3 Representing and describing data: descriptive statistics

Statistics is concerned with the collection, analysis and interpretation of quantitative data. Statistical representations and measures allow us to represent data in many different forms to aid interpretation. Both statistics and probability provide important representations which enable us to make predictions, valid comparisons and informed decisions.

Concepts

- Representations
- Validity

Microconcepts

- Population
- Bias
- Samples, random samples, sampling methods
- Outliers
- Discrete and continuous data
- Histograms
- Box-and-whisker plots
- Cumulative frequency graphs
- Measures of central tendency and dispersion
- Skewness
- Scatter graphs
- Correlation

How can scientists determine whether a new drug is likely to be a successful cure?

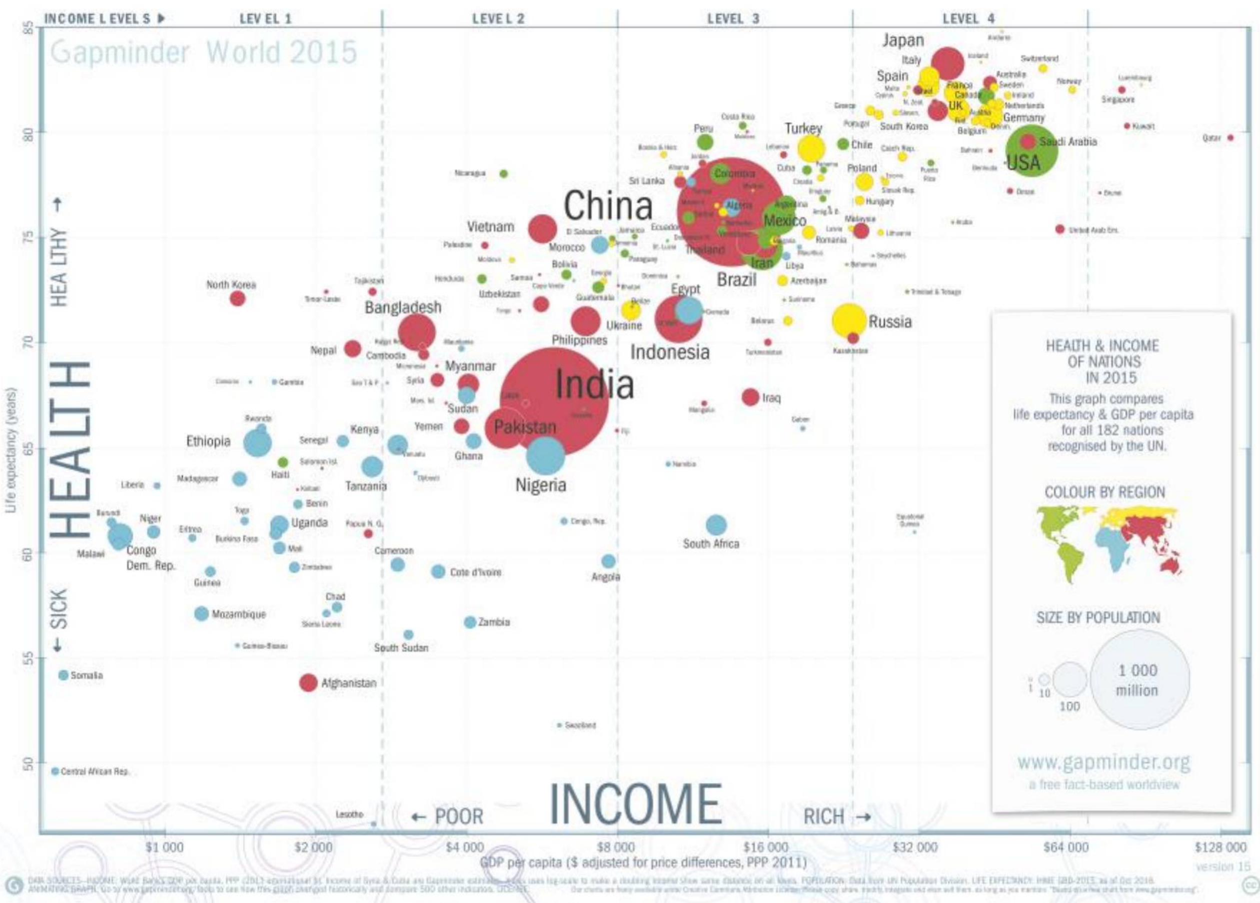
How can a headteacher determine whether teaching in the school has been effective?

How can a football coach determine whether a particular strategy is likely to be successful?

How can we tell if the oceans are warming?

How can you persuade a potential customer that your product is better than the competition?

Below is a graph of GDP per capita (gross domestic product per person) and life expectancy taken from Gapminder (www.gapminder.org). Click the icon to access the complete set of data.



Name four pieces of information represented in this graph.

How do you think this data could have been collected? How exact do you think it might be?

Identify any relationships in the graph.

Do you find anything surprising in the graph?

Do you need to use all the data for analysis or can you just use a sample of the data?

Describe the scale on the *x*-axis. Why do you think it has been done like that?

Developing inquiry skills

Write down any similar inquiry questions you might ask to investigate the relationship between two different quantities, for example, GDP per capita and infant mortality or life expectancy and population.

How are these questions different from the ones used to investigate life expectancy and income?

Think about the questions in this opening problem and answer any you can. As you work through the chapter, you will gain mathematical knowledge and skills that will help you to answer them all.

Before you start

You should know how to:

1 Collect data and represent it in bar charts, pie charts, pictograms and line graphs.

eg The numbers of children in 20 families are shown in the table:

Number of children	Frequency
1	4
2	8
3	5
4	2
5	1

Represent this information in:

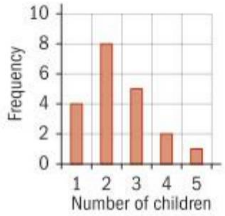
a a pictogram **b** a bar chart

c a pie chart.

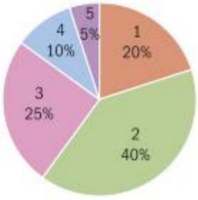
a ☺ = 1 child

1	☺ ☺ ☺ ☺
2	☺ ☺ ☺ ☺ ☺ ☺ ☺ ☺
3	☺ ☺ ☺ ☺ ☺
4	☺ ☺
5	☺

b



c



2 Obtain simple statistics from discrete data, including the mean, median, mode and range.

eg Find the mean, median, mode and range of the following data:

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

Mean = 5.3 Median = 6

Mode = 6 Range = 7

3 Set up axes for graphs using a given scale.

Skills check

Click here for help with this skills check

1 The ages of 25 children at a nursery are shown in the table.

Age	Frequency
0	4
1	5
2	8
3	6
4	2

Represent this information in:

a a pictogram

b a bar chart

c a pie chart.

2 Find the mean, median, mode and range of the following data:

3, 3, 5, 7, 8, 8, 9, 9, 9, 9

3 Draw a set of axes such that 1 cm represents 5 units on the x-axis and 1 cm represents 2 units on the y-axis.

3.1 Collecting and organizing univariate data

Univariate data has only one variable.

Rosa works in a restaurant. The tips that the customers give to the waiters and waitresses are placed in a large jar. It is Rosa's job to count the tips every week, record the amount in a notebook and place the tips and notebook in the safe. At the end of each month, the tips are divided equally amongst the staff.

The manager is interested to see how the tips varied from week to week, and also how many of each type of note or coin there are (5¢, 10¢, 20¢, 50¢, €1, €5, €10, €20). He asks Rosa to make a presentation for the staff.

How can she best make this presentation? Which diagrams can she use? How can the staff use this information to improve service? How can the manager use the information to improve the restaurant?

International-mindedness

Ronald Fisher (1890–1962) lived in the UK and Australia and has been described as "a genius who almost single-handedly created the foundations for modern statistical science". He used statistics to analyse problems in medicine, agriculture and the social sciences.

There are two main types of data: qualitative and quantitative.

Qualitative data is data that is not given numerically, for example favourite ice cream flavour.

Quantitative data is numerical and can be classified as **discrete** or **continuous**.

Investigation 1

1 The numbers of cherries in 24 boxes are shown below.

44 43 42 42 43 41 45 42 40 43 44 41
42 41 43 42 42 40 45 43 43 41 45 43

This data is **discrete**.

Complete the frequency table for this data.

Number of cherries	Frequency
40	2
41	
42	
43	
44	
45	

Continued on next page

2 The lengths, in minutes, of 20 telephone calls are shown below.

4.2	6.8	10.4	8.2	11.5	1.6	5.9	7.6	3.1	21.5
13.5	5.8	4.1	22.8	13.6	11.2	9.5	1.8	12.4	4.9

This data is **continuous**.

Complete the frequency table for this data.

Length, t (minutes)	Frequency
$0 \leq t < 5$	
$5 \leq t < 10$	
$10 \leq t < 15$	
$15 \leq t < 20$	
$20 \leq t < 25$	

3 Explain why the examples in parts **1** and **2** are different.

4 **Factual** What type of values can discrete data take? How is discrete data collected? What type of values can continuous data take? How is continuous data collected?

5 Men's jeans are sized by waist measurement, eg 28 inches, 30 inches, 32 inches and so on. Here are the jeans sizes of 10 men:

28 30 28 34 32 30 36 28 30 30

This is discrete data.

Do you need to change your answer to "What type of values can discrete data take?"

6 Here are the waist measurements, in cm, of 10 people:

24.3 27.2 22.1 28.3 27.0 29.6 32.4 23.8 21.7 35.2

Is this discrete or continuous data?

How does it differ from the previous example?

7 **Conceptual** What are the differences between discrete and continuous data?

Discrete data is either data that can be **counted**, for example the number of cars in a car park, or data that can only take specific values, for example shoe size.

Continuous data can be **measured**, for example height, weight and time.

Discrete and continuous data can be organized into a frequency table or a grouped frequency table.

For continuous data, the classes must cover the full range of the values and they must not overlap.

Example 1

The ages of boys in a football club are:

10 11 11 10 12 13 11 10 12 14 15 15 16 10 11 15 10 11 11 12
12 12 13 16 16 14 15 12 12 10 11 11 14 14 15 16 16 11 10 13

a State whether this data is discrete or continuous.

b Construct a grouped frequency table for this data. Let x represent age.

a The data is continuous.

b

Age, x	Frequency
$10 \le x < 11$	7
$11 \le x < 12$	9
$12 \le x < 13$	7
$13 \le x < 14$	3
$14 \le x < 15$	4
$15 \le x < 16$	5
$16 \le x < 17$	5

Remember that you are 10 for a whole year!

Notice that all the possible ages are included in the classes.

International-mindedness

The 19th-century German psychologist Gustav Fechner popularized the median, although the French mathematician Pierre Laplace had used it earlier.

Exercise 3A

1 State whether the following data sets are discrete or continuous.

a the number of apples in a bag

b the weights of students in Grade 6

c the number of blue cars in a parking lot

d the football boot sizes of a football team

e the number of visitors to the Tower of London each week

f the weights of 20 puppies

g the depth of snow on a ski slope

h the number of sixes when you throw a die 25 times

i the time it takes to run 100 metres

j the lengths of 20 worms.

2 Construct a frequency table for this data.

The number of sweets in 25 packets:

21 23 22 24 21 22 23 25 24 24
22 23 25 21 23 23 24 26 25 25
21 22 22 24 22

3 Construct a grouped frequency table for the following data.

The heights, in metres, of 20 trees in a garden:

5.8 3.6 3.9 4.1 4.4 3.2 2.4 2.6 5.1 2.5
4.5 3.6 2.4 5.2 4.7 3.5 3.3 2.8 4.1 2.1

4 The following data shows the weights of 25 dogs, in kilograms.

2 5 31 22 16 7 12 35 9 18 5 11 15
6 3 14 8 10 12 25 27 34 7 1 5

Construct a suitable table for this data.

When faced with lots of numbers, how do you know which "average" is best to use?

Measures of central tendency (or averages)

- The most common measures of central tendency are the mean, median and mode.
- The **mode** of a data set is the value that occurs most frequently. There may be no mode or several modes.
- The **median** of a data set is the value that lies in the middle when the data is arranged in size. When there are two middle values, the median is the midpoint between the two values.
- The **mean** of a data set is the sum of all the values divided by the number of values. For a discrete data set of n values the formula is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n$ and Σ means "the sum of". For example, the mean of the numbers 3, 4, 8, 12, 16 is the sum (Σ) of the numbers divided by 5.
- For a frequency data set, the formula is $\overline{x} = \frac{1}{\sum_{i=1}^{n} f_i}\sum_{i=1}^{n} f_i x_i$, where $\sum_{i=1}^{n} f_i x_i = f_1x_1 + f_2x_2 + f_3x_3 + \cdots + f_nx_n$.

  When there is a frequency table, you need to use the data values and the corresponding frequencies to calculate the mean.

Example 2

1 The grades in a history test for 14 students were as follows:

58 67 66 58 79 83 76 49 35 58 88 91 47 69

a Find the mode, median and mean.

When a 15th student took the test, the mean became 66.2.

b Calculate the grade for the 15th student.

2 Mindy opens some bags of candy and counts how many pieces are in each bag.

Her results are:

Number of pieces of candy	Frequency
23	2
24	3
25	9
26	5
27	1

Find the mean number of candies in a bag.

1 a Mode is 58 — 58 appears three times.

Median is 66.5 — Arranging the data in order:

35 47 49 58 58 58 $\boxed{66\quad 67}$ 69 76 79 83 88 91

The middle number will be between 66 and 67.

Mean is 66 — The mean is

$$\frac{58+67+66+58+79+83+76+49+35+58+88+91+47+69}{14}=66$$

b The mark is 69 — If the new mean is 66.2 then the total for all 15 students will be 66.2 × 15 = 993. Subtracting the total for the 14 students: 993 − 924 = 69.

