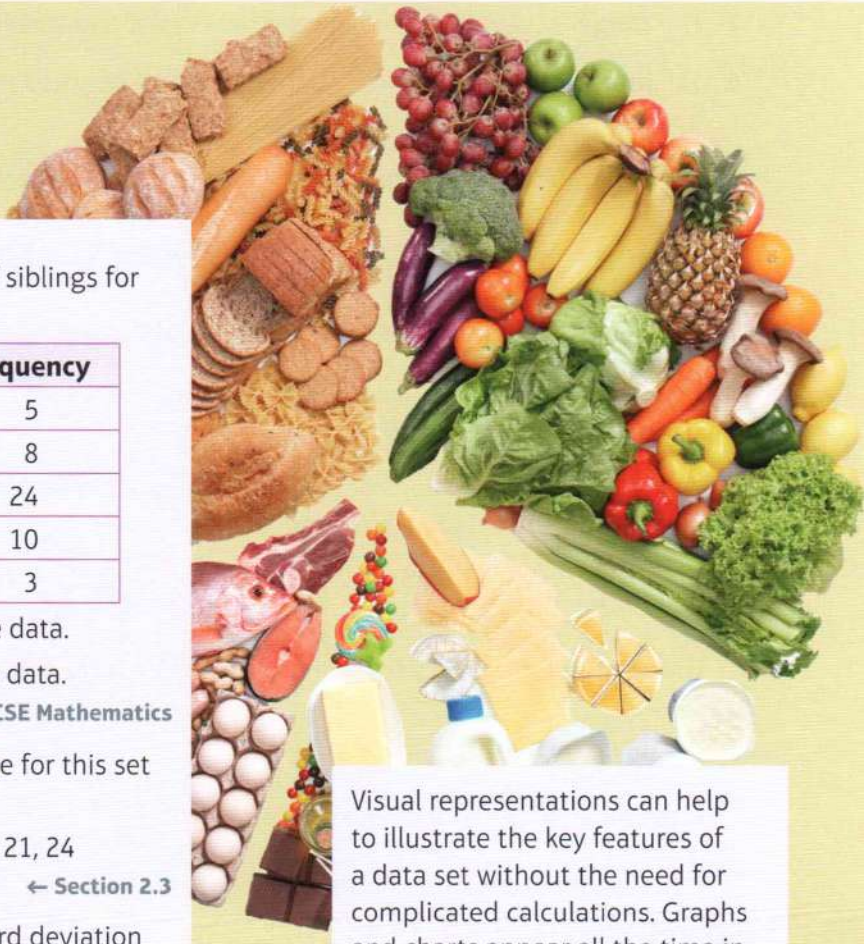
3 Representations of data

Objectives

After completing this chapter you should be able to:

- Identify outliers in data sets → pages 41–43
- Draw and interpret box plots → pages 43–45
- Draw and interpret cumulative frequency diagrams → pages 46–48
- Draw and interpret histograms → pages 48–52
- Compare two data sets → pages 53–54



Prior knowledge check

1 The table shows the number of siblings for 50 year 12 students:

Number of siblings	Frequency
0	5
1	8
2	24
3	10
4	3

a Draw a bar chart to show the data.

b Draw a pie chart to show the data.

← GCSE Mathematics

2 Work out the interquartile range for this set of data:

3, 5, 8, 8, 9, 11, 14, 15, 18, 20, 21, 24

← Section 2.3

3 Work out the mean and standard deviation for this set of data:

17, 19, 20, 25, 28, 31, 32, 32, 35, 37, 38

← Sections 2.1, 2.4

Visual representations can help to illustrate the key features of a data set without the need for complicated calculations. Graphs and charts appear all the time in newspapers and magazines, often stylised to suit the nature of the article.

3.1 Outliers

An outlier is an extreme value that lies outside the overall pattern of the data.

There are a number of different ways of calculating outliers, depending on the nature of the data and the calculations that you are asked to carry out.

- **A common definition of an outlier is any value that is:**
  - **either greater than $Q_3 + k(Q_3 - Q_1)$**
  - **or less than $Q_1 - k(Q_3 - Q_1)$**

In the exam, you will be told which method to use to identify outliers in data sets, including the value of k.

Notation Q_1 and Q_3 are the first and third quartiles.

Example 1

The blood glucose of 30 females is recorded. The results, in mmol/litre, are shown below:

1.7, 2.2, 2.3, 2.3, 2.5, 2.7, 3.1, 3.2, 3.6, 3.7, 3.7, 3.7, 3.8, 3.8, 3.8,
3.8, 3.9, 3.9, 3.9, 4.0, 4.0, 4.0, 4.0, 4.4, 4.5, 4.6, 4.7, 4.8, 5.0, 5.1

An outlier is an observation that falls either 1.5 × interquartile range above the upper quartile or 1.5 × interquartile range below the lower quartile.

a Find the quartiles. **b** Find any outliers.

a Q_1: $\frac{30}{4} = 7.5$ pick the 8th term = 3.2

Work out $n \div 4$ and round up. ← Section 2.2

Q_3: $\frac{3(30)}{4} = 22.5$ pick the 23rd term = 4.0

Work out $3n \div 4$ and round up. ← Section 2.2

Q_2: $\frac{30}{2} = 15$ pick the 15.5th term = 3.8

Work out $n \div 2$ and go halfway to the next term. ← Section 2.1

b Interquartile range = 4.0 − 3.2 = 0.8

Outliers are values less than

$3.2 - 1.5 \times 0.8 = 2$

and greater than $4.0 + 1.5 \times 0.8 = 5.2$

Use the definition of an outlier given in the question.

Therefore 1.7 is the only outlier.

1.7 < 2 so it is an outlier.

Example 2

The lengths, in cm, of 12 giant African land snails are given below:

17, 18, 18, 19, 20, 20, 20, 20, 21, 23, 24, 32

a Calculate the mean and standard deviation, given that $\Sigma x = 252$ and $\Sigma x^2 = 5468$.

b An outlier is an observation which lies ±2 standard deviations from the mean. Identify any outliers for this data.

Notation Σx is the sum of the data and Σx^2 is the sum of the square of each value.

a Mean $= \frac{\Sigma x}{n} = \frac{252}{12} = 21$ cm

Variance $= \frac{\Sigma x^2}{n} - \bar{x}^2 = \frac{5468}{12} - 21^2$

$= 14.666...$

Standard deviation $= \sqrt{14.666...}$

$= 3.83$ (3 s.f.)

