STATISTICS AND PROBABILITY

8 STATISTICS

8A Understanding and representing data

8B Grouped data

8C Summary statistics

8D Summary statistics from displays

8E Collecting data

8F Describing data

8G Comparing data

ESSENTIAL QUESTION

How can you find and use statistics to understand your local community better?

Are you ready?

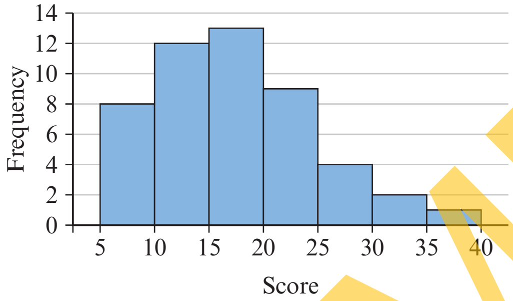
8A **1** Look at this table.

Colour	Frequency
Red	5
Yellow	4
Green	8
Blue	7
Purple	5

a Does it contain numerical or categorical data?

b How many people were surveyed?

8B **2** Look at this graph.



a What type of graph is this?

A column graph **B** bar graph

C line graph **D** histogram

b What is the size of the class intervals?

c How many people were surveyed?

8B **3** What is the most common score in this stem-and-leaf plot?

Key: 1 | 4 = 14

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

8C **4** Consider this data set.

13 4 6 1 2 5 4 10 9

a What is the range of this data set?

b What are the mean, median and mode of this data set?

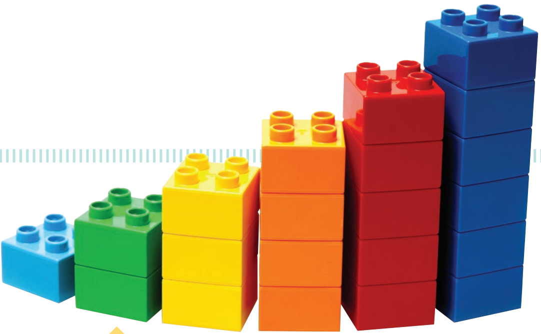
A mean = 6, median = 5, mode = 4

B mean = 46, median = 5, mode = 1

C mean = 6, median = 2, mode = 4

D mean = 46, median = 2, mode = 4

8A Understanding and representing data



Start thinking!

Data that can be counted or measured are called **numerical data**.

1 Name five different examples of numerical data (for example, height).

Data that can be put into categories or groups are called **categorical data**.

2 Name five different examples of categorical data (for example, hair colour).

Numerical data can be further split into two groups: **discrete data** can be counted (whole numbers only) and **continuous data** can be measured (includes decimal numbers).

3 Classify your examples from question **1** as either discrete or continuous.

Categorical data can be further split into two groups: **nominal data** can be arranged into unrelated groups and **ordinal data** can be arranged into groups that have an order.

4 Classify your examples from question **2** as either nominal or ordinal. If you did not include any ordinal data, think of at least two examples now.

5 Brainstorm with a classmate and list as many different types of graphs and visual displays as you can.

6 For each graph or visual display that you have listed, decide if it can be used to display numerical data, categorical data or both. Provide a reason for each decision.

7 Explain why **line graphs**, **scatterplots**, **histograms** and **stem-and-leaf plots** can only be used for numerical data.

8 Explain why **column graphs**, **bar graphs**, **pie graphs** and **dot plots** can be used to display both numerical and categorical data but are best suited to categorical data.

9 Why is it useful to collate data into a frequency table before drawing any graph?

KEY IDEAS

- Numerical data can be classified as either discrete (whole numbers only) or continuous (includes decimal numbers).
- Numerical data are best represented by visual displays such as frequency tables, histograms, stem-and-leaf plots, line graphs and scatterplots.
- Categorical data can be classified as either nominal (unrelated groups) or ordinal (groups that can be put in an order).
- Categorical data are best represented by visual displays such as frequency tables, column and bar graphs, dot plots and pie graphs.
- All graphs should include a title, clearly labelled axes with an even scale and a legend if necessary.

Check the glossary for definitions and examples of each of these graph types.

EXERCISE 8A Understanding and representing data

EXAMPLE 8A-1 Classifying data

Classify these data.

a how much people like chocolate
b number of people at a cinema

THINK	WRITE
a 1 Decide if the data is categorical or numerical.	a
2 Are the categories unrelated (nominal) or do they have an order (ordinal)?	How much people like chocolate is categorical, ordinal data.
b 1 Decide if the data is categorical or numerical.	b
2 Is the data in whole numbers (discrete) or decimal numbers (continuous)?	The number of people in a cinema is numerical, discrete data.

UNDERSTANDING AND FLUENCY

1 Classify these data.

a eye colour
b how much you like winter
c number of pets at home
d favourite movie type
e length of arm span
f distance between home and school
g shoe size
h number of girls in class
i mass of a car
j type of computer
k how fit somebody is
l number of planets in the solar system

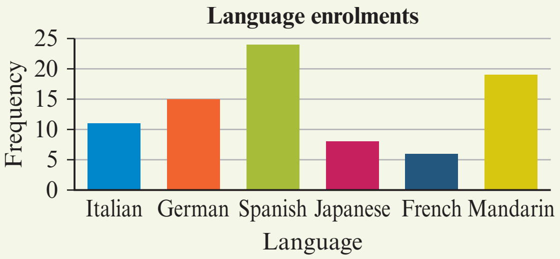
EXAMPLE 8A-2 Reading graphs

Consider this column graph.

a What is it showing?

b What is the least popular language?

c How many people were surveyed in total?



THINK	WRITE
a Look at the axes and the title of the graph.	a The graph is showing enrolment numbers for six languages.
b Look at the smallest column.	b The least popular language is French.
c Add all the frequencies together.	c Number of people surveyed $= 11 + 15 + 24 + 8 + 6 + 19 = 83$ 83 people were surveyed.

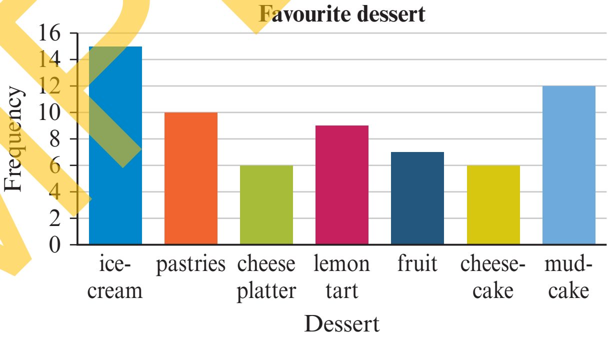
UNDERSTANDING AND FLUENCY

2 Consider this column graph.

a What is it showing?

b How many people were surveyed in total?

c If you were to offer three choices of dessert, which would you choose and why?



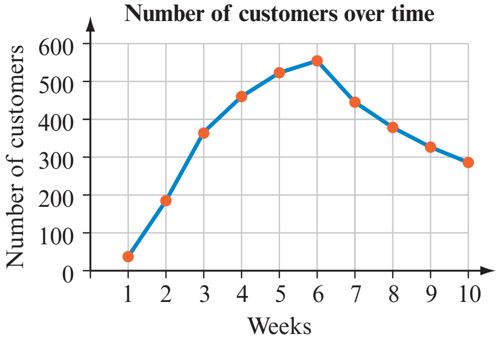
3 Consider this line graph.

a What is it showing?

b What time period does it cover?

c In what week is the maximum number of customers?

d In what week does the number of customers increase the fastest?



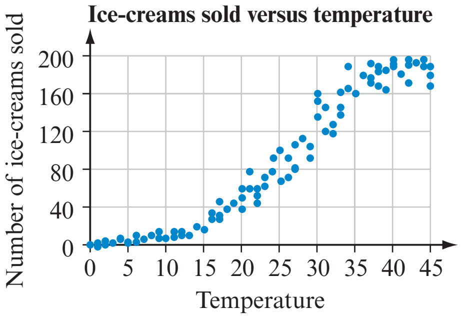
4 Consider this graph.

a What type of graph is it and what is it showing?

b How would you find how many people were surveyed?

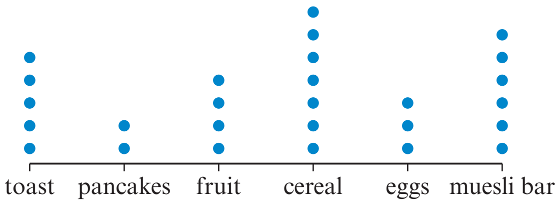
c How might you describe the pattern that you see in the graph?

d If somebody asked what the average number of ice-creams sold on a day in April is, what would you tell them?



UNDERSTANDING AND FLUENCY

5 Consider this dot plot.



a How many people were surveyed?

b Briefly describe what the dot plot shows.

c Explain why dot plots should not be used when a large number of people are surveyed.

6 Consider this stem-and-leaf plot showing the ages of people in a cinema.

a How old is the youngest person at the cinema?

b How old is the oldest person at the cinema?

c What is the most common age bracket at the cinema?

d How does the key help you to read the stem-and-leaf plot?

Key: 2 | 1 = 21

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

EXAMPLE 8A-3 Representing data

Gill collected this data on popular hobbies from a group of students.

reading a book, listening to music, talking to friends, playing digital games, listening to music, exercising, talking to friends, listening to music, playing digital games, playing digital games, listening to music, reading a book, playing digital games, talking to friends, playing digital games, exercising

Classify the data type and present it in an appropriate graph.

THINK

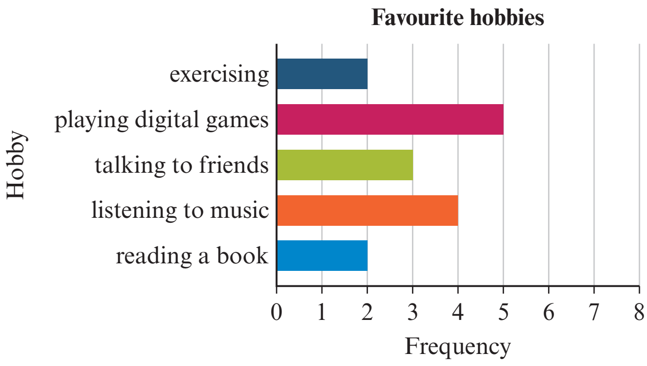
1 This data can be arranged into unrelated groups, so it is categorical nominal data.

2 Decide on an appropriate graph that suits categorical, nominal data. Suitable graph types are a column graph, bar graph, pie graph or a dot plot.

3 Draw your chosen graph, remembering to label both axes and include an appropriate title.

WRITE

It is categorical, nominal data.
A bar graph would best suit this data because the category names are long and it shows the frequencies.



UNDERSTANDING AND FLUENCY

7 Decide which type of data these graphs best represent.

a column graph b histogram c pie graph

d stem-and-leaf plot e line graph f dot plot

8 Joe collected this data on favourite colours.

blue, purple, green, blue, pink, pink, yellow, blue, green, red, pink, blue, red, purple, pink, green, blue, blue, yellow, purple, orange, blue, green, red, blue, red, purple, pink, blue, green, purple, purple, purple, blue, red, pink, green, blue, blue.

Classify the data and present it in appropriate graph.

9 Create an appropriate graph to represent the data in each frequency table.

a

Day	Mass (g)
1	15
2	20
3	22
4	28
5	37
6	45

b

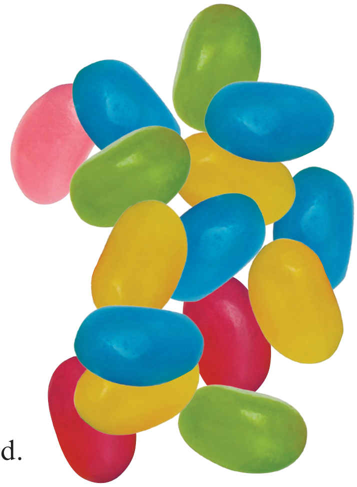
Movie type	Frequency
Action	9
Comedy	14
Drama	7
Horror	4
Animated	10

c

Height (cm)	161	176	154	178	176	185	166	164	155	172	161	172	165	176	172	164
Weight (kg)	54	70	51	76	76	75	62	65	55	61	58	69	57	66	65	57

PROBLEM SOLVING AND REASONING

10 Create an appropriate graphical display to represent the objects shown in the photograph.



11 Create a table that shows which graphs can be used for each type of data. (Hint: the table should contain three columns with the headings ‘Graph type’, ‘Numerical data’ and ‘Continuous data’.) Use ticks to complete the table.

12 For each type of graph listed in the Key ideas:

a give a definition/description of the graph

b draw an example

c explain when and for what type of data the graph is best used.

13 Explain how stem-and-leaf plots can represent numerical, continuous data.

14 Position on a sports ladder is often classified as the wrong data type.

a Find a current sports ladder (or create your own).

b Would you say that it is numerical or categorical data? Why?

PROBLEM SOLVING AND REASONING

c Can you measure or add together the numbers shown in a sports ladder? Explain.

d Use your answer to part c to explain why it cannot be numerical data.

e Is there an order to a sports ladder?

f What type of data is position on a sports ladder?

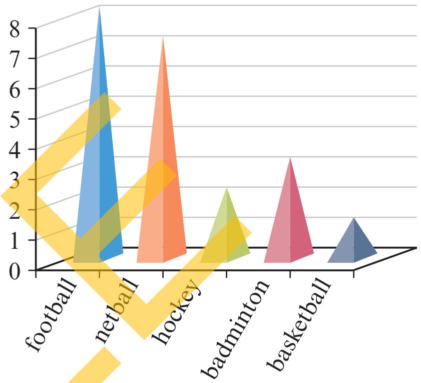
15 Explain why a pie graph is rarely the best graph to represent data, even though it is so commonly used.

