KING EDWARD VI
HANDSWORTH GRAMMAR
SCHOOL FOR BOYS

KING EDWARD VI
ACADEMY TRUST
BIRMINGHAM

Year 12
Statistics 1
Chapter 3 – Representations of Data

HGS Maths

Dr Frost Course

Name: _____

Class: _____

Contents

3.1 Outliers

Notes

Outliers

Outliers are extreme values (either small or large) that can largely influence statistical analysis. They are shown as crosses on a box plot.

Outliers need to be investigated and a decision made as to whether to include them in the analysis. It is not acceptable to drop an observation just because it is an outlier. They can be legitimate observations and are sometimes the most interesting ones. It is important to investigate the nature of the outlier before deciding whether to exclude or include it.

The value may have been recorded incorrectly and it is obvious how to correct it. For example, it may have been recorded in metres instead centimetres.

The value may have been recorded incorrectly but it is possible to get the correct value. For example, the height of a famous sports person.

The value may be due to natural variation in the data.

If the outlier is obviously an error, then it should be excluded.

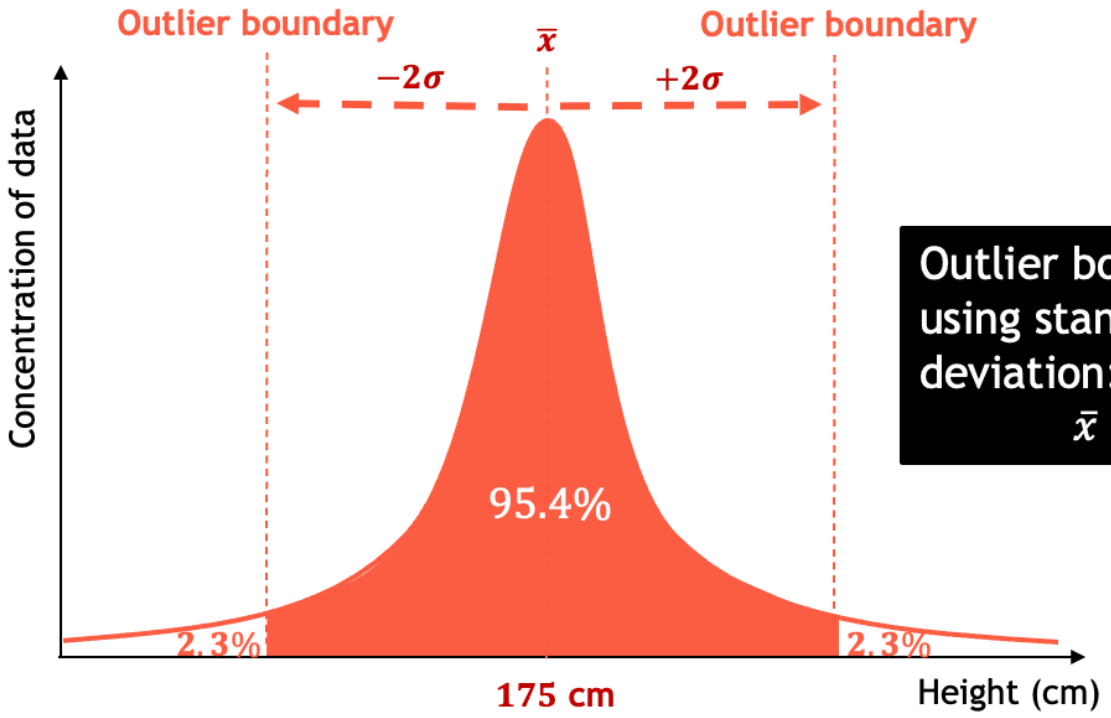
There are many tests to decide if data are outliers. Two examples are:

	Outlier lower limit Outlier if value <	Outlier upper limit Outlier if value >
Using quartiles and IQR	$Q_1 - 1.5 \times \text{IQR}$	$Q_3 + 1.5 \times \text{IQR}$
Using mean and standard deviation	Mean $- 3 \times$ standard deviation	Mean $+ 3 \times$ standard deviation

Outliers with Interquartile Range

Outlier boundaries, beyond which data values are considered outliers, are often defined as 1.5 interquartile ranges beyond the lower and upper quartiles.

Outliers with Standard Deviation



Outlier boundaries using standard deviation:

$$\bar{x} \pm 2\sigma$$

An alternative outlier boundary is **2 standard deviations above and below the mean**.

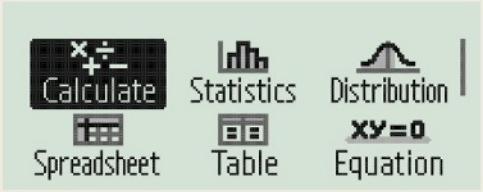
If the data was assumed to be normally distributed as above with a 'bell-curve' shape (which is assumed by default in statistics), then outliers would correspond to the top 2.3% and bottom 2.3% of values, which seems suitably outlier-ey.

Mean and Standard Deviation on a Calculator

Time (secs)	30	34	35	39

These are instructions for the **Casio fx-570/991CW**

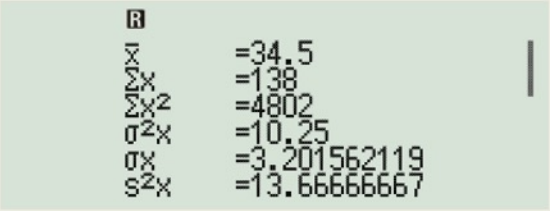
1 Use the arrows and OK to select Statistics.



2 Choose 1-Variable.

3 Enter your values, pressing = after each value, and an additional = after the last value to end your list.

4 Choose '1 Variable Results'. Read off σx (σ) and $\bar{x}$.



Worked Example	Your Turn
The scores of 10 students are recorded: $1, 8, 10, 9, -7, 21, 11, 10, 35, 0.3$ An outlier is an observation that falls either $1.5 \times$ interquartile range above the upper quartile or $1.5 \times$ interquartile range below the lower quartile. Find any outliers.	The scores of 10 students are recorded: $5, 12, 14, 13, 8, 9, 51, -4, 59, 0.2$ An outlier is an observation that falls either $1.5 \times$ interquartile range above the upper quartile or $1.5 \times$ interquartile range below the lower quartile. Find any outliers.

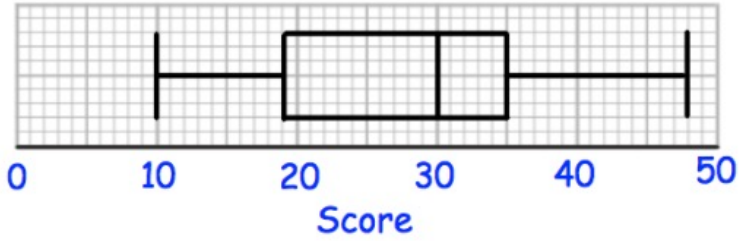
Worked Example	Your Turn
The scores of 10 students are recorded: $1, 8, 10, 9, -7, 21, 11, 10, 35, 0.3$ An outlier is an observation that falls outside ± 2 standard deviations from the mean. Find any outliers.	The scores of 10 students are recorded: $5, 12, 14, 13, 8, 9, 51, -4, 59, 0.2$ An outlier is an observation that falls outside ± 2 standard deviations from the mean. Find any outliers.