Use the summary statistics given to work out the mean and standard deviation quickly.

b Mean $- 2 \times$ standard deviation

$= 21 - 2 \times 3.83 = 13.34$

Mean $+ 2 \times$ standard deviation

$= 21 + 2 \times 3.83 = 28.66$

32 cm is an outlier.

Use the definition of an outlier given in the question.

Watch out Different questions might use different definitions of outliers. Read the question carefully before finding any outliers.

Sometimes outliers are legitimate values which could still be correct. For example, there really could be a giant African land snail 32 cm long.

However, there are occasions when an outlier should be removed from the data since it is clearly an error and it would be misleading to keep it in. These data values are known as **anomalies**.

- **The process of removing anomalies from a data set is known as cleaning the data.**

Watch out Be careful not to remove data values just because they do not fit the pattern of the data. You must justify why a value is being removed.

Anomalies can be the result of experimental or recording error, or could be data values which are not relevant to the investigation.

Here is an example where there is a clear anomaly:

Ages of people at a birthday party: 12, 17, 21, 33, 34, 37, 42, 62, 165

$\bar{x} = 47$ $\sigma = 44.02$ $\bar{x} + 2\sigma = 135.04$

The data value recorded as 165 is significantly higher than $\bar{x} + 2\sigma$, so it can be considered an outlier. An age of 165 is impossible, so this value must be an error. You can clean the data by removing this value before carrying out any analysis.

Notation You can write $165 \gg 135.04$ where $\gg$ is used to denote 'much greater than'. Similarly you can use $\ll$ to denote 'much less than'.

Exercise 3A

1 Some data is collected. $Q_1 = 46$ and $Q_3 = 68$.

A value greater than $Q_3 + 1.5 \times (Q_3 - Q_1)$ or smaller than $Q_1 - 1.5 \times (Q_3 - Q_1)$ is defined as an outlier.

Work out whether the following are outliers using this rule:

a 7 **b** 88 **c** 105

2 The masses of male and female turtles are given in grams. For males, the lower quartile was 400 g and the upper quartile was 580 g. For females, the lower quartile was 260 g and the upper quartile was 340 g.

An outlier is an observation that falls either 1 × (interquartile range) above the upper quartile or 1 × (interquartile range) below the lower quartile.

a Which of these male turtle masses would be outliers?

400 g 260 g 550 g 640 g

b Which of these female turtle masses would be outliers?

170 g 300 g 340 g 440 g

c What is the largest mass a male turtle can be without being an outlier?

Hint The definition of an outlier here is different from that in question 1. You will be told which rule to use in the exam.

3 The masses of arctic foxes are found and the mean mass was 6.1 kg. The variance was 4.2.

An outlier is an observation which lies ±2 standard deviations from the mean.

a Which of these arctic fox masses are outliers?

2.4 kg 10.1 kg 3.7 kg 11.5 kg

b What are the smallest and largest masses that an arctic fox can be without being an outlier?

E **4** The ages of nine people at a children's birthday party are recorded. $\Sigma x = 92$ and $\Sigma x^2 = 1428$.

a Calculate the mean and standard deviation of the ages. **(3 marks)**

An outlier is an observation which lies ±2 standard deviations from the mean.

One of the ages is recorded as 30.

b State, with a reason, whether this is an outlier. **(2 marks)**

c Suggest a reason why this age could be a legitimate data value. **(1 mark)**

d Given that all nine people were children, clean the data and recalculate the mean and standard deviation. **(3 marks)**

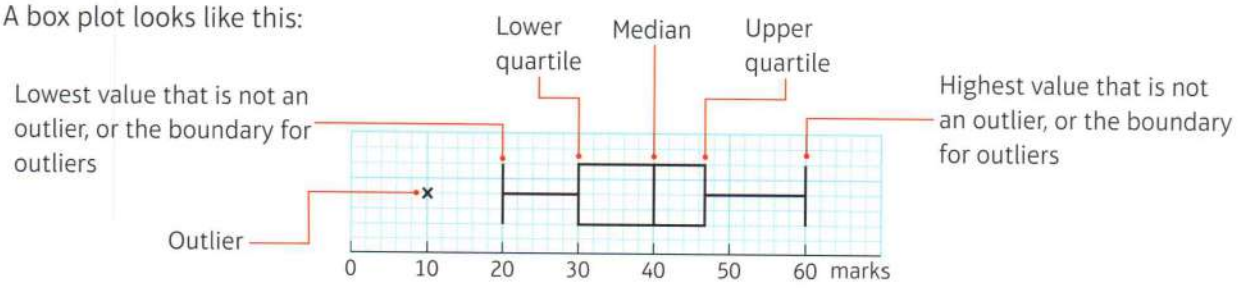
Problem-solving

After you clean the data you will need to find the new values for n, Σx and Σx^2.

3.2 Box plots

A box plot can be drawn to represent important features of the data. It shows the quartiles, maximum and minimum values and any outliers.

A box plot looks like this:



Two sets of data can be compared using box plots.

Example 3

a Draw a box plot for the data on blood glucose levels of females from Example 1.

The blood glucose level of 30 males is recorded. The results, in mmol/litre, are summarised below:

Lower quartile	= 3.6
Upper quartile	= 4.7
Median	= 4.0
Lowest value	= 1.4
Highest value	= 5.2

An outlier is an observation that falls either 1.5 × interquartile range above the upper quartile or 1.5 × interquartile range below the lower quartile.

b Given that there is only one outlier for the males, draw a box plot for this data on the same diagram as the one for females.

c Compare the blood glucose levels for males and females.

a

The quartiles and outliers were found in Example 1. ← page 41

The outlier is marked with a cross. The lowest value which is not an outlier is 2.2.

Always use a scale and label it. Remember to give your box plot a title.

Online Explore box plots and outliers using technology.

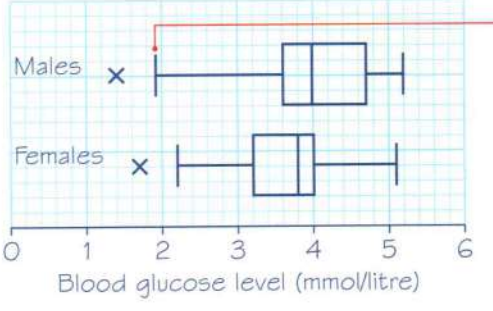
b Outliers are values less than

$3.6 - 1.5 \times 1.1 = 1.95$

and values greater than

$4.7 + 1.5 \times 1.1 = 6.35$

There is 1 outlier, which is 1.4.