2 Mean is 25 — $\text{Mean} = \dfrac{23\times 2+24\times 3+25\times 9+26\times 5+27\times 1}{2+3+9+5+1}=25$

Example 3

Answer the following questions, and in each case interpret the meaning of the values calculated, and discuss why extreme values or an extreme mode affect the mean more than the median.

a The number of ice creams sold over a period of 13 weeks is as follows:

146 151 158 158 161 149 160 147 158 160 216 225 238

Write down the mode, and use technology to find the mean and median for this data set.

b Two dice are thrown 100 times and their total score is recorded in the table:

Score	2	3	4	5	6	7	8	9	10	11	12
Frequency	21	9	8	4	7	20	13	9	6	2	1

Write down the mode, and use technology to find the mean and median for this data set.

c The weights, w kg, of 50 cats are recorded in the table:

Weight (kg)	Frequency
$2 \le w < 3$	5
$3 \le w < 4$	19
$4 \le w < 5$	17
$5 \le w < 6$	5
$6 \le w < 7$	3
$7 \le w < 8$	1

Continued on next page

Find an approximation for the median and mean, and write down the modal class.

a	Mean = 171.3; this is the average number of ice creams sold during the 13 weeks. Median = 158; this is the middle value. Half of the amounts are above this value and half are below it. Mode = 158; this is the number of ice creams that occurs the most frequently. The three large values have the effect of making the mean value larger, but the median is not affected since it is the middle value.	Put the numbers into a list on your GDC. Go to Statistics, Stat calculations, one-variable statistics, number of lists 1, enter the name of the list and press Enter. The GDC gives you a lot of data. The first is the mean. Then scroll down until you reach the median. Mode: 158 occurs the most often.
b	Mean = 5.82; this is the average score for the 100 throws of the dice. Median = 7; this is the middle value. Half of the scores are above this value and half are below it. Mode = 2; this is the score that occurs the most often. Since the mode is 2, it makes the mean value smaller, but the median is not affected since half of the values are less than 7.	Put the scores into one list on your GDC and the frequencies into a second list. Go to Statistics, Stat calculations, one-variable statistics, number of lists 1, enter the names of the two lists and press Enter. The GDC gives you a lot of data. The first is the mean. Then scroll down until you reach the median. The mode is 2 since this occurs 21 times.
c	Approximation for the mean = 4.2 kg; this is the approximate mean weight of the 50 cats. Approximation for the median = 4.5 kg; approximately half the weights are above and half below this middle value. Modal class = $3 \le w < 4$; this group has more of the cats' weights than any of the other groups. The modal class is the second smallest group, but this does not have much effect on the mean in this example because the middle group is also quite large, whereas the top three groups are very small in comparison.	Using your GDC, enter the midpoints of the groups into one list and the frequencies into a second list and proceed as above. These values only give an estimate since we do not have the original data on the cats' weights and only know that five cats weigh between 2 kg and 3 kg, etc.

Investigation 2

1 Using technology, complete the table for the following data sets.

A The dress sizes of 15 females:

0 0 2 2 2 4 4 6 6 8 10 12 14 16 16

B The shoe sizes of 19 children:

23 23 23 23 26 28 35 35 36 36 36 37 39 41 43 40 38 37 41

TOK

Why have mathematics and statistics sometimes been treated as separate subjects?

C The number of times that 20 commuters travelled by train in one month:

40 50 41 28 51 52 49 50 51 28 48 33 35 28 45 40 51 62 28 49

D The ages of boys in a basketball club:

Age	10	11	12	13	14	15	16	17	18
Frequency	5	6	8	10	12	11	12	35	32

	Mode	**Median**	**Mean**
Data set A			
Data set B			
Data set C			
Data set D			

2 For each set of data, list the advantages and disadvantages of the mean, median and mode, and decide which best represents the data in each case. Explain your choice.

3 How do you decide which measure of central tendency best represents the data?

4 Why do we need more than one measure of central tendency?

This grouped frequency table shows data set E, the scores that 60 students gained in an entrance test:

Score	**Frequency**
$20 < x \leq 30$	3
$30 < x \leq 40$	5
$40 < x \leq 50$	7
$50 < x \leq 60$	8
$60 < x \leq 70$	9
$70 < x \leq 80$	12
$80 < x \leq 90$	10
$90 < x \leq 100$	6

5 What is the modal group (or class) for data set E?

6 Use your GDC to find approximations for the mean and median.

7 Why are these values only approximations?

8 Discuss which value is more appropriate to use in this case.

Data set F shows the weights, in kilograms, of 20 different breeds of dogs:

9 What is the modal group (or class) for data set F?

Weight (kg)	$0 < w \leq 10$	$10 < w \leq 20$	$20 < w \leq 30$	$30 < w \leq 40$	$40 < w \leq 50$	$50 < w \leq 60$
Frequency	6	5	4	3	2	1

10 Use your GDC to find approximations for the mean and median.

11 Why are these values only approximations?

12 Discuss which value is more appropriate to use in this case.

Exercise 3B

1 For the following sets of data find the mean, median and mode. State which of these measures is most appropriate to use in each case, giving a reason for your answer.

a The times, in minutes, to run 1500 metres:

7.2 7.3 7.5 7.8 8.0 8.3 8.6
8.6 8.6 9.0 9.2 9.5 10.0 10.5
10.6 11.1 15.3 16.8 17.2

b The weights, in kg, of 13 pumpkins:

2.6 2.9 4.7 6.8 6.9 7.2 8.5 8.9
10.1 11.5 12.5 14.7 15.0

c The monthly amounts of pocket money, in euros, for 21 Grade 6 students:

10 10 10 15 15 15 15 20 25
25 30 30 35 35 35 40 0 50
50 80 100

2 For the following sets of data, find

i the modal class

ii an approximation for the mean

iii an approximation for the median.

Comment on the meaning of these values and state which one is most appropriate to use in each case, giving a reason for your answer.

a

Number of cars (n)	Frequency
$0 \le n < 30$	12
$30 \le n < 60$	28
$60 \le n < 90$	39
$90 \le n < 120$	42
$120 \le n < 150$	54
$150 \le n < 180$	65

b

Speed of cars (s mph)	Frequency
$40 \le s < 45$	4
$45 \le s < 50$	8
$50 \le s < 55$	23
$55 \le s < 60$	15
$60 \le s < 65$	6
$65 \le s < 70$	4

c

Time to complete a puzzle (t minutes)	Frequency
$2 \le t < 3$	2
$4 \le t < 4$	5
$4 \le t < 5$	3
$5 \le t < 6$	7
$6 \le t < 7$	4
$7 \le t < 8$	9
$8 \le t < 9$	3

Example 4

The ages of 15 cats are:

10 10 11 11 11 12 12 12 12 13 13 14 14 24 25

Find the median, mean and mode for this data.

Comment on whether there are any data points that distort the calculation of the mean.

Remove these values and recalculate the mean. Discuss your answer.

The median is 12.	Enter the data into a list on the GDC.
The mean is 13.6. The mode is 12.	Go to Statistics, Stat calculations, one-variable statistics, number of lists 1, enter the name of the list and press Enter. The GDC gives you a lot of data. The first is the mean. Then scroll down until you reach the median.
	The mode is the number that appears the most.
24 and 25 are much larger than the other numbers. If they are removed, the mean becomes 11.9, which is much closer to the median and the mode.	24 and 25 are called **outliers**. Outliers are extreme data values that can distort the results of statistical processes.

Investigation 3

1 Find the mean, median and mode for the following sets of numbers:

a The monthly salaries, in Australian Dollars (AUD), of 12 employees in a factory:

4000 4200 4200 4250 4400 4400 4400 4450 4500 4550 4600 20 000 42 000

b The ages of students on a chemistry course at university:

19 18 18 21 22 19 20 17 20 21 22 19 19 19 20 17 55 63

c The lengths of time, in seconds, for which 15 people can hold their breath:

20 22 23 23 23 58 61 61 65 74 79 80 81 83 92

2 Which data entries do not appear to fit with the rest of the data?

3 Do you think that these entries are a result of an error in the recording of the data or not? Explain your answer.

4 Calculate the mean, median and mode of each data set without the entries you identified in part **2**. Do the values change?

Extreme data values that distort the mean are called **outliers**; they do not “fit” with the rest of the data.

5 **Conceptual** How can outliers affect measures of central tendency?

6 **Conceptual** How can identifying outliers help you decide which measure of central tendency to use to represent the data?

Outliers are extreme data values, or the result of errors in reading data, that can distort the results of statistical processes.

Outliers can affect the mean by making it larger or smaller, but most likely they will not affect the median or the mode.

TOK

Is there a difference between information and data?

Statistics and probability

Exercise 3C

1 Find the mean, median and mode for the following data sets and comment on any pieces of data that you think may be outliers.

a The times of 25 telephone calls in minutes:

1.0 1.5 2.3 2.6 2.8 3.0 3.4 3.8 4.1
4.5 4.6 4.8 5.2 5.3 5.5 5.8 6.0 6.3
6.6 7.3 7.5 7.5 7.5 17.8 25.0

b The heights, in metres, of 15 sunflowers:

1.1 2.2 2.5 2.5 2.5 3.1 3.5 3.6
3.9 4.0 4.1 4.4 4.6 4.9 6.1

c The results of a geography test:

22 39 45 46 46 52 54 58 62 62
62 67 70 75 78 82 89 91 95 98

Measures of dispersion

- Measures of dispersion measure how spread out a data set is.
- The simplest measure of dispersion is the **range**, which is found by subtracting the smallest number from the largest number.
- The standard deviation, σ_x, gives an idea of how the data values are spread in relation to the mean. The standard deviation is also known as the root-mean-squared deviation; its formula is

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2}$$

In examinations you will use technology to find the standard deviation.

Investigation 4

In this investigation you will find the means and standard deviations for two sets of data and compare the results.

A quiz has 10 questions with one mark for each correct answer. Ten boys and ten girls took this test and their results are shown in the table:

Girls' scores	Boys' scores
2	4
3	5
4	5
5	6
5	6
6	6
8	7
8	7
9	7
10	7

Find the mean of the girls' scores and the mean of the boys' scores.

The standard deviation is called the **root-mean-squared deviation**, and to calculate this you work **backwards**.

First find the **deviation** of each score from the mean, then **square** these answers:

Girls' scores – girls' mean	Boys' scores – boys' mean	(Girls' scores – girls' mean)^2	(Boys' scores – boys' mean)^2
		$\Sigma =$	$\Sigma =$

Next you have to find the **mean** of (girls' scores – girls' mean)^2 and the **mean** of (boys' scores – boys' mean)^2.

Lastly you find the **square root** of each of these two values.

This gives you the **standard deviation**.

If you compare the value of the standard deviation to the mean in each case, what can you say about the spread of the data?

Conceptual What does the standard deviation represent?

Example 5

For each of the three data sets in Example 3, find the standard deviation and compare it with the mean.

a Standard deviation = 30.8; this would indicate that the data points are not all close to the mean.	Enter the data into a list on your GDC. Go to Statistics, Stat calculations, one-variable statistics, number of lists 1, enter the name of the list and press Enter. The GDC gives you a lot of data. The first is the mean. Then scroll down until you reach the standard deviation, given by the symbol σx.
b Standard deviation = 2.80; this is a small value and so most of the points will be close to the mean.	Enter the data into two lists on your GDC. Go to Statistics, Stat calculations, one-variable statistics, number of lists 1, enter the names of the two lists and press Enter.

Continued on next page

	The GDC gives you a lot of data. The first is the mean. Then scroll down until you reach the standard deviation, given by the symbol σx.
c The standard deviation is 1.1, which is quite small and would suggest that most of the weights are close to the mean. This is only an approximate value because the original data has not been given, only the grouped data.	Using your GDC, enter the midpoints of the groups into one list and the frequencies into a second list and proceed as above.

The **variance** is the standard deviation squared: $(\sigma_x)^2$.

While the standard deviation is useful for interpreting the spread of data about the mean, other statistical processes such as least squares regression, probability theory and investments use the variance.

TOK

Is standard deviation a mathematical discovery or a creation of the human mind?

The **interquartile range** (IQR) is the **upper quartile**, Q_3, minus the **lower quartile**, Q_1.

When the data values are arranged in order, the lower quartile is the data point at the 25th percentile and the upper quartile is the data point at the 75th percentile.

The interquartile range is another method of interpreting the spread of data. It is more reliable than the range because it is not affected by outliers.

Consider the following scores in a biology exam, arranged in order:

18, 22, 26, 39, 45, $\boxed{46}$, 46, 52, 54, 58, 62, $\boxed{62}$, 62, 67, 70, 71, 75, $\boxed{78}$, 82, 89, 91, 95, 98

The median is the middle value, 62, since half the numbers are above 62 and half the numbers are below 62.

To find Q_1, locate the number that is in the $\left(\frac{n+1}{4}\right)$th place. Here it is the number in the $\frac{23+1}{4} = 6$th place, so the lower quartile is 46, since one-quarter of the numbers are below 46 and three-quarters of the numbers are above 46.

To find Q_3, locate the number that is in the $\left(\frac{3(n+1)}{4}\right)$th place. Here it is the number in the $\frac{3(23+1)}{4} = 18$th place, so the upper quartile is 78, as three-quarters of the numbers are below 78 and one-quarter of the numbers are above 78.

The interquartile range is then $Q_3 - Q_1 = 78 - 46 = 32$.

Example 6

For the data sets in Example 3, find:

i the variance to 2 dp **ii** the range **iii** the IQR.

a **i** Variance = $(30.838\ldots)^2 = 950.98$ **ii** Range = $238 - 146 = 92$	The variance is the square of the standard deviation.
iii IQR = $188.5 - 150 = 38.5$	The maximum and minimum values as well as Q_1 and Q_3 are all found on the GDC as described above.
b **i** Variance = $(2.801\ldots)^2 = 7.85$ **ii** Range = $12 - 2 = 10$ **iii** IQR = $8 - 3 = 5$	Note that different calculators use different methods for quartiles, so you may get slightly different answers from other students' and from the formula.
c **i** Variance = $(1.1)^2 = 1.21$ **ii** Range = $7.5 - 2.5 = 5$ **iii** IQR = $4.5 - 3.5 = 1$ Once again, these values are only approximate values since this is a grouped frequency table.	

Investigation 5

1 For the data sets A–D in Investigation 2, complete the table:

	Mean	Standard deviation	Variance	Range	Lower quartile	Upper quartile	IQR
A							
B							
C							
D							

2 For each data set, compare the value of the standard deviation with that of the mean.

Discuss whether you think that the data is close to the mean or has a wide spread.

Discuss whether you think there are any outliers.

Continued on next page

3 Discuss the difference between the range and the interquartile range, and which one best represents the spread of the data.

4 **Factual** What is spread?

5 **Conceptual** Which value do you think gives a better representation of the spread: the range or the IQR? Why do you think this?

6 For data set D, do you think that the values from the GDC are exact or approximate? Explain.

7 **Conceptual** How does using technology save time and increase accuracy?

8 **Conceptual** What does the standard deviation represent?

Investigation 6

Complete the table for the following sets of numbers:

A: Find the mean and standard deviation of the numbers 3, 4, 6, 8, 9, 10, 15, 17.

B: Add 3 to each of the numbers in A and then find the mean and standard deviation.

