CHAPTER

19

Statistics and Probability

Statistics

Statistics is concerned with the collection and analysis of **data**. We encounter statistics in many aspects of everyday life. It would be hard to find an area of business or industry that does not use some form of statistics in its decision-making processes.

Questions such as 'Is the Government's economic policy having an effect on youth unemployment?', 'Is the State Road Authority's advertising campaign having an effect on the road toll?', 'How does Australia's standard of living compare with that of other countries?' or 'Is there a relationship between marks in English and marks in Mathematics?' all require data to be collected and analysed before they can be answered in an intelligent and justifiable way.

The **data** has to be **collected**, **recorded** and **represented** in a way that is appropriate to the questions we wish to answer. We then **analyse** the data and draw conclusions. This process is discussed at length in the final section of this chapter.

In this chapter we will deal only with small data sets. This will enable us to do the statistical calculations and graphical representations by hand to allow us to understand the ideas clearly. It is important to be aware that when larger data sets are involved, statistical software is often used.

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

19A Stem-and-leaf plots and back-to-back stem-and-leaf plots

In this chapter we provide a summary of all of the representations that you have met so far in your study of statistics, and then introduce other representations.

We first revise stem-and-leaf plots.

Stem-and-leaf plots

Numerical data (quantitative data) consists of values in which there is a definite numerical order. For example, scores in a test or heights of students in a class.

A **stem-and-leaf plot** can be used to represent numerical data.

Stem-and-leaf plots

- A stem-and-leaf plot represents the values in a data set in the form of a stem and a leaf.
- The **stem** is the first digit of a two-digit number, or the first two digits of a three-digit number, and so on.
- The **leaf** is usually the last digit of the value. For example:
  - For the number 47, 4 is the stem and 7 is the leaf.
  - For the number 251, 25 is the stem and 1 is the leaf.
  - For the number 11.6, 11 is the stem and 6 the leaf.

Stem-and-leaf plots are useful for displaying the shape of the data and giving the reader a quick overview. They retain most of the raw numerical data. They are also useful for highlighting outliers and finding the mode. However, stem-and-leaf plots are only useful for data sets between 15 and 150 values.

Example 1

The marks out of 50 obtained by 16 students in a Mathematics test are:

43 24 29 19 11 14 25 17 32 27 29 7 14 19 39 49

Represent this information on a stem-and-leaf plot.

Solution

We use the first digit of each mark as the stem, writing 7 as 07.

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

3 | 2 represents a mark of 32

ICE-EM MATHEMATICS YEAR 9
ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

Example 2

The approximate heights of 20 students, in centimetres, are given below.

164 158 152 167 146 149 167 171 181 154
167 158 164 172 176 180 178 165 159 153

Represent this information on a stem-and-leaf plot and give the frequency of each group.

Solution

We use the first two digits of each height as the stem.

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

17 | 2 means 172

Back-to-back stem-and-leaf plots

Back-to-back stem-and-leaf plots are used to compare two similar sets of numerical data.

Example 3

A class of 25 students sit for two Mathematics tests, each out of 100. Their results are recorded in the following back-to-back stem-and-leaf plot.

Test 1		Test 2
5 5	4	
9 6 6 4	5	5 6 6
9 8 8 7 6 5 5 4 3 3	6	5 7
8 8 5 2 2 2 1	7	5 9
1	8	0 0 4 4 5 5 5 6 7 9 9
0	9	0 0 1 3 5 7 7

a How many students scored 70 or more for each test?

b How many students scored less than 50 for each test?

Solution

a Nine students scored 70 or more in test 1 and 20 students scored more than 70 in test 2.

b Two students scored less than 50 in test 1 and no student scored less than 50 in test 2.

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

Example 4

The back-to-back stem-and-leaf plot shows the city and country highway fuel consumption of 16 cars. The fuel consumption is measured in litres per 100 km.

City					Country				
				5	8	9			
				6	5	5	6	7	
8	7	7	5	7	4	5	5	6	7
	7	6	3	8	0	0	6	8	8
				9					
5	5	4	4	10					
		5	3	11					
	8	7	6	12					

5 |7| 4 means 7.5 L per 100 km in the city and 7.4 L per 100 km in the country.

a Find the maximum and minimum fuel consumption (in litres per 100 km) for:

i country driving **ii** city driving

b Find the percentage of cars tested with fuel consumption less than 8 litres per 100 km for:

i country driving **ii** city driving

Solution

a i 8.8 L per 100 km and 5.8 L per 100 km are the maximum and minimum fuel consumptions for country driving respectively.

ii 12.8 L per 100 km and 7.5 L per 100 km are the maximum and minimum fuel consumptions for city driving respectively.

b i $\left(\frac{11}{16}\right) \times 100\% = 68.75\%$

68.75% of cars tested for country driving have a fuel consumption that is less than 8 litres per 100 km.

ii $\left(\frac{4}{16}\right) \times 100\% = 25\%$

25% of cars tested for city driving have a fuel consumption that is less than 8 litres per 100 km.

Exercise 19A

Example 1

1 The marks for a mathematics test done by a particular class of 25 students were:

48 43 29 36 37 21 15 24 35 44 37 35 25
39 28 25 46 37 24 26 42 45 33 47 29

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

a Present this information on a stem-and-leaf plot.

b Which 10-mark group contains the most students?

Example 2

2 The approximate weights in kilograms of 30 people are:

85	78	94	86	104	93	76	84	95	91	106	89	94	97	91
82	76	93	84	86	79	96	94	81	77	87	82	96	102	86

a Present this information on a stem-and-leaf plot.

b How many of the 30 people weigh:

i more than 85 kg?

ii less than 90 kg?

iii strictly between 80 and 95 kg?

3 The stem-and-leaf plot below displays the approximate time (in minutes) taken to travel to school for a class of students.

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

1 | 4 means 14

a How many students are in the class?

b How many students take:

i less than 25 minutes?

ii more than 40 minutes?

iii between 20 and 40 minutes?

c The two students who take the longest and shortest time to get to school arrive at the same time. How much later does one leave home than the other?

4 The stem-and-leaf plot opposite displays the approximate percentage inflation rate for a number of countries, correct to 1 decimal place.

2	8 9
3	1 4 4 7
4	2 4 5 5 7 9
5	0 1 1 4 5 7 9
6	2 3 5 6 6
7	2 4

2 | 4 means 2.8%

a How many countries are included in the sample?

b How many countries in the sample have an inflation rate that is:

i greater than 5.5%?

ii less than 3.4%?

c If countries with an inflation rate greater than 6.4% are classified as having an unstable economy, what percentage of countries have an unstable economy

Example 3

5 A class of 16 students sits for two Mathematics tests. Their results are recorded in the following back-to-back stem-and-leaf plot.

Test 1		Test 2
9 8 5 5	4	
7 5 3 2	5	8
9 8 8 7 6 6 6 5	6	1
	7	5
	8	3 3 3 4 4 5 5 5 6 7 9
	9	4 4

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

a What percentage of students scored 50 or more for:

i test 1? **ii** test 2?

b What percentage students scored less than 50 for:

i test 1? **ii** test 2?

c What percentage of students scored between 60 and 70 (inclusively) for:

i test 1? **ii** test 2?