The end of the whisker is plotted at the outlier boundary (in this case 1.95) as we do not know the actual figure.

Problem-solving

When drawing two box plots, use the same scale so they can be compared. Remember to give each a title and label the axis.

c The median blood glucose for females is lower than the median blood glucose for males. The interquartile range (the width of the box) and range for blood glucose are smaller for the females.

When comparing data you should compare a measure of location and a measure of spread. You should also write your interpretation in the context of the question.

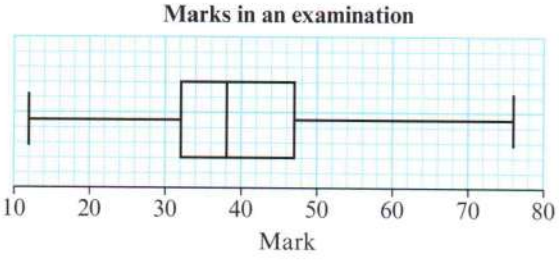
Exercise 3B

1 A group of students did a test. The summary data is shown in the table.

Lowest value	Lower quartile	Median	Upper quartile	Highest value
5	21	28	36	58

Given that there were no outliers, draw a box plot to illustrate this data.

2 Here is a box plot of marks in an examination.



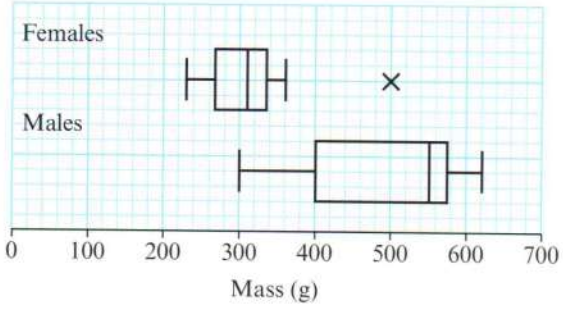
a Write down the upper and lower quartiles.

b Write down the median.

c Work out the interquartile range.

d Work out the range.

(P) **3** The masses of male and female turtles are given in grams. Their masses are summarised in the box plots.



a Compare and contrast the masses of the male and female turtles.

b A turtle was found to have a mass of 330 grams. State whether it is likely to be a male or a female. Give a reason for your answer.

c Write down the size of the largest female turtle.

(E) **4** Data for the maximum daily gust (in knots) in Camborne in September 1987 is taken from the large data set:

13	17	19	20	21
21	22	23	24	25
25	25	26	26	26
27	29	30	30	30
33	35	38	46	78

a Calculate Q_1, Q_2 and Q_3. **(3 marks)**

An outlier is defined as a value which lies either 1.5 × the interquartile range above the upper quartile or 1.5 × the interquartile range below the lower quartile.

b Show that 46 and 78 are outliers. **(1 mark)**

c Draw a box plot for this data. **(3 marks)**

3.3 Cumulative frequency

If you are given data in a grouped frequency table, you are not able to find the exact values of the median and quartiles. You can draw a **cumulative frequency diagram** and use it to help find estimates for the median, quartiles and percentiles.

Example 4

The table shows the heights, in metres, of 80 giraffes.

Height, h (m)	Frequency
$4.6 \leqslant h < 4.8$	4
$4.8 \leqslant h < 5.0$	7
$5.0 \leqslant h < 5.2$	15
$5.2 \leqslant h < 5.4$	33
$5.4 \leqslant h < 5.6$	17
$5.6 \leqslant h < 5.8$	4

a Draw a cumulative frequency diagram.

b Estimate the median height of the giraffes.

c Estimate the lower quartile and the 90th percentile.

d Draw a box plot to represent this data.

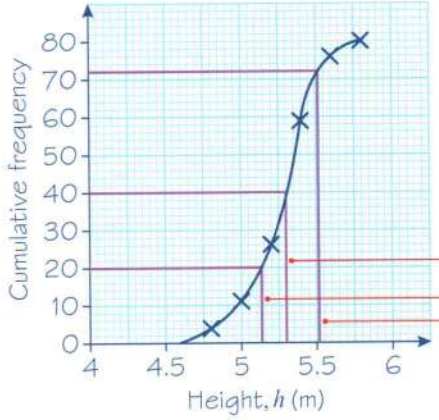
a Add a column to the table to show the cumulative frequency:

Height, h (m)	Frequency	Cumulative frequency
$4.6 \leqslant h < 4.8$	4	4
$4.8 \leqslant h < 5.0$	7	11
$5.0 \leqslant h < 5.2$	15	26
$5.2 \leqslant h < 5.4$	33	59
$5.4 \leqslant h < 5.6$	17	76
$5.6 \leqslant h < 5.8$	4	80

4 + 7 = 11

11 + 15 = 26

This represents the number of data values that are in the range $4.6 \leqslant h < 5.4$.



The lowest possible value for the height is 4.6 m so plot (4.6, 0).

Plot each point using the upper class boundary for x and the cumulative frequency for y: coordinates (4.8, 4), (5.0, 11), (5.2, 26), (5.4, 59), (5.6, 76) and (5.8, 80). Join the points with a smooth curve.

b The median is the 40th data point.
An estimate for the median is 5.3 m.

For part **b**, draw a line across from 40 on the cumulative frequency axis and then down to the height axis.

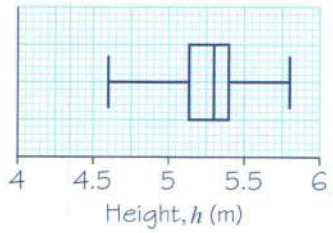
c The lower quartile is the 20th data point.
The 90th percentile is the 72nd data point.
An estimate for the lower quartile is 5.15 m.
An estimate for the 90th percentile is 5.52 m.

For part **c**, draw lines to estimate the lower quartile and the 90th percentile.

The data is continuous so you do not add 1.

← Section 2.2

d Height of giraffes



The minimum possible data value is 4.6 and the maximum possible data value is 5.8. The median and lower quartile are taken from parts **b** and **c**. The upper quartile is at 5.4.