C: Subtract 2 from each of the numbers in A and then find the mean and standard deviation.

D: Add 5 to each of the numbers in A and then find the mean and standard deviation.

E: Multiply the numbers in A by 3 and then find the mean and standard deviation.

F: Multiply the numbers in A by −2 and then find the mean and standard deviation.

G: Multiply the numbers in A by 0.5 and then find the mean and standard deviation.

	Mean	Standard deviation
A		
B		
C		
D		
E		
F		
G		

1 **Conceptual** What happens to the mean when you add or subtract a number from each data value?

2 **Conceptual** What happens to the standard deviation when you add or subtract a number from each data value?

3 **Conceptual** What happens to the mean when you multiply each data value by a constant?

4 **Conceptual** What happens to the standard deviation when you multiply each data value by a constant?

5 The mean of a set of numbers is 10 and the standard deviation is 1.5.

- **a** If you add 3 to each number, write down the new mean and standard deviation.
- **b** If you multiply each number by 4, write down the new mean and standard deviation.

The mean of a set of numbers is $\bar{x}$ and the standard deviation is σ_x.
If you add k to or subtract k from each of the numbers then the mean is $\bar{x} \pm k$ and the standard deviation is σ_x.
If you multiply each number by k then the mean is $k \times \bar{x}$ and the standard deviation is $|k| \times \sigma_x$.

TOK

Do different measures of central tendency express different properties of the data?

How reliable are mathematical measures?

Exercise 3D

1 Stan divided the lawn into 30 equal plots. He counted the number of daisies in each plot:

12 15 8 16 24 5 13 2 34 21 18 15 12 8 4

22 15 6 15 3 13 25 9 17 11 6 15 12 26 16

- **a** State whether the data is discrete or continuous.
- **b** Find the mean, the median and the mode, and comment on which is more appropriate to use.
- **c** Find the standard deviation and comment on your result.
- **d** Find the range and interquartile range.

2 Gal asked 60 people how much money they spent the last time they had eaten in a restaurant. The table shows his results.

Cost of dinner, UK£	Frequency
$10 \le c < 20$	6
$20 \le c < 30$	12
$30 \le c < 40$	28
$40 \le c < 50$	10
$50 \le c < 60$	4

- **a** Write down the modal class.
- **b** Find estimates for the mean and the median.
- **c** Find an estimate for the standard deviation and comment on the result.
- **d** Find estimates for the variance, the range and the interquartile range, and explain why these are all estimates.

3 The monthly salaries of the employees in a retail store had a mean value of US$3500 and standard deviation US$250. At the end of the year they all received an increase of US$100. Write down the new mean and the new standard deviation.

4 The table shows the number of orthodontist visits per year made by the students in Grade 10.

Number of visits	0	4	6	8	10	12	14
Frequency	3	2	8	4	2	12	5

- **a** Find the mode, the median and the mean, and comment on which is the most appropriate to use.
- **b** Find the standard deviation and comment on the result.
- **c** Find the range and interquartile range, and comment on the spread of the data.

5 The number of sweets in 25 bags has a mean of 30 and a standard deviation of 3. In a special promotion, the manufacturer doubles the number of sweets in each bag. Write down the new mean and the new standard deviation of the number of sweets in a bag.

6 The table shows the heights of 50 wallabies.

Height (x cm)	Frequency
$150 \le x < 160$	3
$160 \le x < 170$	5
$170 \le x < 180$	13
$180 \le x < 190$	23
$190 \le x < 200$	4
$200 \le x < 210$	2

a Write down the modal class.

b Find estimates for the mean and standard deviation; comment on your results.

7 Mrs Ginger's Grade 8 class sat an English test. The grade was out of 40 marks. The mean grade was 32 marks and the standard deviation was 8 marks.

In order to change this to a mark out of 100, Mrs Ginger thinks that it would be acceptable to multiply all the grades by 2 and then add 20 to each one.

Mr Ginger thinks that it would be fairer to multiply all the grades by 2.5.

Miss Ginger suggests multiplying by 3 and subtracting 20 from each grade.

a Write down the new mean and the new standard deviation for each suggestion.

Matty had an original grade of 12, Zoe had an original grade of 25 and Ans had an original grade of 36.

b Find their new grades under all three suggested changes.

8 The heights in centimetres of 15 basketball players are:

175 183 191 196 198 201 203
203 204 206 207 209 211 212 213

The heights of 15 randomly chosen males are:

154 158 158 162 165 168 171
176 178 180 181 182 182 183 186

a Find the mean and standard deviation for each group.

b Compare your results and comment on any similarities or differences.

9 The table shows the monthly salaries of all the staff at Mount High College.

Monthly salary ($$x$)	Number of males	Number of females
$1000 < x \le 1500$	4	9
$1500 < x \le 2000$	8	14
$2000 < x \le 2500$	14	11
$2500 < x \le 3000$	16	10
$3000 < x \le 3500$	7	3
$3500 < x \le 4000$	2	1
$4000 < x \le 4500$	3	0

a Estimate the mean and standard deviation for male staff and for female staff.

b Compare your results and comment on any similarities or differences.

Developing inquiry skills

Click the icon for the full set of life expectancy and GDP data. What can you say about the spread of data in both lists?

What are the standard deviations for life expectancy and GDP per capita?

Do these values imply that the points are all close to the mean values or not?

3.2 Sampling techniques

Here is some of the data on the number of airports in various countries from the CIA factbook for 2013. You can find the full table in the ebook.

Country	Number of airports	Country	Number of airports	Country	Number of airports
United States	13 513	Somalia	61	Lebanon	8
Brazil	4093	Chad	59	Turks and Caicos Islands	8
Mexico	1714	Ethiopia	57	Togo	8
Canada	1467	Yemen	57	Sierra Leone	8
Russia	1218	Suriname	55	Burundi	7
Argentina	1138	Morocco	55	Equatorial Guinea	7
Bolivia	855	French Polynesia	54	Rwanda	7
Colombia	836	Nigeria	54	Kuwait	7
Paraguay	799	Uzbekistan	53	Moldova	7
Indonesia	673	Austria	52	Falkland Islands (Islas Malvinas)	7
South Africa	566	Afghanistan	52	Benin	6
Papua New Guinea	561	Belize	47	Kosovo	6
Germany	539	Uganda	47	Micronesia, Federated States of	6
China	507	Israel	47	Western Sahara	6
...	...	...	...	...	...

It is possible to use all this data for analysis. However, it would be easier if you could just take a **sample** of the data instead.

A **population** is the whole group from which you may collect data.

A **sample** is a small group chosen from the population.

Simple random sampling is selecting a sample completely at random. For example, using a random number generator or picking numbers from a hat.

Systematic sampling is, for example, taking every fifth entry starting at a random place.

In the table of airports, all the data is from a website and is called the **population**. How can you take a **random** sample of this data to use for analysis?

Investigation 7

1 Using the data in the table, use technology to find the mean number of airports.

Using the spreadsheet of data in the ebook, click on the arrow next to the Σ AutoSum icon and select Average. Then highlight all the entries and press Enter. If you are using your GDC then you need to enter all the data into a list, select Statistics, start calculations, one-variable statistics, 1 list, enter the name you gave the list and press OK.

2 Which of the following methods do you think will give you a random sample? Using the spreadsheet or your GDC, find the mean in each case and give a reason why you think it does or does not give a random sample.

a Take the first 50 countries. Which type of sample do you think this is? Discuss whether it will be a good representation for the mean.

b Take the first 25 countries and the last 25 countries. Which type of sample do you think this is? Discuss whether it will be a good representation for the mean.

c Take every fifth country. Which type of sample do you think this is? Discuss whether it will be a good representation for the mean.

d Use the random number generator on your GDC to pick out 30 countries. Which type of sample do you think this is? Discuss whether it will be a good representation for the mean.

To use the random number generator on the GDC, select Probability, Random, Integer(1, total number of countries, number of countries you want to select), eg Integer(1,450,9) will give you a list of nine random integers between 1 and 450. Be careful to check that none of the integers are repeated. If so, you will have to select some more numbers at random to make up your total.

e Put the names of all the countries into a hat and pick out 50 countries at random.

f Ask your friends which country they are from and use those countries.

3 Which methods in part **2** do you think will give the most reliable estimates of the population mean? Explain your answer.

4 How do you know whether the data is biased?

5 **Factual** What is biased data? What is a reliable result?

6 **Conceptual** How can you decide whether or not the results are biased or reliable?

Convenience sampling is getting data by selecting people who are easy to reach, for example people at a school, club, etc. It does not include a random sample of participants and so the results could be biased.

A **biased** sample is one that is not random—for example, researching spending habits on cars and only interviewing people exiting a garage.

Example 7

The following data shows the IQs of 200 people:

56	62	65	68	69	70	71	71	75	77	79	79	81	81	81	83	84	85	85	85
86	86	86	87	87	87	87	87	87	87	88	88	88	88	88	89	89	89	89	89
89	89	89	89	89	89	89	91	92	92	92	92	93	93	93	93	93	93	94	94
94	94	94	94	94	95	95	95	95	95	95	96	96	96	96	96	96	96	97	97
97	98	98	98	98	98	98	98	99	99	99	99	99	99	99	99	99	99	99	99
100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	101	101	101	101	101	101	101	101	101	101	102	102	102
102	103	103	103	103	104	104	104	104	104	105	106	106	107	107	107	107	107	107	107
107	107	107	108	108	108	109	110	110	110	110	112	112	113	113	113	114	114	115	115
117	118	119	121	121	125	128	129	129	131	134	135	136	137	140	141	143	145	148	156

a Find the mean of the IQs.

b Find the mean of the first 20 numbers and the last 20 numbers.

Comment on the type of sample this is and any advantages and/or disadvantages it may have.

c Find the mean of the subset of the data consisting of every fifth IQ.

Comment on the type of sample this is and any advantages and/or disadvantages it may have.

d Find the mean of a random sample of 30 IQs.

Comment on the type of sample this is and any advantages and/or disadvantages it may have.

e Comment on which of these methods gives the best approximation to the mean of all 200 IQs.

Continued on next page

a 99.74	Put the data into a frequency table and find the mean.
b 104 This is a biased sample since it is not random. It is easy to use but unreliable.	
c 98.725 This is systematic sampling since every fifth entry is selected. It gives a good representation of the mean. It is easy to use and is not time-consuming.	You can start at any number and choose every fifth number until you have 40 numbers in total. Everyone will have a slightly different answer for this depending on the starting place.
d 101.7 This is a random sample. Each time a random sample is chosen, different numbers will be selected. The advantage is that the selection is truly random. The disadvantage is that it will select different numbers each time and can be time-consuming.	For example, using randomly generated integers, the selection that appeared was: 26, 194, 38, 77, 142, 174, 27, 153, 34, 176, 36, 40, 67, 122, 84, 148, 162, $\boxed{38}$, 43, 132, 90, 98, 133, 166, 175, 103, 136, 185, 196, 64 38 is repeated, so another random number must be generated. This is 10. Find the corresponding number in the list and then find the mean of the selected numbers.
e In this case the systematic sample gave the closest value to the mean. It is also quite a simple method to use.	

Quota sampling is setting certain quotas for your sample, for example selecting a sample of eight boys and eight girls.

For example, the school canteen is considering introducing a new lunch menu and would like feedback from the students. The school has 250 boys and 300 girls and so the canteen manager decides to interview 25 boys and 30 girls to find out their opinion of the new menu. He stands at the entrance to the canteen and interviews the first 25 boys and 30 girls who come into the canteen.

This is called a quota sample. It is not random. It can be biased and unreliable. The advantage is that it is inexpensive, easy to perform and saves time.

However, it is more reliable than convenience sampling where people are selected based on availability and may not be representative of the population. This type of sampling produces a non-probability sample and can also be biased and unreliable.

Stratified sampling is selecting a random sample where numbers in certain categories are proportional to their numbers in the population.

For example, if 20% of students in a school were in Grade 7, then you would choose 20% of your sample from Grade 7. The 20% must be a random sample and not a convenience sample.

Example 8

Mandy asks all the students in her school to take a memory test. The students have to remember as many objects as they can from the 20 that Mandy shows them. The results are shown in the table.

Class 7 (20 students)	16, 15, 13, 15, 12, 8, 18, 16, 12, 11, 14, 17, 16, 9, 11, 10, 17, 13, 14, 13
Class 8 (27 students)	19, 15, 16, 14, 11, 16, 18, 15, 13, 12, 10, 8, 20, 14, 17, 12, 10, 7, 19, 20, 13, 17, 16, 16, 16, 15, 11
Class 9 (23 students)	17, 14, 15, 8, 7, 13, 15, 19, 16, 13, 11, 10, 17, 17, 20, 15, 11, 10, 7, 13, 16, 15, 15,
Class 10 (26 students)	9, 10, 10, 12, 18, 16, 17, 15, 11, 11, 14, 16, 19, 19, 11, 15, 17, 13, 13, 14, 13, 13, 9, 10, 8, 15
Class 11 (30 students)	16, 15, 15, 16, 16, 18, 11, 12, 13, 9, 10, 11, 16, 17, 15, 12, 12, 15, 15, 15, 18, 20, 16, 17, 17, 15, 14, 14, 14, 14
Class 12 (24 students)	9, 11, 16, 14, 13, 13, 18, 19, 12, 10, 11, 9, 16, 16, 18, 14, 15, 15, 16, 13, 13, 12, 18, 19

a In order to take a stratified sample of 40 students from the 150 in total, show that Mandy needs to select five students from Class 7.

b Determine how many students Mandy needs to select from each of the other classes.

a $\frac{20}{150} \times 40 = 5.333... \approx 5$ students	To select the 5 students, Mandy needs to use a random number generator to pick 5 numbers from the list for Class 7.
b From class 8 Mandy needs to select 7 students.	$\frac{27}{150} \times 40 = 7.2$, so 7 students from Class 8 Here again Mandy needs to select the 7 students using a random number generator.
From class 9 Mandy needs to select 6 students.	Similarly for the other classes.
From class 10 Mandy needs to select 7 students. From class 11 Mandy needs to select 8 students. From class 12 Mandy needs to select 6 students.	Note that due to rounding, the total is only 39.

Investigation 8

A dog kennel has 120 dogs. The ages of the dogs are:

1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3

3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5

5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 7

7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 8 9 9 9 9 9 10 10 10

10 11 11 11 11 12 12 12 13 13 13 14 14 15 16 16

1 How many dogs are in the population?

2 Represent the ages of the dogs and the frequencies in a table.

3 Which average can you read from the table?

4 Find the mean age of the dogs.

5 Discuss which method for finding the mean is easier: from the raw data or from the frequency table.

6 **Conceptual** Why can it be helpful to organize data in a table?