3.2 Box Plots

Notes

Worked Example

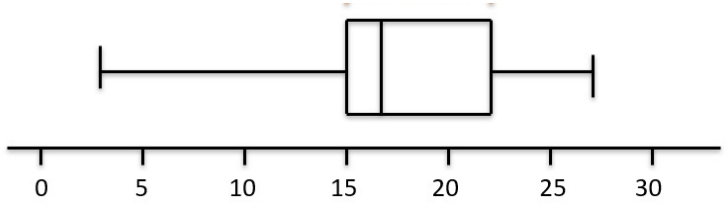
Using the box plot, write down:



a) The minimum
b) The lower quartile
c) The median
d) The upper quartile
e) The maximum
f) The range
g) The interquartile range

Your Turn

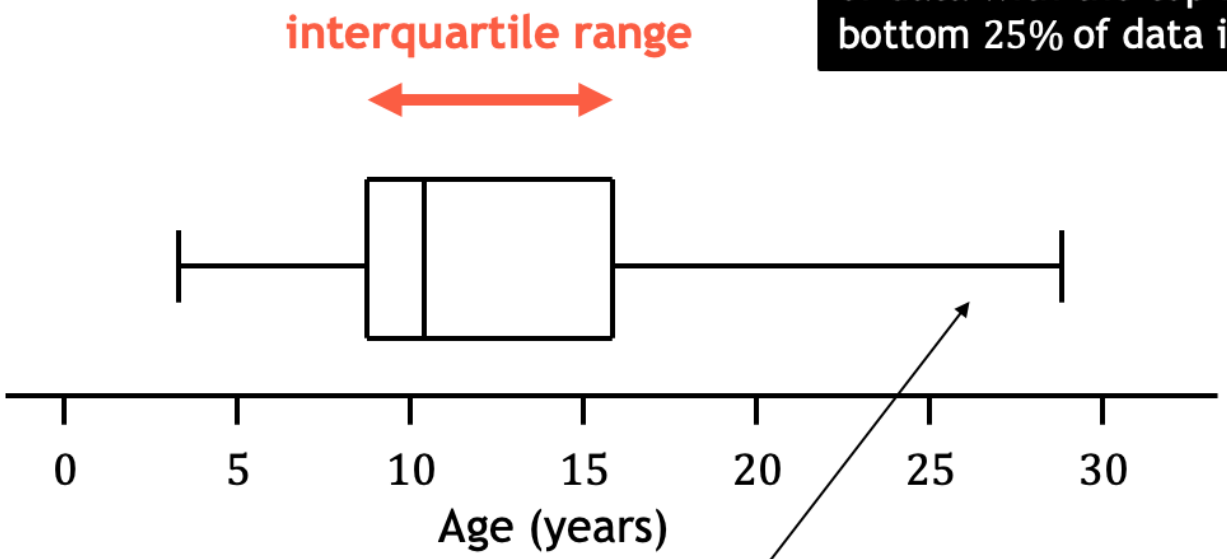
Using the box plot, write down:



a) The minimum
b) The lower quartile
c) The median
d) The upper quartile
e) The maximum
f) The range
g) The interquartile range

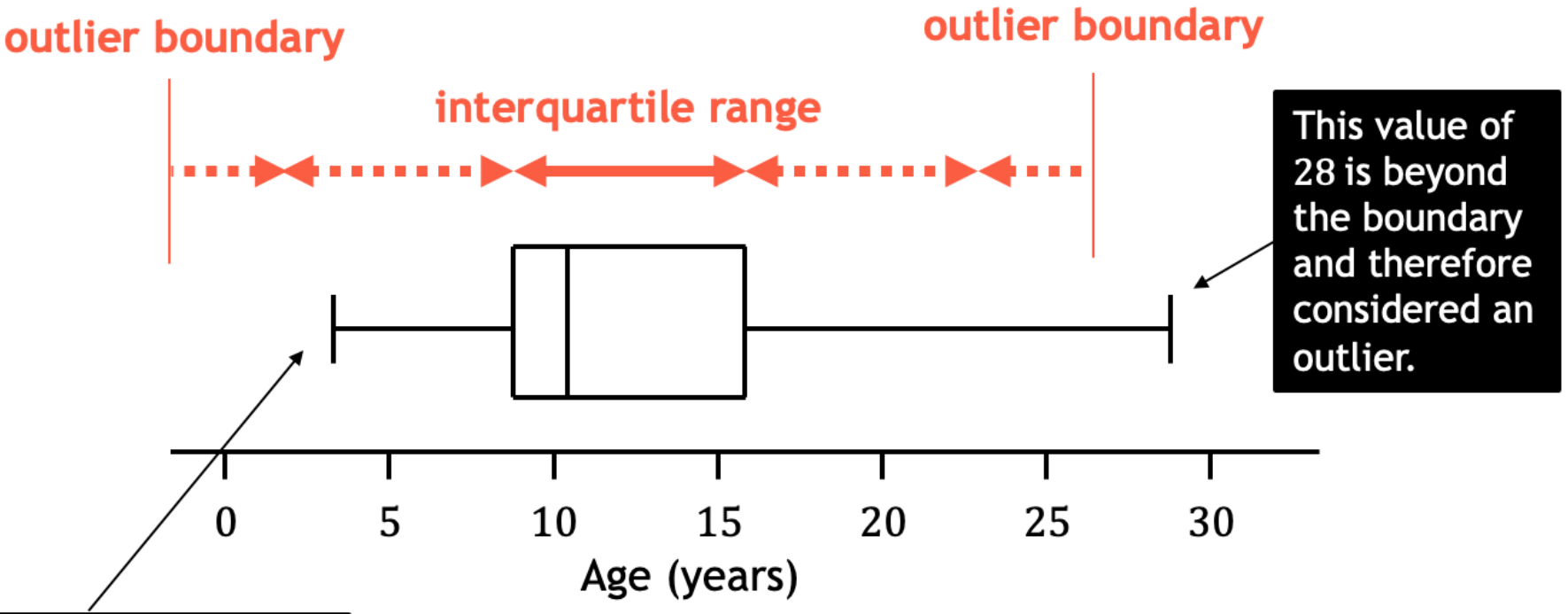
Introducing Outliers

The interquartile range is a way of determining the spread of data with the top 25% and bottom 25% of data ignored.



But while this 28-year-old might be considered an **outlier** (i.e. an **extreme value**), other values in the top 25% shouldn't be considered extreme, given that half the data, by definition, lies in the outside two quartiles!

Outliers with Interquartile Range



One way of defining the boundary at which values are considered outliers/extreme values, is <u>1.5 IQRs beyond the lower or upper quartiles*</u>.

* We will cover why the choice of '1.5' later.

Outliers on Box Plots

If there are outliers at one end of a box plot, use a × for each outlier, with the whisker extending to the maximum/minimum value that is not an outlier (or use the outlier boundary).

Worked Example

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

Sketch a box plot for this data, marking any outliers.

Smallest values	Largest values	Lower quartile	Median	Upper quartile
$0, 4$	$22, 26$	9	11	15

Your Turn

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

Sketch a box plot for this data, marking any outliers.

Smallest values	Largest values	Lower quartile	Median	Upper quartile
$0, 3$	$21, 27$	8	10	14

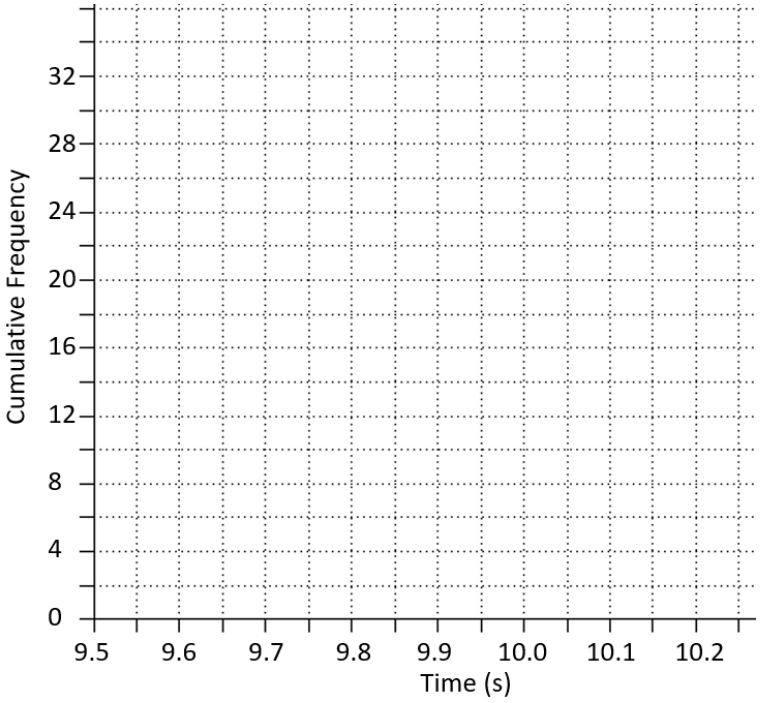
3.3 Cumulative Frequency

Notes

Worked Example

Draw a cumulative frequency diagram for the data:

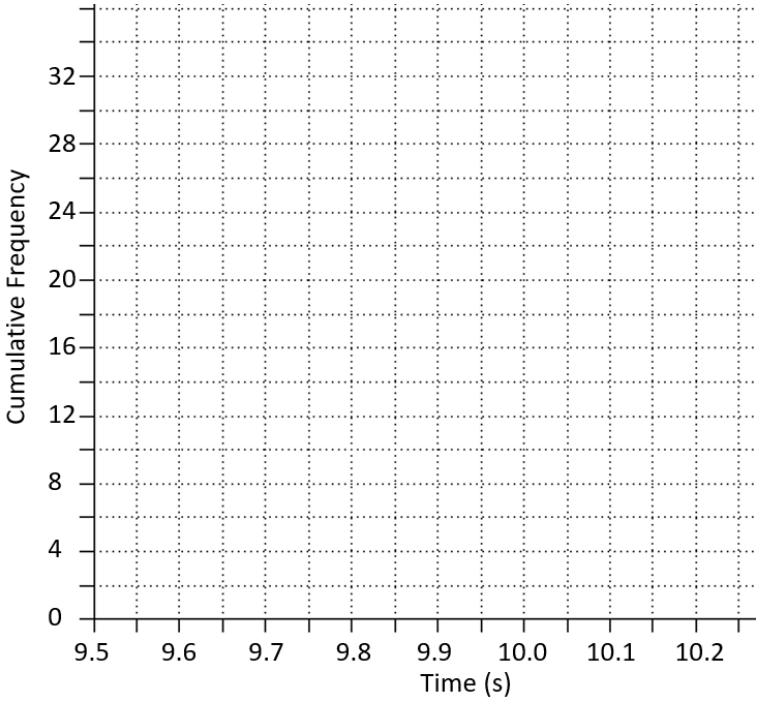
Time (s)	Frequency
$9.6 < t \leq 9.8$	3
$9.8 < t \leq 10.05$	7
$10.05 < t \leq 10.15$	8
$10.15 < t \leq 10.2$	14



Your Turn

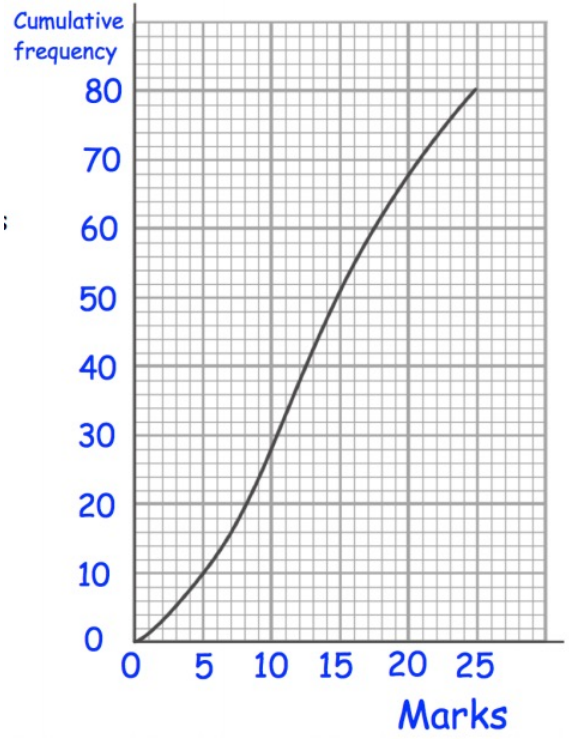
Draw a cumulative frequency diagram for the data:

Time (s)	Frequency
$9.6 < t \leq 9.7$	1
$9.7 < t \leq 9.9$	4
$9.9 < t \leq 10.05$	10
$10.05 < t \leq 10.2$	17



Worked Example

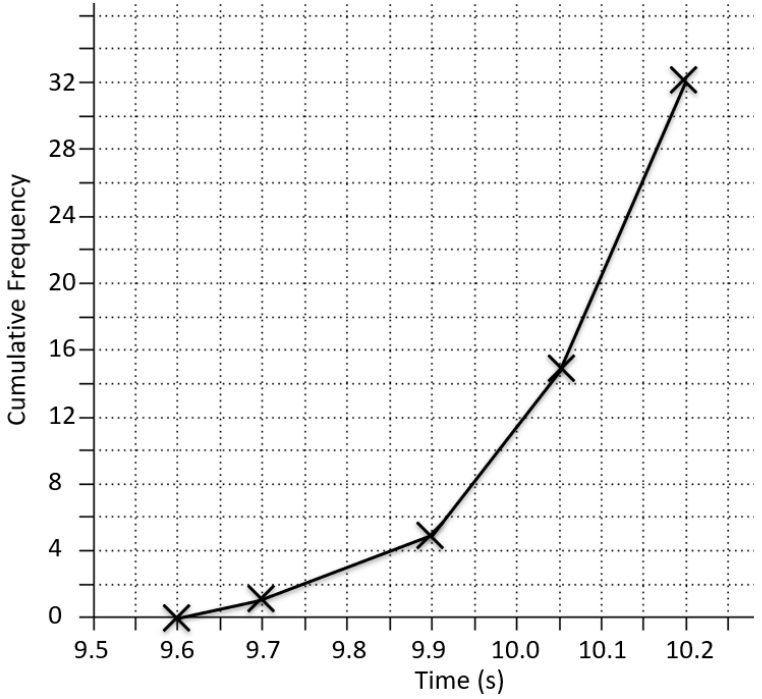
Use the cumulative frequency diagram to estimate the:



a) Lower quartile
b) Median
c) Upper quartile
d) 60th percentile
e) Interquartile range
f) 10th – 90th interpercentile range

Your Turn

Use the cumulative frequency diagram to estimate the:



a) Lower quartile
b) Median
c) Upper quartile
d) 90th percentile
e) Interquartile range
f) 20th – 80th interpercentile range

Dr Frost 398e and 398f

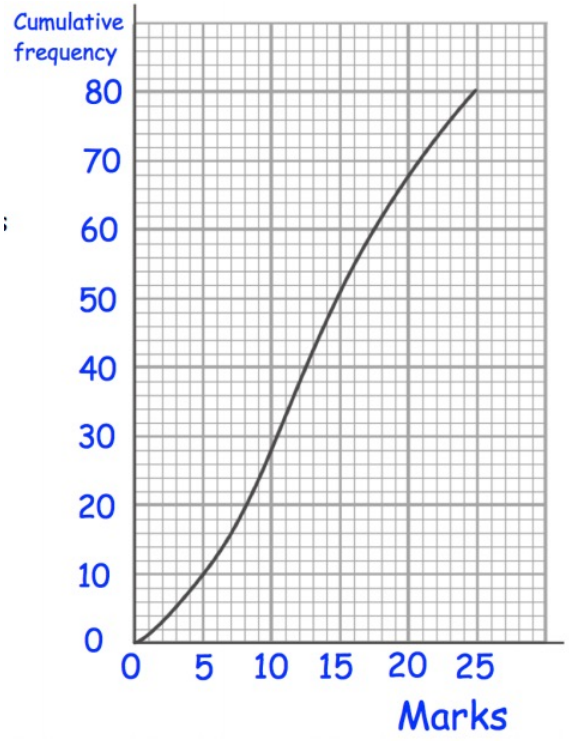
Worked Example

Use the cumulative frequency diagram to estimate the number of students who achieved:

a) Fewer than 23 marks.

b) More than 12 marks.

c) Between 7 and 21 marks.



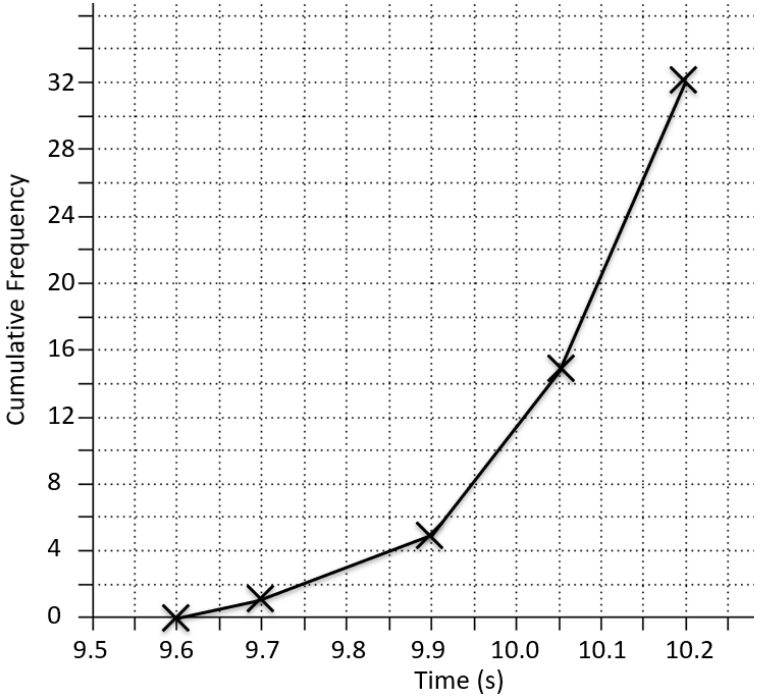
Your Turn

Use the cumulative frequency diagram to estimate the number of runners who had a time:

a) Less than 10.15 seconds.