Exercise 3C

1 The table shows the masses, in kilograms, of 120 Coulter pine cones.

Mass, m (kg)	Frequency
$1.0 \leqslant m < 1.2$	7
$1.2 \leqslant m < 1.4$	18
$1.4 \leqslant m < 1.6$	34
$1.6 \leqslant m < 1.8$	41
$1.8 \leqslant m < 2.0$	15
$2.0 \leqslant m < 2.2$	5

a Draw a cumulative frequency diagram for this data.

b Estimate the median mass.

c Find the interquartile range and the 10th to 90th interpercentile range.

d Draw a box plot to show this data.

2 The table shows the lengths, in cm, of 70 earthworms.

Length, l (cm)	Frequency
$6.0 \leqslant l < 6.5$	3
$6.5 \leqslant l < 7.0$	13
$7.0 \leqslant l < 7.5$	14
$7.5 \leqslant l < 8.0$	26
$8.0 \leqslant l < 8.5$	10
$8.5 \leqslant l < 9.0$	4

a Draw a cumulative frequency diagram for this data.

b Estimate the median and quartiles.

c Estimate how many earthworms are
i longer than 8.2 cm **ii** shorter than 7.3 cm.

d Draw a box plot to show this data.

(P) **3** The table shows the times taken by 80 men and 80 women to complete a crossword puzzle.

Time, t (min)	Frequency (men)	Frequency (women)
$5 \leqslant t < 6$	2	3
$6 \leqslant t < 7$	14	15
$7 \leqslant t < 8$	17	21
$8 \leqslant t < 9$	40	35
$9 \leqslant t < 10$	7	6

a Draw cumulative frequency diagrams for both sets of data on the same axes.

b Which gender had the lower median time?

c Which gender had the bigger spread of times?

d The qualifying time for the next round of a national competition is 7.5 minutes. Estimate the numbers of men and women who qualified for the competition.

Problem-solving

To compare spread you could use the interquartile range or the 10th to 90th percentile range.

4 A vet measures the masses, in kg, of male and female domestic shorthair cats. Her results are given in the table.

Mass, w (kg)	Frequency (male)	Frequency (female)
$2.5 \leqslant w < 3.0$	1	5
$3.0 \leqslant w < 3.5$	12	17
$3.5 \leqslant w < 4.0$	20	32
$4.0 \leqslant w < 4.5$	27	12
$4.5 \leqslant w < 5.0$	7	4
$5.0 \leqslant w < 5.5$	3	0

a Draw cumulative frequency diagrams for both sets of data on the same axes.

b Which gender has the greater spread of masses?

A female domestic shorthair cat is considered underweight if its mass is below 3.2 kg.
A male domestic shorthair cat is considered underweight if its mass is below 3.8 kg.

c Which gender has fewer underweight cats?

3.4 Histograms

Grouped continuous data can be represented in a **histogram**.

Generally, a histogram gives a good picture of how the data is distributed. It enables you to see a rough location, the general shape and how spread out the data is.

In a histogram, the **area** of the bar is proportional to the frequency in each class. This allows you to use a histogram to represent grouped data with unequal class intervals.

- **On a histogram, to calculate the height of each bar (the frequency density) use the formula**

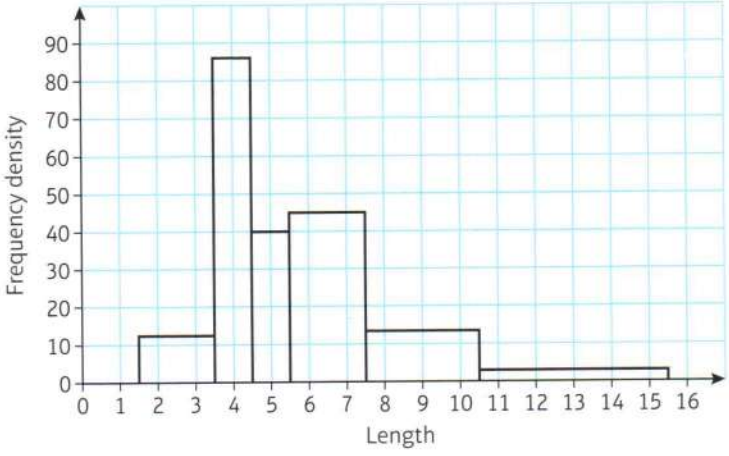
  area of bar = k × frequency.

  $k = 1$ is the easiest value to use when drawing a histogram.

  If $k = 1$, then

  $$\text{frequency density} = \frac{\text{frequency}}{\text{class width}}$$

- **Joining the middle of the top of each bar in a histogram with equal class widths forms a frequency polygon.**



Example 5

A random sample of 200 students was asked how long it took them to complete their homework the previous night. The time was recorded and summarised in the table below.

Time, t (minutes)	$25 \leqslant t < 30$	$30 \leqslant t < 35$	$35 \leqslant t < 40$	$40 \leqslant t < 50$	$50 \leqslant t < 80$
Frequency	55	39	68	32	6

a Draw a histogram to represent the data.

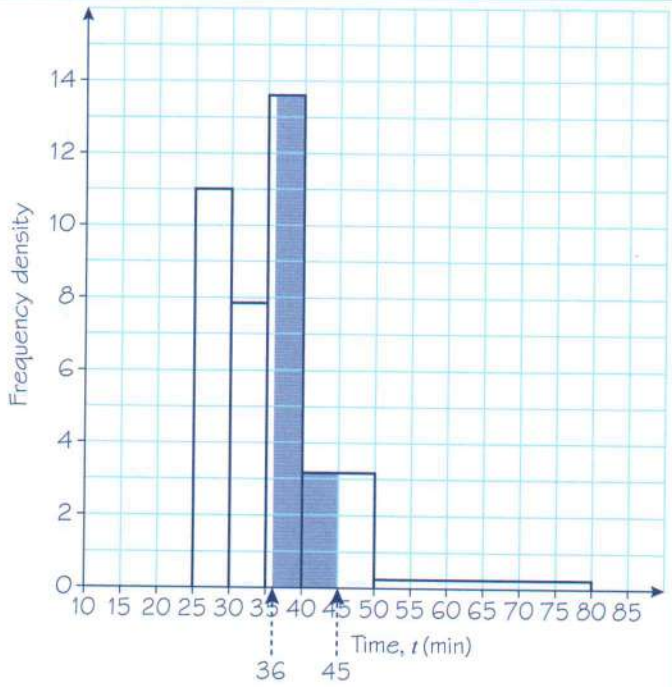
b Estimate how many students took between 36 and 45 minutes to complete their homework.

a

Time, t (minutes)	Frequency	Class width	Frequency density
$25 \leqslant t < 30$	55	5	11
$30 \leqslant t < 35$	39	5	7.8
$35 \leqslant t < 40$	68	5	13.6
$40 \leqslant t < 50$	32	10	3.2
$50 \leqslant t < 80$	6	30	0.2

Frequency density $= \frac{55}{5} = 11$

Class width $= 30 - 25 = 5$



b Shaded area $= (40 - 36) \times 13.6 + (45 - 40) \times 3.2$
$= 70.4$ students

To estimate the number of students who spent between 36 and 45 minutes, you need to find the area between 36 and 45.