16 Graphs can often be used in a misleading way in order to support a person's point of view. For each of these graphs:

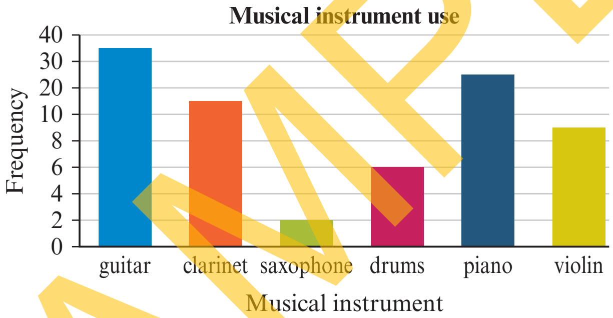
i describe how the graph is or could be misleading

ii describe what would need to be done in order to make the data in the graph clearer.

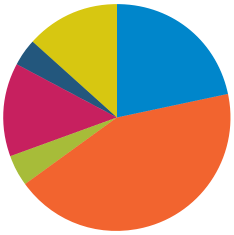
a



b



c



CHALLENGE

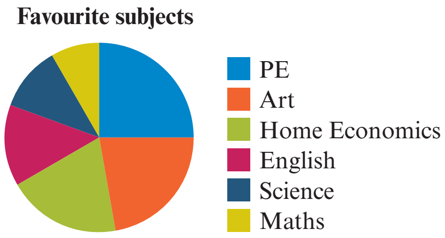
17 Consider this pie graph.

a What is it showing?

b What is the most popular subject?

c Explain why you cannot tell how many people chose Maths.

d If 36 people were surveyed, use your understanding of angles within a circle to find the number of people who chose each category. (Hint: you will need a protractor.)



Another 14 people were surveyed.
Three chose Maths, five chose Art, four chose Information technology and two chose PE.

e Redraw the pie graph to include these new people. Does this change the most popular subject? (Hint: use your answers from part d and the new information to first create a frequency table.)

Reflect

If all types of graphs can be used to represent numerical data, why are some better suited than others?

8B Grouped data

Start thinking!

Joseph wanted to collect data on the heights of the people in his class. Before drawing a frequency table, he measured the heights of the shortest person (145 cm) and the tallest person (183 cm) in the class.

1 How many rows would a simple frequency table require to cover this range of data?

When data covers a large range, you can group it into class intervals. Each table row contains a spread of data.

2 Explain why there should be no less than 5 groups and no more than 10 groups in a frequency table.

3 What class intervals should Joseph use for his frequency table?

Joseph collected this data:

145, 183, 167, 172, 161, 158, 153, 168, 165, 174, 157, 152, 173,
166, 158, 159, 160, 171, 171, 161, 165, 172, 165, 158, 154.

4 Arrange this data into a frequency table with your chosen class intervals from question **5**. Include a row that gives the total frequency.

When grouping data it is important to consider data type.

5 What are the two types of numerical data?

6 What type of data is height?

7 If you used class intervals such as 140–144, 145–149, where would you place 144.6 cm?

When grouping continuous data, use 'open' class intervals such as 140–<145.

8 Redraw your frequency table if necessary so that it uses open class intervals.

9 Why is it important to use open class intervals for this scenario?

KEY IDEAS

- Tables displaying grouped numerical data make use of class intervals to group the data.
- Class intervals should be chosen so that a table contains 5–10 groups.
- Identify the data type before constructing a table: continuous data need class intervals such as 0–<10; discrete data can be shown this way or in class intervals such as 0–9.
- A histogram can be used to represent grouped numerical data.
- There are no gaps between the columns of a histogram (but there is a small gap between the vertical axis and the first column) and category marks should be on the edges of the columns.

EXERCISE 8B Grouped data

EXAMPLE 8B-1 Using a frequency table to represent data

Draw an appropriate frequency table to represent this data.

4.5, 11.6, 67.3, 33.7, 28.1, 36.4, 22.6, 54.8, 1.4, 66.8, 36.4, 29.3, 37.8, 42.3, 52.1, 38.3

THINK

1 A frequency table should have 5–10 groups. The minimum score is 1.4 and the maximum score is 66.8. This gives a range of 65.4. Class intervals of 10 could be used, giving 7 groups.

2 This data is continuous, so the class intervals must in the form of 0–<10.

3 Draw the frequency table using the raw data. You may wish to include a tally column to ensure that you don't miss any scores.

WRITE

Class	Frequency
0–<10	2
10–<20	1
20–<30	3
30–<40	5
40–<50	1
50–<60	2
60–<70	2

UNDERSTANDING AND FLUENCY

1 Draw an appropriate frequency table to represent these data.

a 5, 16, 28, 24, 31, 39, 3, 18, 13, 11, 25, 33, 8, 12, 19, 21, 31, 28

b 14.5, 73.2, 22.1, 43.9, 42.0, 58.4, 19.8, 37.6, 62.1, 29.4, 34.5, 72.1, 59.1, 52.3, 63.1, 26.3, 34.0, 41.9, 48.5, 16.4, 31.2, 52.9

c 1.2, 5.4, 9.3, 11.4, 3.3, 4.7, 3.3, 3.9, 4.8, 6.6, 2.9, 1.9, 10.6, 9.7, 10.8, 3.6, 4.8, 2.7, 2.1, 1.7, 1.9, 11.9, 6.7, 5.4, 5.1, 1.6, 1.8

d 42, 79, 56, 49, 77, 50, 51, 46, 48, 72, 61, 78, 63, 45, 58, 53, 73, 58, 49, 61, 68, 67, 43, 49, 75, 77, 58, 54, 67, 72, 51, 56, 53, 48, 76, 78, 72, 42, 48, 53

2 Use each table to draw a histogram.

a

Class	Frequency
0–<5	8
5–<20	6
15–<20	7
20–<25	2
25–<30	3
30–<35	1
35–<40	5
40–<45	6
45–<50	9

b

Class	Frequency
0–<20	14
20–<40	21
40–<60	18
60–<80	13
80–<100	8
100–<120	2

c

Class	Frequency
0–<10	5
10–<20	8
20–<30	12
30–<40	3
40–<50	11
50–<60	9
60–<70	6

EXAMPLE 8B-2 Drawing a histogram

Use this data to draw an appropriate histogram.

34, 22, 29, 16, 12, 16, 26, 32, 39, 20, 23, 19, 36, 25, 11, 16, 13, 13, 19, 28, 12, 14, 32, 10

THINK

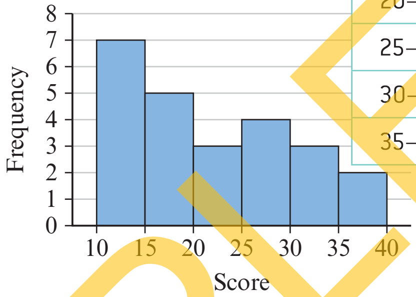
1 Collate the data into a frequency table. Ensure that there are 5–10 class intervals.

2 Draw the axes with an even scale that allows the minimum and maximum values to be shown. Ensure that there is a half space between the vertical axis and first category mark.

3 Draw the histogram and remember to label both axes and give it a title.

WRITE

Class	Frequency
10–<15	7
15–<20	5
20–<25	3
25–<30	4
30–<35	3
35–<40	2



UNDERSTANDING AND FLUENCY

3 Use these data to draw an appropriate histogram.

a 13, 46, 13, 17, 35, 9, 22, 15, 8, 2, 35, 42, 42, 17, 16, 22, 29, 31, 47, 29, 13, 20, 36, 47, 28, 23, 30, 38

b 18.1, 24.5, 32.1, 15.6, 22.5, 29.1, 34.6, 16.7, 19.4, 17.5, 21.8, 27.5, 29.2, 30.1, 20.0, 33.1, 32.8, 31.9, 33.8, 14.3

c 64, 18, 120, 7, 29, 40, 145, 38, 72, 38, 18, 29, 2, 56, 49, 87, 99, 104, 59, 5, 29, 112, 118, 34, 59, 29, 19, 13

d 125, 726, 632, 465, 428, 257, 283, 399, 619, 402, 132, 196, 183, 743, 120, 703, 336, 652, 349, 402, 560, 144, 759, 717, 588, 185, 464, 685, 268, 352, 310, 408, 114, 782, 189

e 1.25, 1.89, 1.09, 1.76, 1.15, 1.36, 1.55, 1.67, 1.99, 1.32, 1.08, 1.14, 1.17, 1.62, 1.88, 4.9, 1.68, 1.49, 1.08, 1.16, 1.24, 1.19, 1.26, 1.83, 1.52, 1.18, 1.07, 1.42, 1.01, 1.19

f 25, 58, 48, 33, 26, 53, 42, 49, 58, 53, 46, 24, 58, 53, 46, 41, 38, 47, 44, 58, 53, 57, 39, 21, 48, 46, 42, 58, 52, 43, 42, 37, 36, 27, 46, 42, 49, 53, 57, 59

4 Data was collected on the number of hours spent listening to music per week, as shown below.

7	10	2	4	24	3	7	9	5	15
17	16	19	20	5	3.5	5	10	14	7
7	5	9	10	17	7	4	10	11	16
4	5.5	12	14	6	7	12	14	16	3

a Create an appropriate frequency table to collate the data.

b Draw a histogram to represent this data.

EXAMPLE 8B-3 Drawing a stem-and-leaf plot

Draw a stem-and-leaf plot for this data set.

1.8, 2.6, 1.9, 3.4, 5.2, 1.8, 2.7, 4.2, 4.9, 5.1, 7.6, 3.1, 4.1,
3.0, 2.8, 2.1, 1.9, 1.3, 2.8, 2.9, 4.3, 4.9, 5.1, 3.3, 2.0.

THINK

1 The minimum value is 1.3 and the maximum value is 7.6, so show stems ranging from 1 to 7.

2 Place each piece of data into the plot. Start with 1.8, which has a stem of 1 and a leaf of 8. Write the digit 8 in the stem 1 row. Continue until all values have been considered.

3 Rearrange the leaves so that they are in order. Include a key.

WRITE

Key: 1 | 3 = 1.3

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

UNDERSTANDING AND FLUENCY

5 Consider this stem-and-leaf plot.

a Which part of the plot represents the class intervals?

b What is the size of the class intervals?

c How many people were surveyed?

d What advantage does a stem-and-leaf plot have over a histogram?

Key: 1 | 2 = 12

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

6 Draw a stem-and-leaf plot to represent each data set.

a 16, 7, 36, 67, 14, 25, 42, 37, 19, 2, 46, 48, 51, 22, 18, 6, 17, 13, 13, 9, 11, 27, 31, 36, 42, 15, 23, 59, 33, 36, 99.

b 2.2, 2.7, 2.8, 1.6, 5.9, 3.4, 4.8, 6.2, 3.7, 2.8, 1.2, 4.2, 4.8, 5.1, 4.2, 5.3, 1.7, 1.9, 3.3, 2.2, 4.4, 4.8, 4.3, 1.8, 3.4.

c 56, 74, 36, 85, 22, 16, 48, 26, 95, 102, 16, 75, 59, 32, 15, 18, 68, 92, 43, 55, 12, 64, 66, 72, 42, 42, 18, 33, 81, 108, 111, 117, 19, 33, 36, 49, 43, 47, 52, 61, 77, 19, 8, 26, 22, 88, 46, 73, 42, 29.

d 112, 162, 124, 163, 177, 113, 134, 142, 165, 133, 119, 126, 142, 137, 153, 143, 166, 118, 121, 127, 132, 119, 144, 132, 119, 126, 172, 164, 134, 153, 142, 167, 146, 132, 119, 113, 127, 164, 138, 142, 165, 113.

7 Data was collected on the ages of customers in a clothes store in a day (see table on the right).

a Why is this table difficult to read?

b Redraw the table with larger class intervals so that it is easier to read.

Class	Frequency
<10	2
10–<12	6
12–<14	12
14–<16	16
16–<18	14
18–<20	13
20–<22	11
22–<24	9
24–<26	10
26–<28	8
28–<30	6
30–<32	4
32–<34	6
34–<36	5
36–<38	3
38–<40	2
40–<42	3
42–<44	1
44–<46	1
46–<48	0
48–<50	1
>50	6

UNDERSTANDING AND FLUENCY

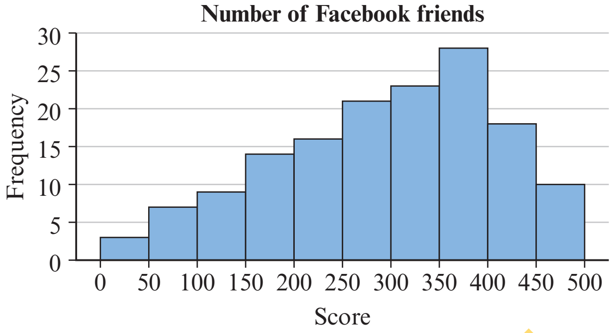
8 Use the stem-and-leaf plot on the right to draw a histogram. Why is this reasonably easy to do?

Key: 1 | 2 = 12

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

9 Explain why you can't accurately draw a stem-and-leaf plot from a histogram.

10 Consider this histogram.



a What does it show?

b What is the size of its class intervals?

c State the most common class and its frequency.

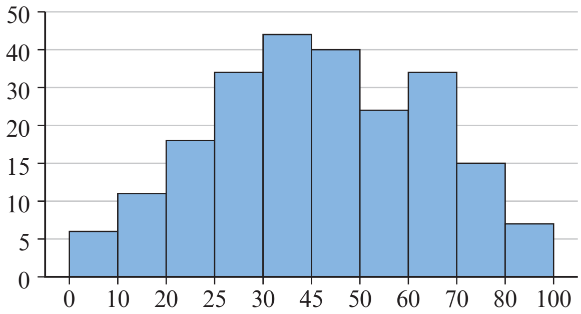
d If you were to add another piece of data, such as 40, to the histogram, to which class interval would you add it? Explain.