7 Describe how you would take a sample of 40 dogs.

8 **Conceptual** How can you decide whether your sample is unbiased?

9 Take a systematic sample of every five dogs, then find the mean of the sample.

10 Calculate the number of dogs of each age in a stratified sample of 40 dogs. Use the same method as in Example 8. Discuss why you do not get exactly 40 dogs using this method.

11 **Conceptual** How do you decide on the best sampling method to use?

Exercise 3E

1 The heights, to the nearest cm, of the students in a school are as follows:

Class 7 (28 students): 153, 149, 155, 148, 151, 150, 156, 154, 149, 152, 155, 154, 152, 156, 150, 151, 154, 155, 158, 147, 154, 155, 155, 156, 149, 151, 152, 153

Class 8 (30 students): 155, 154, 156, 158, 153, 155, 158, 157, 156, 155, 149, 151, 154, 153, 155, 154, 152, 159, 151, 149, 148, 153, 156, 155, 157, 155, 154, 157, 155, 156

Class 9 (26 students): 151, 158, 155, 156, 155, 158, 159, 160, 154, 153, 148, 156, 149, 150, 157, 156, 157, 156, 154, 155, 158, 153, 155, 150, 158, 160

Class 10 (24 students): 161, 158, 156, 148, 155, 156, 149, 159, 155, 156, 157, 158, 158, 161, 151, 159, 155, 156, 153, 160, 158, 155, 156, 158

Class 11 (25 students): 163, 160, 158, 149, 151, 159, 158, 162, 161, 156, 155, 154, 150, 151, 160, 159, 158, 156, 156, 155, 156, 157, 158, 158, 157

Class 12 (27 students): 151, 163, 165, 158, 155, 156, 159, 160, 161, 165, 159, 155, 156, 158, 158, 157, 155, 154, 152, 150, 163, 159, 158, 155, 156, 162, 158

a Find the mean height of the whole school.

b Use an appropriate sampling method to collect a sample of 50 students and find the mean of the sample. Comment on whether or not your sample is unbiased.

2 The ages of 100 people in a family camping site are as follows:

1	1	1	1	1	2	2	3	3	3
4	4	4	5	5	5	5	6	6	7
7	7	7	7	8	8	8	8	9	9
10	10	10	11	11	12	12	12	12	12
13	13	14	15	15	16	16	18	18	18
19	19	19	20	21	21	23	24	24	25
34	35	35	35	36	36	37	38	38	38
39	40	40	40	41	42	42	43	43	45
45	47	49	49	50	50	50	55	57	62
62	63	65	65	67	67	69	70	71	72

The manager decides to charge less for people over the age of 60.

a Find the mean age of the 100 people and decide whether or not the manager will lose much revenue due to this decision.

b Using an appropriate sampling method, pick a random sample of 35 people and find the mean age of the sample.

c Using a systematic sampling method of every third person, find the mean age of your sample.

d Comment on which method from parts **b** and **c** you think will give the better approximation to the population mean.

3 The number of goals scored in 50 hockey matches is as follows:

Girls: 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3,
3, 3, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8, 9
Boys: 0, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3,
3, 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8

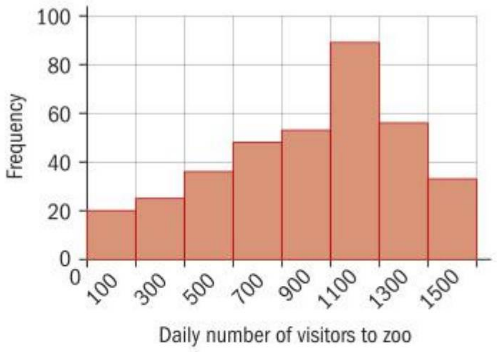
a Find the mean number of goals scored in all 50 matches.

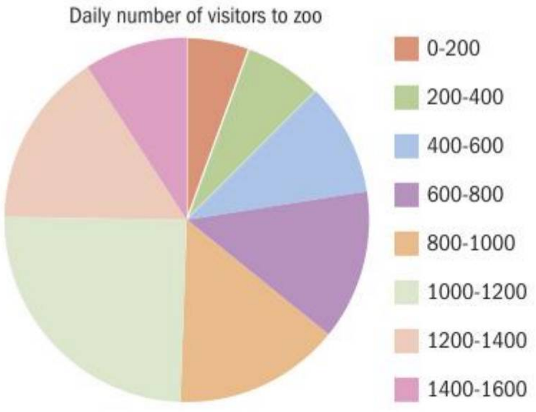
b Taking a random sample of 12 girls and 12 boys, find the mean of these 24 matches.

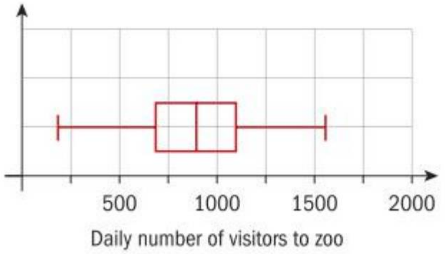
c Comment on whether or not your sample gives a good approximation to the population mean.

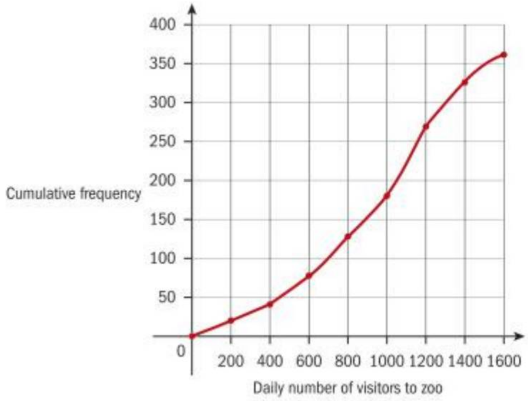
3.3 Presentation of data

A zoo is open 360 days in the year. The number of visitors each day was recorded and displayed in several different types of graph.









Reflect Which graph do you think is the most useful?

Discuss how the different types of graph might be interpreted.

Developing inquiry skills

Returning to the life expectancy and GDP data, what do you think would be the best method to use for taking a random sample of the data to use for finding an estimate of the mean?

Is the data discrete or continuous?

Frequency histograms

A frequency **histogram** is very similar to a **bar chart**. However, in a histogram there are no spaces between the bars.

Bar charts are useful for graphing **qualitative** data such as colour preference, whereas histograms are used to graph **quantitative** data.

In frequency histograms, as in bar charts, the vertical axis represents frequency.

To draw a frequency histogram, you need to find the lower and upper boundaries of the classes and draw the bars between these boundaries.

International-mindedness

What are the benefits of sharing and analysing data from different countries?

Example 9

Belinda collected data on the time in seconds that the girls and boys in her year group took to complete a 100 m run.

The results are:

Girls' times: 13.5, 13.8, 14.1, 14.3, 14.6, 14.7, 14.9, 15.2, 15.2, 15.3, 15.5, 15.5, 15.6, 15.7, 15.9, 16.1, 16.1, 16.3, 16.3, 16.3, 16.4, 16.6, 16.7, 16.7, 16.9, 17.2, 17.2, 17.5, 17.6, 17.8, 17.8, 18.4, 18.5, 18.9, 19, 20.1, 20,7, 21.4, 21.8, 22.5

Boys' times: 11.5, 11.8, 12.1, 12.4, 12.4, 12.6, 13.1, 13.2, 13.2, 13.2, 13.5, 13.6, 13.7, 14, 14.1, 14.1, 14.2, 14.3, 14.3, 14.3, 14.5, 14.6, 14.7, 14.9, 15.3, 15.4, 15.5, 15.5, 15.5, 15.7, 15.8, 16.3, 16.4, 16.6, 16.6, 16.7, 17.1, 17.4, 17.7, 20.5

a Complete the frequency table for this data.

Time (t seconds)	Frequency (girls)	Frequency (boys)
$11 < t \leq 12$		
$12 < t \leq 13$		
$13 < t \leq 14$		
$14 < t \leq 15$		
$15 < t \leq 16$		
$16 < t \leq 17$		
$17 < t \leq 18$		
$18 < t \leq 19$		
$19 < t \leq 20$		
$20 < t \leq 21$		
$21 < t \leq 22$		
$22 < t \leq 23$		

b Draw a frequency histogram for the girls and a frequency histogram for the boys to represent this data.

c The PE teacher was interested in comparing the times of the boys and the girls for the 100 m run. For both the girls and the boys, find the following information:

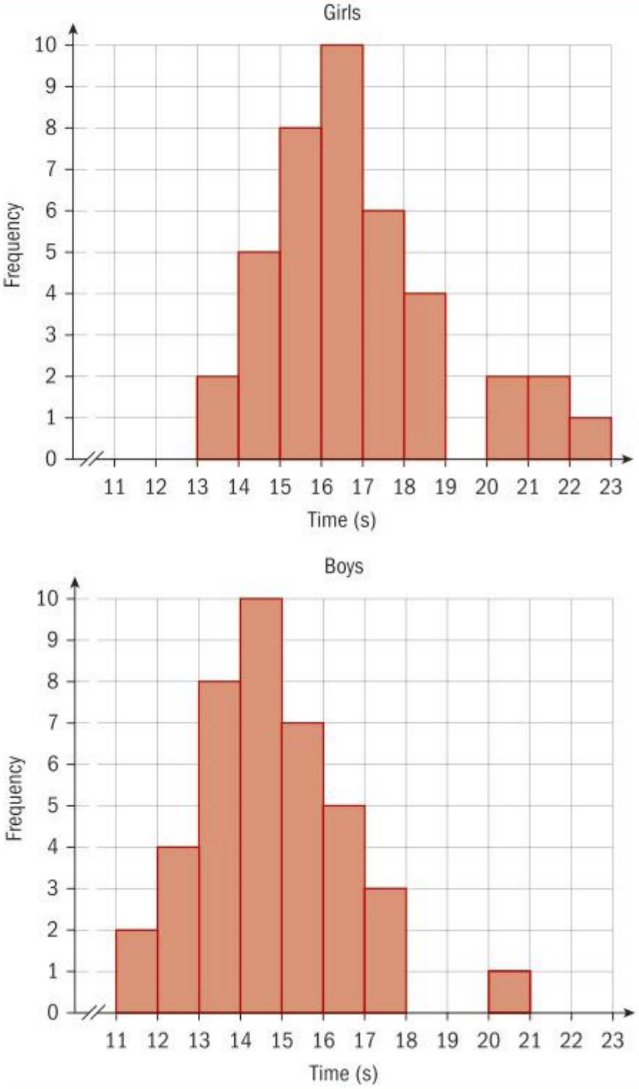
	Fastest time	Lower quartile	Median	Upper quartile	Slowest time
Girls					
Boys					

a

Time (t seconds)	Frequency (girls)	Frequency (boys)
$11 < t \leq 12$	0	2
$12 < t \leq 13$	0	4
$13 < t \leq 14$	2	8
$14 < t \leq 15$	5	10
$15 < t \leq 16$	8	7
$16 < t \leq 17$	10	5
$17 < t \leq 18$	6	3
$18 < t \leq 19$	4	0
$19 < t \leq 20$	0	0
$20 < t \leq 21$	2	1
$21 < t \leq 22$	2	0
$22 < t \leq 23$	1	0

Continued on next page

b



The *x*-axis will have to go from 11 to 23 at least.

The *y*-axis will have to go up as far as 10 at least.

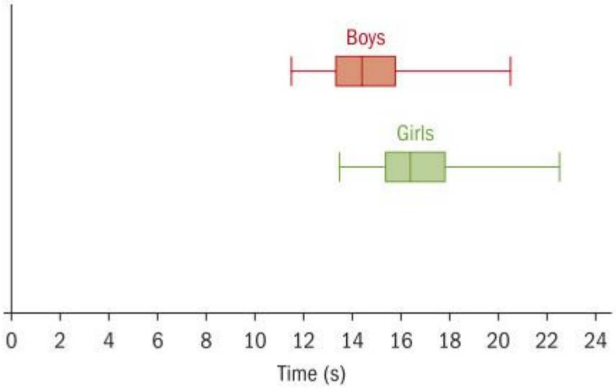
c

	Fastest time	Lower quartile	Median	Upper quartile	Slowest time
Girls	13.5	15.4	16.35	17.8	22.5
Boys	11.5	13.35	14.4	15.75	20.5

You can use the five values from the example to draw a box-and-whisker plot to compare the data sets.

Box-and-whisker plots are very convenient for comparing sets of data. Here you can easily see that boys have a faster time than girls for all five values. However, the spread of data for girls and boys is fairly equal.

You can also see that in both cases the data is not symmetrical about the median. The data between the median and the slowest time is more spread out than the rest of the data.



- To draw a box-and-whisker plot you need five pieces of information, called the five-number summary: the smallest value, the lower quartile (LQ), the median, the upper quartile (UQ) and the largest value.
- An outlier is a value that is much smaller or much larger than the other values. An outlier is a point less than the $\mathrm{LQ} - 1.5 \times \mathrm{IQR}$ or greater than the $\mathrm{UQ} + 1.5 \times \mathrm{IQR}$.

Example 10

Data on the shoe sizes of a group of students is shown in the table.

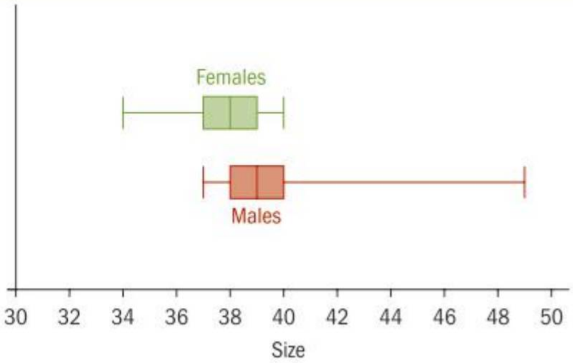
Shoe size	34	36	37	38	39	40	41	42	49
Females	2	2	10	8	7	3	0	0	0
Males	0	0	3	7	12	9	3	1	1

Draw a box-and-whisker plot for the females and for the males and compare the two plots.

State whether the box plots are symmetrical.

Comment on whether there are any outliers.

Draw the box plots again showing any outliers clearly. Outliers are represented by crosses.



The five-number summary for the females is 34, 37, 38, 39, 40.

$\mathrm{IQR} = 39 - 37 = 2$

$\mathrm{LQ} - 1.5(2) = 37 - 3 = 34$

$\mathrm{UQ} + 1.5(2) = 42$

So there are no outliers for the females.