Example 6

A random sample of daily mean temperatures (T, °C) was taken from the large data set for Hurn in 2015. The temperatures were summarised in a grouped frequency table and represented by a histogram.

a Give a reason to support the use of a histogram to represent this data.

b Write down the underlying feature associated with each of the bars in a histogram.

On the histogram the rectangle representing the $16 \leqslant T < 18$ class was 3.2 cm high and 2 cm wide. The frequency for this class was 8.

c Show that each day is represented by an area of $0.8\,\text{cm}^2$.

d Given that the total area under the histogram was $48\,\text{cm}^2$, find the total number of days in the sample.

a Temperature is continuous and the data were given in a grouped frequency table.

b The area of the bar is proportional to the frequency.

c Area of bar = $3.2 \times 2 = 6.4$
$6.4 \div 8 = 0.8\,\text{cm}^2$

An area of $6.4\,\text{cm}^2$ represents a frequency of 8.

d There were 60 days in the sample.

Total number of days × area for one day = total area under histogram

Exercise 3D

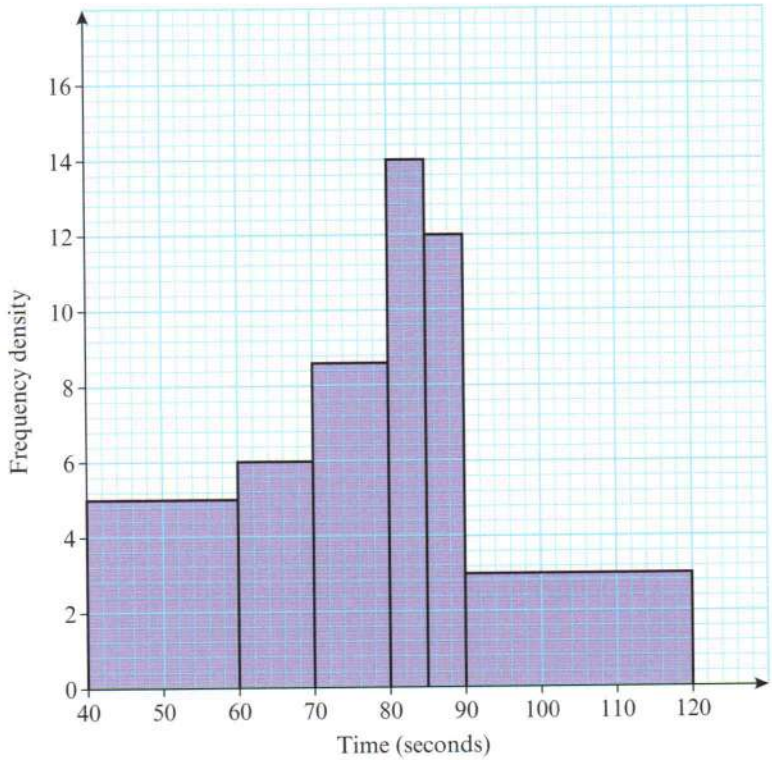
1 The data shows the mass, in pounds, of 50 adult puffer fish.

Mass, m (pounds)	Frequency
$10 \leqslant m < 15$	4
$15 \leqslant m < 20$	12
$20 \leqslant m < 25$	23
$25 \leqslant m < 30$	8
$30 \leqslant m < 35$	3

a Draw a histogram for this data.

b On the same set of axes, draw a frequency polygon.

(P) **2** Some students take part in an obstacle race. The time it took each student to complete the race was noted. The results are shown in the histogram.



a Give a reason to justify the use of a histogram to represent this data.

The number of students who took between 60 and 70 seconds is 90.

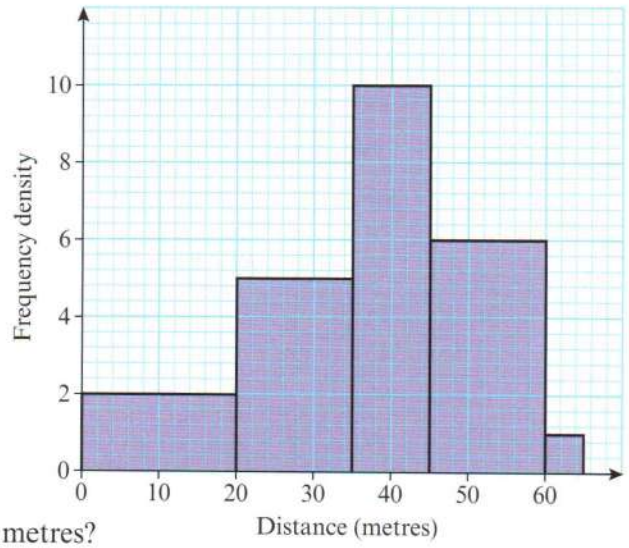
b Find the number of students who took between 40 and 60 seconds.

c Find the number of students who took 80 seconds or less.

d Calculate the total number of students who took part in the race.

Watch out Frequency density × class width is always **proportional** to frequency in a histogram, but not necessarily **equal** to frequency.

(P) **3** A Fun Day committee at a local sports centre organised a throwing the cricket ball competition. The distance thrown by every competitor was recorded. The histogram shows the data. The number of competitors who threw less than 20 m was 40.



a Why is a histogram a suitable diagram to represent this data?