Example 4

6 Two walking clubs record the ages of their members, as shown below. They both have a membership of 22 people.

Club 1: 82, 82, 78, 78, 78, 73, 73, 73, 73, 72, 69, 67, 67, 65, 34, 25, 24, 23, 16, 13, 12, 11

Club 2: 37, 38, 39, 42, 43, 55, 57, 65, 65, 66, 66, 66, 67, 68, 68, 69, 71, 72, 72, 72, 73

a Draw a back-to-back stem-and-leaf plot to represent their ages.

b Compare the ages of the members of the two clubs.

7 The following data gives the number of kicks by members of the Yellowlegs and Blackarm teams in their Australian Rules country grand final.

Blackarm: 8, 8, 10, 10, 11, 11, 12, 12, 12, 12, 13, 13, 14, 15, 16, 17, 18, 19, 19, 20, 21, 26

Yellowlegs: 0, 4, 5, 6, 7, 8, 8, 9, 9, 10, 10, 13, 13, 14, 15, 16, 16, 17, 18, 19, 19, 25

a Prepare a back-to-back stem-and-leaf plot, using two rows per stem.

It is started for you here.

Blackarm		Yellowlegs
	0	0 4
	0	5
	1	0 0
6 5	1	
	2	
	2	

b Compare the number of players of both clubs who had:

i fewer than 10 kicks **ii** 15 or more kicks

8 The weights of 40 male and 40 female students are compared in the back-to-back stem-and-leaf plot shown below. The weights are given correct to the nearest kilogram.

Females		Males
9 8	3	
9 8 8 7 7 7 7 6 6	4	
9 8 7 7 7 5 4 1 1 0 0 0	5	2 8
5 4 4 4 0 0 0 0	6	1 1 1 3 3 4 4 5 5 5 7 7 8 9
5 2 1 0 0	7	0 0 1 3 4 4 5 6 6
1 0 0	8	0 0 1 2 3 3 5 5 7 8
	9	0 7
	10	0 0 9

6 | 1 means 61 kg

a **i** Find the percentage of males who weigh more than 79 kg.

ii Find the percentage of females who weigh less than 50 kg.

b **i** What are the weights of the 20th and 21st heaviest males?

ii What are the weights of the 20th and 21st heaviest females?

c **i** What is the difference in weight between the heaviest and lightest males?

ii What is the difference in weight between the heaviest and lightest females?

9 Collect data in your class comparing differences between boys and girls using back-to-back stem-and-leaf plots. Height, arm length or hours of television watched are just some examples you may chose.

19B Grouped data

Without a lot of practice, little useful information can be gained from looking at unstructured data, such as in Example 5 below.

To understand a data set with a large number of values, it is helpful both to divide the range of values of the data into intervals, called **class intervals** or classes, and to record the **frequency** of each class interval – that is, the number of data items in each class. The resulting summary of the data is called a grouped frequency table or simply a **frequency table**.

Example 5

The marks out of 50 for a mathematics test done by a class of 25 students are:

48 43 29 36 37 21 15 24 35 44 37 35 25
29 39 28 25 46 37 24 26 42 45 33 47

Present this information in a frequency table, using groupings of 15–19, 20–24, 25–29 and so on, up to 45–49.

Solution

Grouping the data produces the following frequency table.

Mark	Tally	Frequency
15–19	|	1
20–24	|||	3
25–29	𝍸 |	6
30–34	|	1
35–39	𝍸 ||	7
40–44	||||	3
45–49	||||	4

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

Choice of classes

In examples involving grouped data, the following points should be considered:

- Interval sizes should be chosen so that a sensible number of classes results. Between 5 and 10 classes are commonly used.
- Classes should be of equal width, except occasionally for the first or last class. For example, if an exam is marked out of 50, and 10 classes are used, the last class would normally go from 45 to 50.
- It is best to use classes such as 0–4, 5–9, 10–14, ... or 0–9, 10–19, 20–29, ... or 0–99, 100–199 since they match our decimal number system. Avoid using classes such as 12–16, 17–21, 22–26, even if the smallest value is 12. Apart from being systematic, this allows classes to be easily created from a stem-and-leaf plot.

The frequency table below shows the amount of sodium in 100 gram samples of different foods.

Amount of sodium (mg)	Tally	Frequency
0–49	𝍸 𝍸 III	13
50–99	IIII	4
100–149	𝍸 𝍸 𝍸	15
150–199	𝍸 𝍸 𝍸 𝍸	20
200–249	𝍸 𝍸 𝍸 𝍸 III	23
250–299	𝍸 𝍸 𝍸 IIII	19
300–349	𝍸 I	6

The amount of sodium is a continuous variable, and it can take any value from 0 to 350 for the foods considered. The amounts here have been stated correct to the nearest milligram.

Consider the class of foods that contain 200–249 mg of sodium (in each 100 g sample) of which there are 23.

This means that there are 23 values from 199.5 up to, but not including, 249.5. The real endpoints – 199.5 and 249.5 – are called **class boundaries**.

Exercise 19B

Example 5

1 The number of runs scored in each innings by a batsman throughout a cricket season was:

42	18	5	73	97	61	47	31	8	1
14	26	71	58	27	11	26	51	4	1
15	92	18	37	40	65	72	3	5	18

a Present this information in a frequency table using class intervals

0–9, 10–19, 20–29, ...

b On how many occasions did the batsman score:

i more than 49 runs? **ii** fewer than 10 runs?

iii fewer than 80 runs? **iv** between 20 and 59 runs?

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

2 The approximate times taken to run 100 m by 25 students were as follows:

13.9	11.1	12.6	14.1	13.8	13.2	14.4	12.8	12.1
12.5	11.8	13.4	14.2	12.6	11.9	12.8	14.7	14.2
13.6	13.9	14.5	13.1	12.7	12.9	13.6		

a Present this information in a frequency table using class intervals 11.0–11.4, 11.5–11.9, 12.0–12.4 and so on.

b Present this information in a frequency table using class intervals 11.0–11.9, 12.0–12.9, 13.0–13.9 and so on.

3 The ages of employees in a factory range from 22 to 64.

a Explain how you might break up the age range into five classes.

b Do the same for nine classes.

4 On a particular Saturday, the prices of houses sold in Geelong, Victoria, ranged from $294 000 to $618 000. Show how to group the selling prices of the houses into eight classes.

5 The maximum daily temperature in Sydney was recorded each day for a month, correct to 0.1°C. The results are given in the table.

Maximum temperature (°C)	Frequency
24.0–24.2	8
24.3–24.5	6
24.6–24.8	4
24.9–25.1	4
25.2–25.4	3
25.5–25.7	3
25.8–26.0	2

a On how many days was the recorded maximum temperature:

i greater than 25.7°C?

ii less than 24.6°C?

iii between 24.3°C and 25.7°C (inclusive)?

iv between 24.9°C and 26.0°C (inclusive)?

b Why is it impossible to tell on how many days the maximum temperature was above 25°C?