The five-number summary for the males is 37, 38, 39, 40, 49.

$\mathrm{IQR} = 40 - 38 = 2$

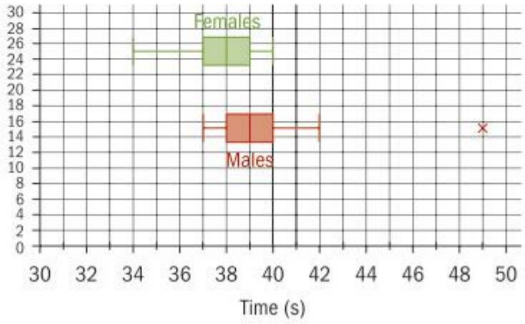
$\mathrm{LQ} - 1.5(2) = 35$

$\mathrm{UQ} + 1.5(2) = 43$

So 49 is an outlier for the males.

Continued on next page

The data for the females is more symmetrical than the data for the males.



Investigation 9

The weights, w kg, of 30 new-born babies in the Town hospital are:

2.6 3.1 1.8 2.5 4.6 3.6 3.4 2.9 4.8 6.9 4.1 5 3.5 1.2 4.4
5.1 9.6 3.3 4.1 3.7 2.8 2.9 3.4 5.1 4.6 3.9 2.7 3.6 4.2 4.9

The weights, w kg, of 30 new-born babies in the Country hospital are:

2.9 4.1 2.6 3.2 5.1 4.7 3.9 3.8 5.6 5.9 4.8 4.5 2.9 2.6 4.8
6.8 9.2 8.3 5.7 6.3 3.8 2.9 4.4 1.8 4.3 4.9 3.5 6.6 3.7 4.6

1 Complete the grouped frequency table:

Town hospital frequencies	Weight (w kg)	Country hospital frequencies
	$1 \le w < 2$	
	$2 \le w < 3$	
	$3 \le w < 4$	
	$4 \le w < 5$	
	$5 \le w < 6$	
	$6 \le w < 7$	
	$7 \le w < 8$	
	$8 \le w < 9$	
	$9 \le w < 10$	

2 Draw a histogram to represent the data.

3 Using the **original data** find the five-number summary for each hospital and draw box-and-whisker plots. You use the original data because it will give you exact answers and not approximations.

4 Discuss whether you can tell whether there are any outliers from the histogram.

5 Compare the two box-and-whisker plots. Discuss whether either of them is skewed.

6 **Conceptual** How do box-and-whisker plots allow you to compare data visually?

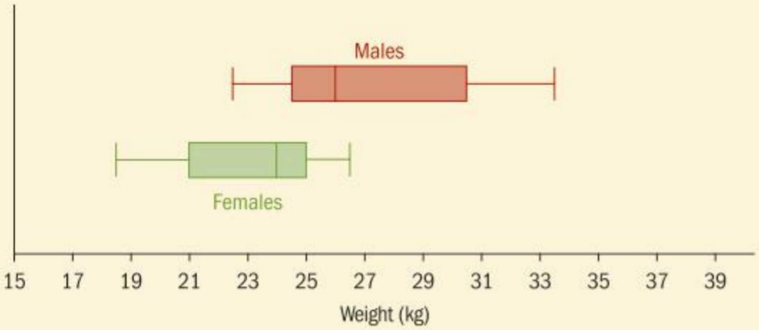
7 **Conceptual** How do you know which diagram to use to represent a data set?

Interpreting a box-and-whisker plot:

- 25% of values are between the smallest value and the LQ.
- 25% of values are between the LQ and the median.
- 25% of values are between the median and the UQ.
- 25% of values are between the UQ and the largest value.

Investigation 10

The box-and-whisker plots show the weights, in kilograms, of 24 male poodles (upper) and 24 female poodles (lower).



1 Write down the median for both groups.

2 Write down the IQR for both groups.

3 Write down the percentage of female poodles that weigh less than 24 kg.

4 Write down the percentage of male poodles that weigh between 26 and 30.5 kg.

5 Compare the two box plots and discuss the differences.

6 **Factual** Which measures of central tendency and dispersion can you read from a box-and-whisker plot and from a histogram?

7 **Conceptual** Is it useful to have more than one way of representing a univariate data set?

TOK

Can you justify using statistics to mislead others?

How easy is it to be misled by statistics?

Exercise 3F

1 Theo threw a die 40 times. The numbers that appeared were:

2 3 3 1 6 6 5 2 4 4 1 1 5 6 3 4 2 2 3 5
1 6 4 2 2 3 1 4 4 5 1 6 6 3 2 2 1 1 4 5

Millie also threw a die 40 times and the numbers that she threw are shown below.

6 5 5 6 1 1 3 4 5 4 3 2 2 2 4 5 4 6 6 1
1 2 2 1 3 3 3 6 5 5 4 1 2 2 3 3 6 5 1 3

a Construct frequency tables for these two sets of data.

b Draw a histogram to represent each set of data and compare the two histograms.

2 The number of goals scored in 25 female hockey matches is as follows:

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

The number of goals scored in 25 male hockey matches is as follows:

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

a Construct frequency tables for these two sets of data.

b Draw a histogram to represent each data set and compare the two histograms.

3 The heights, in cm, of 32 female gymnasts were recorded:

148 152 147 149 150 147 151
142 156 148 148 149 150 152
155 154 151 154 148 150 149
145 147 148 161 152 162 149
146 151 150 157

a Construct a grouped frequency table, using groups of 5 cm.

b Draw a histogram to represent this data.

c Draw a box-and-whisker plot to represent the data.

d State whether the data is symmetrical or not. Give a reason for your answer.

4 The times, in minutes, to complete 200 games of chess are shown in the table.

Time (x minutes)	Frequency
$20 \le x < 30$	36
$30 \le x < 40$	67
$40 \le x < 50$	48
$50 \le x < 60$	27
$60 \le x < 70$	10
$70 \le x < 80$	7
$80 \le x < 90$	5

a Draw a histogram to represent this data.

b Find the mean, median, LQ, UQ and range and determine whether there are any outliers.

c Given that the quickest time was 26 minutes and the longest time was 84 minutes, draw a box-and-whisker plot to represent this data.

d Marcus took 45 minutes to complete his game. Comment on whether you think that he should be satisfied with this result.

5 100 students were given 30 seconds to memorize 10 objects. The results (number of objects remembered) for the 50 boys and 50 girls are shown in the table.

Boys	Score	Girls
0	2	1
1	3	3
3	4	6
10	5	9
18	6	12
12	7	10
3	8	5
2	9	2
1	10	2

a Find the mean, median, LQ, UQ and range for the boys and for the girls, and determine whether there are any outliers.

b Draw box-and-whisker plots for the boys and the girls, and compare the results.

c Comment on whether the data sets are symmetrical.

6 This data shows the time it took in minutes for 35 students to complete a sudoku puzzle.

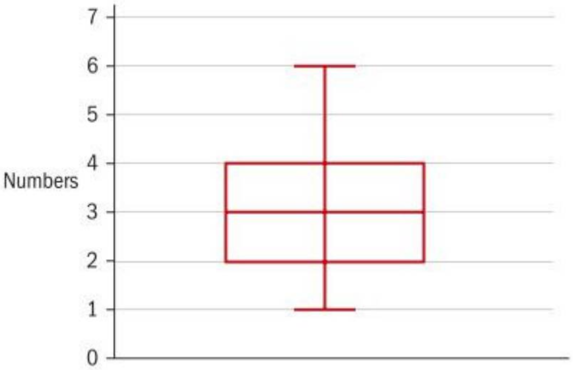
9	12	21	15	15	8	12	11	22
24	17	10	15	6	12	18	35	12
8	19	22	26	24	17	18	21	20
16	9	43	12	16	15	12	18	

a Find the mean, median, LQ, UQ and range for this data and determine whether there are any outliers.

b Draw a box-and-whisker plot.

c If Jin took 16 minutes to complete the sudoku puzzle, find the number of students who took longer than him to complete the puzzle.

7 The box-and-whisker plot shows the results from tossing a die 60 times.

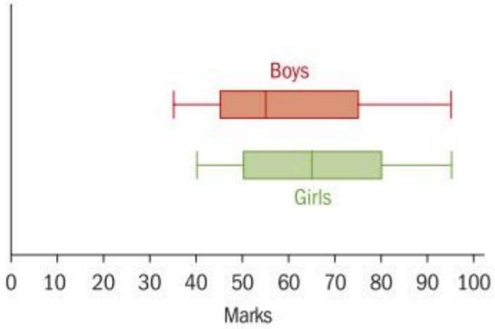


a Write down the median score.

b Find the interquartile range.

c Comment on whether the data is symmetrical.

8 The box-and-whisker plots show the scores in a mathematics test for 60 boys (top) and 60 girls (bottom).



a Write down the median score for the boys and for the girls.

b Find the interquartile range for the boys and for the girls.

c Write down the percentage of boys who scored between 45 and 55.

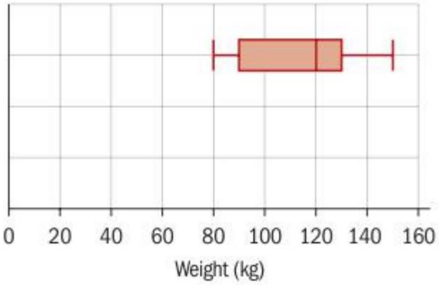
d Write down the percentage of girls who scored between 65 and 95.

e Find the number of boys who scored less than 45.

f Find the number of girls who scored more than 50.

g Comment on whether the data is symmetrical.

9 The box-and-whisker plot shows the weights, in kilograms, of 40 pandas all of the same gender.



a Write down the median weight.

b Write down the range.

c Find the number of pandas that weigh less than 90 kg.

d Write down the percentage of pandas that weigh between 120 and 160 kg.

e Find the number of pandas that weigh between 90 and 130 kg.

f If the average weight of a panda is about 120 kg, state what information you can deduce from the 40 pandas in the sample.

Male pandas weigh, on average, between 80 kg and 140 kg, and females weigh, on average, between 70 kg and 120 kg.

g From the box plot, state the gender of the pandas in the sample.

The **cumulative frequency** is the sum of all the frequencies up to a particular value. To draw a cumulative frequency curve, you need to construct a cumulative frequency table, with the upper boundary of each class interval in one column and the corresponding cumulative frequency in another. Then plot the upper class boundary on the x-axis and the cumulative frequency on the y-axis.

Example 11

The number of visitors, n, to Hailes Castle was noted on 200 separate days of the year.

Number of visitors (n)	Frequency
$0 \le n < 50$	16
$50 \le n < 100$	38
$100 \le n < 150$	50
$150 \le n < 200$	36
$200 \le n < 250$	32
$250 \le n < 300$	19
$300 \le n < 350$	6
$350 \le n < 400$	3

a Explain how you can tell that there were fewer than 100 visitors on 54 of the days.

b Complete this table with the upper boundaries and cumulative frequencies.

Upper boundary	Cumulative frequency
$n < 50$	16
$n < 100$	54
$n < 150$	
$n < 200$	
$n < 250$	
$n < 300$	
$n < 350$	
$n < 400$	

c Draw a cumulative frequency curve for this data.

d Use your cumulative frequency curve to find an estimate for the median, the lower quartile and the upper quartile. (In other words, find the values on the x-axis corresponding to 100, 50 and 150 on the y-axis.)

e Find an estimate for the 85th percentile.

f If you are told that the lowest number of visitors was 25 and the highest number was 370, draw a box-and-whisker plot to represent this data.

g Determine whether there are any outliers.

a On 16 days there were $0 \le n < 50$ visitors and on 38 days there were $50 \le n < 100$ visitors, so in total there were fewer than 100 visitors on $16 + 38 = 54$ days.

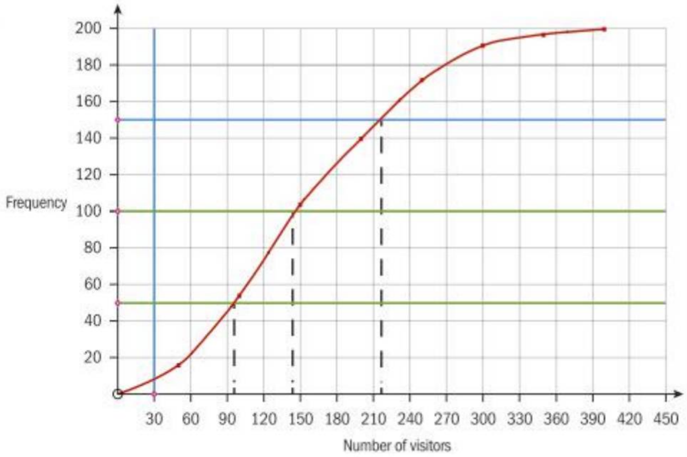
b

Upper boundary	Cumulative frequency
$n < 50$	16
$n < 100$	54
$n < 150$	104
$n < 200$	140
$n < 250$	172
$n < 300$	191
$n < 350$	197
$n < 400$	200

Add 54 to 50 to get 104.

Add 104 to 36 to get 140 and so on.

c



The upper boundary is plotted against the cumulative frequency, and the points are joined up with a smooth curve.

d Approximate values:

Median = 146

LQ = 95

UQ = 215

To find the median, draw a horizontal line from $100\left(=200\times\frac{1}{2}\right)$ on the y-axis to the curve.

From where this line meets the curve, draw a vertical line down to the x-axis and read off the answer.

Similarly for the LQ $\left(200\times\frac{1}{4}\right)$ and the UQ $\left(200\times\frac{3}{4}\right)$.

These are approximate values because you are not told the exact number of visitors each day.

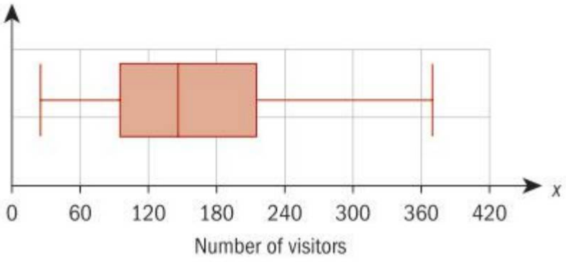
Continued on next page

e Approximately 247

85% of 200 = 170.