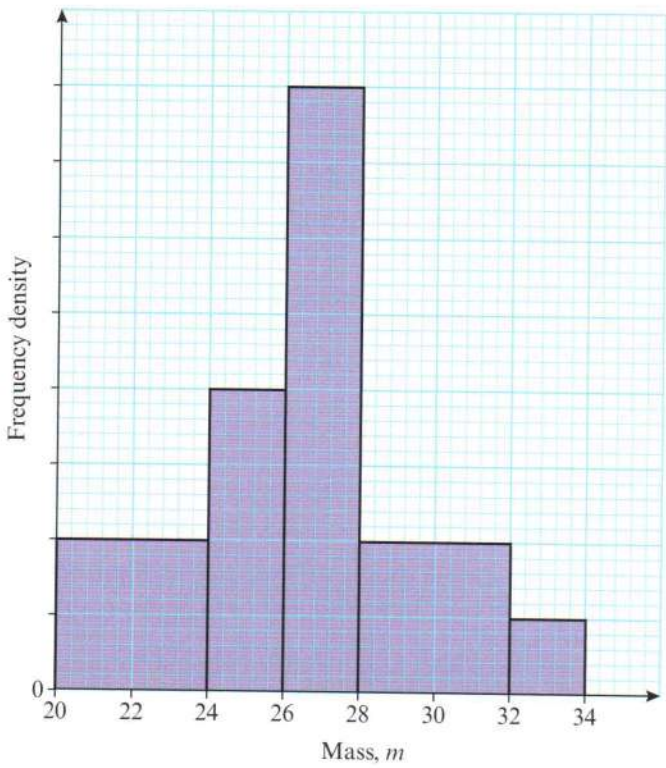
b How many people entered the competition?

c Estimate how many people threw between 30 and 40 metres.

d How many people threw between 45 and 65 metres?

e Estimate how many people threw less than 25 metres.

(P) **4** A farmer found the masses of a random sample of lambs. The masses were summarised in a grouped frequency table and represented by a histogram. The frequency for the class $28 \leqslant m < 32$ was 32.



a Show that 25 small squares on the histogram represents 8 lambs.

b Find the frequency of the $24 \leqslant m < 26$ class.

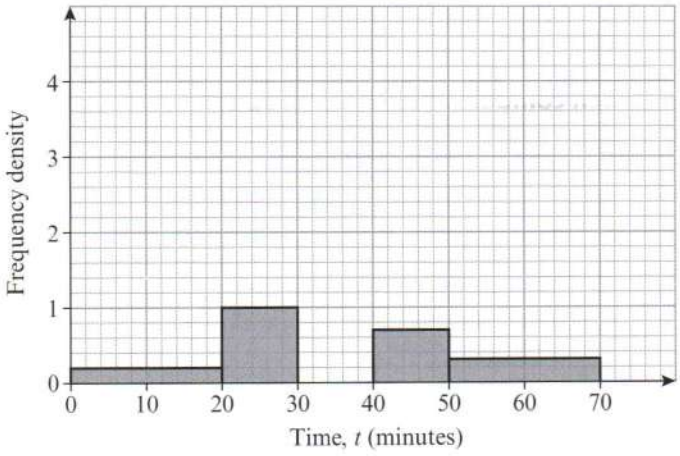
c How many lambs did the farmer weigh in total?

d Estimate the number of lambs that had masses between 25 and 29 kg.

Problem-solving

You can use area to solve histogram problems where no vertical scale is given. You could also use the information given in the question to work out a suitable scale for the vertical axis.

E/P **5** The partially completed histogram shows the time, in minutes, that passengers were delayed at an airport.



a i Copy and complete the table.

Time, t (min)	Frequency
$0 \leqslant t < 20$	4
$20 \leqslant t < 30$	
$30 \leqslant t < 35$	15
$35 \leqslant t < 40$	25
$40 \leqslant t < 50$	
$50 \leqslant t < 70$	

ii Copy and complete the histogram. **(4 marks)**

b Estimate the number of passengers that were delayed for between 25 and 38 minutes. **(2 marks)**

E/P **6** The variable y was measured to the nearest whole number. 60 observations were taken and are recorded in the table below.

y	10–12	13–14	15–17	18–25
Frequency	6	24	18	12

a Write down the class boundaries for the 13–14 class. **(1 mark)**

A histogram was drawn and the bar representing the 13–14 class had a width of 4 cm and a height of 6 cm.

For the bar representing the 15–17 class, find:

b i the width **(1 mark)**

ii the height. **(2 marks)**

Problem-solving

Remember that area is proportional to frequency.

E/P **7** From the large data set, the daily mean temperature for Leeming during May 2015 is summarised in the table.

A histogram was drawn. The $8 \leqslant t < 10$ group was represented by a bar of width 1 cm and a height of 8 cm.

Daily mean temperature, t (°C)	Frequency
$4 \leqslant t < 8$	4
$8 \leqslant t < 10$	8
$10 \leqslant t < 11$	6
$11 \leqslant t < 12$	7
$12 \leqslant t < 15$	5
$15 \leqslant t < 16$	1

a Find the width and height of the bar representing the $11 \leqslant t < 12$ group. **(2 marks)**

b Use your calculator to estimate the mean and standard deviation of temperatures in Leeming in May 2015. **(3 marks)**

c Use linear interpolation to find an estimate for the lower quartile of temperatures. **(2 marks)**

d Estimate the number of days in May 2015 on which the temperature was higher than the mean plus one standard deviation. **(2 marks)**

3.5 Comparing data

- **When comparing data sets you can comment on:**
  - **a measure of location**
  - **a measure of spread**

You can compare data using the mean and standard deviation or using the median and interquartile range.
If the data set contains extreme values, then the median and interquartile range are more appropriate statistics to use.

Watch out Do not use the median with the standard deviation or the mean with the interquartile range.

Example 7

From the large data set, the daily mean temperature during August 2015 is recorded at Heathrow and Leeming.

For Heathrow, $\Sigma x = 562.0$ and $\Sigma x^2 = 10\,301.2$.

a Calculate the mean and standard deviation for Heathrow.

For Leeming, the mean temperature was 15.6 °C with a standard deviation of 2.01 °C.

b Compare the data for the two locations using the information given.

a Mean $= 562.0 \div 31 = 18.12\ldots = 18.1$ °C

Standard deviation

$$= \sqrt{\frac{10\,301.2}{31} - \left(\frac{562.0}{31}\right)^2} = 1.906\ldots$$

$= 1.91$ °C (3 s.f.)

b The mean daily temperature in Leeming is lower than in Heathrow and the spread of temperatures is greater than in Heathrow.

Use $\bar{x} = \frac{\Sigma x}{n}$. There are 31 days in August so $n = 31$.