6 The following set of raw data shows the lengths, recorded to the nearest millimetre, of 40 leaves taken from a particular tree.

42	56	27	52	60	47	49	51	32	30
54	33	54	43	49	46	48	41	43	61
51	40	45	50	45	45	42	53	42	58
33	55	46	39	37	39	34	40	48	38

a Construct a frequency table with classes 25–29, 30–34 and so on.

b In which class is a leaf measured as 29.7 mm included?

c In which class is a leaf measured as 34.3 mm included?

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

19C Histograms and summary of representations

Histograms

Numerical data can be displayed in a **histogram**. Data grouped into classes is often displayed this way. A histogram is a type of graph in which the frequencies of the data are displayed by touching columns. They are particularly useful with data sets containing a large number of values.

Example 6

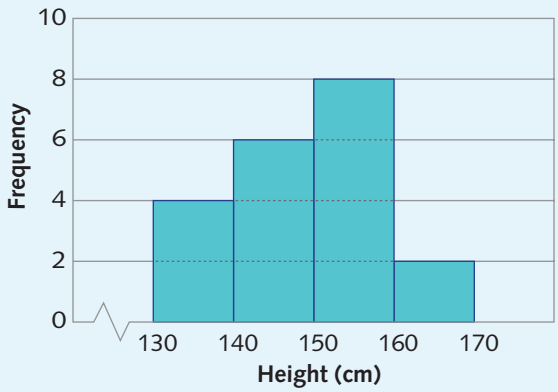
The heights of 20 Year 9 students were measured and the results are shown in the frequency table opposite. Represent this information in a histogram. The heights were measured correct to the nearest centimetre.

Height (in cm)	Frequency
130 – 139	4
140 – 149	6
150 – 159	8
160 – 169	2

Solution

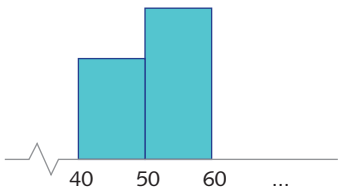
In this graph, the left-most column represents the class 130–139; the second, 140–149; and so on. All our examples follow a similar convention.

From the discussion in the previous section, this means that a height of 139.4 cm would be included in the 130–139 class and that a height of 139.8 cm would be included in the 140–149 class.



Note:

- The values on the horizontal axis in Example 6 indicate, for example, that the first class consists of heights, measured to the nearest centimetre, at least 130 cm and less than 140 cm.
- If the variable is integer-valued, such as a mark, then in the diagram shown the first class is 40–49, the second class is 50–59 and so on.

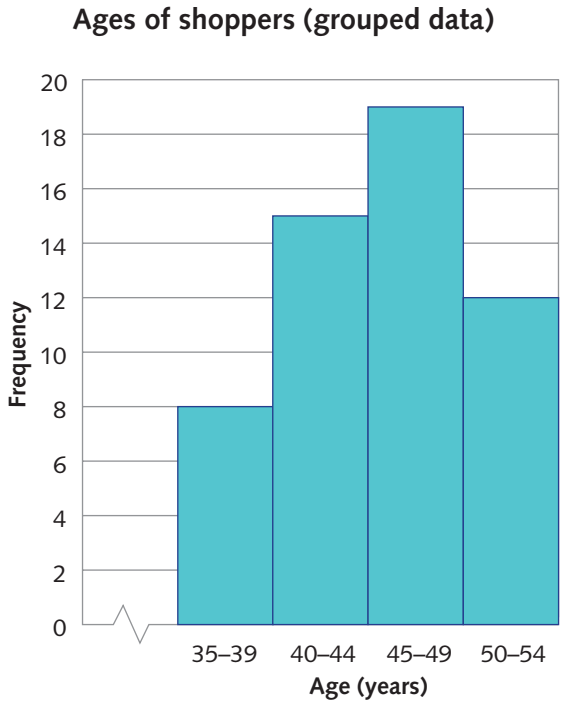


- When working with histograms, another name for class is **bin**.

A number of shoppers were asked to record their ages as they left a department store. The data was recorded as follows.

Age (years)	35	36	37	38	39	40	41	42	43	44
Tally	I	𝍸 I		I		𝍸 II	II	II		IIII
Frequency	1	6	0	1	0	7	2	2	0	4
Age (years)	45	46	47	48	49	50	51	52	53	54
Tally	II	III	𝍸	𝍸	IIII	IIII	III	II	I	II
Frequency	2	3	5	5	4	4	3	2	1	2

Some aspects of the information become clearer after grouping. The final histogram is shown below.



Now we can read off facts such as:

- The 45–49 age range contains the most shoppers.
- Nineteen shoppers fall in the 45–49 age range.

Classes are usually chosen so that the entire data set is broken up into at most 10 classes. The advantage of grouping data is that trends are often easier to recognise.

A disadvantage of grouping data into classes is that we cannot see the individual values that were recorded.

Types of representations

So far we have been considering only numerical data, such as age, height or test scores. However, we are often also interested in data concerning the preferences or attitudes in a community, or the number of items with certain attributes. Such data is called **categorical**.

The following representations have been introduced in this book or previous books.

The table on the next page provides some guidelines for selecting which representation to use.

Type of data	Representation	Qualifications on use
Categorical	Pictograph	Usually fewer categories
	Column graph	Not too many categories
Numerical (or quantitative)	Dot plots	Best for small data sets
	Histogram	Best for medium to large data sets
	Stem-and-leaf plot	Best for small to medium data sets

Pictographs, column graphs and dot plots were introduced in earlier books. We provide an example of each.