Draw a horizontal line from 170 on the y-axis until it meets the curve. Draw a vertical line from that point to the x-axis and read off the answer.

f



g IQR = 215 − 95 = 120

$95 - 1.5 \times 120 = -85$

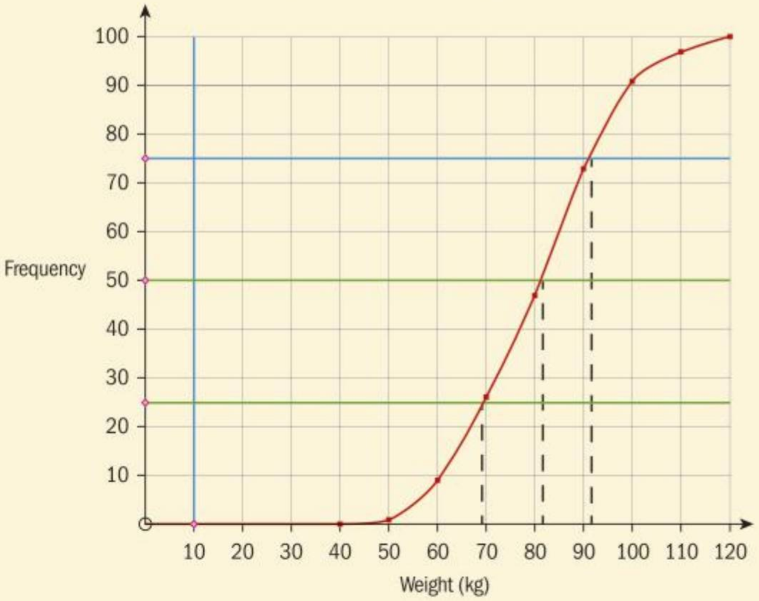
$215 + 1.5 \times 120 = 395$

So there are no outliers.

To find any **percentile**, p%, you read the value on the curve corresponding to p% of the total frequency.

Investigation 11

This is the cumulative frequency curve of the weights of 100 male athletes.



1 **Factual** What do the horizontal lines at 75, 50 and 25 on the vertical axis represent?

2 What percentage of the data values are below 70 kg?

International-mindedness

Hans Rosling (1948 – 2017) was a professor of international health at Sweden's Karolinska Institute. He co-founded the Swedish chapter of Medécins Sans Frontières, and was able to clearly show the importance of collecting and analysing real data in order to understand situations and plan for the future.

3 If 90% of the athletes weigh more than x kg, how could you find the value of x?

4 How accurate are the values you have found?

5 **Factual** How can you find out whether there are any outliers?

6 **Conceptual** How does the cumulative frequency curve allow you to analyse the data?

Exercise 3G

1 The table shows the average times, in minutes, that 100 people waited for a train.

Time (x minutes)	Frequency
$0 \le x < 2$	5
$2 \le x < 4$	11
$4 \le x < 6$	23
$6 \le x < 8$	31
$8 \le x < 10$	19
$10 \le x < 12$	8
$12 \le x < 14$	3

a Construct a cumulative frequency table for this data.

b Sketch the cumulative frequency curve.

c Use your graph to find estimates for the median and interquartile range.

d Find the 10th percentile.

The train company will refund the fare if customers have to wait 11 minutes or more for a train.

e Determine how many customers can claim for a refund of their fare.

2 Nuria recorded the number of words in a sentence in one chapter of her favourite book. The results are shown in the table.

Number of words (x)	Frequency
$0 \le x < 4$	5
$4 \le x < 8$	32
$8 \le x < 12$	41
$12 \le x < 16$	28
$16 \le x < 20$	22
$20 \le x < 24$	12
$24 \le x < 28$	7
$28 \le x < 32$	3

a Construct a cumulative frequency table for this data.

b Sketch the cumulative frequency curve.

c Use your graph to find estimates for the median and interquartile range.

d Determine whether there are any outliers.

e Find the 90th percentile.

f The smallest sentence had 1 word and the longest sentence had 31 words. Draw a box-and-whisker plot to represent this data.

g A children's book has, on average, 8 words in a sentence and an adult book has, on average, 15 words in a sentence. State the type of book you think Nuria is reading, justifying your answer.

3 A tourist attraction is open 350 days of the year. The number of visitors each day for the 350 days was recorded and the results are shown in the table.

Number of visitors (n)	Frequency
$100 \le n < 200$	24
$200 \le n < 300$	36
$300 \le n < 400$	68
$400 \le n < 500$	95
$500 \le n < 600$	73
$600 \le n < 700$	38
$700 \le n < 800$	16

a Draw a suitable graph to represent this data.

b Use your graph or the data to find estimates for the median and interquartile range.

c Determine whether or not there are any outliers.

d The smallest number of visitors was 185 and the largest number was 792. Draw a box-and-whisker plot to represent this data.

If the number of tourists is fewer than 350 in a day, then the attraction loses revenue.

e Determine the number of days on which the attraction loses revenue.

4 The table shows the number of points that 120 students received on their IB diploma.

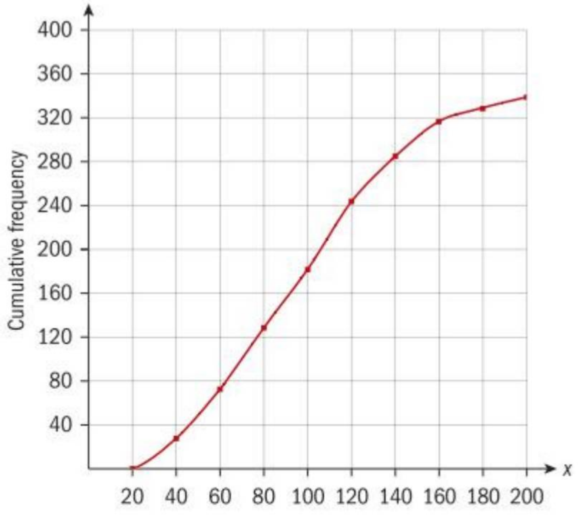
Number of points (n)	Boys	Girls
$21 \le x < 24$	2	1
$24 \le x < 27$	8	5
$27 \le x < 30$	10	8
$30 \le x < 33$	15	18
$33 \le x < 36$	9	12
$36 \le x < 39$	8	5
$39 \le x < 42$	4	8
$42 \le x < 45$	4	3

a Draw suitable graphs to represent this data.

b Use your graphs to compare the results for the boys and the girls.

c Mary and Martin both score 29 points. Compare their points with the other students of their gender.

5



a Using the cumulative frequency curve, write down estimates for:

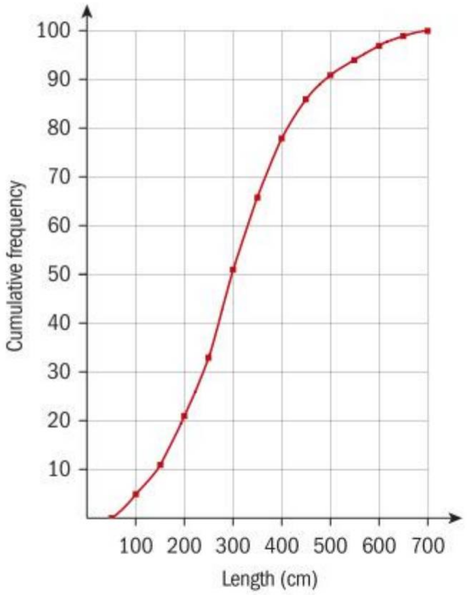
i the median

ii the interquartile range

iii the 90th percentile.

b Determine whether there are any outliers.

6



The cumulative frequency curve shows the lengths, in cm, of 100 snakes in a zoo.

a Write down estimates for the median, the lower quartile and the upper quartile.

b The smallest snake is 9 cm long and the longest is 650 cm long. Draw a box-and-whisker plot to represent this data.

c Construct a frequency table for the lengths of the snakes.

d Find estimates for the mean and standard deviation of the lengths of the snakes.

Developing inquiry skills

Return to the opening problem for the chapter. How can you best present the data on life expectancy and GDP?

3.4 Bivariate data

Monica and her friends are training for a charity run. She is interested to find out whether the height of the runners has any effect on the time taken to complete the race. You can see her data in Investigation 12.

TOK

Why are there different formulae for the same statistical measures, such as mean and standard deviation?

Bivariate data has **two** variables; **univariate** data has only one variable.

With bivariate data you have data on two different variables collected from the same individuals that you want to compare to see whether there is any **correlation** between the two variables.

Mr Price was interested to find out whether the number of past papers that his students completed had an effect on the grade they obtained in their final examination. The data he collected is shown below.

Number of past papers	2	6	5	1	4	8	3	12	7	4	2	8	10	9
Examination grade (%)	48	70	61	45	58	85	55	96	80	56	43	88	92	89

He plots all these points on a graph to see whether there is any correlation between the two variables. The number of past papers is the **independent** variable and this is plotted on the *x*-axis. The examination grade is the **dependent** variable and this is plotted on the *y*-axis.

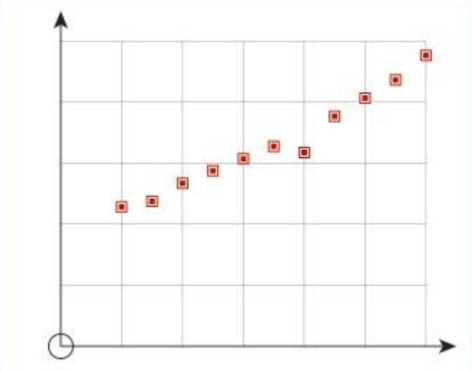
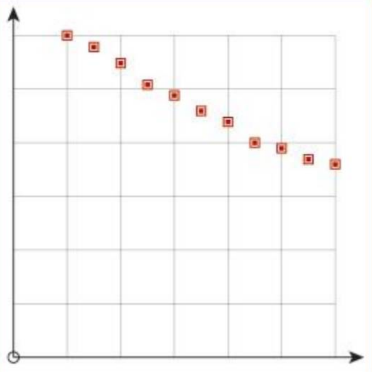
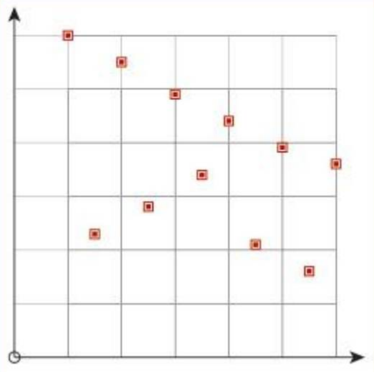
The pattern of dots or crosses will give him an indication of how closely the variables are related.

Do you think that the two variables are related?

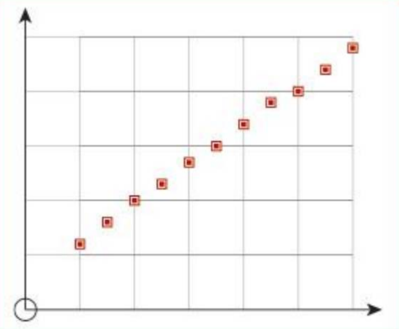
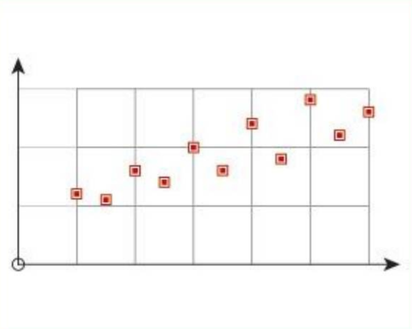
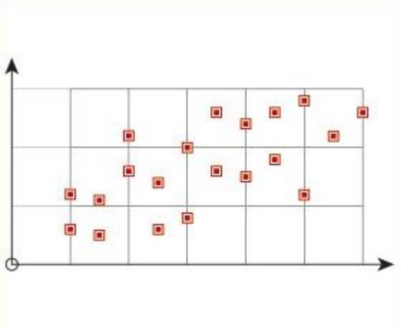
How closely do you think they are related?

What advice would you give to students who have to take examinations?

Types of correlation

		
Correlation can be **positive**. When the independent variable increases, so does the dependent variable.	Correlation can be **negative**. When the independent variable increases, the dependent variable decreases.	There can also be **no correlation**. This occurs when the points are scattered randomly.

Correlation can also be described as strong, moderate or weak.

		
Strong	Moderate	Weak

Example 12

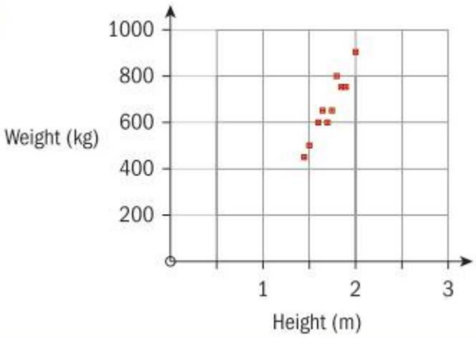
The table gives the heights and weights of 10 camels.

Weight (kg)	450	600	500	750	750	650	900	600	650	800
Height (m)	1.45	1.6	1.5	1.85	1.9	1.75	2.0	1.7	1.65	1.8

a Draw a scatter graph to represent this information.

b Comment on the relationship.

a



TOK

To what extent can we rely on technology to produce our results?

b There appears to be a strong, positive relationship between the height and the weight: the taller the camel, the more it weighs.

Investigation 12

Twelve students trained every week for a 5 km charity run. Their heights, h cm, and the times, t minutes, it took them to complete the run are shown in the table.

Height (h)	150	163	155	148	154	141	162	148	171	152	153	145
Time (t)	22	18	20	25	21	32	19	24	15	22	21	30

1 Draw a scatter graph to show this information. The height is the independent variable and goes on the horizontal axis; the time taken to complete the run is the dependent variable and goes on the vertical axis. Place a cross at each point, eg at (150, 22), (163, 18) etc.

2 Now that you have a visual picture, do you think that these variables are related?

3 If so, how would you describe the relationship?

4 **Conceptual** What are scatter diagrams useful for?

Example 13

The table shows the number of members in each of nine families and the number of pets the family has.

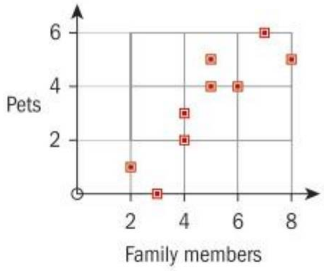
Number of members	2	3	4	4	5	5	6	7	8
Number of pets	1	0	3	2	5	4	4	7	5

a Draw a scatter graph to represent this data.

b Describe the correlation between the two variables.

c State, with a reason, whether you think that one variable "causes" the other.

a



b The correlation is positive and moderate.

c No, the number of members in a family is not caused by the number of pets a family has, and vice versa.

Example 14

The table shows the number of schools and the number of restaurants in a town over a 40-year period.