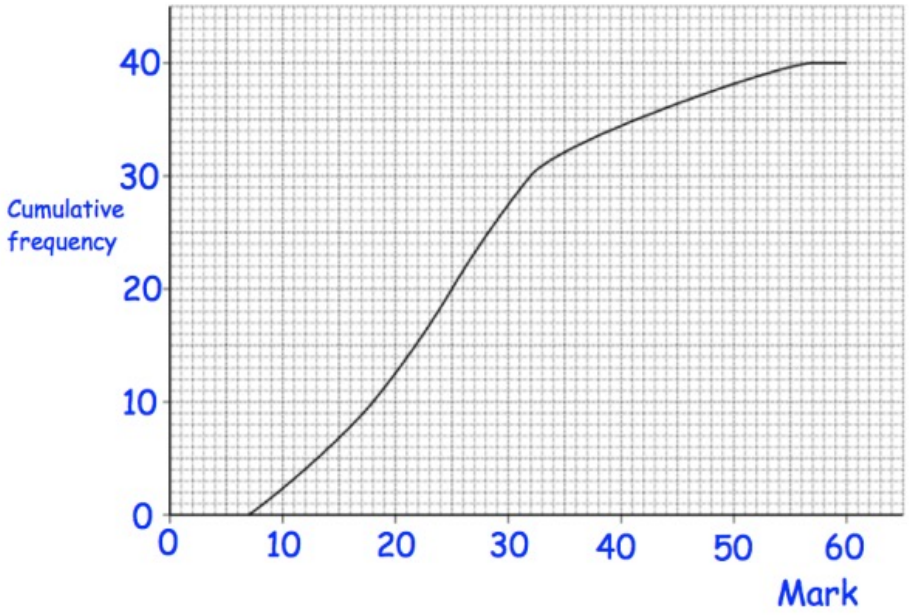
b) Greater than 9.95 seconds.

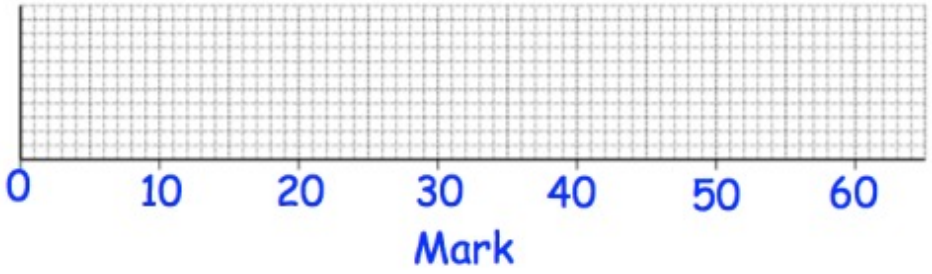
c) Between 9.8 and 10 seconds.



Worked Example

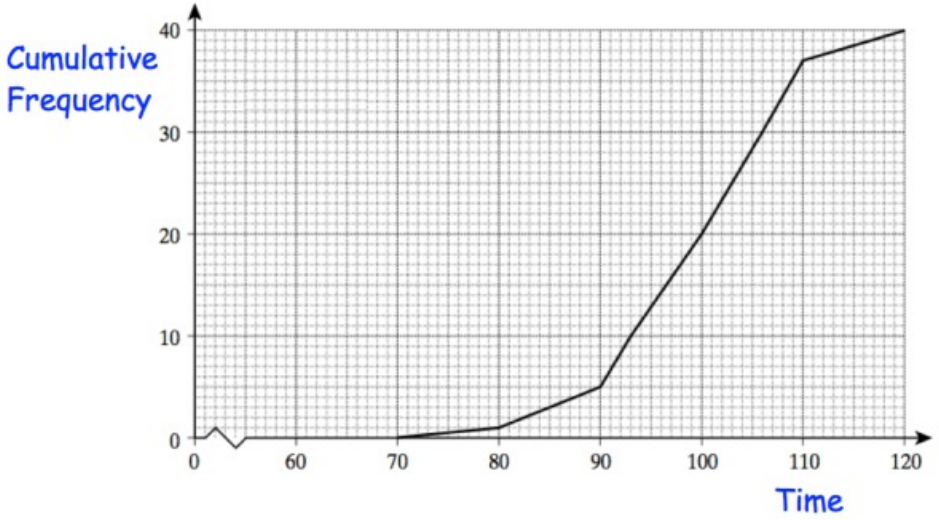
Use the cumulative frequency diagram to draw a box plot:

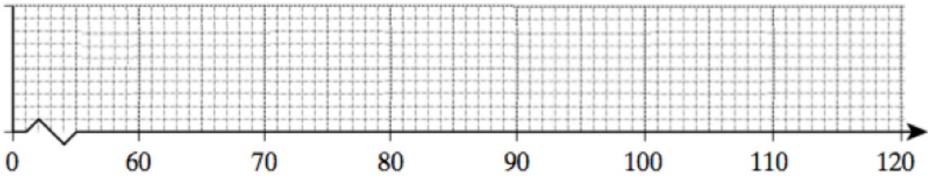




Your Turn

Use the cumulative frequency diagram to draw a box plot:





Dr Frost 399d

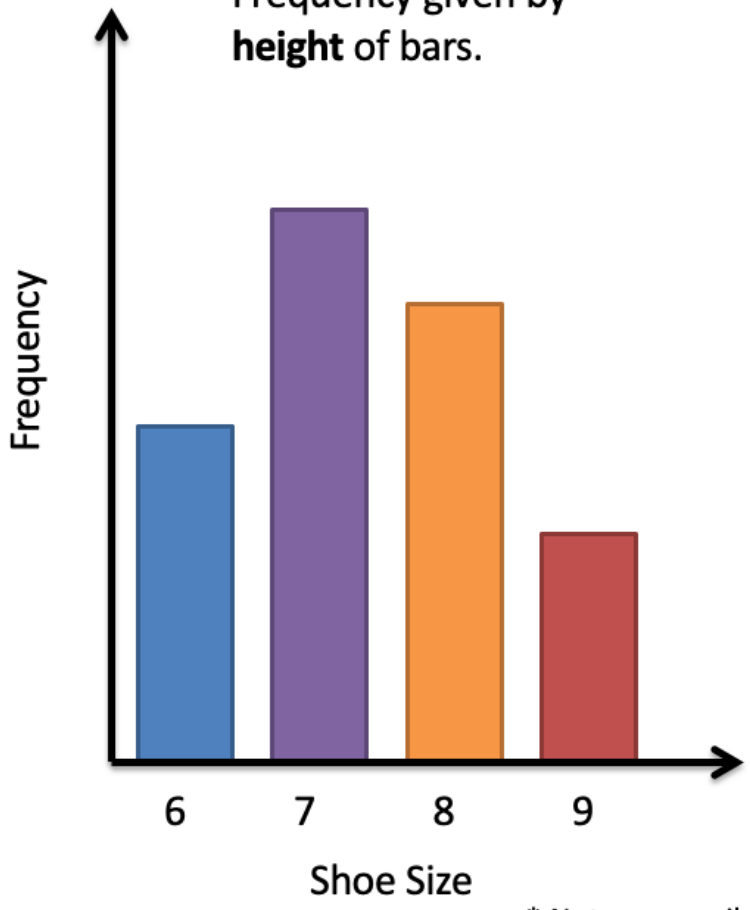
3.4 Histograms

Notes

Bar Charts vs Histograms

Bar Charts

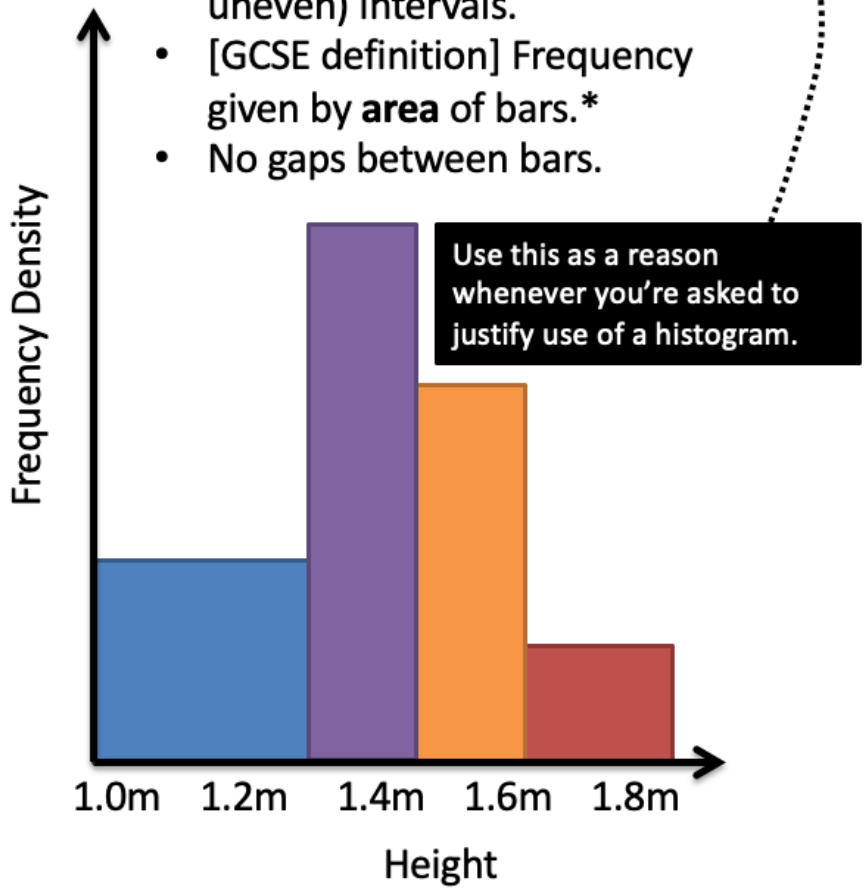
- For **discrete** data.
- Frequency given by **height** of bars.



Histograms

- **For continuous data.**
- Data divided into (potentially uneven) intervals.
- [GCSE definition] Frequency given by **area** of bars.*
- No gaps between bars.

Use this as a reason whenever you're asked to justify use of a histogram.



* Not necessarily true. We'll correct this in a sec.

Histograms

A histogram shows the distribution of a continuous variable. Therefore, there are no gaps between the bars. (However, there could be class intervals with zero frequency.)

In a histogram:

- frequency $\propto$ area of bar;
- frequency $\propto$ height of bar $\times$ width of bar;
- if the classes are not the same width the heights of the bars need to be adjusted;
- the x-axis is a continuous linear scale;
- each bar starts at the lower class boundary and ends at the upper class boundary.

Worked Example

Plot a histogram for the data:

Height, h (nearest cm)	Frequency
$1 - 4$	5
$5 - 7$	4
$8 - 9$	3

Your Turn

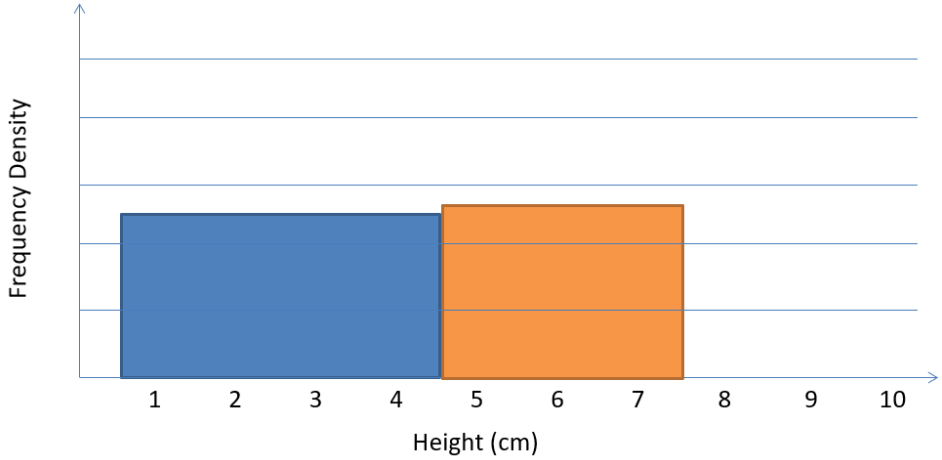
Plot a histogram for the data:

Weight, w (nearest kg)	Frequency
$1 - 2$	4
$3 - 6$	3
$7 - 9$	5

Worked Example

Complete the table and histogram:

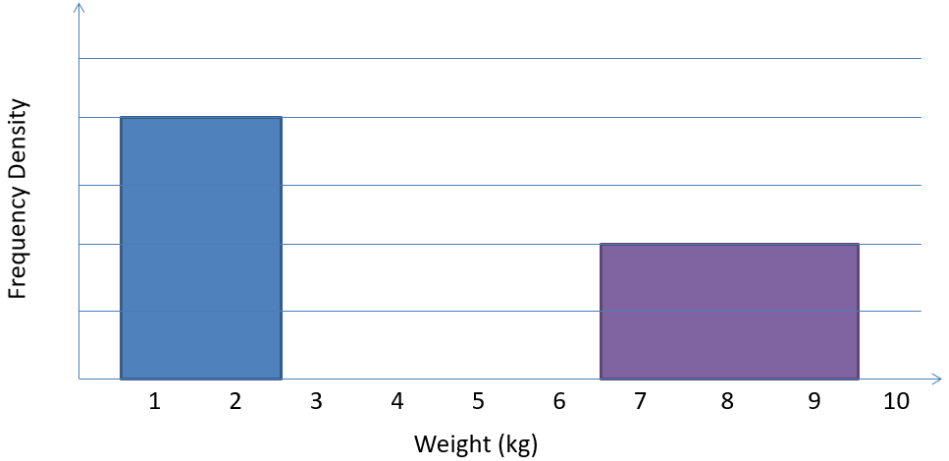
Height, h (nearest cm)	Frequency
$1 - 4$	
$5 - 7$	4
$8 - 9$	3



Your Turn

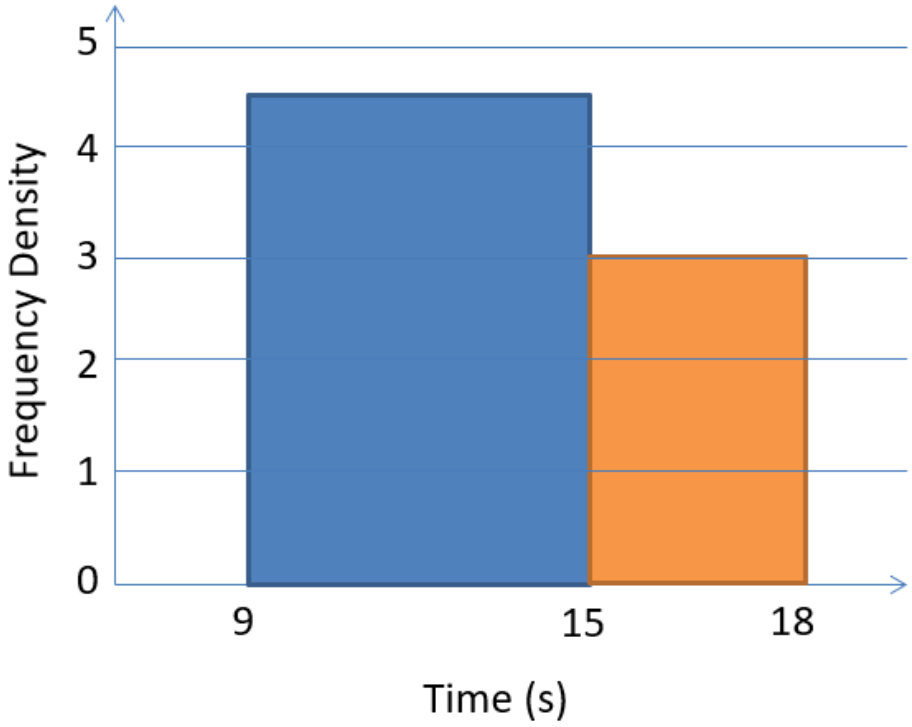
Complete the table and histogram:

Weight, w (nearest kg)	Frequency
$1 - 2$	4
$3 - 6$	3
$7 - 9$	



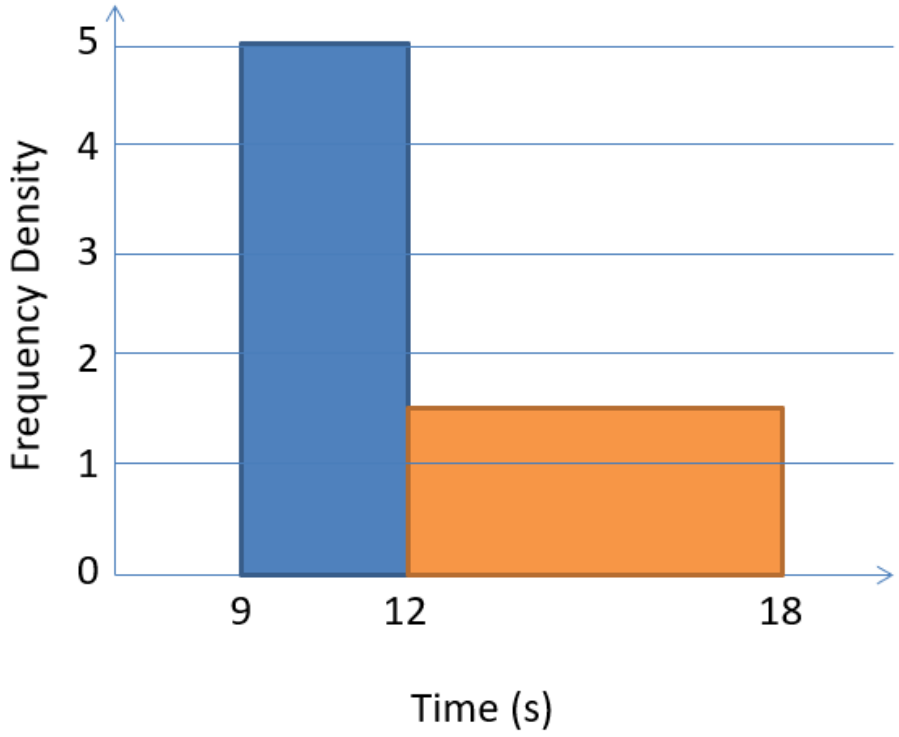
Worked Example

There were 54 runners in a 100 m race.
The following histogram represents their times. Determine the number of runners with times below 13 seconds.



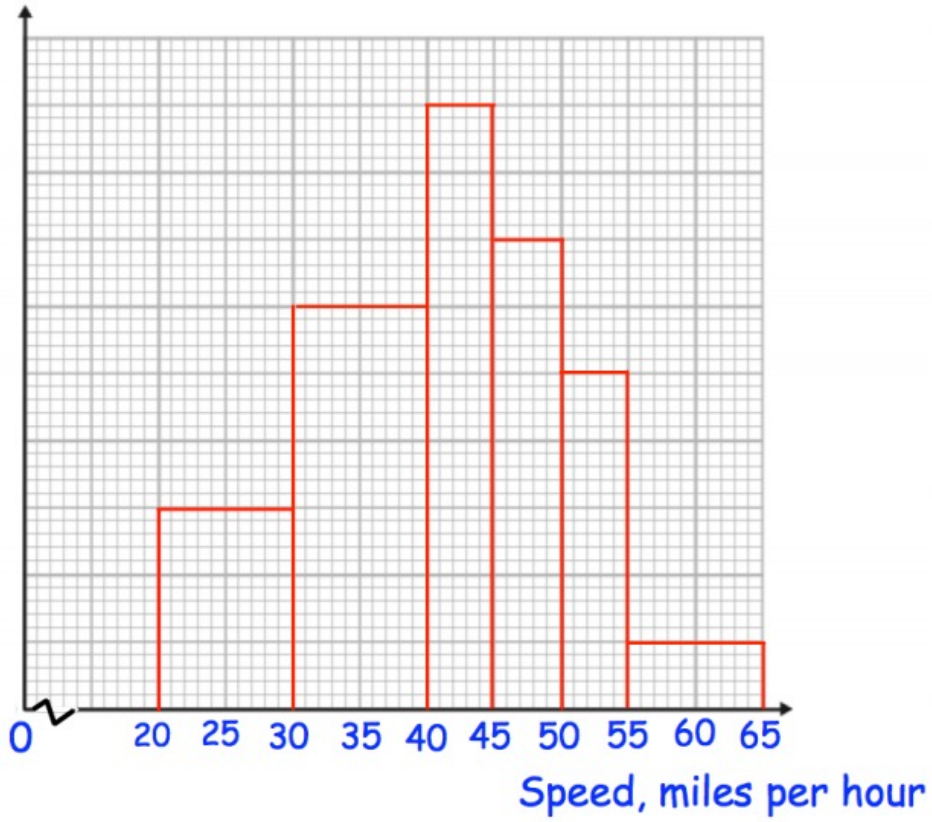
Your Turn

There were 60 runners in a 100 m race.
The following histogram represents their times. Determine the number of runners with times below 14 seconds.



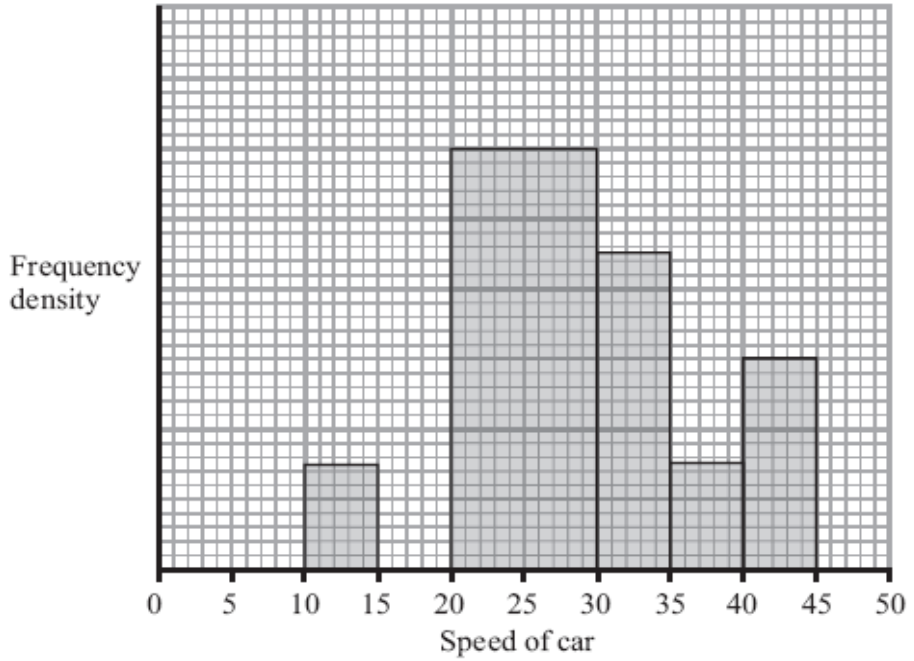
Worked Example

The histogram shows the speeds of 82 cars.
Calculate the number of cars that were driving at speeds of at least 50 miles per hour.