Pictograph

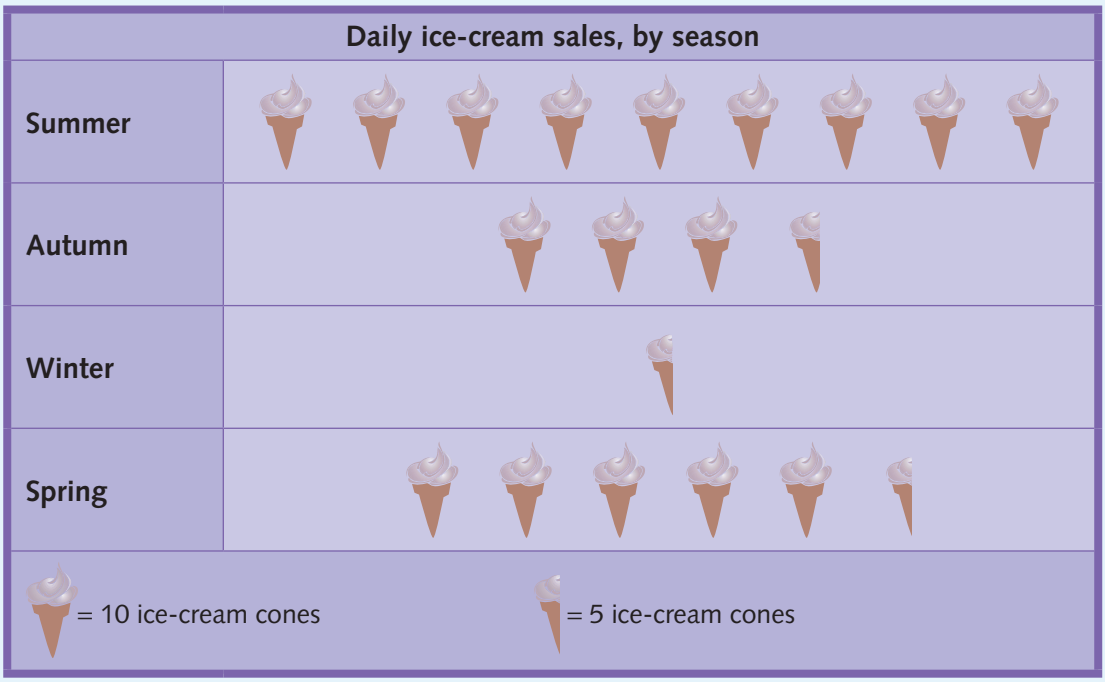
Example 7

The table below shows the average daily sales of ice-cream cones from a kiosk in each season over a 12-month period.

Season	Summer	Autumn	Winter	Spring
Ice-cream sales	90	35	5	55

Use a pictograph to represent these data.

Solution



ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

Column graph

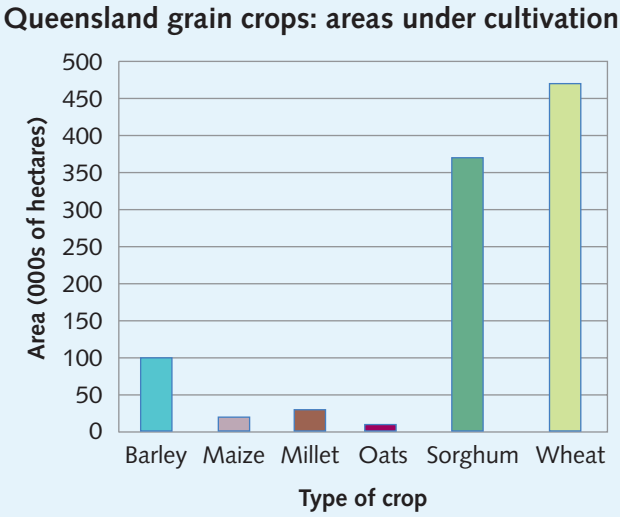
Example 8

The table below gives the approximate areas under cultivation, in thousands of hectares, for different grain crops in a region of Queensland.

Grain crop	Area under cultivation (in thousands of hectares)
Barley	100
Maize	20
Millet	30
Oats	10
Sorghum	370
Wheat	470

Draw a column graph representing the data.

Solution



Dot plot

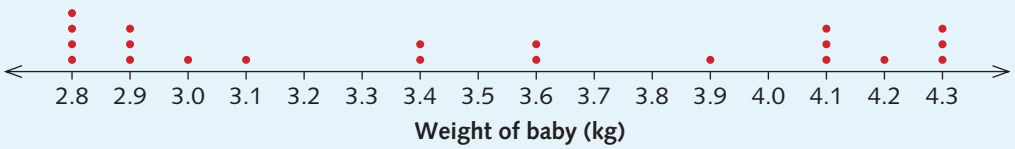
Example 9

Twenty-one babies were born at a hospital on one weekend last year. Their approximate birth weights, in kilograms, are given here.

2.8 2.8 2.8 2.8 2.9 2.9 2.9 3.0 3.1 3.4 3.4
3.6 3.6 3.9 4.1 4.1 4.1 4.2 4.3 4.3 4.3

Draw a dot plot for the data.

Solution



ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

Exercise 19C

The first five questions are review questions.

Example 8

1 In a class of 25 students, 10 have green eyes, 8 have blue eyes, 4 have brown eyes and 3 have grey eyes. Present this information in a column graph.

2 The table opposite lists the percentage of viewers watching each television channel on a particular night. Present this information in a column graph.

Channel	Percentage of viewers
1	26
2	34
3	25
4	15

Example 9

3 A group of 30 people in the building trade were asked how many times in the last week they had visited a particular hardware shop. Their responses were:

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

Draw a dot plot for the data.

4 A group of students were asked to select their favourite fast food. The results are in the table below.

Food type	Number of students
Hamburgers	8
Chicken	8
Fish and chips	15
Pizza	22

Draw a column graph to illustrate the results.

5 The ages in years of the 15 members of a sporting team are:

23 19 18 19 23 25 22 22 19 22 18 23 24 21

Construct a dot plot.

Questions on histograms follow.

Example 6

6 The frequency table shown opposite gives information regarding the test results of a group of 23 students. Present this information in a histogram.

Score	Frequency
15–19	1
20–24	3
25–29	5
30–34	2
35–39	6
40–44	3
45–49	3

7 In a medical test, a group of students had the distance from hip to heel measured. The measurements were made correct to the nearest centimetre. The results were as follows.

85 86 91 87 77 88 83 86 74 89 85 85 80
94 82 84 89 84 94 84 76 93 86 84 94 84

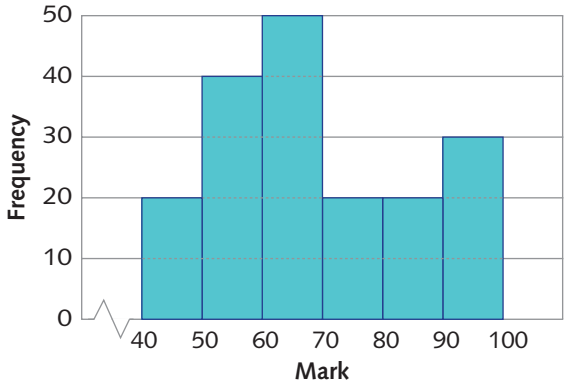
ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

Present this information in a histogram using the classes:

a 70–79, 80–89, 90–99

b 70–74, 75–79, 80–84, 85–89, 90–94

8 The histogram below gives information about the results of Year 9 students in a history examination.



a How many students sat for the examination?

b If the pass mark was 60, how many students passed?

c What percentage of students obtained 90 or more?

d What percentage of students obtained less than 70?

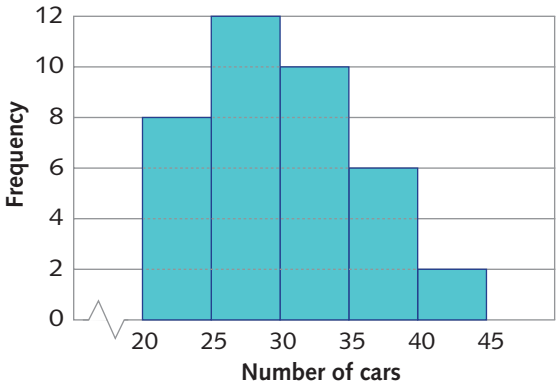
9 **a** Present the information given in the stem-and-leaf plot opposite using a histogram.

0	2 3 4
1	1 5 7 9 9
2	3 4 6
3	
4	5 7 8 8
5	2 5

2 | 4 means 24

b Why is it impossible to present the information in a histogram using a stem-and-leaf plot?