Year	1980	1985	1990	1995	2000	2005	2010	2015	2020
Number of schools	12	13	15	15	16	17	17	18	19
Number of restaurants	28	30	33	34	36	36	38	39	40

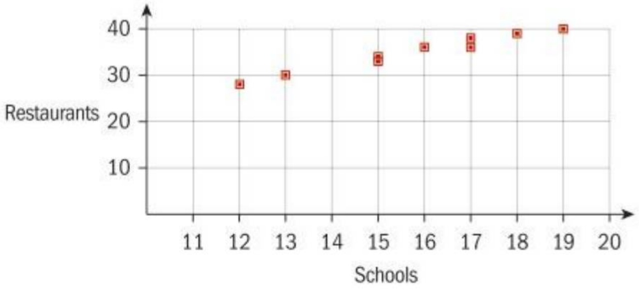
a Draw a scatter graph of the number of schools and the number of restaurants.

b Describe the correlation between the two variables.

c State whether you think that one set of variable "causes" the other.

d State a possible reason why the number of schools and the number of restaurants increased over the 40-year period.

a



b There is a strong positive correlation.

c Not really.

d The population in the town could be increasing over the years, which could require more schools and more restaurants.

Example 15

The table shows the temperature in °C and the time in days taken for cream to turn sour.

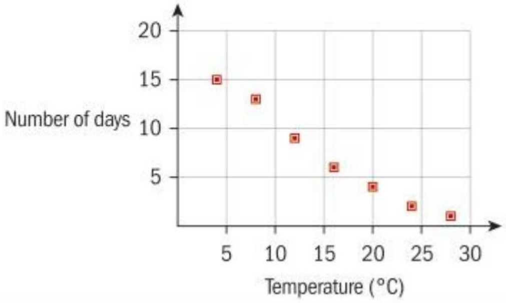
Temperature (°C)	4	8	12	16	20	24	28
Time in days	15	13	9	6	4	2	1

a Draw a scatter graph to represent this data.

b Describe the correlation between the two variables.

c State whether one variable "causes" the other.

a



b There is a strong negative correlation: the higher the temperature, the lower the number of days for the cream to sour.

c The increase in temperature could very well be the reason that the cream turns sour more quickly.

Investigation 13

Consider the following data sets and the correlations that were found:

A: The number of hours spent training for a race and the time taken to complete the race has a strong negative correlation.

B: The age of participants in a race and the time taken to complete the race has a moderate positive correlation.

C: The number of fish in a garden pool and the number of trees in the garden has a strong positive correlation.

D: The speed of a car and its horsepower has a moderate positive correlation.

E: The temperature and the number of hats sold has a weak negative correlation.

1 In which sets do you think that one variable has an influence on or "causes" the other?

2 In which examples is there a moderate or strong correlation but one variable does not cause the other?

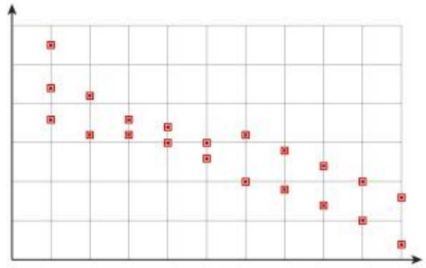
3 **Factual** What is causation?

4 **Conceptual** Does correlation imply causation?

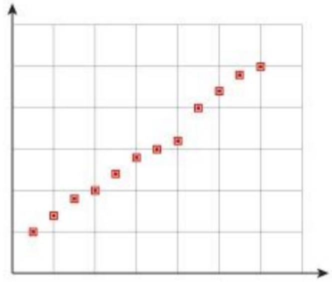
Exercise 3H

1 For the following scatter graphs, describe the type of correlation and the strength of the relationship.

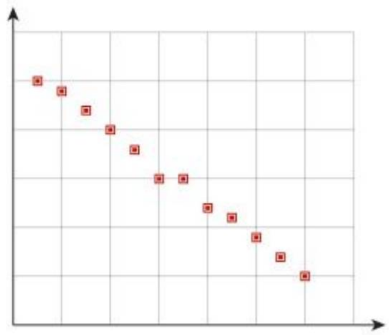
a



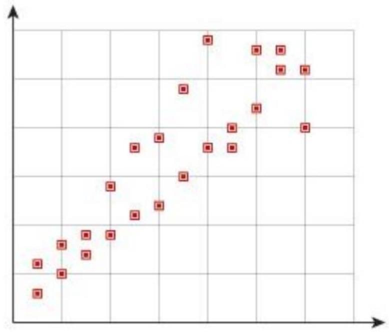
b

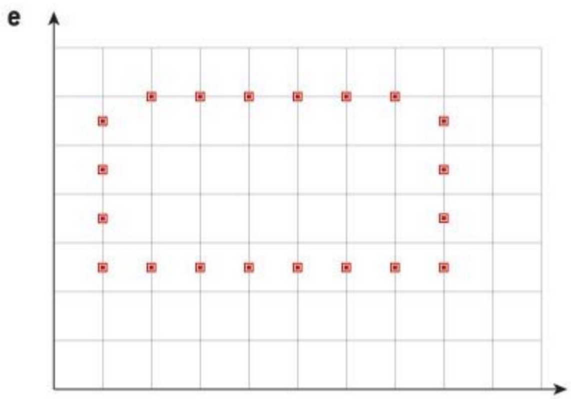


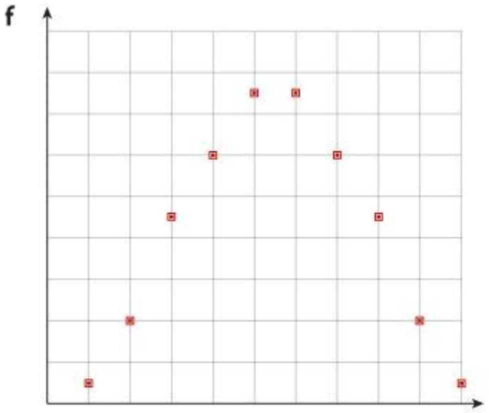
c



d







2 The table gives the heights, in cm, and weights, in kg, of 11 football players selected at random.

Height (h cm)	161	173	154	181	172	184	176	169	165	180	173
Weight (w kg)	74	76	61	80	76	88	79	76	75	83	75

a Plot the points on a scatter diagram.

b Comment on the type of correlation. Interpret what this means in terms of the football players.

c State whether the correlation might indicate a causation in this instance. Justify your answer.

3 The table shows the size, in inches, of 10 laptop screens and the cost, in euros, of the laptop.

Size (inches)	11.6	11.6	13.3	14	14	14	15	15.6	15.6	15.6
Cost (euros)	145	170	700	450	370	175	320	500	420	615

a Plot the points on a scatter diagram.

b Describe and interpret the correlation.

c State whether you think that the size has an influence on the cost.

4 Twelve students took tests in English and mathematics. The results are shown in the table.

English	44	66	71	33	87	90	55	76	65	95	40	58
Mathematics	71	75	58	63	55	87	54	58	77	54	56	51

a Plot the points on a scatter diagram.

b Describe the correlation.

c State whether you think that the grade for the English test has an influence on the grade for the mathematics test.

5 The data in the table shows the position in the league and the number of goals scored for each team in a hockey league.

Position	1	2	3	4	5	6	7	8	9	10	11	12
Goals scored	52	50	47	44	43	37	36	24	16	12	10	7

a Plot the points on a scatter diagram.

b Describe the correlation.

c State whether you think that the position in the league has an influence on the number of goals scored.

Developing inquiry skills

For the life expectancy and GDP data, use technology to draw a scatter graph of the data.

Is there any correlation between the two variables?

Is there any causal connection between life expectancy and GDP?

Developing your toolkit

Now do the Modelling and investigation activity on page 146.

Chapter summary

- **Univariate** data has only one variable.
- There are two main types of data: **qualitative** and **quantitative**.
- Qualitative data is data that is not given numerically, for example favourite ice cream flavour. Quantitative data is numerical and is classified as **discrete** or **continuous**.
- Discrete and continuous data can be organized into frequency tables or grouped frequency tables.
- For continuous data, the classes must cover the full range of the values and must not overlap.
- Discrete data is either data that can be counted, for example the number of cars in a car park, or data that can only take specific values, for example shoe size.
- Continuous data can be measured, for example height, weight and time.
- The most common measures of central tendency are the mean, median and mode.
- The **mode** of a data set is the value that occurs most frequently. There may be no mode or several modes.
- The **median** of a data set is the value that lies in the middle when the data values are arranged in size. When there are two middle values, the median is the midpoint between the two values.
- The **mean** of a data set is the sum of all the values divided by the number of values.

  For a discrete data set of n values the formula is $\overline{x} = \sum_{i=1}^{n} x_i$.

  For a frequency data set the formula is $\overline{x} = \dfrac{1}{\sum_{i=1}^{n} f_i} \sum_{i=1}^{n} f_i x_i$.
- **Outliers** are extreme data values that can distort the results of statistical processes.
- Measures of dispersion measure how spread out a data set is.
- The **range** is found by subtracting the smallest number from the largest number.
- The **standard deviation**, σ_x, gives an idea of how the data values are related to the mean. The standard deviation is also known as the root-mean-squared deviation and its formula is

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2}$$

- The **variance** is the standard deviation squared, $(\sigma_x)^2$.

Continued on next page

- The **interquartile range** (IQR) is the upper quartile, Q_3, minus the lower quartile, Q_1.
- When the data is arranged in order, the **lower quartile** is the data point at the 25th percentile and the **upper quartile** is the data point at the 75th percentile.
- The mean of a set of numbers is $\bar{x}$ and the standard deviation is σ_x. If you add k to or subtract k from each of the numbers, then the mean becomes $\bar{x} \pm k$ and the standard deviation remains σ_x. If you multiply each number by k then the mean becomes $k \times \bar{x}$ and the standard deviation becomes $|k| \times \sigma_x$.
- The **population** is the whole group from which you can collect data.
- A **sample** is a small group chosen from the population.
- **Simple random sampling** is selecting a sample completely at random, for example by using a random number generator or picking numbers from a hat.
- **Systematic sampling** is, for example, taking every fifth entry starting at a random place.
- **Convenience sampling** is getting data from people who are easy to reach, for example the members of a school, club, etc. It does not select a random sample of participants and so the results could be biased.
- A **biased** sample is one that is not random, for example researching spending habits on cars and only interviewing people exiting a garage.
- **Quota sampling** is setting certain quotas for your sample, for example selecting a sample of eight boys and eight girls.
- **Stratified sampling** is selecting a sample where the numbers in certain categories are proportional to their numbers in the population. For example, if 20% of students in a school were in Grade 7, then you would choose 20% of your sample from Grade 7.
- To draw a box-and-whisker plot you need five pieces of information: the smallest value, the lower quartile (LQ), the median, the upper quartile (UQ) and the largest value.
- An outlier is a point less than the LQ $- 1.5 \times$ IQR or greater than the UQ $+ 1.5 \times$ IQR.
- Interpreting a box-and-whisker plot:
  - 25% of the values are between the smallest value and the LQ.
  - 25% of the values are between the LQ and the median.
  - 25% of the values are between the median and the UQ.
  - 25% of the values are between the UQ and the largest value.
- The **cumulative frequency** is the sum of all the frequencies up to a particular value. To draw a cumulative frequency curve, you need to construct a cumulative frequency table, with the upper boundary of each class interval in one column and the corresponding cumulative frequency in another. Then plot the upper boundary on the x-axis and the cumulative frequency on the y-axis.
- To find any **percentile**, p%, you read the value on the curve corresponding to p% of the total frequency.
- **Bivariate** data has two variables; univariate data has only one variable.
- With bivariate data you have paired data on two variables that you want to compare to see whether there is any **correlation** between the two variables.
- Correlation can **positive** or **negative**, or there may be **no correlation**, and correlation can also be described as **strong**, **moderate** or **weak**.

Developing inquiry skills

Return to the opening problem.

- Has what you have learned in this chapter helped you to answer the questions?
- What information did you manage to find?
- What assumptions did you make?
- How will you be able to construct a model?
- What other things did you wonder about?

Thinking about the inquiry questions from the beginning of this chapter:

- Has what you have learned in this chapter helped you to think about an answer to most of these questions?
- Are there any questions that you would like to explore further, perhaps for your internal assessment topic?

Chapter review

Click here for a mixed review exercise

1 State whether the following sets of data are discrete or continuous, and, in each case, construct a frequency table.

a The number of apples in a 1 kg bag:

8 7 9 7 8 10 9 8 7
11 9 9 10 12 7 8 10

b The lengths of pencils, in cm:

7.4 8.5 9.6 7.1 14 13.5 8.8
7.4 11.2 13.6 12.8 14.2 9.8

c The shoe sizes of Grade 6:

34 35 34 33 36 37 36 38 35
36 37 38 35 37 37 38 35

2 Find the mean, median and mode for the following data sets. State which measure of central tendency is best to use in each case.

a The heights of 15 dogs, in cm:

7 23 32 41 32 56 64 67
88 91 110 78 56 45 32

b The price of a pair of shoes in dollars:

46 54 58 62 62 79 96
120 135 185 270 300

c The number of hours Grade 12 students sleep:

4 7 6 6 8 6 9 8 6 5
4 5 5 6 8 8 8 6 7

3 The data table shows the lengths of 120 pike fish.

Length of pike (l cm)	Frequency
$20 \le l < 30$	2
$30 \le l < 40$	12
$40 \le l < 50$	23
$50 \le l < 60$	46
$60 \le l < 70$	28
$70 \le l < 80$	9

a Write down the modal class.

b Find estimates for the median, mean and standard deviation.

c Draw a histogram to represent the data.

4 The marks, out of 50, for a history test have a mean of 38 and a standard deviation of 7. To get a percentage mark, Mr Thoughtful doubles all the marks. Write down the new mean and the new standard deviation.

Statistics and probability

5 Mr Pringle sells vegetables at the market. The number of tomatoes in a bag has a mean of 10 and a standard deviation of 1. On his birthday, Mr Pringle gives everyone who buys a bag of tomatoes three extra tomatoes. Write down the new mean and the new standard deviation on Mr Pringle's birthday.

6 Ursula measures the heights of 35 tulips in her garden. The data she gathered is:

20	20	21	22	22	22	24	25	27
28	28	29	30	31	32	33	33	34
34	34	35	35	36	37	39	39	39
40	41	41	42	43	43	44	45	

a Find the mean and standard deviation and comment on your answer.

b Find the range and interquartile range.

c Write down the median, LQ, UQ, smallest value and largest value and check whether there are any outliers.

d Draw a box-and-whisker plot to represent the data.

7 The grouped frequency table shows the number of hours of voluntary service completed by the 200 students at a community high school.