PROBLEM SOLVING AND REASONING

11 Explain why you can't accurately decrease the size of the class intervals in this table.

Class	Frequency
0–<20	16
20–<40	48
40–<60	42

12 Consider this histogram.



a How would you improve this histogram?

b Redraw the histogram so that it is accurate.

c What is the most common class interval in this improved histogram? What is its frequency?

d How many people were surveyed for this histogram?

13 Stem-and-leaf plots are not limited to class intervals of 10. You can split the stems of plots so that they can be more easily read. Consider this split stem-and-leaf plot, showing the ages of people buying a cinema ticket for a particular film.

Key: 1 | 2 = 12

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

a Look at the values of the leaves. What is the size of the class intervals?

b What is the most common class interval?

c To which class interval would you add the value 25?

PROBLEM SOLVING AND REASONING

14 Use this data to draw a split stem-and-leaf plot and comment on what pattern you see.

8	24	18	17	2	13	22	8	9	11	16	10	7	16	12
13	13	22	19	6	5	9	7	14	12	11	10	20	16	13
11	20	17	19	8	9	12	13	7	14	16	19	21	9	12

15 How might you represent the stems of a stem-and-leaf plot that has class intervals of 2?

16 This stem-and-leaf plot is missing its key. For the possible keys below, state:

a the minimum score

b the maximum score

c the size of the class intervals

d the range of the plot.

i Key: 1|2 = 12 ii Key: 1|2 = 1.2

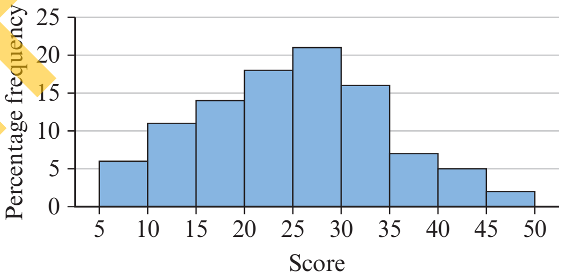
iii Key: 1|2 = 1200 iv Key: 1|2 = 0.12

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

17 Draw a histogram to represent the heights of people in your class. Be sure to first draw a frequency table with appropriate class intervals. Which class interval do you belong in?

CHALLENGE

18 Consider this percentage frequency histogram.



a How is it different from a normal histogram?

b What percentage of scores are between 30 and 35?

c What percentage of scores are greater than 35?

d Without performing a calculation, state the sum of the percentage frequency columns. Explain how you know.

e If there were 400 scores in total, calculate:

i the number of scores between 15 and 20

ii the number of scores less than 25

iii the number of scores between 20 and 40.

19 Create a percentage frequency histogram to represent this data set.

Weights of newborn babies at a particular hospital in one week (in kg)

3.25, 4.15, 2.75, 3.60, 3.95, 3.05, 2.85, 4.20,
1.95, 3.50, 3.65, 3.15, 3.70, 3.95, 4.10, 4.85,
2.90, 3.10, 3.30, 3.25, 3.50, 4.05, 3.45, 3.85,
3.75, 3.15, 3.45, 3.20, 3.25, 4.25, 2.55, 2.95,
3.40, 3.85, 3.80, 3.55, 3.20, 3.00, 3.20, 3.75,
4.00, 4.15, 3.80, 3.75, 3.40, 3.25, 3.15, 3.05,
3.85, 2.95.

Reflect

In which situations would you use a histogram to represent data and in which would you use a stem-and-leaf plot? Why?

8C Summary statistics

Start thinking!

A numerical data set is usually summarised by its **centre** and **spread**.

Kylie collected data on ages of students in a canteen, shown below.

14 12 17 15 14 16 16 18 12 13 14 13 15
16 14 12 13 15 16 14 15 13 16 14 15

1 How many pieces of data (scores) are there?

2 Arrange the data in order. What is the most common number?
This measure of centre is called the **mode**.

Another measure of centre is to find the exact middle of a data set, called the **median**.

3 Which score is the median of this data set?

4 Why is it important to ensure that the scores are ordered before finding the median?

The final measure of centre is the **mean** and is commonly known as the average. The mean is found by adding together all scores, then dividing by the number of scores.

5 What is the mean of this data set?

The most common measure of spread is the range; the difference between the minimum and maximum scores.

6 Find the range and explain why it is easier to do so when the data set is ordered.

KEY IDEAS

- A piece of data is often called a score rather than a number.
- There are three main measures of centre: the mode (most common number or numbers); the median (the middle of the ordered set) and the mean (the average of the set).
- One measure of spread is the range: the difference between the maximum and minimum scores.
- To calculate the mean from a table, add a column multiplying the score by the frequency. Divide the total of this new column by the total of the frequencies. In this table, the mean is 30 ÷ 20 = 1.5.
- To calculate the median from a table, draw in a column adding together the frequencies. The row containing the $\frac{n+1}{2}$th score is then easily identified. In this table, the median is the 10.5th score, which is in the row '1', so the median is 1.
- An **outlier** is a piece of data that is very different from the rest of the data set.

Score (x)	Frequency (f)	Score × frequency ($x \times f$)	Cumulative frequency
0	3	$0 \times 3 = 0$	3
1	8	$1 \times 8 = 8$	11
2	7	$2 \times 7 = 14$	18
3	1	$3 \times 1 = 3$	19
4	0	$4 \times 0 = 0$	19
5	1	$5 \times 1 = 5$	20
Total	20	30	

EXERCISE 8C Summary statistics

EXAMPLE 8C-1 Finding summary statistics from raw data

Find the mean, median, mode and range for this data set.

5, 7, 8, 3, 4, 6, 2, 4, 9, 3

THINK	WRITE
1 To find the mean, add all the scores together and divide them by how many scores there are.	$\frac{5 + 7 + 8 + 3 + 4 + 6 + 2 + 4 + 9 + 3}{10} = 5.1$ The mean is 5.
2 To find the median, rearrange the scores in order and find the middle number. If the set contains an even number of scores, find the average of the two middle scores.	2, 3, 3, 4, 4, 5, 6, 7, 8, 9 The median is $\frac{4 + 5}{2}$
3 To find the mode, look at the ordered number list and state the most common scores(s).	The modes are 3 and 4.
4 To find the range, subtract the lowest number from the highest number.	$9 - 2 = 7$ The range is 7.

UNDERSTANDING AND FLUENCY

1 For each data set, find:

i the mean ii the median

iii the mode iv the range.

a 12, 4, 8, 2, 9, 5, 2

b 5, 8, 1, 4, 7, 10, 2, 5, 3

c 12, 16, 12, 7, 8, 11, 14, 6, 13, 18, 4

d 20, 21, 28, 15, 32, 19, 25, 38, 22

2 For each data set, find:

i the mean ii the median

iii the mode iv the range.

a 3, 11, 16, 8, 4, 7, 12, 9

b 2, 8, 5, 9, 7, 4, 3, 6, 2, 4

c 15, 12, 6, 35, 7, 8, 9, 10, 10, 8

d 100, 125, 148, 122, 76, 118, 142, 148, 109, 122

EXAMPLE 8C-2 Finding the mean from a table

Find the mean for the data shown in this table.

Score (x)	Frequency (f)
1	4
2	7
3	8
4	8
5	2
6	1

THINK

1 Add a 'score × frequency' column to the table and complete it.

2 Divide the two totals to find the mean.

WRITE

x	f	$x \times f$
1	4	4
2	7	14
3	8	24
4	8	32
5	2	10
6	1	6
Total	30	90

mean = 90 ÷ 30 = 3

UNDERSTANDING AND FLUENCY

3 Find the mean for the data shown in each table, correct to two decimal places.

a

Score (x)	Frequency (f)
1	6
2	7
3	5
4	3
5	1

b

Score (x)	Frequency (f)
10	8
20	6
30	8
40	2

c

Score (x)	Frequency (f)
13	3
14	4
15	8
16	11
17	12
18	4

d

Score (x)	Frequency (f)
0	11
1	13
2	6
3	3
4	1

e

Score (x)	Frequency (f)
15	29
20	41
25	58
30	72

f

Score (x)	Frequency (f)
1	6
2	11
3	9
4	4
5	3
19	1

EXAMPLE 8C-3 Finding the median from a table

Find the median for the data shown in Example 8C-2.

THINK

1 Add a cumulative frequency column to the table.

2 Find the $\frac{n+1}{2}$th score, where n is 30.
The 15.5th score will be in the row containing scores of 3.

WRITE

x	f	cf
1	4	4
2	7	11
3	8	19
4	8	27
5	2	29
6	1	30
Total	30	

median = 3

4 Find the median for the data shown in each table from question 3.

EXAMPLE 8C-4 Finding the mode and range from a table

Find the mode and range for the data shown in Example 8C-2.

THINK

1 The mode is the score with the highest frequency. The highest frequency in the table is 8. Which scores have this frequency?

2 The range is the difference between the highest score and the lowest score.

WRITE

modes = 3, 4

range = 6 – 1 = 5

UNDERSTANDING AND FLUENCY

5 Find the mode(s) and range for the data shown in each table from question 3.

6 Find the mean, median, mode and range for the data shown in each table.

a

Score (x)	Frequency (f)
1	3
2	5
3	8
4	7
5	2

b

Score (x)	Frequency (f)
5	8
6	6
7	4
10	2

c

Score (x)	Frequency (f)
10	2
11	4
12	7
13	8
14	3
15	1

PROBLEM SOLVING AND REASONING

7 A number of people were surveyed on how many pairs of shoes they bought in a year. Use the results shown in the table to find the mean, median, mode and range.

Score (x)	Frequency (f)
1	9
2	13
3	21
4	18
5	13
6	11
7	7
8	4
9	1
10	3

8 Explain why the only summary statistic that you can find for categorical data is the mode.

9 Consider this data obtained on ages of students in a sports club.

14 15 17 13 14 15 16 17 16 15 14 15 15
16 16 16 17 16 17 15 17 14 16 16 16

a Find the mean, median, mode and range for the data set.

One student included the age of the coach (62) in their data set.

b Recalculate the summary statistics for the data set to include the coach.

c How does the inclusion of this piece of data affect the statistics?

A piece of data that is very different from the rest of the data set is called an outlier.

d How does an outlier affect the mean? How does it affect the median?

e Which measure of centre (mean or median) would you use to describe this data set? Why?

10 Would you use the mean or median as the measure of centre for these data sets.

a 4, 8, 4, 2, 6, 9, 5, 23, 8, 5, 2, 6, 9, 4, 2, 6, 9, 5, 6, 3, 8, 6, 5, 7, 9, 7, 3, 5, 2, 4, 3

b 11, 14, 19, 29, 46, 23, 18, 8, 33, 38, 22, 27, 13, 16, 19, 37, 42, 49, 35, 28, 25

c 87, 99, 123, 145, 134, 98, 106, 114, 128, 32, 148, 133, 88, 107, 111, 135, 122

d

Score	Frequency
10	1
20	6
30	9
40	7
50	5

11 Summary statistics can also be calculated on tables that use class intervals. Consider this raw data list showing the ages of people at an all-ages festival.

16 48 22 28 31 27 19 18 17 20 24 23 35 42
16 18 18 21 22 19 27 26 19 17 18 21 22 38
19 18 31 24 16 18 34 27 21 20 18 17

a Use the raw list to calculate the mean, median, mode and range.

b Construct a frequency table with class intervals of 5 to represent the data.

c What is the modal class? How does this relate to the mode of the raw data?

d What is the median class? How does this relate to the median of the raw data?

e What is the range of class intervals? How does this relate to the range of the raw data?

PROBLEM SOLVING AND REASONING

To find the mean, use the midpoint of each class interval.

f Add a column to your frequency table that gives the midpoint (or halfway point) of each class interval.

g Add another column to your table and calculate midpoint × frequency for each class interval. (Hint: this is the same process as 'score × frequency' but using the midpoints instead.)

h Use the sum of the midpoint × frequency column and the total number of scores to find the mean.

i How does this mean compare to the mean you found using the raw data?

12 Dinith collected data on the number of ice-creams sold per day in January, shown below.

52 46 13 16 21 29 33 59 46 43 47 46
42 38 8 22 19 27 31 38 44 42 58 55
52 53 47 36 48 42 45

a Calculate the mean, median, mode and range on the raw data.

b Construct a frequency table with class intervals of:

i 5 **ii** 10 **iii** 15.

c Calculate the mean, median, modal class and range for each table in part **b**.

d What can you say about the effect of class intervals (in particular the size of the class intervals) on the accuracy of summary statistics (in particular the mean)?

13 Another measure of spread is **standard deviation**. It measures the average spread from the mean. A small standard deviation means that most values are close to the mean and a large standard deviation means that the values are spread far from the mean. Standard deviation is best found using a calculator with the appropriate function. It can also be found (for a sample) using the formula

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

where s represents standard deviation, Σ means 'the sum of', $\bar{x}$ represents the mean, x represents an individual score and n represents the number of scores.

a Use your calculator or the standard deviation formula to find the standard deviation for each data set, correct to two decimal places.

i 4, 8, 2, 6, 4, 9, 7, 7, 4, 1, 2, 4, 6, 9, 7, 4, 6, 7, 8, 1, 2, 3, 1, 6, 2, 6

ii 22, 27, 35, 64, 12, 74, 37, 93, 27, 33, 11, 71, 64, 42, 81, 37, 13, 19, 88, 50

iii 103, 118, 109, 111, 117, 116, 117, 117, 105, 107, 116, 113, 108, 109, 112, 113

b Use your results to state if each data set from part **a** has a large or small spread from the mean.