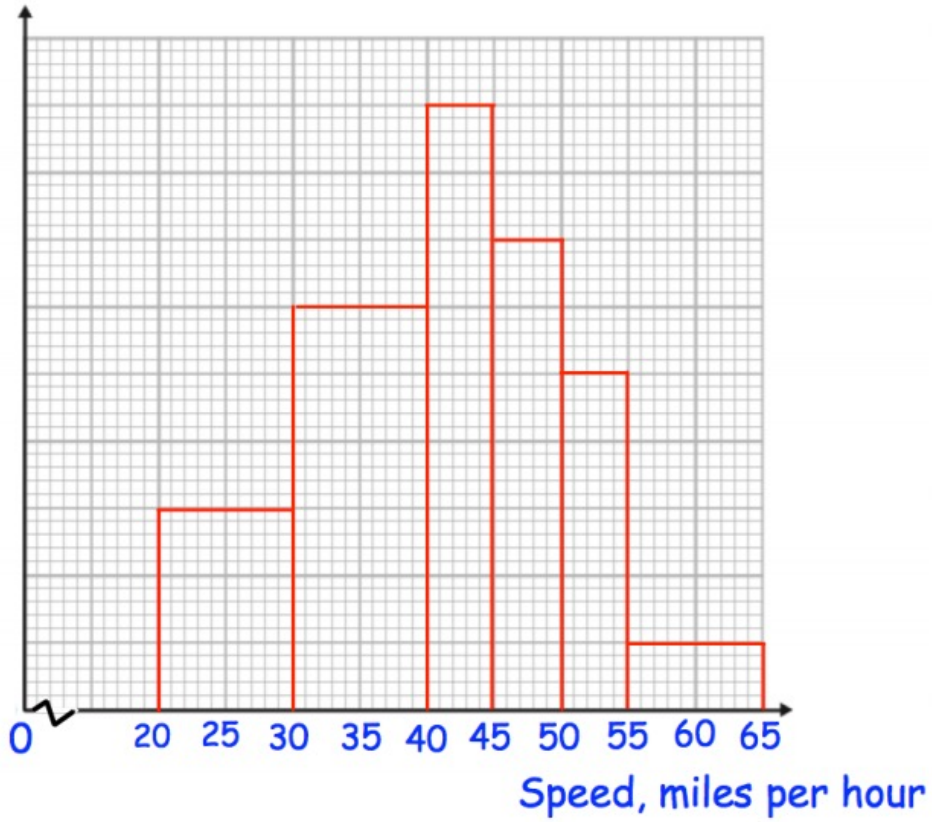
Your Turn

The histogram shows the speeds of 450 cars.
Calculate the number of cars that were driving at speeds of at least 35 miles per hour.



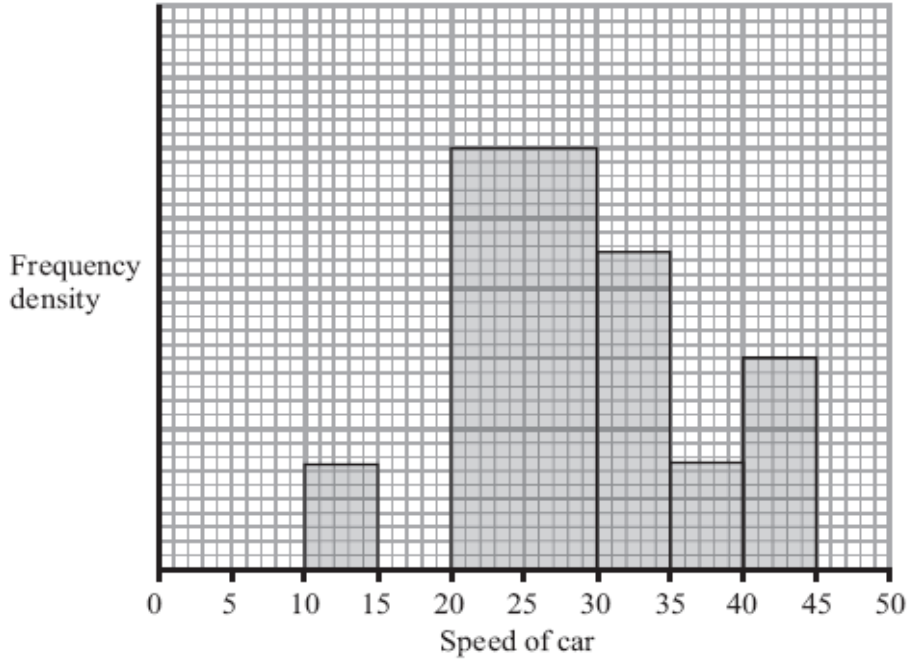
Worked Example

The histogram shows the speeds of 82 cars.
Estimate the mean speed.



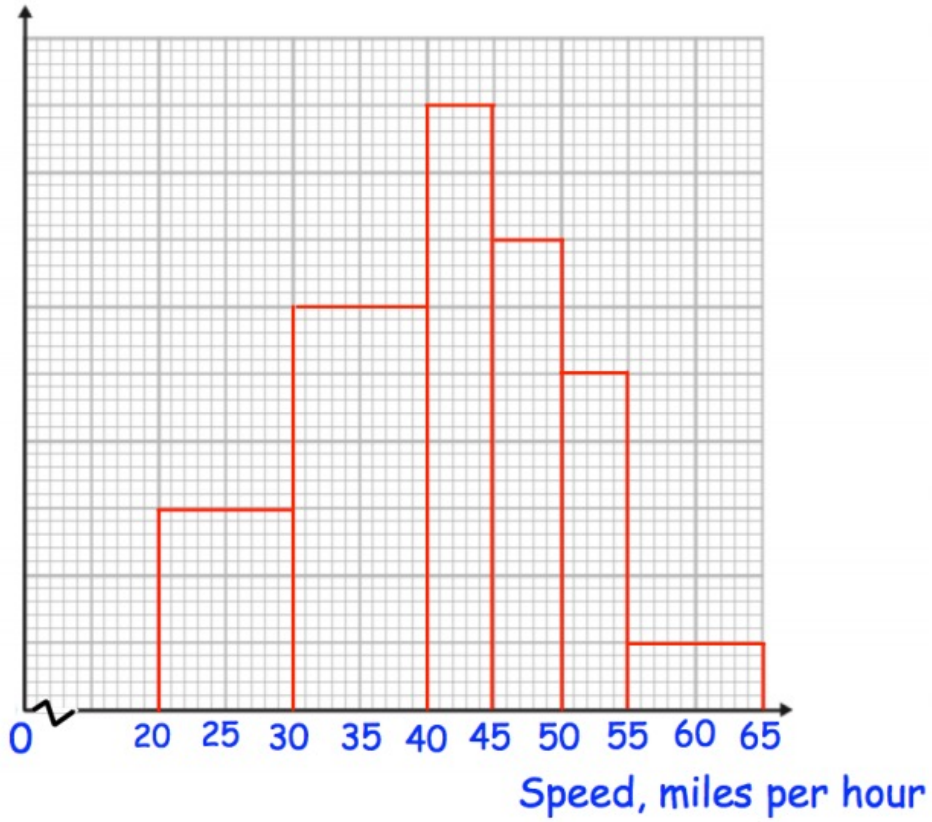
Your Turn

The histogram shows the speeds of 450 cars.
Estimate the mean speed.



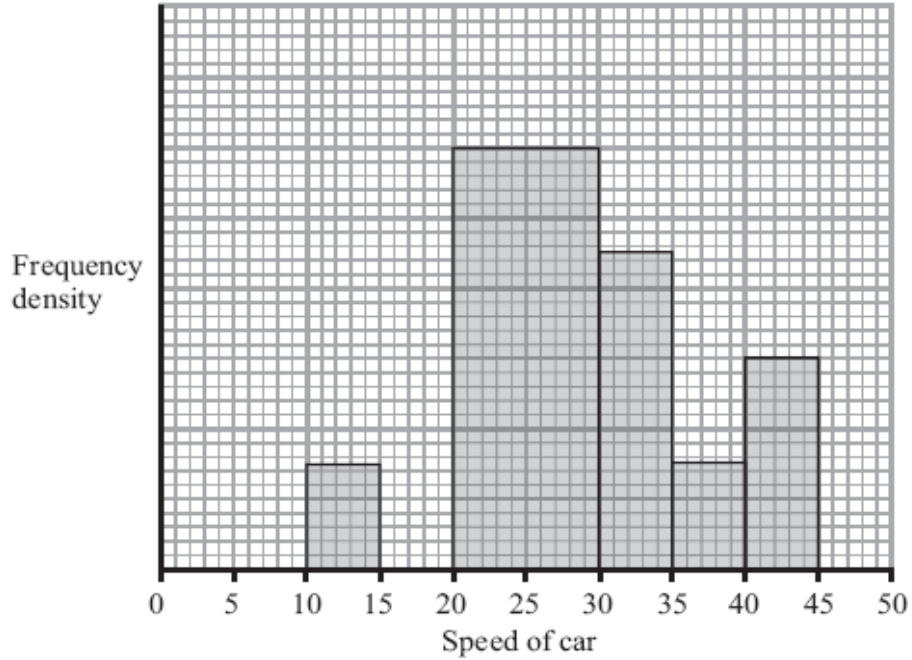
Worked Example

The histogram shows the speeds of 82 cars.
Estimate the median speed.



Your Turn

The histogram shows the speeds of 450 cars.
Estimate the median speed.



Worked Example

The frequency table shows some running times.

