu = 4.4 \)  \( \sigma = 0.3 \)
   FR: \( \mu = 4.1 \)  \( \sigma = 0.2 \)
b 1.18
c 84%
d

<table>
<thead>
<tr>
<th></th>
<th>Cage</th>
<th>Barn</th>
<th>Free range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min_( x )</td>
<td>4.7</td>
<td>3.9</td>
<td>3.8</td>
</tr>
<tr>
<td>( Q_1 )</td>
<td>5</td>
<td>4.1</td>
<td>4</td>
</tr>
<tr>
<td>med</td>
<td>5.15</td>
<td>4.35</td>
<td>4.1</td>
</tr>
<tr>
<td>( Q_3 )</td>
<td>5.5</td>
<td>4.6</td>
<td>4.2</td>
</tr>
<tr>
<td>Max_( x )</td>
<td>5.8</td>
<td>4.9</td>
<td>4.4</td>
</tr>
</tbody>
</table>

i Cage: 5.15
   Barn: 4.35
   Free: 4.1

ii It could be concluded that the more space a chicken has, the fewer eggs it lays because the median is greatest for cage eggs.