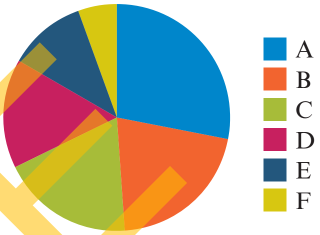
Reflect

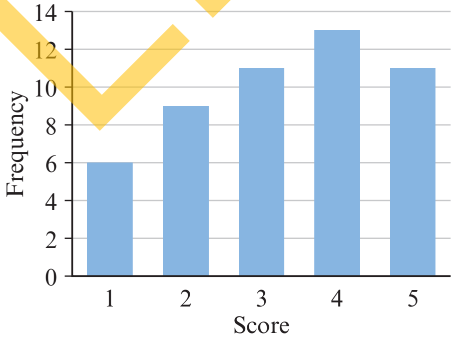
Which summary statistics do you think are most useful to find and use? Explain.

8D Summary statistics from displays

Start thinking!

Consider these two graphical displays.





1 Explain why the mode is the only summary statistic that can be found using the pie graph.

2 How does this relate to the type of data usually displayed in a pie graph?

3 Which two summary statistics are easy to read from the column graph?

4 Find the summary statistics you identified in question 3 for the column graph.

5 How would you find the mean and median from this column graph?

To make finding summary statistics from a graph easier, create a table from the graph. This is not necessary, but it can make the process simpler and more obvious.

6 Create a frequency table from the column graph.

7 Add a score × frequency column to your table and calculate the mean.

8 Add a cumulative frequency column to your table and find the median.

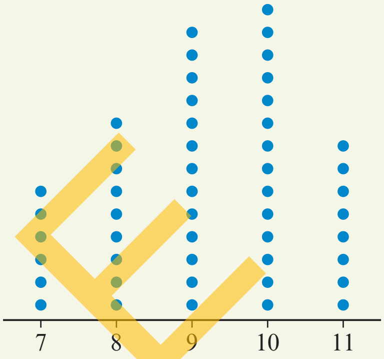
KEY IDEAS

- To calculate summary statistics from a display that individually lists scores (such as a stem-and-leaf plot), the data can be treated either as a raw list (most accurate) or as a table with class intervals (approximate).
- To calculate summary statistics from other displays (such as dot plots and column graphs), a table can be created to help with the calculations.
- The mode and range are usually very easily read from any display.
- If a display shows categorical data, only the mode can be found.

EXERCISE 8D Summary statistics from displays

EXAMPLE 8D-1 Finding summary statistics from a dot plot

Find the mean, median, mode and range for the data displayed in this dot plot.



THINK

WRITE

1 Create a frequency table by counting the number of dots in each column.

Score (x)	Frequency (f)
7	6
8	9
9	13
10	14
11	8

2 Add a total row, a 'score × frequency' column and a cumulative frequency column to the table.

x	f	$x \times f$	cf
7	6	42	6
8	9	72	15
9	13	117	28
10	14	140	42
11	8	88	50
Total	50	459	

3 Find the mean by dividing the $x \times f$ total by the f total.

$$\text{mean} = 459 \div 50 = 9.18$$

4 Find the median by locating the $\frac{n+1}{2}$th score, where $n = 50$. This is the 25.5th score.

median = 9

5 Find the mode by locating the score with the highest frequency.

mode = 10

6 Find the range by subtracting the minimum score from the maximum score.

$$\text{range} = 11 - 7 = 4$$

UNDERSTANDING AND FLUENCY

1 Find the mean, mode, median and range for the data displayed in each table.

a

Score (x)	Frequency (f)
1	3
2	6
3	8
4	2

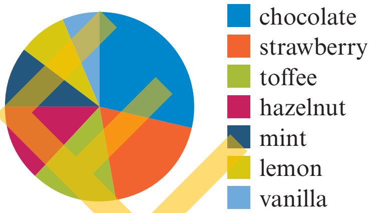
b

Score (x)	Frequency (f)
10	3
20	6
30	4
40	14
50	18

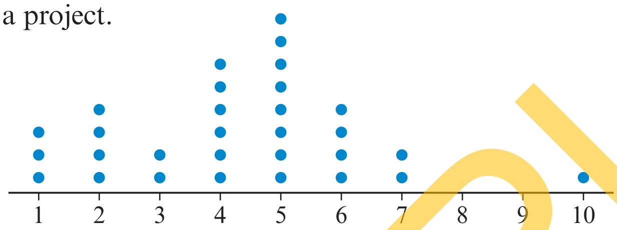
c

Score (x)	Frequency (f)
5	9
10	7
15	5
20	7
25	3

2 Find the mode of the data set from this pie graph.



3 This dot plot shows the time in hours to complete a project.

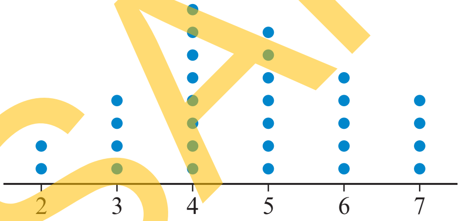


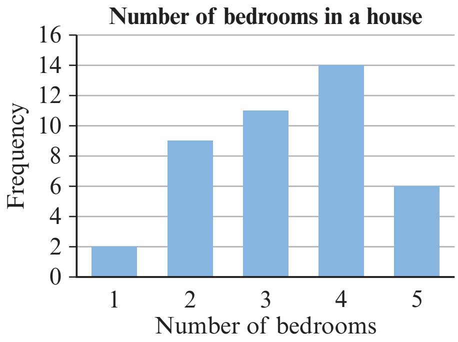
a How many people were surveyed? **b** State the mode and range.

c Create a frequency table.

d Use the frequency table to calculate the median and mean of project completion time.

4 Find the mean, median, mode and range for the data displayed in this dot plot.





5 Data was collected on the number of bedrooms in a house and is shown in this column graph.

a What is the most common number of bedrooms?

b What is the range of the number of bedrooms?

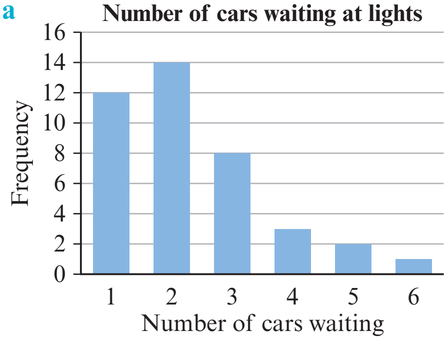
c Create a frequency table to represent the data shown in the column graph.

d Use the frequency table to calculate the median and the mean of the number of bedrooms.

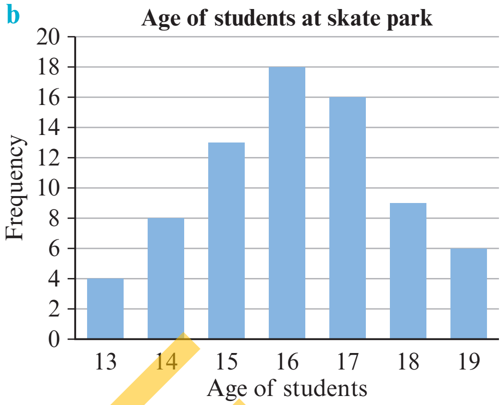
UNDERSTANDING AND FLUENCY

6 Find the mean, median, mode and range for the data shown in each column graph.

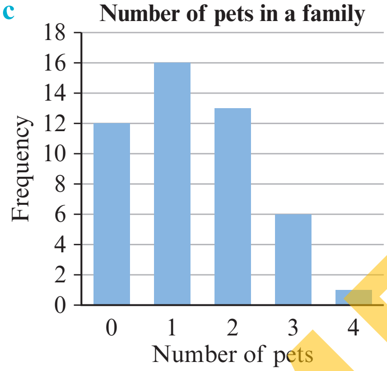
a



b



c



EXAMPLE 8D-2 Finding summary statistics from a stem-and-leaf plot

Find the mean, median, mode and range for the data displayed in this stem-and-leaf plot.

Key 2|1 = 21

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

THINK	WRITE
1 Find the mean by dividing the sum of scores by the number of scores. There are 25 scores in this stem-and-leaf plot. Add these scores and divide by 25.	mean = 459 ÷ 50 = 9.18
2 Find the median by locating the $\frac{n+1}{2}$th score where $n = 25$. This is the 13th score.	median = 39
3 Find the mode by locating the most common score(s).	mode = 46
4 Find the range by calculating the difference between the minimum score (15) and the maximum score (55).	range = 55 – 15 = 40

PROBLEM SOLVING AND REASONING

7 Find the mean, median, mode and range for the data displayed in each stem-and-leaf plot.

a

Key 2|1 = 21

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

b

Key 1|3 = 1.3

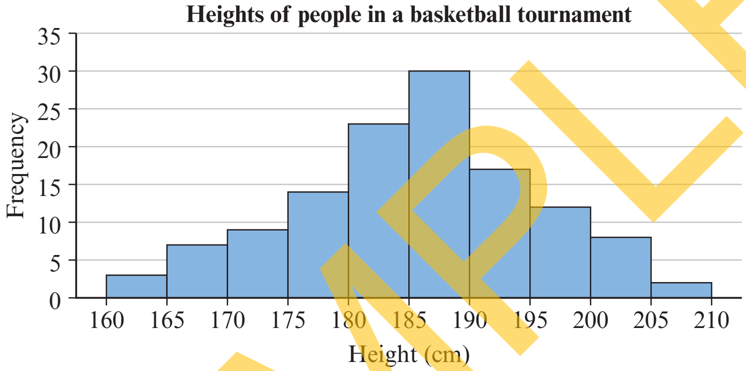
Stem	Leaf
1	2 3 8 8 8 9
2	0 0 1 2 3 6 7 8
3	2 3 7 9
4	4 6
5	2

c

Key 1|2 = 120

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

8 Consider this histogram.



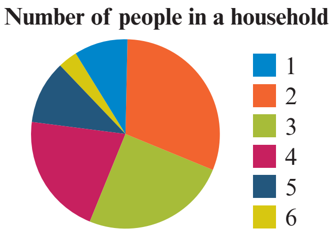
a Create a frequency table with appropriate class intervals to represent the histogram.

b Find the modal class, the median class and the range.

c Use the frequency table to find the mean.

9 a Why is calculating summary statistics from a histogram less accurate than calculating summary statistics from a column graph?

b When might it still be advantageous to use a histogram rather than a column graph?



10 This pie graph was produced by surveying 120 people on the number of people in their household.

a Create a frequency table to represent the pie graph. (Hint: you will need to use a protractor to find each sector size.)

b Find the mean, median, mode and range.

c Why do you think pie graphs are not generally used to represent numerical data?

11 Draw a better graphical display for the data from question 10. How does this show the centre and spread of data better?

PROBLEM SOLVING AND REASONING

12 Match these summary statistics with these graphs.

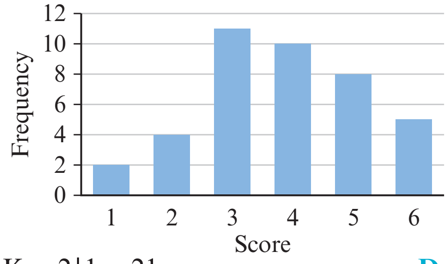
a mean = 3.3, median = 3, mode = 3, range = 5

b mean = 59.3, median = 62, mode = 74, range = 97

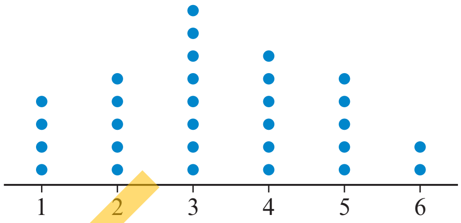
c mean = 3.825, median = 4, mode = 3, range = 5

d mean = 40.74, median = 36.5, mode = 74, range = 97

A



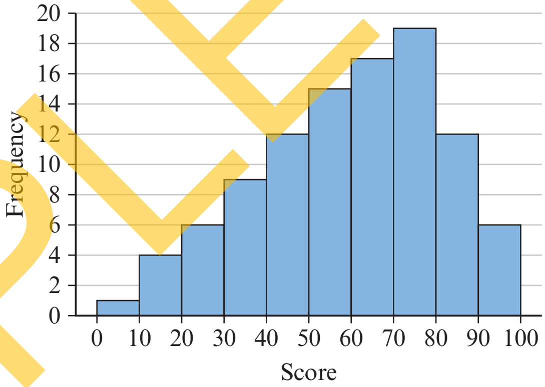
B



C Key 2|1 = 21

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

D



CHALLENGE

13 Consider this stem-and-leaf plot.

a Find the mean, median, mode and range.

b Calculate the standard deviation.

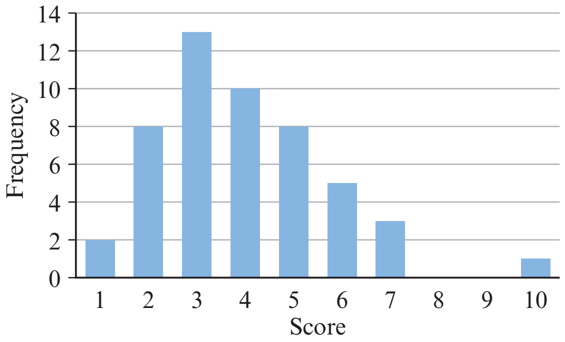
c Write a sentence comparing the measures of centre. Which would you choose to represent the data?

d Write a sentence comparing the measures of spread. Which would you choose to represent the data?

Key 1|4 = 14

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

14 Use this column graph to calculate the measures of centre (mean, median and mode) and measures of spread (range and standard deviation). Which measures would you use to represent the centre and spread of the data? Why?



Reflect

How is calculating summary statistics from a graphical display different from calculating summary statistics from a raw list?

8E Collecting data

Start thinking!

Finn was investigating popular pets in his suburb and wrote down five questions on a survey sheet.