10 A student counted the number of cars that passed through an intersection on each cycle of the traffic lights. The results of her investigation are shown in the histogram opposite.



a For how many cycles did the student record the number of cars passing through the intersection?

b In how many cycles did:

i fewer than 30 cars pass through the intersection

ii at least 35 cars pass through the intersection?

c Why is it impossible to determine from the histogram how many cars in total passed through the intersection during the survey period?

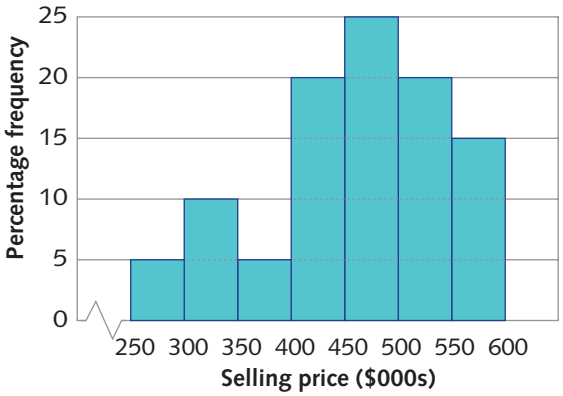
d Determine:

i the minimum number of cars that could have passed through the intersection during the survey period

ii the maximum number of cars that could have passed through the intersection in the survey period

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

11 The percentage frequency histogram opposite gives the auction prices of houses sold in Wollongong on a particular day.



a What percentage of houses sold for:

i less than $500 000

ii at least $400 000

iii between $300 000 and $550 000?

b If 420 houses were sold, how many houses sold for:

i less than $350 000

ii at least $300 000

iii between $500 000 and $600 000?

12 Thirty calls were made by an employee of a telephone call centre. The lengths, correct to the nearest minute, are summarised in the following table.

Length of call	0–5	6–10	11–15	16–20
Number of calls	15	8	5	2

a If a call went for 5 minutes and 40 seconds, in which class is it recorded?

b Draw a histogram of the data.

Activity

Reaction times (left and right)

Group 1 (*possibly 20 students*)

1 Stand next to a wall with another student.

2 Hold a ruler against a wall with the 0 cm mark at the bottom.

3 Ask the other student to close their left eye and align their index finger of their right hand with the 0 cm mark. (Their finger is to be 5 cm from the wall.)

4 Explain that you will let go of the ruler without warning, and they must try to pin it against the wall with their index finger of their right hand.

5 Measure and record the distance dropped.

Group 2 (*possibly 20 students*)

1 Repeat steps 1–5 above, closing the right eye and using the index finger of the left hand.

2 Draw a stem-and-leaf plot diagram for both sets of data and compare.

3 Draw two histograms.

4 Think of other ways of carrying out this activity.

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

19D Mean, median, mode and range

Mean

The **mean** of a numerical data set is a measure of its centre. It is calculated by first adding together all the data values and then dividing the sum by the number of data values.

Mean

We use the following formula to find the mean of a set of data:

$$\text{mean} = \frac{\text{sum of values}}{\text{number of values}}$$

The more common name for mean is 'average'.

Example 10

Allen obtained the following marks in 8 tests:

43 35 41 29 33 39 47 42

Calculate the mean.

Solution

$$\text{Mean} = \frac{\text{sum of values}}{\text{number of values}}$$

$$= \frac{43 + 35 + 41 + 29 + 33 + 39 + 47 + 42}{8}$$

$$= \frac{309}{8}$$

$$= 38.625$$

Note that the mean was not a mark obtained by Allen in any of his tests.

Example 11

The following table gives the number of children in each of 20 families. Calculate the mean number of children per family.

Number of children	0	1	2	3
Frequency	4	5	7	4

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

Solution

There are 4 families with 0 children. This gives a total of 0 children.

There are 5 families with 1 child. This gives a total of $1 \times 5 = 5$ children.

Continuing in this way, we have:

$$\text{average} = \frac{0 \times 4 + 1 \times 5 + 2 \times 7 + 3 \times 4}{4 + 5 + 7 + 4}$$
$$= \frac{31}{20}$$
$$= 1.55$$

It is obviously impossible for a family to have 1.55 children. In general, the average of a data set is not one of the original values.

Median

We often see the median value used to describe the housing market in a city. The **median** is the middle value when all values are arranged in numerical order. Here are some numbers arranged in numerical order.

2 2 3 3 3 4 [5] 11 13 18 18 19 21

Median

This data set has an odd number of values. The middle value is 5 since it has the same number of values on either side of it. Hence the median of this data set is 5.

Here is another set of numbers arranged in numerical order.

1 3 4 4 5 7 | 9 11 13 13 19 21

Median

The above data set has an even number of values. The middle values are 7 and 9. We take the average of 7 and 9 to calculate the median.

$$\text{Median} = \frac{7 + 9}{2} = 8$$

Hence the median of this data set is 8. Note that this value does not occur in the data set.

Median

- When the data set has an *odd* number of values and they are arranged in numerical order, the median is the *middle value*.
- When the number of values is *even* and they are arranged in numerical order, the median is the *average of the two middle values*.
- In general, if there are n items in the *ordered* data set, the median lies in the $\left(\frac{n+1}{2}\right)$th position.

Example 12

Calculate the median of the following data sets.

a 43, 35, 41, 29, 33, 39, 42 **b** 4, 6, 8, 5, 12, 10

Solution

a To locate the median, first put the values in numerical order:

29 33 35 [39] 41 42 43

Median = 39

b Again, place the values in numerical order.

4 5 6 ⋮ 8 10 12

$$\text{Median} = \frac{6+8}{2} = 7$$

Since the data values need to be written in numerical order to locate the median, a stem-and-leaf plot helps find the median.

Example 13

The stem-and-leaf plot below shows the approximate weights of the students in a class. Determine the median weight of the students.

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

6 | 4 means 64 kg

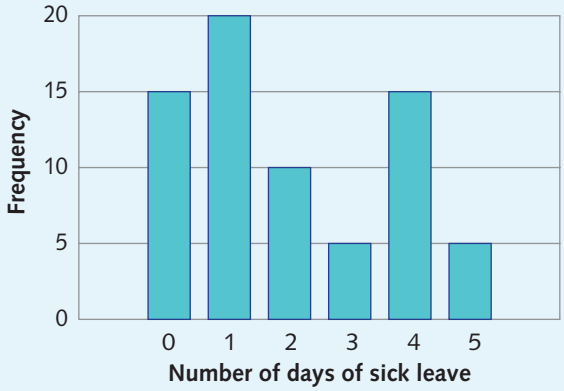
Solution

From the stem-and-leaf plot, 21 weights have been recorded. So the median will be the weight in position 11 (giving 10 weights on either side). This will be the sixth value in the second class. Hence the median weight is 58 kg.

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

Example 14

The column graph below gives the number of days of sick leave taken by employees at a factory during a particular month. Locate the median.