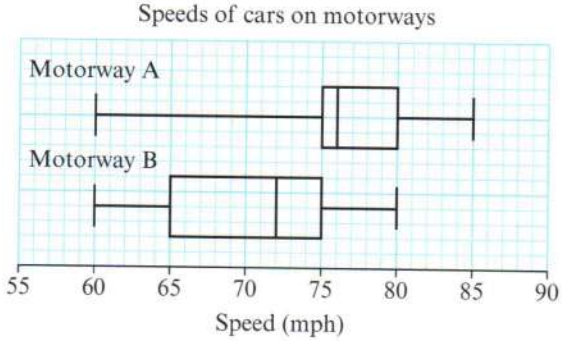
Use your calculator to do this calculation in one step. Round your final answer to 3 significant figures.

Compare the mean and standard deviation as a measure of location and a measure of spread.

Exercise 3E

P **1** The box plots show the distribution of speeds of cars on two motorways.

Compare the distributions of the speeds on the two motorways.



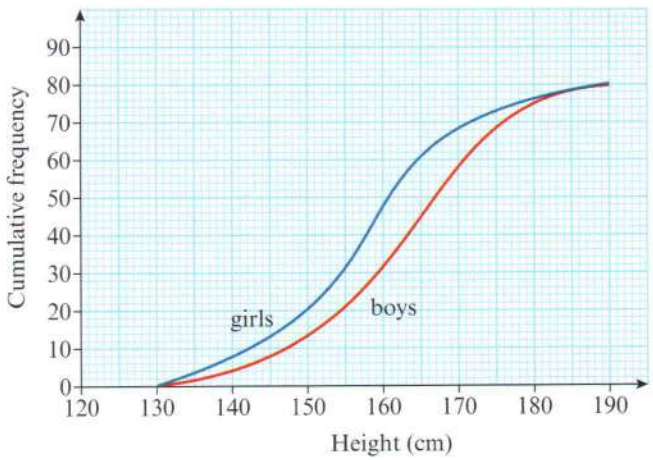
P **2** Two classes of primary school children complete a puzzle. Summary statistics for the times, in minutes, the children took are shown in the table.

Calculate the mean and standard deviation of the times and compare the distributions.

	n	Σx	Σx^2
Class 2B	20	650	22 000
Class 2F	22	598	19 100

(P) **3** The cumulative frequency diagram shows the distribution of heights of 80 boys and 80 girls in a basketball club.

Compare the heights of boys and girls in the club.



(E) **4** A sample of the daily maximum relative humidity is taken from the large data set for Leuchars and for Camborne during 2015. The data is given in the table.

Leuchars	100	98	100	100	100	100	100	100	94	100	91	100	100	89	100
Camborne	92	95	99	96	100	100	90	98	81	99	100	99	91	98	100

a Find the median and quartiles for both samples. **(4 marks)**

b Compare the two samples. **(2 marks)**

Large data set

You will need access to the large data set and spreadsheet software to answer this question. Look at the data on daily mean windspeed in Hurn in 1987 and in 2015.

1 For each year, calculate:

a the mean **b** the mode **c** the standard deviation.

2 Compare the daily mean windspeeds in Hurn in 1987 and 2015.

Hint You can use the MODE function in a spreadsheet to find the modal value from a range of cells.

Mixed exercise 3

1 Jason and Perdita decided to go on a touring holiday on the continent for the whole of July. They recorded the distance they travelled, in kilometres, each day:

155, 164, 168, 169, 173, 175, 177, 178, 178, 178, 179, 179, 179, 184, 184, 185, 185, 188, 192, 193, 194, 195, 195, 196, 204, 207, 208, 209, 211, 212, 226

a Find Q_1, Q_2 and Q_3

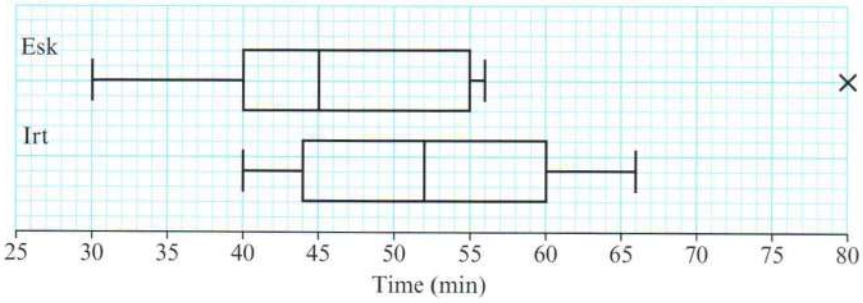
Outliers are values that lie outside $Q_1 - 1.5(Q_3 - Q_1)$ and $Q_3 + 1.5(Q_3 - Q_1)$.

b Find any outliers.

c Draw a box plot of this data.

(P) **2** Fell runners from the Esk Club and the Irt Club were keen to see which club had the faster runners overall. They decided that all the members from both clubs would take part in a fell run. The time each runner took to complete the run was recorded.

The results are summarised in the box plot.



a Write down the time by which 50% of the Esk Club runners had completed the run.

b Write down the time by which 75% of the Irt Club runners had completed the run.

c Explain what is meant by the cross (×) on the Esk Club box plot.

d Compare and contrast these two box plots.

e What conclusions can you draw from this information about which club has the faster runners?

f Give one advantage and one disadvantage of comparing distributions using box plots.

(P) **3** The table shows the lengths, in cm, of 60 honey badgers.

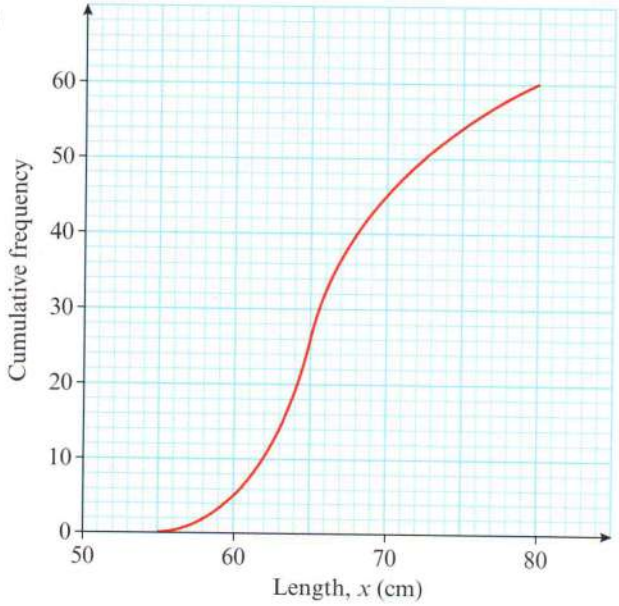
Length, x (cm)	Frequency
$50 \leqslant x < 55$	2
$55 \leqslant x < 60$	7
$60 \leqslant x < 65$	15
$65 \leqslant x < 70$	31
$70 \leqslant x < 75$	5

a Draw a cumulative frequency diagram for this data.

b Find the median length of a honey badger.

c Find the interquartile range.