Q1. What pet do you have at home?
Q2. What do you think is the most popular pet?
Q3. What pet would you most like to have at home?
Q4. How many pets do you have at home?
Q5. What do you think is the average number of pets?

1 What type of data is collected for each question?

2 Why might it be important to consider what type of data you get for each response? (Hint: what summary statistics can you use for each data type? How can you display them?)

3 What is the difference between Finn's first and third questions?

4 Which question (**1** or **3**) do you think gives fairer results? That is, which question will tell you more about the most popular pet? Explain.

It is important that all questions in a survey give fair results that are useful to the investigator.

5 Decide and explain which of the five questions would give fair results in an investigation.

Once you have decided what questions to ask, how will you collect the information?

The first thing to decide is who to survey.

The **population** of an investigation is the entire group of people or objects under consideration.

6 What is the population of Finn's investigation?

Usually it is too time-consuming or difficult to survey the entire population (a **census**), so instead you take a **sample** by surveying only some of the population.

KEY IDEAS

- In statistics, a population is the entire group that is important to an investigation.
- A census is a survey of the entire population. Surveying only some of the population is called a sample.
- If a sample does not reflect the population it is said to be **biased**.
- Common sampling methods include **random sampling** (for example, pulling names out of a hat); systematic sampling (sampling at fixed intervals such as every fifth person); and **stratified sampling** (dividing the population into categories such as males and females and taking a random sample from each category that is proportional to its size).
- Using your own collected data is called using **primary data**. Using data that somebody else has collected is called using **secondary data**.

EXERCISE 8E Collecting data

EXAMPLE 8E-1 Classifying a survey as a sample or a census

Eden surveys everybody in town to find out the most popular sport in the district. Decide whether this is a census or sample.

THINK	WRITE
1 Identify the population.	The population is Eden's district.
2 Identify who is surveyed.	Eden's town has been surveyed.
3 A census is an entire population – is the survey taken the same as the population?	The survey taken is not the same as the population, so this is only a sample.

UNDERSTANDING AND FLUENCY

1 Classify each survey as a sample or a census.

a Peter surveys everybody in his class to find out the favourite movie of the entire class.

b Gaylia surveys 40 people at random from her year level to find out the favourite food of the year level.

c Zoë surveys everybody in her class to find out the favourite song in the year level.

d Matt surveys everybody in his family to find what should be the family pet.

e Joel surveys the 25 people in his football club to find out the club's most popular fundraiser.

f Silvia surveys everybody in her street to find out the town's average age.

2 For each situation:

i identify the target population

ii decide whether a census or sample would be more appropriate.

a finding the opinion of the students in your school of a new school rule

b deciding who will be the next Prime Minister of the country

c cooking a meal for your friends and checking for allergies

d finding the favourite music genre of teenagers in your town or suburb

e finding the average electricity usage in Australian households

f finding which local cinema is screening a movie at the best time

EXAMPLE 8E-2 Classifying sampling techniques

Lukas selects a sample of 100 people by surveying every 1000th person in the phone book. Classify this sampling technique as random, systematic or stratified.

THINK	WRITE
1 Look at the sampling definitions in the Key ideas. Which sampling method takes a sample that is in proportion to the population?	
2 Write your answer.	This sample is systematic.

UNDERSTANDING AND FLUENCY

3 Classify each sampling technique as random, stratified or systematic.

a Renee asks every fourth person she sees at a shopping centre their opinion on a local issue.

b Adrian asks 10 boys and 10 girls from town that he sees on a particular day their opinion on a new school uniform rule.

c Oliver asks every fifth person on the school roll of Year 9s who should be the Year 9 school captain.

d Luisa asks 12 boys and 13 girls from her school of 120 boys and 130 girls what they think the school canteen should offer at lunch times.

e Bridget uses a 'Lucky dip' type system in her class to select people to survey.

f Carlos surveys everybody he sees what their favourite movie is.

EXAMPLE 8E-3 Classifying results as biased or fair

Hayden was investigating the Australian public's opinion on which was better out of AFL or rugby. He asked the question: 'Which do you prefer: AFL or rugby?' of every 10th person on the electoral roll from his hometown in Victoria. Classify the results he would obtain as biased or fair, providing a reason.

THINK	WRITE
1 Consider the question that he asks – would this provide fair or biased results? The question relates directly to the topic, so the results should be fair.	
2 Consider his sampling method – would this provide fair or biased results? He uses systematic sampling but the sample he takes only considers one town in Victoria, which may be biased towards one opinion.	The results that he would obtain would be biased because his sample is not representative of the entire population.

UNDERSTANDING AND FLUENCY

4 Classify each sample from questions 1 and 3 as either biased or fair, providing a reason for your answer.

5 For each of these questions, classify the results that would be obtained as biased or fair, providing a reason.

a To investigate average hourly wage rate for high school students in her area, Cynthia asks the question 'What are you paid per hour?' of 200 adults at random from her area.

b To investigate the most popular TV channel in his year level, Luke asks every 20th person on the school roll the question 'What TV channel do you watch most?'.

c To investigate the average number of people in households in his local area, Juan asks everybody he knows the question, 'How many people are in your household?'.

d To investigate the most popular movie genre in her year level, Jasmine asks everybody in the year level the question, 'Do you prefer action or comedy films?'.

6 For each of these scenarios, write three questions that would provide fair answers. Include at least one question that provides numerical data and one question that provides categorical data.

a investigating movie preferences in your school

b investigating opinions about graffiti in your community

c investigating family structure in your community

d investigating technology within the family home

7 For each scenario in question 6, write a question that would provide biased answers. Include an explanation as to why the answers would be biased.

8 The size of a sample is important when conducting an investigation.

a If you wanted to sample from a group of 1000 people, would it be better to sample 10 people or 100 people? Explain.

b Which is a better sample: 10 people from a group of 50 or 20 people from a group of 200? Explain.

c Explain why the size of a population should be considered when deciding on a sample size.

There is no set minimum sample size in any study. It is obvious that the larger the sample taken, the more likely it is to reflect the population. However, it can be too time-consuming or difficult to always sample large numbers of people.

d For each scenario, state how trustworthy you think the results are.

i A survey finds that Channel 7 is the most popular channel in Tasmania, based on polling 190 homes.

ii A study finds that 86% of women see an improvement in wrinkles after using a moisturiser (36 women are surveyed).

iii Recent data shows that of 421 road deaths in NSW in 2010, speeding contributed to 160.

iv A study surveying 1000 couples worldwide finds that one in three marriages end in divorce.

e Why is it important to consider sample size when looking at the results of an investigation?

9 Explain why small sample sizes can often lead to biased results.

10 Rebecca said that collecting primary data was always better than using secondary data. Explain why it is more important that the sampling method is fair than where the data comes from.

11 A useful website to collect secondary data from is the Australian Government census website. Use the Internet to access http://www.abs.gov.au/websitedbs/censushome.nsf/home/Census for these questions.

a Enter your postcode into the 'QuickStatsSearch' box. This will give you data on your local area and how it compares to the rest of Australia.

b Choose one of these quick statistics and write down the data for your local area. You may wish to include a sentence on how it compares to other places in Australia.

To access more detailed data, click on 'Data' at the left-hand side. You can choose to find data by location or topic.

c Choose to search for data by topic. Use the 2011 census data.

d Leave the count method as 'Place of usual residence', as this sorts people based on where they usually live. Select a topic from the drop-down box for investigation.

e Select a more detailed topic from the list that appears. If the topics that appear seem confusing, select a different main topic from the drop-down box.

f Click 'Select location' and then enter in the name of your town or suburb. Select your town or suburb from the list that appears and click 'Select product'.

g Click to view the census table and download the Excel file from the 'Downloads' tab.

h Use the Excel document to collect the data and write a paragraph summarising the results found and how this compares to the rest of Australia.

CHALLENGE

12 Ellen wanted to collect some data from her school using stratified sampling. Her year level consists of 60 boys and 40 girls.

a How many students in her year level in total?

b What fraction of her year level are boys?

c What fraction of her year level are girls?

d If she wanted a sample of 10 people, how many boys and girls should she randomly select?

e If she wanted a sample of 25 people, how many boys and girls should she randomly select?

f Explain how you got your answers to parts **d** and **e**.

g Explain why, if she wanted a sample of n people, that the stratified sample can be found using $n \times \frac{3}{5}$ boys and $n \times \frac{2}{5}$ girls.

13 Determine the structure of the stratified sample if a group of:

a 10 people are to be chosen from 120 girls and 80 boys

b 50 people are to be chosen from 350 adults and 150 children

c 25 people are to be chosen from 180 Year 8s and 195 Year 9s

d 9 animals are to be chosen from 38 birds, 57 cats and 76 dogs.

14 Random sampling often sounds like an easy way of carrying out fair sampling, but in practice it can lead to biased results because people are not good at choosing truly random samples. For small samples a process such as drawing names out of a hat can be used. But what about larger samples? To get unbiased results, random sampling works best when using a pre-gathered list and a random number generator. Microsoft Excel is one program that has a random number generator. Say you want to generate 20 random numbers from a list of 100.

a Open a new Microsoft Excel document.

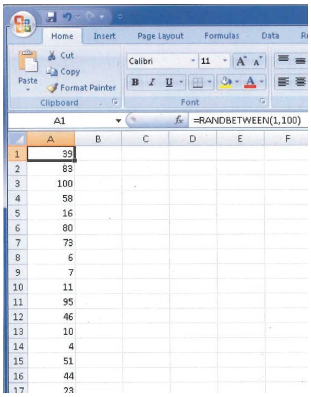
b Type =RANDBETWEEN(1,100) into cell A1. What do you think the numbers 1 and 100 represent in this formula?

c Either use the 'Fill down' function or copy and paste the cell contents of A1 down the column to fill the first 20 cells. This will generate 20 numbers for you.

d Write down the 20 numbers that you generate.

e How can this list of numbers be used to represent a list of 100 people?

f What would you do if the same number was generated twice? Would you survey that same person twice?



A1 =RANDBETWEEN(1,100)

	A	B	C	D	E	F
1	39					
2	83					
3	100					
4	58					
5	16					
6	80					
7	73					
8	6					
9	7					
10	11					
11	95					
12	46					
13	10					
14	4					
15	51					
16	44					
17	23					

Most modern calculators also have a function that will generate random numbers.

g Investigate your calculator and generate another 20 random numbers between 1 and 100. Remember to continue generating until you have 20 unique numbers. See your teacher if you need help.

Reflect

What needs to be considered in order to gain fair results from an investigation?

8F Describing data

Start thinking!

There are many ways to describe the distribution of data, such as using summary statistics to describe the data's centre and spread. A simpler way that provides a quick overview is to describe the shape of the data's distribution. Consider these three graphs.

1 Which of these graphs would you describe as roughly **symmetric**? Why?

2 Do you think it is important for a graph to be perfectly symmetric? Explain.

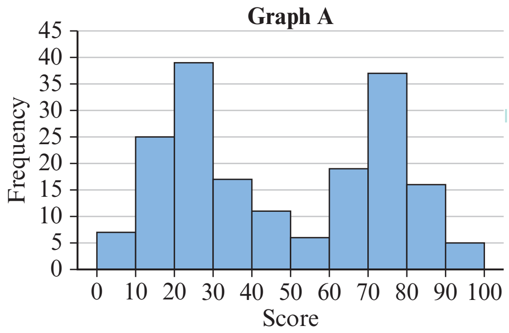
3 Which of these graphs would you describe as **skewed**? Explain.

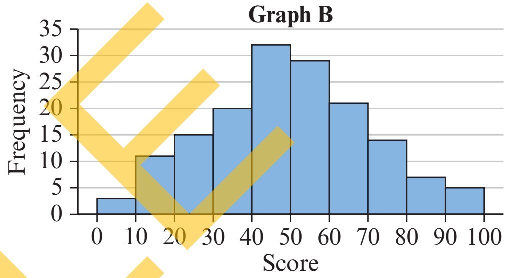
Skewed graphs can be further described as **positively skewed** or **negatively skewed**. A graph that is positively skewed is skewed towards the vertical axis. A graph that is negatively skewed is skewed away from the vertical axis.

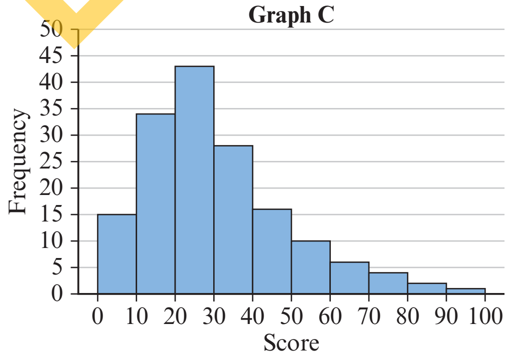
4 Describe the skewed graph as either positively or negatively skewed.

The remaining graph can be described as **bimodal**.

5 Explain why you think it gets this name.

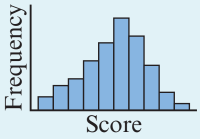




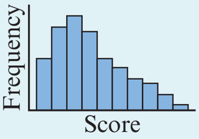


KEY IDEAS

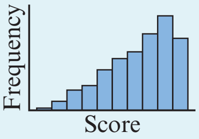
- Symmetric distributions have a middle peak and a roughly even spread on either side.



- Positively skewed distributions have a centre closer to the left of the distribution.



- Negatively skewed distributions have a centre closer to the right of the distribution.

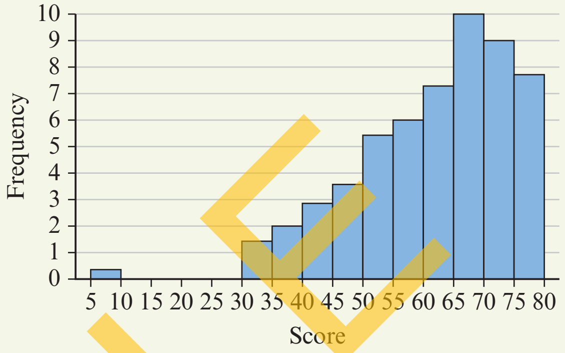


- An outlier is an unusual piece of data that is far away from the rest of the distribution.
- Any distribution that is skewed or has an outlier should have its centre described using the median rather than the mean.
- Bimodal distributions are more difficult to describe with statistics, and are best described using the mode of each peak.