Solution

From the column graph, there are $15 + 20 + 10 + 5 + 15 + 5 = 70$ data values, so the median will the average of the 35th and 36th positions.

The 35th position is a 1 (there are 15 zeroes and 20 ones) and the 36th position is a 2, so

$$\text{median} = \frac{1+2}{2} = 1.5 \text{ days of sick leave}$$

Mode

One of the questions we often use statistics to answer is 'Which is the most popular?' The most popular or most common value will be the most frequently occurring value in a data set. A value with the highest frequency is called the **mode**. There may be more than one mode.

For example, in a survey of 'favourite sports' the following results were obtained:

swimming, golf, golf, badminton, swimming, cricket, golf, swimming, cricket, golf, cricket, golf, badminton, swimming, swimming, cricket, cricket, golf, swimming, cricket, swimming, swimming

We can arrange the data into a frequency table.

Favourite sport	Tally	Frequency
Golf	𝍸 I	6
Cricket	𝍸 I	6
Swimming	𝍸 III	8
Badminton	II	2

In the survey above, the sport with the highest frequency is swimming. Hence the mode is swimming. It should be noted that in this example of categorical data it is senseless to refer to mean or median.

Example 15

The number of emails Annie sent each day was recorded for 30 days. The results are shown below. Find the mean, median mode.

Number of emails	Frequency
12	8
13	6
14	4
15	8
16	4

Solution

i Mean $= \dfrac{12\times 8+13\times 6+14\times 4+15\times 8+16\times 4}{30} = \dfrac{414}{30} = 13.8$

ii The median lies in the $\dfrac{30+1}{2} = 15.5$th position (between 15th and 16th). Median $= 14$

iii Mode $= 12$ and 15 (both have the highest frequency)

There are two modes for this data, so the data are said to be **bimodal**.

Mode

The **mode** is the value (or values) with the highest frequency.

Range

The range is one of the simplest, and easiest to calculate, measures of the spread of a data set.

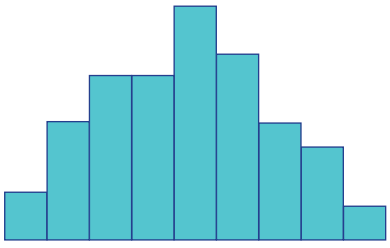
The **range** of a set of numerical data is the difference between the largest data value and the smallest data value.

For example, the range of data in Example 15 is $16 - 12 = 4$ emails.

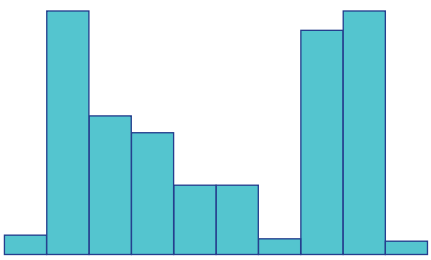
Comparing data displays

Compare the following histograms.

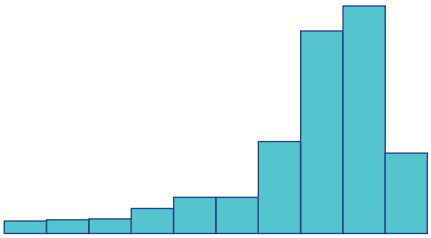
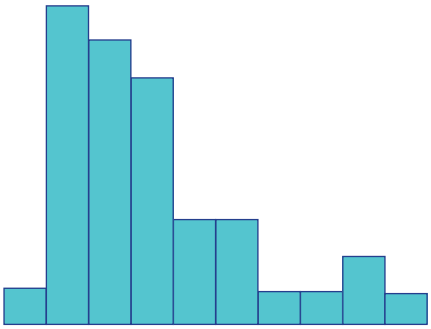
Symmetrical



Bimodal



There are two modes.

Negatively skewed	Positively skewed
	
It has a long tail to the left and a very short tail to the right.	It has a long tail to the right and a very short tail to the left.

These properties can also be seen through other representations, such as with the stem-and-leaf plot given here. The stem-and-leaf plot below displays a positively skewed distribution of data.

1	0 0 1 1 2 2
2	4 6 6
3	1 2 2 3 3 4 5 5 5 7
4	
5	6 6
6	7
7	
8	9
18	7
19	2

5 | 6 means 56

Sometimes there are one or two values very different from the others. These values are called **outliers**. For example, the values 187 and 192 are outliers in the data set above.

The mean, median and mode convey different information and each has its advantages and disadvantages. Consider the data set below, giving monthly salaries of the 16 employees in a small firm.

$7000 $7000 $7000 $7000 $7000 $7700 $7700 $8400 $8400
$9800 $9800 $11200 $11200 $13900 $20000 $42000

Median = $8400

Mode = $7000

Mean = $11568.75

The mean is much higher than both the median and mode. Normally, the median is a preferable measure of 'centre' to the mean in situations such as this. The values $20000 and $42000 are much larger than the others and have a large effect on the mean.

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

Exercise 19D

All answers are to be given correct to 1 decimal place, unless otherwise specified.

Example 10

1 Calculate the mean and mode of each data set.

a 1, 1, 3, 5, 5, 5, 10 **b** 4, 4, 4, 7, 8, 8, 10, 11

2 In a football season, the number of kicks obtained by a player week by week was:

22 16 18 31 10 8 19 16 18 12 10 9 16

Calculate the mean number of kicks per week obtained by the player.

3 Twelve students sat for a test and their results are displayed in the stem-and-leaf plot opposite.

a Calculate the mean of the marks.

b How many students obtained a mark higher than the mean?

c If a 13th student obtained a mark of 32 for the test, would the mean for the 13 students be higher or lower than the answer for part **a**?

1	8 9
2	2 4 5 6
3	1 4 9
4	2 3 6

4 | 3 means 43

4 The number of strokes scored on the 18th hole of a golf course was recorded for a number of golfers. The results are shown opposite.

a How many players had their score recorded?

b What is the average score?

c How many players took fewer strokes than the average?

d What number of strokes is the mode?

Number of strokes	Number of players
2	1
3	6
4	27
5	20
6	10

5 Five people have an average weight of 67 kg. If a child of weight 25 kg is added to the group, what is the average weight?

6 Part-way through a cricket season, a batsman has had scores of 15, 76, 42 and 27. Assume that the batsman is dismissed in each innings.

a Calculate the batsman's average.

b If his average after the next innings is 42, how many runs did he score in that innings?

c The batsman has 12 innings in a season. He wants to have an average of 50 at the end of the season. How many runs does he need to score in the remaining 7 innings?

7 During a term, a student has an average of 46 after the first 4 tests and his average for the next 6 tests is 38. What is his average for the 10 tests?

8 A data set has a mean of 15. What will happen to the mean (that is, will it decrease or increase) if:

a a data value of 24 is added to the set?

b a data value of 15 is added to the set?

c data values of 6 and 25 are added to the set?

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

Example 12

9 Find the median for each of the following data sets.

a $8, 6, 12, 4, 1, 9, 15, 3$

b $18, 26, 47, 13, 18$

c $1.6, 1.9, 2.4, 1.8, 3.7, 0.9, 2.6, 1.7$

d $647, 326, 849, 586, 710, 694$

10 Copy and complete the following table, which relates the number of data values to the position of the median. (A position of 5.5 means that the median is the average of the 5th and 6th data values.)