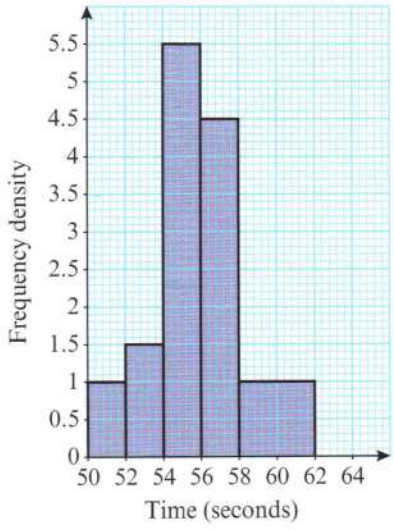
This diagram shows the distribution of lengths of European badgers.



d Compare the distributions of lengths of honey badgers and European badgers.

e Comment on the suitability of using cumulative frequency diagrams to compare these distributions.

(P) **4** The histogram shows the time taken by a group of 58 girls to run a measured distance.



a Work out the number of girls who took longer than 56 seconds.

b Estimate the number of girls who took between 52 and 55 seconds.

(E/P) **5** The table gives the distances travelled to school, in km, of the population of children in a particular region of the United Kingdom.

Distance, *d* (km)	$0 \leqslant d < 1$	$1 \leqslant d < 2$	$2 \leqslant d < 3$	$3 \leqslant d < 5$	$5 \leqslant d < 10$	$10 \leqslant d$
Number	2565	1784	1170	756	630	135

A histogram of this data was drawn with distance along the horizontal axis. A bar of horizontal width 1.5 cm and height 5.7 cm represented the 0–1 km group.

Find the widths and heights, in cm, to one decimal place, of the bars representing the following groups:

a $2 \leqslant d < 3$ **b** $5 \leqslant d < 10$ **(5 marks)**

6 The labelling on bags of garden compost indicates that the bags have a mass of 20 kg.
The masses of a random sample of 50 bags are summarised in the table opposite.

Mass, *m* (kg)	**Frequency**
$14.6 \leqslant m < 14.8$	1
$14.8 \leqslant m < 18.0$	0
$18.0 \leqslant m < 18.5$	5
$18.5 \leqslant m < 20.0$	6
$20.0 \leqslant m < 20.2$	22
$20.2 \leqslant m < 20.4$	15
$20.4 \leqslant m < 21.0$	1

a On graph paper, draw a histogram of this data.

b Estimate the mean and standard deviation of the mass of a bag of compost.

[You may use $\Sigma fy = 988.85$, $\Sigma fy^2 = 19\,602.84$]

c Using linear interpolation, estimate the median.

7 The number of bags of potato crisps sold per day in a bar was recorded over a two-week period. The results are shown below.

20 15 10 30 33 40 5 11 13 20 25 42 31 17

a Calculate the mean of this data.

b Find the median and the quartiles of this data.

An outlier is an observation that falls either 1.5 × (interquartile range) above the upper quartile or 1.5 × (interquartile range) below the lower quartile.

c Determine whether or not any items of data are outliers.

d On graph paper draw a box plot to represent this data. Show your scale clearly.

(E) **8** The daily maximum gust is measured at Hurn for a period of 57 days. The data is summarised in the table.

Daily maximum gust, g (knots)	Frequency
$10 \leqslant g < 15$	3
$15 \leqslant g < 18$	9
$18 \leqslant g < 20$	9
$20 \leqslant g < 25$	20
$25 \leqslant g < 30$	9
$30 \leqslant g < 50$	7

A histogram is drawn to represent this data. The bar representing the $10 \leqslant g < 15$ class is 2.5 cm wide and 1.8 cm high.

a Give a reason to support the use of a histogram to represent this data. **(1 mark)**

b Calculate the width and height of the bar representing the $18 \leqslant g < 20$ class. **(3 marks)**

c Use your calculator to estimate the mean and standard deviation of the maximum gusts. **(3 marks)**

d Use linear interpolation to find an estimate for the number of days the maximum gust was within one standard deviation of the mean. **(4 marks)**

(E) **9** From the large data set, data was gathered in September 1987 and in September 2015 for the mean daily temperature in Leuchars. Summary statistics are given in the table.

	Min	Max	Median	Σx	Σx^2
1987	7.0	17.0	11.85	356.1	4408.9
2015	10.1	14.1	12.0	364.1	4450.2

a Calculate the mean of the mean daily temperatures in each of the two years. **(2 marks)**

b In 2015, the standard deviation was 1.02. Compare the mean daily temperatures in the two years. **(2 marks)**

c A recorded temperature is considered 'normal' for the time of year if it is within one standard deviation of the mean. Estimate for how many days in September 2015 a 'normal' mean daily temperature was recorded. State one assumption you have made in making the estimate. **(3 marks)**

Challenge

The table shows the lengths of the films in a film festival, to the nearest minute.

Length (min)	Frequency
70–89	4
90–99	17
100–109	20
110–139	9
140–179	2

A histogram is drawn to represent the data, and the bar representing the 90–99 class is 3 cm higher than the bar representing the 70–89 class.

Find the height of the bar representing the 110–139 class.

Summary of key points

1 A common definition of an outlier is any value that is:
- either greater than $Q_3 + k(Q_3 - Q_1)$
- or less than $Q_1 - k(Q_3 - Q_1)$

2 The process of removing anomalies from a data set is known as cleaning the data.

3 On a histogram, to calculate the height of each bar (the **frequency density**) use the formula

area of bar = k × frequency.

4 Joining the middle of the top of each bar in a histogram with equal class widths forms a frequency polygon.

5 When comparing data sets you can comment on:
- a measure of location
- a measure of spread