EXERCISE 8F Describing data

EXAMPLE 8F-1 Describing the distribution

Describe the distribution of this histogram.



THINK

1 Is the distribution symmetric, skewed or bimodal?

2 Does the distribution have an outlier?

3 Write your answer.

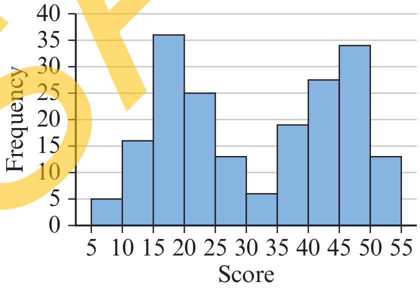
WRITE

The distribution is negatively skewed with an outlier.

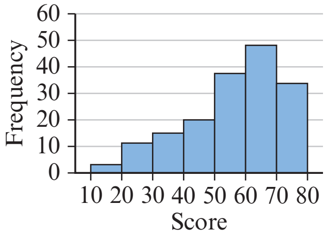
UNDERSTANDING AND FLUENCY

1 Describe the distribution of each histogram.

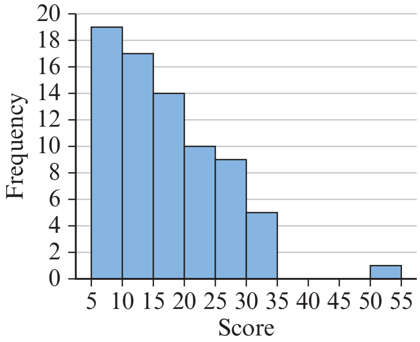
a



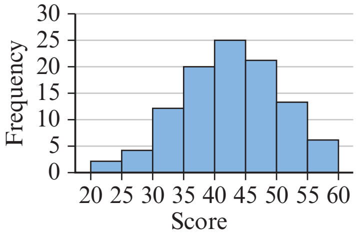
b



c



d



UNDERSTANDING AND FLUENCY

2 Describe the distribution of each stem-and-leaf plot.

a

Key 1|3 = 13

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

b

Key 1|3 = 13

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

c

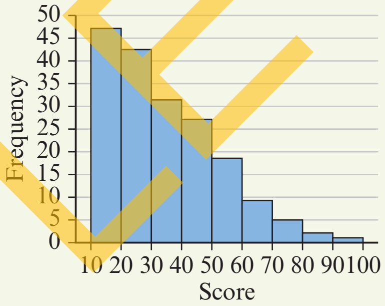
Key 1|3 = 13

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

EXAMPLE 8F-2

Deciding which measure of centre best describes a distribution

Decide which measure of centre would best describe this distribution.



THINK

1 Describe the distribution.

2 The mean is affected by skew and outliers. In these cases it is better to use the median.

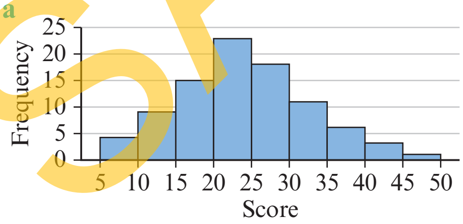
WRITE

The distribution is positively skewed.

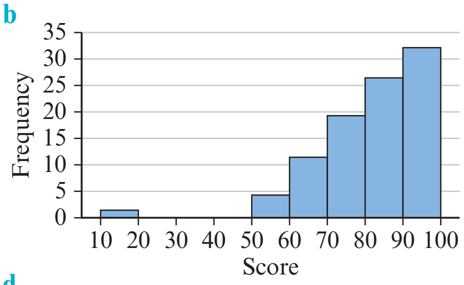
The centre of the distribution would be best described by the median, as it is skewed.

3 Decide which measure of centre would best describe each distribution.

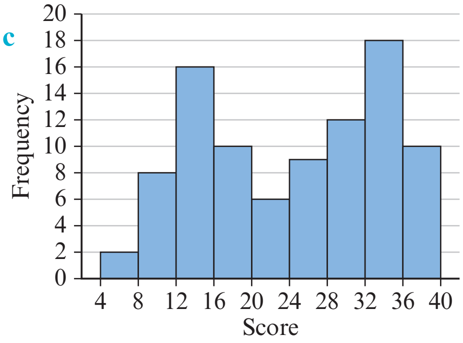
a



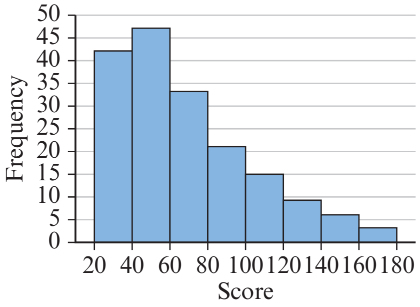
b



c



d



4 Decide which measure of centre would best describe each distribution in question **1**.

EXAMPLE 8F-3 Describing a histogram

Use this data to:

a draw a histogram

b describe its distribution

c state the best measure of centre, providing a reason.

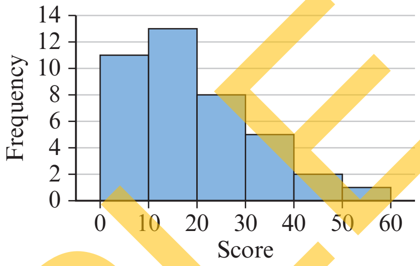
49, 3, 16, 12, 20, 49, 22, 37, 32, 18, 34, 13, 4, 7, 17, 9, 13, 59, 25, 1,
15, 30, 23, 27, 4, 17, 26, 3, 10, 5, 8, 2, 31, 11, 8, 20, 27, 13, 16, 11

THINK

a Draw a histogram with an even scale and label on both axes. The data has a range of 58, so class intervals of 10 would be appropriate.

WRITE

a



THINK

b Decide if the data is skewed.

WRITE

b The distribution is positively skewed.

THINK

c Identify the best measure of centre. When the data is not symmetric, the median should always be used.

WRITE

c The median should be used as the measure of centre because the distribution is skewed and therefore the mean may not be an accurate measure of centre.

UNDERSTANDING AND FLUENCY

5 For each data set:

i draw a histogram

ii describe its distribution

iii state the best measure of centre, providing a reason.

a 35, 33, 42, 99, 54, 68, 4, 91, 97, 55, 99, 86, 40, 58, 41, 95, 38, 62, 35, 88, 82, 98, 77, 69, 78, 78, 82, 81, 98, 57, 88, 41, 60, 85, 82, 85, 91, 90, 80, 49, 58, 66, 97, 95, 82, 84, 78, 91, 62, 42

b 9, 2, 3, 9, 43, 8, 2, 15, 12, 17, 10, 10, 14, 9, 34, 7, 12, 18, 18, 47, 2, 12, 24, 34, 19, 1, 12, 18, 35, 47, 6, 14, 8, 35, 7, 9, 4, 17, 2, 20, 8, 12, 21, 24, 48, 6, 7, 8, 17, 41

c 49, 23, 45, 23, 31, 77, 62, 21, 52, 51, 60, 54, 46, 69, 27, 60, 142, 41, 32, 80, 52, 21, 80, 65, 37, 33, 74, 45, 48, 78, 70, 21, 55, 64, 33, 42, 59, 67, 32, 79, 30.

6 This stem-and-leaf plot does not seem to have much of a noticeable pattern to it.

Key 1|3 = 13

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

a Redraw it as a split stem-and-leaf plot with:

i four stems

ii ten stems.

b Describe the distributions that you see.

PROBLEM SOLVING AND REASONING

7 Ava collected data on the weight of dogs (in kilograms) in her community and put it into the table as shown at right.

a Draw a histogram to represent the data.

b What pattern can you see?

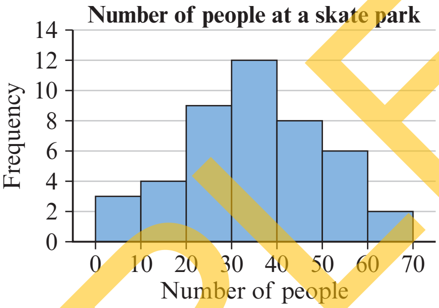
c Reorder the data into more appropriate class intervals and redraw your histogram.

d What pattern can you now see?

e Write a sentence describing the weight distribution of dogs as shown in your second histogram.

Class interval	Frequency
0–<2	1
2–<4	6
4–<6	3
6–<8	5
8–<10	6
10–<12	5
12–<14	8
14–<16	6
16–<18	7
18–<20	3
20–<22	7
22–<24	2
24–<26	5
26–<28	3
28–<30	4
30–<32	5
32–<34	4
34–<36	2
36–<38	1
38–<40	4

8 This histogram shows the number of people at a skate park recorded at various times.



a Describe the shape of the distribution.

b What measure of centre would be the most appropriate to use?

c Create a frequency table that represents the histogram.

d Use the frequency table to calculate the appropriate measure of centre.

e Write a sentence that describes the distribution in terms of the number of people at the skate park.

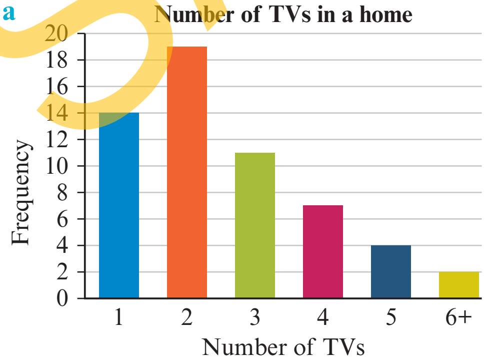
9 For each data distribution:

i describe its shape **ii** state the appropriate measure of centre

iii calculate the chosen measure of centre

iv write a sentence that summarises the graph in its context.

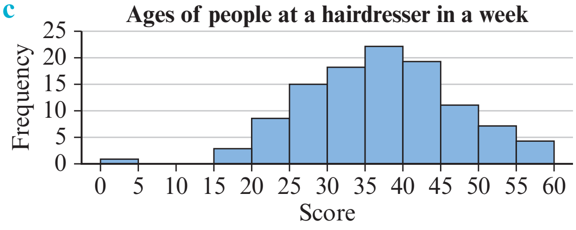
a



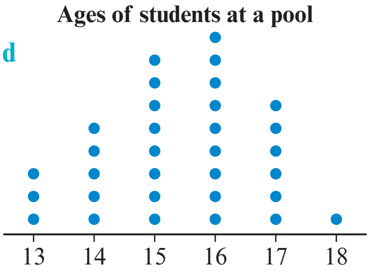
b Length of hair in Year 9 students (cm)
Key 1|3 = 13

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

c



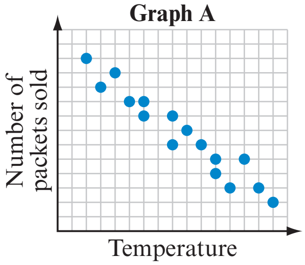
d

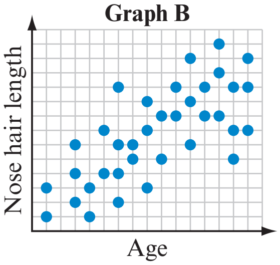


<1 LINE TOO DEEP>

CHALLENGE

10 The relationships represented in scatterplots can be described in terms of their direction (positive or negative) and strength (strong, moderate, weak or none). Consider these two graphs.



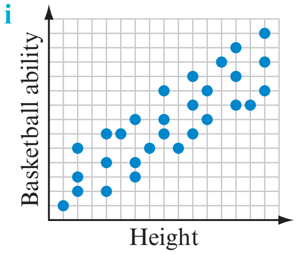


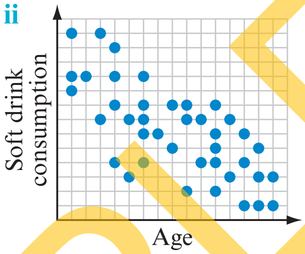
a Which would you consider to show a positive trend? Explain.

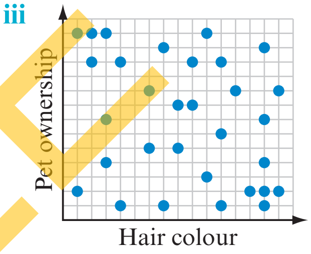
b Which would you consider to show a strong relationship? Explain.

c Can you describe a graph that shows no relationship in terms of its direction? Explain.

d Describe each of these graphs in terms of both their direction and strength.







11 Describe each of the scatterplots from question **10** in context. For example, the first scatterplot could have the description 'the number of jackets sold decreases strongly as temperature increases'.

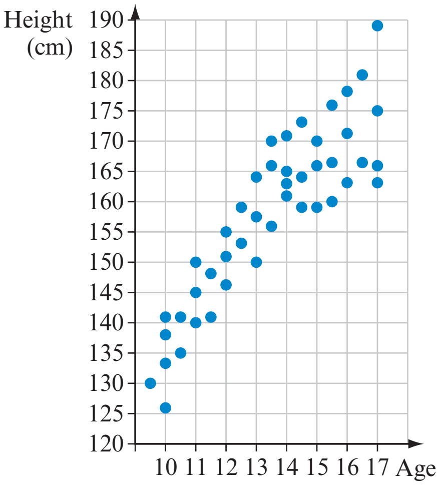
12 Scatterplots that have moderate to strong relationships can be used to make predictions.

a Consider Graph A shown in question **10**. If the temperature was cold, would you predict that you would sell many or few jackets?

b Consider the first scatterplot shown in question **10d**. If somebody was tall, what kind of basketball ability would you predict that they had?

c Consider your answer to part **b**. Does this mean that if somebody is tall that they will be good at basketball? Explain.