Number of hours (x)	Frequency
$0 \le x < 10$	8
$10 \le x < 20$	16
$20 \le x < 30$	41
$30 \le x < 40$	54
$40 \le x < 50$	36
$50 \le x < 60$	22
$60 \le x < 70$	17
$70 \le x < 80$	6

a Construct a cumulative frequency table for this data.

b Plot the points and draw the cumulative frequency curve.

c Use your curve to find approximate values for the median and the interquartile range.

The lowest number of hours was 8 and the greatest number was 76.

d Draw a box-and-whisker plot to represent the data.

8 Mr Farmer has 50 chickens. He collects data on the temperature and the average number of eggs that the chickens lay.

Temperature (°C)	Number of eggs
14	43
15	44
16	48
17	46
18	50
19	48
20	50
21	52
22	53
23	55

a Draw a scatter graph to represent this information.

b Describe the correlation.

c Comment on whether the temperature has an effect on the number of eggs laid.

Exam-style questions

9 **P1:** The grouped frequency table below shows the results of a statistics test taken by 70 students.

Test result x%	Frequency
$0 \le x < 20$	8
$20 \le x < 40$	17
$40 \le x < 60$	25
$60 \le x < 80$	13
$80 \le x < 100$	7

a State the modal class for the data. (1 mark)

b Find an estimate for the mean. (3 marks)

c Find an estimate for the standard deviation. (2 marks)

d A similar class took the same test. Their mean mark was 45% and the standard deviation was 19.5.

Compare the marks of the two classes, stating your conclusions (2 marks)

10 **P2:** The weights (in grams) of 25 mice were recorded as follows.

10, 11, 12, 12, 13, 14, 14, 14, 14, 15, 15, 15, 16, 16, 16, 17, 18, 18, 19, 19, 19, 20, 20, 20, 21

a Find the mean weight of the mice. (2 marks)

b Find the median weight of the mice. (2 marks)

c Find the interquartile range. (4 marks)

d The weight of another mouse was added to the data but found to be an outlier.

Find the least possible weight this new mouse could be, given that it is heavier than each of the others. (2 marks)

11 **P2:** The following tables show the mean daily temperatures, by month, in both Tenerife and Malta.

Tenerife	
Month	**Mean daily temperature (°C)**
January	19
February	20
March	21
April	21
May	23
June	25
July	28
August	29
September	28
October	26
November	23
December	20

Malta	
Month	**Mean daily temperature (°C)**
January	16
February	16
March	17
April	20
May	24
June	28
July	31
August	31
September	28
October	25
November	21
December	17

a Find the mean temperature over the course of the year for Tenerife. (2 marks)

b Find the standard deviation of temperatures in Tenerife. (2 marks)

c Find the mean temperature over the course of the year for Malta. (2 marks)

d Find the standard deviation of temperatures in Malta. (2 marks)

e By referring directly to your answers from parts **a–d**, make contextual comparisons about the temperatures in Tenerife and Malta throughout the year. (4 marks)

12 **P1:** A population of ferrets has mean age 5.25 years and standard deviation 1.2 years.

a Find the mean age of the same population of ferrets 3 years later. (2 marks)

b Find the standard deviation of the same population of ferrets 2 years later, justifying your answer. (2 marks)

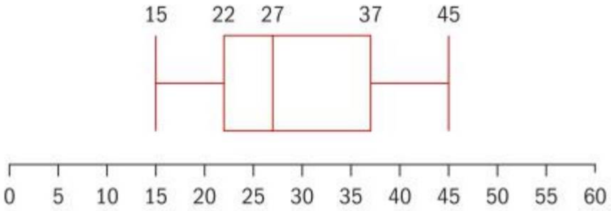
13 **P1:** The following table shows the population sizes of England, Wales, Scotland and Northern Ireland.

Region	Population (millions)
England	54.8
Wales	3.10
Scotland	5.37
Northern Ireland	1.85

A polling company wishes to ask questions of a stratified sample of the UK population, and decides that a sample size of 5000 people would be appropriate.

Determine how many people from each region they should choose. (6 marks)

14 **P1:** The box-and-whisker diagram shows the average times taken for a class of students to walk to school.



a Find the range. (2 marks)

b Find the interquartile range. (2 marks)

c Find the percentage of students who took between 15 and 37 minutes to walk to school. (1 mark)

d It was found that another student who took 60 minutes to walk to school.

Determine whether this time would be counted as an outlier. (3 marks)

15 **P1:** Ben practises playing the Oboe daily.

The time (in minutes) he spends on daily practice over 28 days is as follows:

10, 15, 30, 35, 40, 40, 45, 55, 60, 62, 64, 64, 66, 68, 70, 70, 72, 75, 75, 80, 82, 84, 90, 90, 105, 110, 120, 180

a Find the median time. (2 marks)

b Find the lower quartile. (2 marks)

c Find the upper quartile. (2 marks)

d Find the range. (2 marks)

e Determine whether there are any outliers in the data. (4 marks)

f Draw a box-and-whisker diagram for the above data, marking any outliers as required. (3 marks)

16 **P2:** Ava practises the piano daily.

The time (in minutes) she spends on daily practice over 75 days is as follows.

Time (t minutes)	Frequency
$0 \le x < 15$	4
$15 \le x < 30$	5
$30 \le x < 45$	12
$45 \le x < 60$	24
$60 \le x < 75$	18
$75 \le x < 90$	7
$90 \le x < 100$	5

a State the modal class. (1 mark)

b Find the class in which the median time lies. (2 marks)

c Construct a cumulative frequency table for this data. (3 marks)

d Sketch the cumulative frequency curve. (2 marks)

e Use your curve to find estimates for the median and interquartile range. (4 marks)

17 **P2:** The following table shows the salaries of the members of a small private business.

Position	Salary ($)
Director	120 000
Line Manager 1	80 000
Line Manager 2	80 000
Analyst 1	25 000
Analyst 2	25 000
Analyst 3	25 000
Analyst 4	25 000
Analyst 5	25 000
Analyst 6	25 000
Analyst 7	25 000
Analyst 8	25 000

a Calculate the mean salary. (2 marks)

b Find the median salary. (2 marks)

c Calculate the interquartile range. (2 marks)

Analyst 8 decides to argue for a pay rise.

d Suggest which measure of average (mean, median or mode) Analyst 8 should use to support their case. Justify your answer. (2 marks)

e Suggest which measure of average (mean, median or mode) the managing director might use to counter the claim that Analyst 8 should be paid more. Justify your answer. (2 marks)

f Comment on which measure of average would be fairest as a representative salary for employees in this company. Justify your answer. (2 marks)

18 P1: a Define, as fully as you can, the terms random sampling, stratified sampling and systematic sampling. (5 marks)

b A researcher wishes to investigate the size of rats in a London Underground station.

i Suggest one reason why systematic sampling should not be used. (2 marks)

ii Determine whether random or stratified sampling would be more appropriate in this investigation. Justify your answer. (2 marks)

19 P2: The following raw data is a list of the height of flowers (in cm) in Eve's garden.

26.5, 53.2, 27.5, 33.6, 44.6, 39.5, 24.9, 45.1, 47.8, 39.3, 33.1, 38.7, 44.1, 22.3, 44.1, 30.5, 25.5, 35.9, 37.1, 40.2, 23.3, 36.2, 34.8, 37.3

a Copy and complete the following grouped frequency table.

Height (x cm)	Frequency
$20 \le x < 25$	
$25 \le x < 30$	
$30 \le x < 35$	
$35 \le x < 40$	
$40 \le x < 45$	
$45 \le x < 50$	
$50 \le x < 55$	

(3 marks)

b Find an estimate for the mean height, using the frequency table. (2 marks)

c Find an estimate for the variance, using the frequency table. (2 marks)

d Find an estimate for the standard deviation, using the frequency table. (2 marks)

e Eve's neighbour's garden was also surveyed.

It was found that the flowers in the neighbour's garden had a mean height of 32.1 cm and standard deviation 7.83 cm.

Compare the heights of the flowers in the two gardens, drawing specific conclusions. (3 marks)

20 P1 The population of Frankfurt in Germany was found to be 718 824.

A company chose a random sample of 1200 residents of Frankfurt to ask for comments on the city's proposed integrated transport system.

611 of the sample chosen were female.

a Calculate an estimate for the number of females in Frankfurt. (2 marks)

b The company decided to repeat their survey, but this time chose to use a stratified sample rather than a random sample.

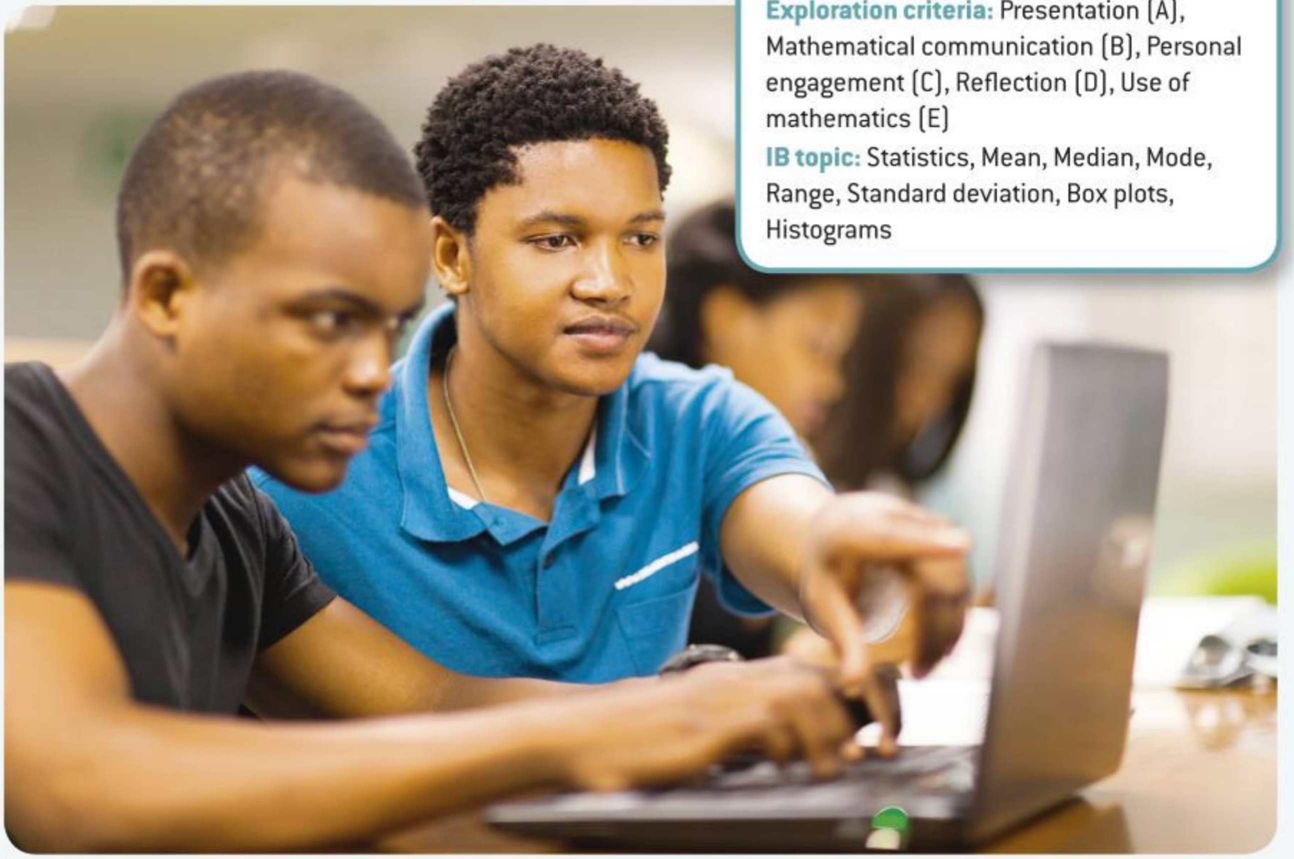
c Suggest two possible types of strata (apart from gender) that would be sensible for the company to use. (2 marks)

What's the difference?

Approaches to learning: Thinking skills, Communicating, Collaborating, Research

Exploration criteria: Presentation (A), Mathematical communication (B), Personal engagement (C), Reflection (D), Use of mathematics (E)

IB topic: Statistics, Mean, Median, Mode, Range, Standard deviation, Box plots, Histograms



Example experiment

Raghu does an experiment with a group of 25 students.

Each member of the group does a reaction test and Raghu records their times.

Raghu wants to repeat the experiment, but with some change.

He then wants to compare the reaction times in the two experiments.

Discuss:

How could Raghu change his experiment when he does it again?

With each change, is the performance in the group likely to improve, stay the same or get worse?

Alternatively, Raghu could use a different group when he repeats the experiment.

What different group could he use?

With each different group, is the performance likely to improve, stay the same or get worse?

Your experiment

Your task is to devise an experiment to test your own hypothesis.

You will need to do your experiment two times and compare your results.

Step 1: What are you going to test? State your aim and hypothesis.

Write down the aim of your experiment and your hypothesis about the result.

Why do you think this is important?

What are the implications of the results that you may find?

Make sure it is clear what you are testing for.

Step 2: How are you going to collect the data? Write a plan.

- What resources/sites will you need to use?
- How many people/students will you be able to/need to collect data from to give statistically valid results?
- Exactly what data do you need to collect? How are you going to organize your data? Have you done a trial experiment?
- Are there any biases in the way you present the experiment? How can you ensure that everyone gets the same instructions?
- Is your experiment a justifiable way of testing your hypothesis? Justify this. What are the possible criticisms? Can you do anything about them?
- Is the experiment reliable? Is it likely that someone else would reach a similar conclusion to yours if they used the same method?

Step 3: Do the experiment and collect the data.

Construct a results sheet to collect the data.

Give clear, consistent instructions.

Step 4: Present the data for comparison and analysis.

How are you going to present the data so that the two sets can be easily compared?

How are you going to organize the summary statistics of the two data sets so that you can compare them?

Do you need to find all of the summary statistics covered in this chapter?

Step 5: Compare and analyse.

Describe the differences between your two sets of data.

Make sure that your conclusion is relevant to your aim and hypothesis stated at the beginning.

Step 6: Conclusions and implications.

What are the conclusions from the experiment?

Are they different from or the same as your hypothesis? To what extent? Why?

How confident are you in your results? How could you be more certain?

What is the scope of your conclusions?

How have your ideas changed since your original hypothesis?

Extension

- How could you test whether the spread (rather than the average) of the data has changed significantly?
- How could you analyse changes in individual results, rather than whole class changes?
- Investigate the “difference in means test”.