Number of data values	5	11	21	10	20	100				
Position of median	3			5.5			15	22	13.5	24.5

Example 13

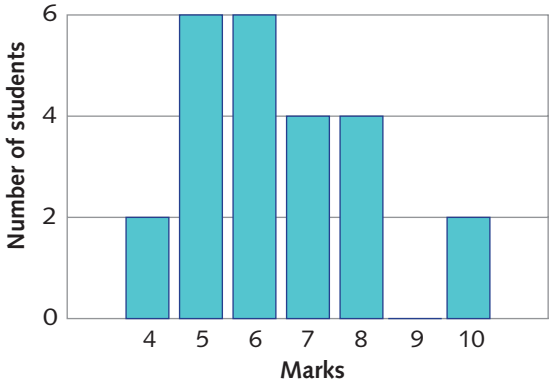
11 Fill in the frequency column, and hence find the median of the following data set.

		Frequency
5	4 5 6 8 9	
6	2 2 4 5 7 9	
7	1 3 5 5 8	
8	2 4 7 9	
9	3 5	

7 | 1 means 71

Example 14

12 The marks obtained for a quiz by a group of students are displayed in the column graph opposite.



a How many students had their marks recorded?

b What is the median of the marks?

c What is the mode of the marks?

13 The distinct values a, b, c, d, e and f are arranged here in numerical order, with a having the least value. Describe the effect on the median and the mean if:

a f is increased by 12

b the value a is deleted from the list

c a is decreased by 6 and f is increased by 6

d a is decreased by 16 and f is increased by 6

e b and e are both increased (but b is still less than c)

f c is increased by 4 (c is still less than d) and b is decreased by 4

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

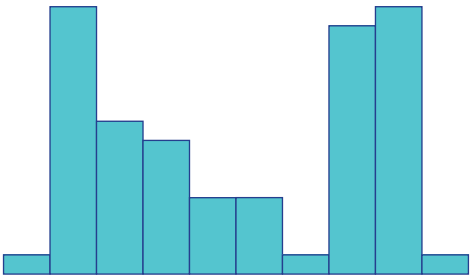
14 The following statement in a newspaper seems to be incorrect.

According to latest enrolment figures, half the student population of the university are over the age of 24. The 'average' student is 29 years old.

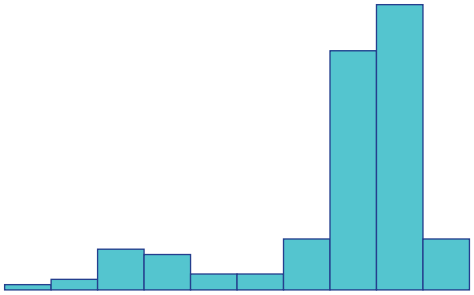
Can you give an example of a set of 4 students with mean age 24 and median age of 29?

15 Describe each set of data.

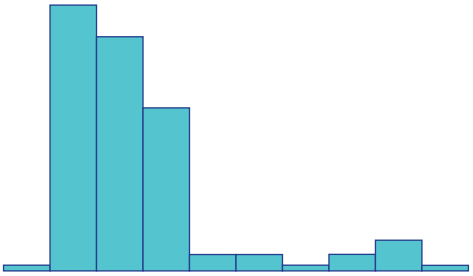
a



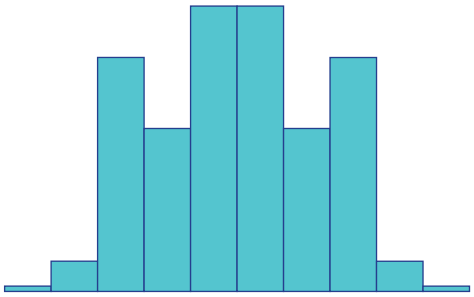
b



c



d



16 For the data in each stem-and-leaf plot, find the range, median and mean.

a

0	
1	0 0 1 1 2 2
2	4 6 6
3	1 2 2 3 3 4 5 5 5 7
4	
5	6 6
6	5 7

5 | 6 means 56

b

0	8
1	0 0
2	1 1
3	2 4
4	2 2
5	2 2 4 5 6 8 8
6	3 4 4 5 5 7 8 9

3 | 1 means 31

c Describe the distribution in part **b**.

17 Two professional baseball teams in the United States have a total of 63 players on their lists of players. The median salary of these players is \$1500000, and the mean salary is \$3382202. The range of salaries is \$31506000. What would you expect the histogram of this data to look like?

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

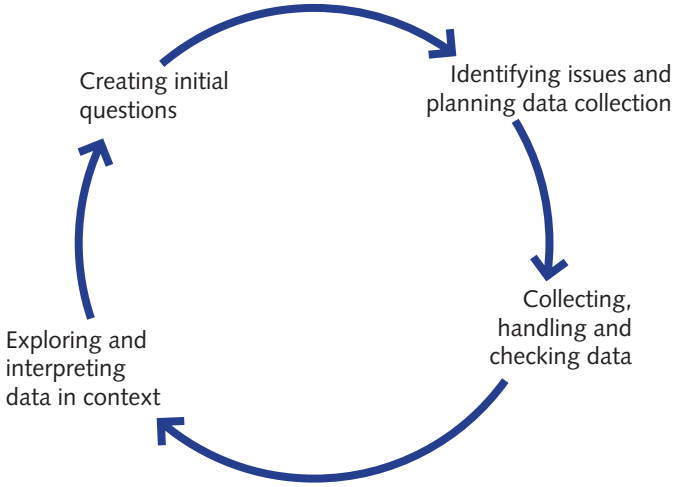
19E The statistical investigation process

The best way to understand how data is collected and presented is to do data-collection and presentation activities yourself. Some of the different ways that you can organise, present and discuss data have been explained throughout this chapter.

In this section we discuss the statistical data investigation process. Although there are many ways to collect and organise data, we will use the following steps:

- creating initial questions
- identifying issues and planning data collection
- collecting, handling and checking data
- exploring and interpreting data in context.

The statistical data investigation process is illustrated in the following diagram.



You can see that the statistical process never really ends. Once we have interpreted the data, we might discover that there are more questions we want to ask and more data to collect.

As an example, consider this process in the context of the suggested Activity, 'Reaction times (left and right)'.

Initial questions	• How long is the reaction time? • Is there a difference between left and right side reaction times? • Are left-handed students consistently different to right-handed students in these recorded times?
Issues and planning	• What height should the ruler be held in relation to the person? • Should several attempts be recorded and the average taken? • How accurate is the measurement of reaction distance?
Collecting, handling and checking data	• Should all data be recorded? • What should be recorded if the participant fails to stop the ruler?
Exploring and interpreting data	• Display with back-to-back stem-and-leaf plots. • Interpret results by measuring mean, median and range. • Address initial questions.

Further suggestions for statistical investigations appear on the AMSI website.