Consider the scatterplot on the right.



d Describe the scatterplot in context: what is it showing?

e A friend said that the graph must be wrong because it implies that a 30-year-old has a height of about 270 cm. Explain why their reasoning is incorrect.

f Use your answer to part **e** to explain why a graph should only be used to make predictions within the data range that it shows.

Reflect

In what ways can you describe data? How is this useful?

8G Comparing data sets

Start thinking!

When you want to compare two data sets, it is useful if they are in the same format so that the comparison is easier. One graphical display that can be used to easily compare two data sets is a back-to-back stem-and-leaf plot. Consider this plot, showing the ages of males and females in a gaming store.

Key: 1|3 = 13

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

1 How is this plot different from and how is it the same as a standard stem-and-leaf plot?

2 What is the minimum and maximum age of females in the store?

3 What is the minimum age and the maximum age of males in the store?

4 What is the overall range of ages in the store?

5 How would you describe the distribution of the females in the store?

6 How is this different from the distribution of the males in the store?

7 Write a sentence comparing the ages of males and females in the store. Is one gender more likely to be older?

8 Why is it easier to place the two data sets into the same plot rather than drawing a separate plot?

9 For each data set, find the mean, median and range and place them into a table.

10 Compare the ranges of the two data sets. Does this support your answer to question **7**?

11 Write a sentence comparing the centres of the two data sets. How does this compare to your answer to question **10**?

KEY IDEAS

- To briefly compare two data sets, it is easiest to place the data into a graphical display where the shape of the distributions can easily be compared.
- One way to do this is to construct a back-to-back stem-and-leaf plot and compare the distribution of the leaves.
- Other graphs, such as column graphs, histograms and dot plots can be used to do this as well.
- To make a more thorough comparison of two data sets, summary statistics should be calculated for each data set and the centre and spread of each set compared.

EXERCISE 8G Comparing data sets

EXAMPLE 8G-1 Comparing data sets by drawing a back-to-back stem-and-leaf plot

Use this data to draw a back-to-back stem-and-leaf plot and make a brief comparison of the two data sets.

Age of people at a pool:

Winter: 45, 23, 15, 36, 57, 31, 9, 38, 44, 56, 52, 13, 36, 27, 48, 44, 48, 14, 27, 45

Summer: 31, 16, 14, 15, 23, 56, 24, 18, 17, 8, 11, 13, 16, 21, 17, 36, 20, 17, 14, 15

THINK

1 Both sets of data share the same stem, so locate the minimum and maximum numbers across both sets.

2 The youngest age is 8 and the oldest age is 57, so you should have six stems: from 0 to 5.

3 Draw the back-to-back stem-and-leaf plot with these stems, placing the winter leaves on the left of the stems and the summer leaves on the right. Rearrange the leaves so that they are in order and include a key.

4 Look at the centre and spread of the two data sets. Where does the centre appear to be for each set? What does this mean when you think about ages?

WRITE

Key: 1|2 = 12

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

The two data sets are spread roughly over the same age brackets, but the data sets are skewed in different directions. The pool seems to attract younger people in the summer and older people in the winter.

1 Use this back-to-back stem-and-leaf plot to answer these questions.

a What is the maximum score:

i in group A? ii in group B?

iii overall?

b What is the most common score:

i in group A? ii in group B?

iii overall?

c How would you describe the distribution of:

i group A? ii group B?

Key: 1|2 = 12

Leaf	Stem	Leaf
Group A		Group B
9 7 4 3 1	0	7 9
9 8 8 6 4 3 2 1 0	1	3 4 6 8
7 5 3 0	2	0 1 1 4 4 4 5 8 9
8 3	3	2 2 3 4 7 9 9
1	4	0 3 5

UNDERSTANDING AND FLUENCY

2 For each back-to-back stem-and-leaf plot, make a brief comparison of the two data sets.

a

Key: 1|2 = 12

Leaf	Stem	Leaf
Group A		Group B
1	1	0 1 3 5 8 8
8 7 2	2	2 4 5 6 7 8 9 9
9 7 6 3 1	3	4 6 7 8
9 8 7 7 6 5 3 1 1	4	0 1 2
9 7 6 4 3 1 0	5	1 3

b

Key: 3|2 = 3.2

Leaf	Stem	Leaf
Group A		Group B
	3	0 0 1 2 4 5 6
9 8 7 6 5 4 4 4	4	1 2 5 5 6 7 8 9 9
8 7 6 6 5 4 3 1 1 0	5	0 3 4 5 5 6 8
8 7 6 5 4 3 3 0	6	
	7	
	8	1

c

Key: 1|2 = 120

Leaf	Stem	Leaf
Group A		Group B
	0	2 3 4
9 7 6 5 5 5	1	1 7 8 9 9
9 8 7 7 6 5 4 4 3	2	0 2 3 5 5 7 9
9 7 6 5 4 4 4 3	3	0 4 6 7 8 8
3 2 2 1	4	0 1 5 6
	5	0 6

d

Key: 1|2 = 12

Leaf	Stem	Leaf
Group A		Group B
	5	1
9 8 7 1	6	
7 6 5 4 3 3 3 2 1	7	8 9 9
9 8 7 5 4 2 2 1 0 0	8	1 2 3 4 6 7 8 8
	9	0 0 5 6 7 8 8 8 9

3 For each of the following, draw a back-to-back stem-and-leaf plot and use it to make a brief comparison of the two data sets.

a Number of people at a cinema over 6 months of weekends:

Friday: 65, 48, 67, 55, 32, 92, 64, 51, 49, 57, 76, 61, 29, 46, 61, 59, 53, 67, 72, 88, 71, 58, 54, 57, 56, 42, 30

Saturday: 67, 78, 84, 26, 37, 42, 99, 84, 75, 68, 64, 55, 75, 85, 59, 66, 77, 78, 83, 81, 92, 94, 77, 76, 79, 89

b Daily maximum temperature in February

Darwin: 32.3, 32.1, 32.9, 33.2, 33.5, 33.8, 33.4, 32.8, 32.5, 33.3, 32.5, 29.5, 32.8, 33.4, 33.1, 32.7, 33.4, 31.4, 33.0, 31.3, 29.9, 30.8, 30.0, 32.2, 32.1, 32.6, 31.8, 30.4

Adelaide: 20.6, 23.7, 25.4, 29.4, 34.2, 38.4, 27.0, 28.3, 28.0, 26.3, 28.3, 31.2, 33.8, 34.3, 36.0, 37.1, 39.2, 40.5, 26.6, 28.4, 32.1, 33.8, 35.6, 38.2, 28.7, 33.4, 21.5, 22.7

c Mass of dogs of two breeds (to nearest hundred grams)

Boston Terriers: 5100, 6800, 7800, 10500, 9200, 5500, 4900, 11200, 8600, 9200, 10500, 4900, 5500, 7600, 8400, 6800, 9200, 8600, 7500, 4600, 8300, 10500, 11000, 8000, 7600, 5100, 6700, 8300, 9200, 10000, 4900, 7500, 7900, 6200, 8200, 4600, 7800, 8100, 6400, 9900.

French Bulldogs: 9800, 12900, 8600, 9900, 9800, 12500, 11200, 9100, 12900, 11200, 9500, 9200, 10500, 11900, 12500, 11800, 9400, 9800, 10800, 10600, 11500, 9600, 12300, 13000, 11000, 12000, 10600, 11100, 12700, 10900, 9500, 11500, 12900, 10100, 11100, 9200, 11500, 10900, 9800, 11300.

EXAMPLE 8G-2 Comparing data sets using summary statistics

Rachel was investigating the average age of customers in two different cafés in her local area. Use her results to make a comparison of the two different cafés.

Café	Mean	Median	Range
Gumtree	21	22	17
Jinx	27	21	45

THINK	WRITE
1 Compare the given measures of spread (the range).	Jinx covers a larger range of ages than Gumtree.
2 Compare the given measures of centre: the medians are similar but Jinx has a higher mean than Gumtree. This difference combined with Jinx's large range suggests that Jinx's distribution may be skewed or affected by an outlier. The median should be used as the measure of centre.	The average age for both cafés is similar, though Jinx may have a skewed distribution and possibly an outlier present.
3 Summarise your findings.	The average customer age at the cafés is similar but Jinx has a larger range of ages.

UNDERSTANDING AND FLUENCY

4 For each table, use the given statistics to make a comparison of the two data sets.

a

Heights in class	Mean	Median	Range
9A	172 cm	168 cm	46 cm
9B	166 cm	167 cm	22 cm

b

Average jeans price	Mean	Median	Range
Store A	$120	$122	$35
Store B	$124	$122	$30

c

Average Internet usage per week	Mean	Median	Range
Primary school	4.5 hours	4 hours	8 hours
Secondary school	11.2 hours	12 hours	17 hours

d

Average SD card capacity	Mean	Median	Range
Store A	20 GB	16 GB	60 GB
Store B	16 GB	16 GB	60 GB

e

Goals per game	Mean	Median	Range
Player A	4.2	4	9
Player B	4.8	4	7

5 Consider the column graph on the right.

a What is it showing?

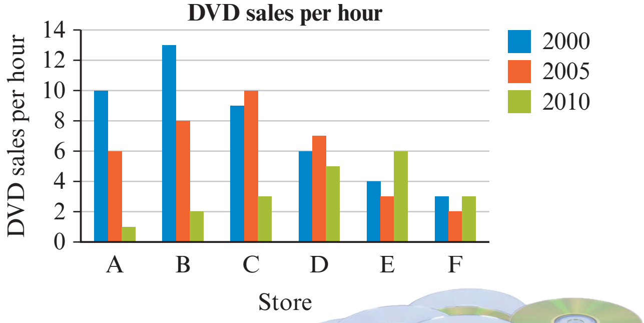
b How many data sets does it display?

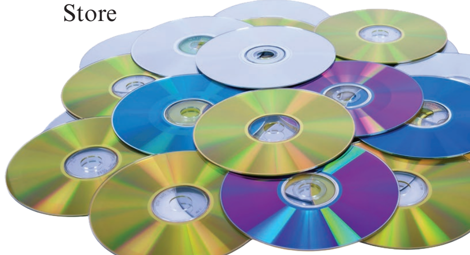
c Describe the pattern shown in:

i 2000 **ii** 2005 **iii** 2010.

d What patterns can you see?

e Write a statement that summarises the graph in terms of the trends throughout each year and from 2000 to 2010.





PROBLEM SOLVING AND REASONING

6 Draw a back-to-back stem-and-leaf plot that you imagine would compare heights of Year 7s and Year 9s. Describe and explain the patterns that you choose to display.

7 Carly was investigating the average age of people visiting a new auction website. Using a fair sampling method, she took three separate samples (A, B and C, as shown). Explain which statistics Carly should use to summarise her findings.

Age of people visiting website			
	Sample A	Sample B	Sample C
Sample size	20	40	60
Mean	22	25	29
Median	20	26	25
Range	10	13	60

8 The local council was investigating how much people spend on computer games every year, and called for submissions. Only three submissions were made, shown in these tables.

Sample A	200 people at a gaming store
Mean	$450
Median	$395
Range	$200

Sample B	150 people from the state
Mean	$200
Median	$190
Range	$250

Sample C	50 local people
Mean	$210
Median	$165
Range	$400

Explain which sample and which statistic the council should use to predict the characteristics of the population.

9 Consider this back-to-back stem-and-leaf plot.

Heights of Year 9 students (cm)

Key: 1|2 = 120

Leaf **Boys**	**Stem**	**Leaf** **Girls**
	14	8
9	15	6 7 9 9
9 9 8 7 5 4 3 1	16	0 0 1 1 2 2 2 3 4 6 7 8
9 8 8 7 6 5 5 4 3 3 3 1	17	0 1 2 3 4
7 3 2 1	18	2
2	19	

a Make a brief comparison of the two data sets, writing your answer in context.

b Calculate the mean, median and range for both data sets displayed in the stem-and-leaf plot.

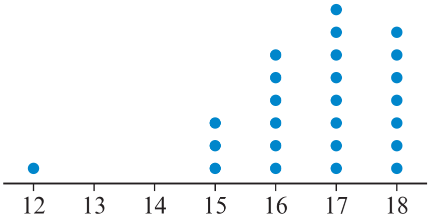
c Use these statistics to make a more detailed comparison of the two data sets.

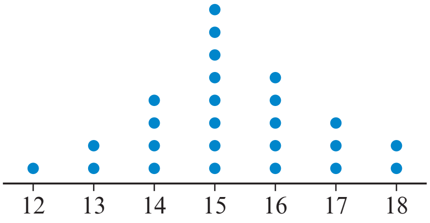
d Does your answer to part c support your answer to part a? Explain.

10 Use data from the 2011 Census on the Australian Bureau of Statistics website to compare a characteristic of your choice between your local area and another area. See the process outlined in Exercise 8E question 11 (page 377) if you need help.

PROBLEM SOLVING AND REASONING

11 Consider these two dot plots.


Ages of students in term 1 school play


Ages of students in term 4 school play

a Compare the shape of these two dot plots.

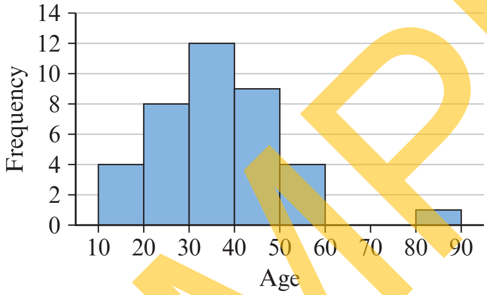
b Find the mean, median and range of each dot plot.

c Write a couple of sentences comparing the two data sets with reference to their summary statistics. Be sure to write in terms of what data is being presented.

d Can you think of a way to represent both dot plots using the same 'axis'?

12 **a** Calculate the mean, median and range for these two data displays.

Age of customers in a milk bar



Age of customers in a milk bar

Key 1|4 = 14

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

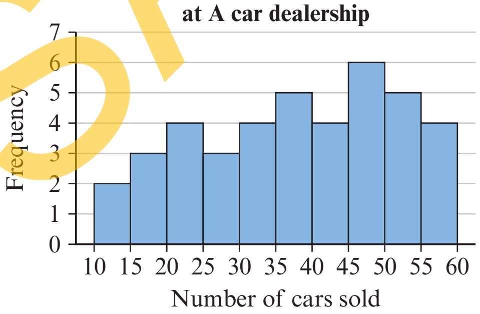
b Compare these two data sets using their shape and summary statistics.

c Explain why you can easily compare the sets even though the displays are different.

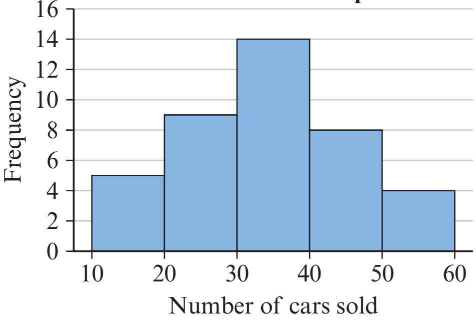
CHALLENGE

13 Consider these two histograms.

Number of cars sold per month at A car dealership



Number of cars sold per month at B car dealership



a Why is it difficult to make a quick comparison of the two histograms as they are?

b Redraw the first histogram so that its class intervals match those of the second.

c Use your answer to part **b** to make a comparison of the two data sets.

d Why can't you redraw the second histogram to match the first histogram?

14 Calculate the measures of centre for the histograms shown (as they are originally) in question **13** and use them to make a thorough comparison of the two data sets. How does this compare to your answers from question **13**?

Reflect

What needs to be considered when comparing two data sets?

<2 LINES TOO DEEP>

CHAPTER REVIEW

SUMMARISE

Create a summary of this chapter using the key terms below. You may like to write a paragraph, create a concept map or use technology to present your work.

numerical
continuous
line graphs
scatterplots
histograms
column graphs
bar graphs
pie graphs
dot plots

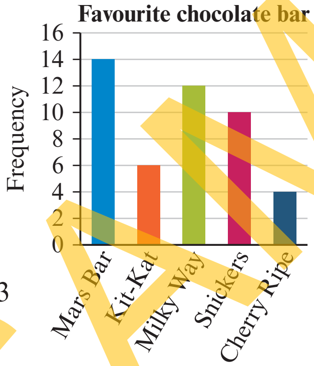
stem-and-leaf plots
frequency table
class intervals
histogram
mean
mode
median
centre
spread

range
standard deviation
census
sample
fair
systematic sampling
stratified sampling
random sampling
biased

primary data
secondary data
symmetric
skewed
positively skewed
negatively skewed
bimodal
back-to-back stem-and-leaf plot

MULTIPLE-CHOICE

Questions **1** and **2** refer to this column graph.



8A **1** The number of people surveyed is:

A 46 **B** 23
C 14 **D** 6

8A **2** Which statement is not supported by the graph?

A The most popular chocolate bar for the group surveyed was Mars Bar.

8F

B The least popular chocolate bar for the group surveyed was Cherry Ripe.

C Six people said that Kit-Kat was their favourite.

D Ten people said that Milky Way was their favourite.

3 If the score of 33 is removed from the data set below, which of these statements is not true?

12, 15, 12, 17, 19, 25, 15, 11, 13, 18, 12, 16, 12, 19, 33, 20

A The mode is unchanged.

B The mean and median are both reduced in value.

<THIS PAGE 3 LINES TOO DEEP>

C The range is smaller.

D The measures of centre are unchanged.

8E **4** Which of these is an example of stratified sampling?

A pulling names out of a hat

B asking every 10th person in a crowd

C dividing a crowd of 100 into males and females, then surveying 10 people from each group

D interviewing your parents about their political views

8C **5** Which statement is true if describing distribution of data?

A Symmetric distributions have a centre closer to the right of the distribution, away from the y-axis.

B Any distribution that is skewed or has an outlier should have its centre described using the median rather than the mean.

C Positively skewed distributions have a middle peak and a roughly even spread on either side of this peak.

D Negatively skewed distributions have a centre closer to the left of the distribution, towards the y-axis.

SHORT ANSWER

8B **1** Data was collected on the number of hours Year 9 students spent using various forms of social media over the course of a week.

45 21 37 21 20 17 31 32 11 15 17 18
20 31 48 32 21 5 11 7 19 18 27 42
40 21 32 23 24 19 38 37 14 15 19 28
22 35 41 33 27 2 10 5 18 18 25 43

- **a** Select an appropriate class interval and construct a frequency table.
- **b** How many people were surveyed?
- **c** What is the most common class interval?
- **d** Draw a histogram to represent this data.
- **e** Construct a stem-and-leaf plot to represent this data.
- **f** Present this data in a stem-and-leaf plot using class intervals of 5.
- **g** Which interval is the most common?

8C 8G **2** Find the mean, median, mode and range for this data set (correct to two decimal places where appropriate).

Score	Frequency
3	10
4	12
5	23
6	5

8C **3** **a** Calculate the mean, median, mode and range (correct to one decimal place where appropriate) for the data set:

25.7 23.3 24.6 26.7 58.9 17.6
25.7 21.0 29.6 20.9 23.6

- **b** Identify the outlier in this data set.
- **c** If the outlier was removed from the data set, what changes do you predict will occur to the measures of centre?
- **d** Remove the outlier. Recalculate the measures of centre to test your prediction.

8D **4** Calculate the mean, mode, median and range for the data represented in this stem-and-leaf plot (correct to two decimal places where appropriate).

Key 2|4 = 24

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

8G **5** The results achieved in Maths tests for two different classes are recorded in a back-to-back stem-and-leaf plot. Find the mean, median, mode and range for Class 9A.

Key: 4|3 = 43

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

6 The following data comparing the number of hours spent at football training per fortnight were recorded for Year 9 boys from two different secondary schools.

School 1: 35 11 27 11 10 7 21 22 1 5 7 8 10 21 38 22 11 5 1 7 9 8 17 32

School 2: 12 14 16 17 20 11 22 11 11 16 19 14 2 4 10 19 17 6 4 6 2 7 16 31

- **a** Collate the data in a back-to-back stem-and-leaf plot using class intervals of 5.
- **b** Calculate the mean, median and range for each school (correct to one decimal place).
- **c** Which measure of centre would you use to discuss the results for each school?

NAPLAN-STYLE PRACTICE

1 Which of these is classified as ordinal data?

○ eye colour (hazel, blue, green)

○ pizza sizes (small, medium, family)

○ the number of SMS messages sent in a month

○ the time taken to walk from home to school

2 Which of these would provide primary data for you to work with?

○ visiting a website and using the data about television viewing habits

○ surveying your class about their favourite television show

○ using information from social media forums about television programs

○ using your friend's survey results

Questions 3 and 4 refer to the following data, which is listed below and also displayed in the stem-and-leaf plot.

25, 27, 29, 41, 32, 35, 20, 27, 49

Stem	Leaf
2	5 7 7 9
3	2 5
4	1 9

3 Which score is missing in the stem-and-leaf plot?

□

4 What is an appropriate key for this stem-and-leaf plot?

□

Questions 5 and 6 refer to the frequency table below.

Score	Frequency
24	97
25	111
26	378
27	246
28	
29	301
Total	1325

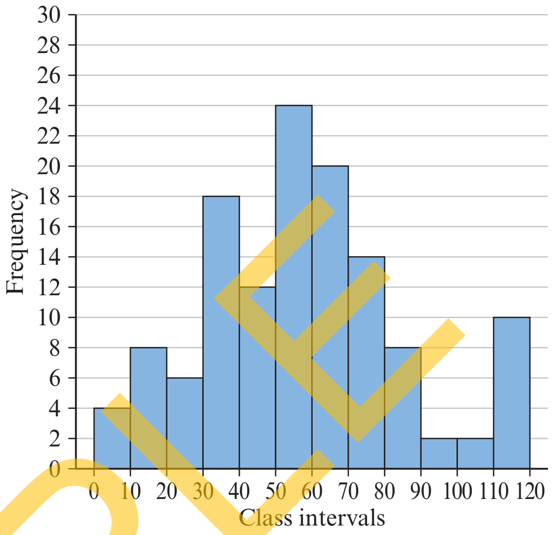
5 What is the missing value in the frequency table?

□

6 Find the percentage of scores that are 25 or less (correct to one decimal place).

□

Questions 7 and 8 refer to the histogram below. The histogram displays the results of research where the heights of plants were measured.



7 How many plants are less than 70 cm tall?

92	46	20	14
○	○	○	○

8 What percentage of plants (correct to two decimal places) are greater than 100 cm tall?

○ 1.56% ○ 12%

○ 9.38% ○ 90.62%

Questions 9 and 10 refer to the data set below.

15 17 18 45 13 15 15 16 15

9 Which value is closest to the mean of this data set?

○ 18.78 ○ 15

○ 32 ○ 2.5

10 A score was recorded incorrectly in the list of data. If the score of 45 should be 15, which statement is true?

○ The range will be unchanged.

○ The mean will be unchanged.

○ The mode will be unchanged.

○ The median will change.

11 A friend purchased 12 concert tickets for a total of $1188. What is the average price per ticket?

12 The heights of a group of Year 9 students were recorded, as follows:

145 cm, 152 cm, 147 cm, 1.35 m, 165 cm, 170 cm.

Which statement is false?

□ The mean height is 1.52 m.

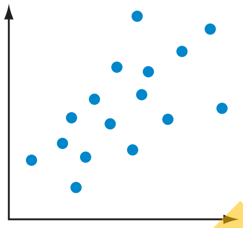
□ The standard deviation is large, indicating a significant spread from the mean.

□ The range is 35 cm.

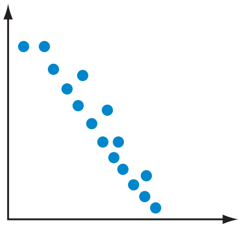
□ The median height is 152 cm.

13 Which graph shows a strong negative linear relationship?

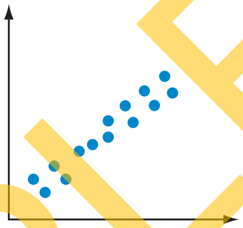
□ Graph A



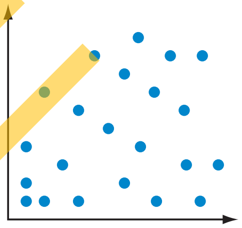
□ Graph B



□ Graph C



□ Graph D



ANALYSIS

This stem-and-leaf plot shows the measurements of shrubs taken at different nurseries.

Key: 15|4 = 15.4 cm

Leaf Brisbane	Stem	Leaf Sydney
7 4 4 4	11	4 4 4 5 7 8 9
7 6 4 3 2 2	12	1 3 6 7
8 7 6 2 1	13	2 3 9
1 0	14	1 4 6 7 8 9
9 8 5 3 2	15	1 2 5 5

a Compare the number of shrubs measured at each location.

b Calculate the mean, median, mode and range of shrub heights at each location (correct to two decimal places).

c Calculate the standard deviation of shrub heights at each location.

d Write a short comparison of the height of shrubs at the two locations.

The manager of the nursery company wanted to collate the data for all of his businesses.

e Generate lists of raw data for Brisbane and Sydney. Using this data and the lists below, create a frequency table using suitable class intervals and collate all of the data about shrub height.

Victoria: 12.1, 11.7, 18.3, 11.4, 11.4, 14.5, 15.6, 17.1, 16.5, 18.6, 13.0, 12.6, 12.4, 10.9, 11.4, 14.0, 16.9, 17.1, 16.5, 18.6

Western Australia: 10.2, 10.6, 19.3, 11.4, 11.4, 15.9, 14.7, 18.3, 17.7, 13.4, 10.0, 12.5

f How many plants were measured in total?

g What is the modal class?

h Represent this data as a histogram and comment on the shape of the distribution.

i Calculate the mean shrub height for Victoria and Western Australia.

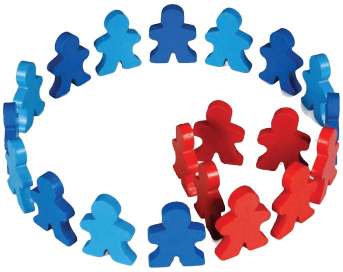
j Hence calculate the average height of shrub across all locations.

k Write a statement that could be used in a marketing campaign, which includes information about the smallest and largest shrubs and the average shrub height.

CONNECT

Investigating your local community

When a local council wants to make changes or seek input from the community, they will often survey a sample of the population in order to help them make their decisions. What issues do you think are important in your local community at the moment?



Your task

You need to:

- choose a topic relevant to your local community to investigate
- decide what information you want to discover
- formulate at least five questions that provide numerical and categorical data and that provide fair, unbiased results
- produce a survey sheet that is easy to fill out and provides information from which you can collate results into displays
- decide on a fair sampling method and the number of people you will survey
- perform your survey
- display your results in at least three forms (for example, table, histogram, stem-and-leaf plot)
- describe your numerical results according to its shape
- calculate summary statistics and use these to provide a more thorough analysis
- access a secondary source of data (for example, census data) in order to compare to your results for your local area and another community (another town, the state, the country)
- write a summary that describes your results and compares them to your second chosen locale.

You may like to present your findings as a report. Your report could be in the form of:

- a poster
- a digital presentation
- an information pamphlet
- other (check with your teacher).