Exercise 19E

1 Here are some questions to initiate a discussion on techniques for collecting data.

a What proportion of Australian households recorded a television program during the past week?

b What percentage of a large shipment of components of a machine can stand up to a stress test?

c Is a particular component of a particular model of car safe? One hundred and fifty thousand of the cars have been distributed worldwide.

2 Here are some issues for which statistical information would help. Discuss how to go about collecting it.

a A coffee-making machine manufacturer wants to study people's colour preference for their machines.

b A council wants to determine the best resources for their library to hold.

c An employer of 10 000 people wants to determine their employees attitudes towards the company.

Review exercise

1 A maths test was given to two different classes and results were recorded in the back-to-back stem-and-leaf plot below.

Class A		Class B
	3	2 4
3	4	5 5 8
9 6 6 3	5	4 9
8 6 5 4 1	6	3
8 7 5 4 2 2 0	7	2 3 3 8
7 6 3 0	8	0 1 4 4 6 7
4 1	9	1 3 6
	10	0

Note: 0 | 7 | 2 means 70% in class A and 72% in class B.

a What percentage of students scored at least 80 in:

i class A? **ii** class B?

b What percentage of students scored less than 50 in:

i class A? **ii** class B?

c What was the range of marks in:

i class A? **ii** class B?

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

d What was the median mark in:

i class A? **ii** class B?

e What was the mean mark in:

i class A? **ii** class B?

f Describe the distribution of data in:

i class A? **ii** class B?

2 T-shirts sold in a shop were tallied over a week with their sizes noted.

T-shirt size	Tally	Frequency
S	𝍸 𝍸 𝍸 III	
M	𝍸 𝍸 𝍸 𝍸 I	
L	𝍸 𝍸 II	
XL	𝍸 II	
XXL	III	

a Complete the frequency column of the table.

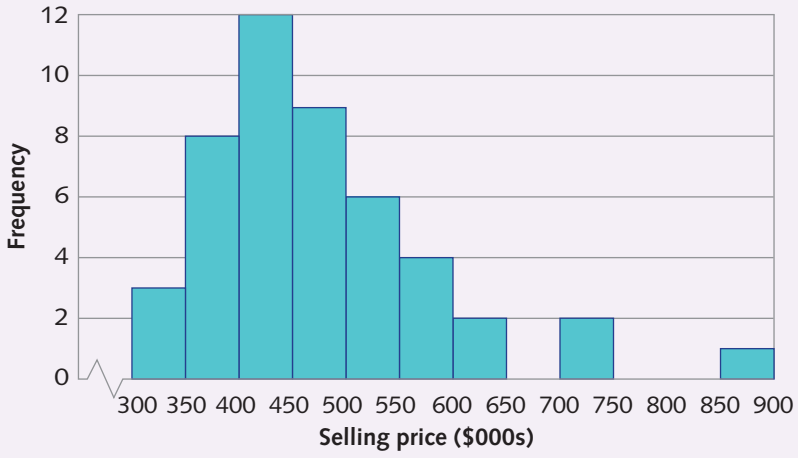
b Determine the mode of the size of the T-shirts sold.

c Determine the percentage of T-shirts sold that were small (S) or medium (M).

d Create a bar chart of the data.

e Explain why calculations of mean or median makes no sense in this situation.

3 Auction prices of houses sold in Blackpool over the month of September 2016 are shown in the histogram below.



a Determine the total number of houses sold in Blackpool over the month of September 2016.

b Determine the percentage of houses, sold in Blackpool over the month of September 2016, that are cheaper than $500 000.

c Describe the distribution of data.

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

d Determine the class of prices in which the median occurred.

e Identify whether the mean or median house price would be greater and explain why.

4 A student received the following marks for five tests: 75, 83, 74, 66 and 90.

a Calculate the mean.

b Determine the minimum mark this student will need to receive on her sixth test in order to average at least 80.

c If the student's median mark is 78 after six tests, determine the mark she received on her sixth test.

Challenge exercise

Locating a country's centre of population

The **population centre** of a country is defined as the point whose latitude and longitude are the average of the latitudes and longitudes of all people in the country. By calculating the population centre of a country, trends about population movement and growth can be obtained.

The following questions investigate the population centre of Australia in 1990. To make the calculations easier, it might be useful to use the statistics facility of a graphics calculator, a spreadsheet or a statistical software package. (Give your answers correct to 1 decimal place.)

1 The city of Melbourne has latitude 38°S and longitude 145°E, and the city of Brisbane has latitude 27°S and longitude 153°E.

a Suppose that one person lives in Melbourne and another person lives in Brisbane. What is the average of the two people's:

i latitude **ii** longitude?

b Find the average of the latitudes and longitudes if:

i 10 people lived in Melbourne and 5 people lived in Brisbane

ii 20 people lived in Melbourne and 30 people lived in Brisbane

iii 1000 people lived in Melbourne and 800 people lived in Brisbane

2 Consider the following table, which gives the latitude, longitude (both rounded to the nearest whole number) and the population in millions (rounded to the nearest million) of the five largest Australian capital cities in 1990.

City	Latitude	Longitude	Population (in millions)
Melbourne	38°S	145°E	3
Sydney	34°S	150°E	4
Perth	32°S	116°E	1
Adelaide	35°S	139°E	1
Brisbane	27°S	153°E	1

ICE-EM Mathematics 9 3ed ISBN 978-1-108-40432-7

a Calculate the population centre of the cities Melbourne and Sydney taken together (that is, the average of the 7 million latitudes for these two cities and the 7 million longitudes for these two cities).

b Calculate the population centre of Melbourne and Perth taken together.

c Calculate the population centre of Melbourne, Perth and Sydney taken together.

d Calculate the population centre of Melbourne, Adelaide and Brisbane taken together.

e Calculate the population centre of the five capital cities taken together. Locate the population centre on a map of Australia.

3 The following table gives the population (in 1000s) of the eight Australian capital cities for a number of years throughout the 20th century.

Year	Sydney	Melbourne	Brisbane	Adelaide	Perth	Hobart	Darwin	Canberra
1901	497	502	121	162	71	36	1	–
1920	885	763	206	255	152	50	1	–
1930	1191	1000	280	311	212	59	2	7
1940	1294	1083	336	330	230	69	2	12
1950	1557	1302	445	434	313	84	5	22
1960	2133	1831	578	577	409	112	12	50
1970	2752	2448	847	827	672	151	33	129
1980	3232	2760	1029	934	902	170	51	246
1990	3657	3081	1302	1050	1193	184	73	310

Using the information about latitude and longitude given in Question **2**, together with the fact that Hobart has latitude 43°S and longitude 147°E, Darwin has latitude 12°S and longitude 131°E, and Canberra has latitude 35°S and longitude 149°E, calculate the population centre of the eight capital cities for each year given above.

Use your calculations to answer the following questions.

a Describe the direction in which the Australian population centre has moved since 1901. Can you give some reasons for this?

b Estimate where the population centre for Australia was in the year 2005.

c Why is it important to recognise these shifts in population? What groups in the community need to know such information?

Note that a more realistic model would include population centres outside the capital cities.