On a histogram the bar for $0 - 2$ seconds is drawn with width 8 cm and height 12 cm

Find the width and height of the bar for $2 - 6$ seconds.

Time (seconds)	Frequency
$0 \leq t < 2$	12
$2 \leq t < 6$	3

Your Turn

The frequency table shows some running times.

On a histogram the bar for $0 - 4$ seconds is drawn with width 6 cm and height 8 cm

Find the width and height of the bar for $4 - 6$ seconds.

Time (seconds)	Frequency
$0 \leq t < 4$	8
$4 \leq t < 6$	9

Worked Example

The variable x was measured to the nearest whole number.

On a histogram the bar representing the $2 - 7$ class has a width of 4 cm and a height of 12 cm.

Find the width and height of the $8 - 10$ class.

x	Frequency
$2 - 7$	18
$8 - 10$	6
$12 -$	4

Your Turn

The variable x was measured to the nearest whole number.

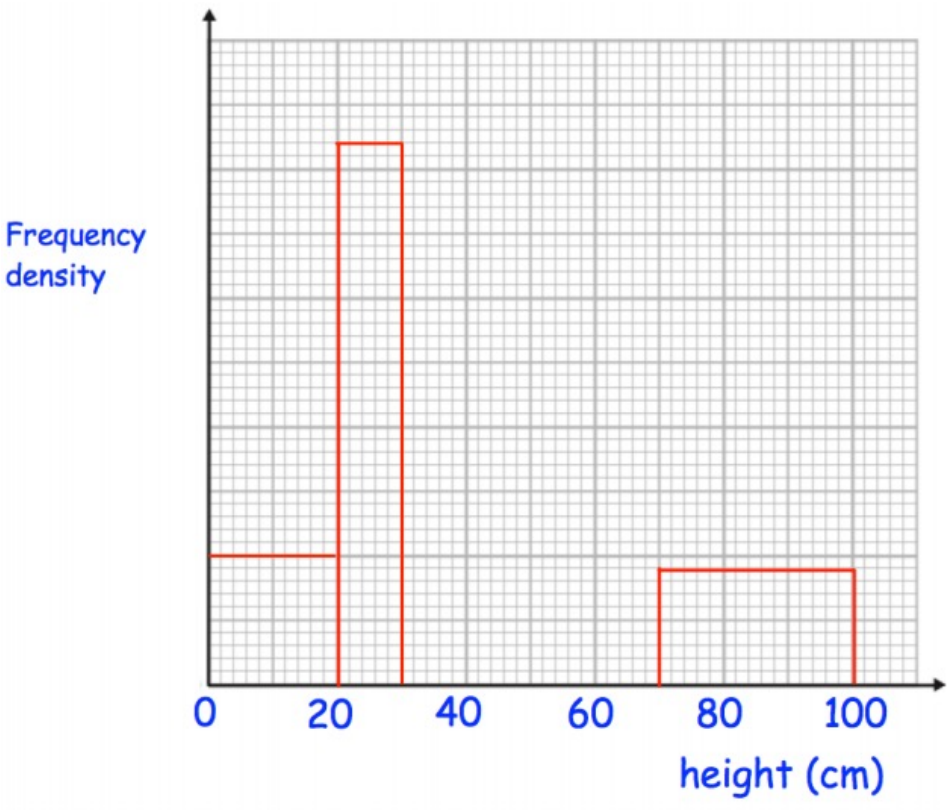
On a histogram the bar representing the $10 - 15$ class has a width of 2 cm and a height of 5 cm.

Find the width and height of the $16 - 18$ class.

x	Frequency
$10 - 15$	15
$16 - 18$	9
$19 -$	16

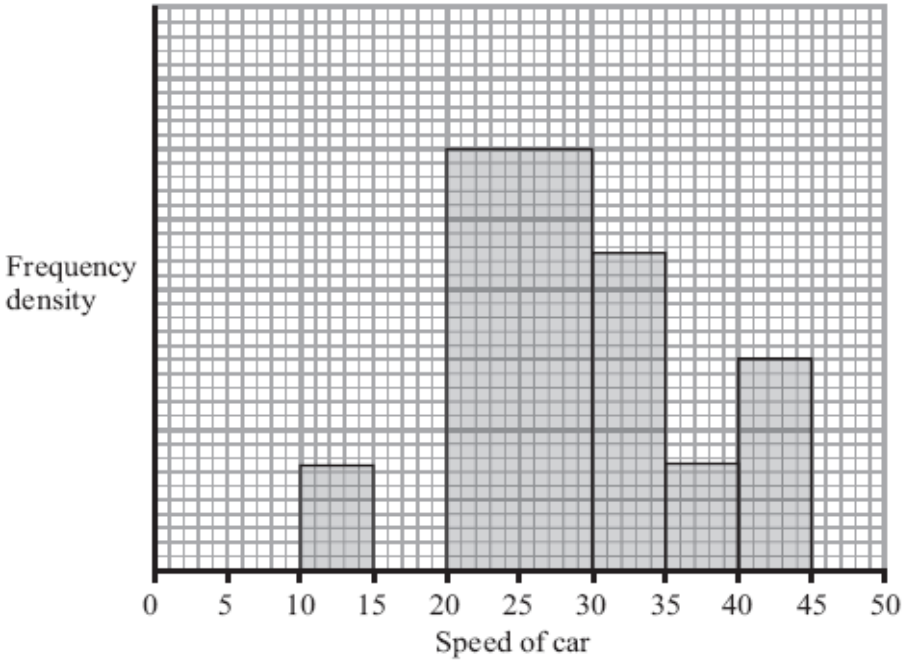
Worked Example

Draw a frequency polygon.



Your Turn

Draw a frequency polygon.



Dr Frost 315d

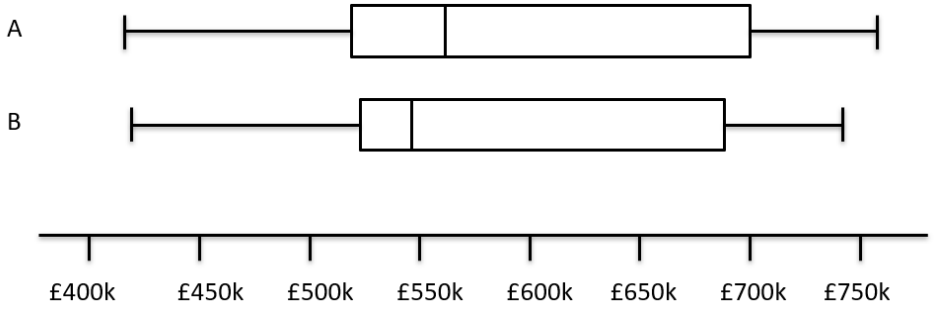
3.5 Comparing Data

Notes

Worked Example	Your Turn
From the large data set, the daily mean temperature during June 1987 is recorded at Camborne and Leuchars. For Camborne, $\sum x = 377.1$ and $\sum x^2 = 4939.45$ For Leuchars, the mean temperature was $10.9\ °C$ with a standard deviation of $2.10\ °C$ Compare the data for the two locations.	From the large data set, the daily mean temperature during August 2015 is recorded at Heathrow and Leeming. For Heathrow, $\sum x = 562.0$ and $\sum x^2 = 10301.2$ For Leeming, the mean temperature was $15.6\ °C$ with a standard deviation of $2.01\ °C$ Compare the data for the two locations.

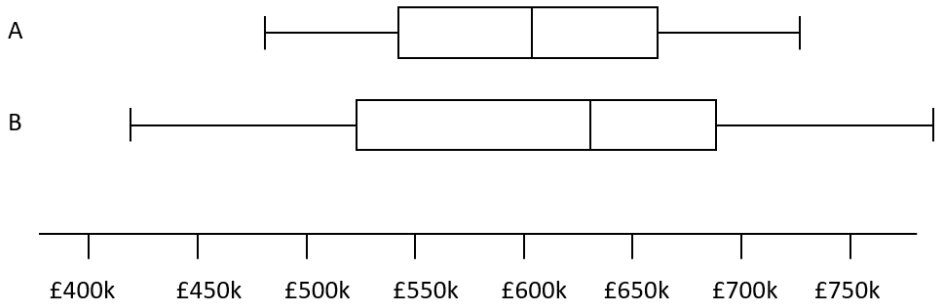
Worked Example

Compare the house prices of locations A and B



Your Turn

Compare the house prices of locations A and B



Worked Example

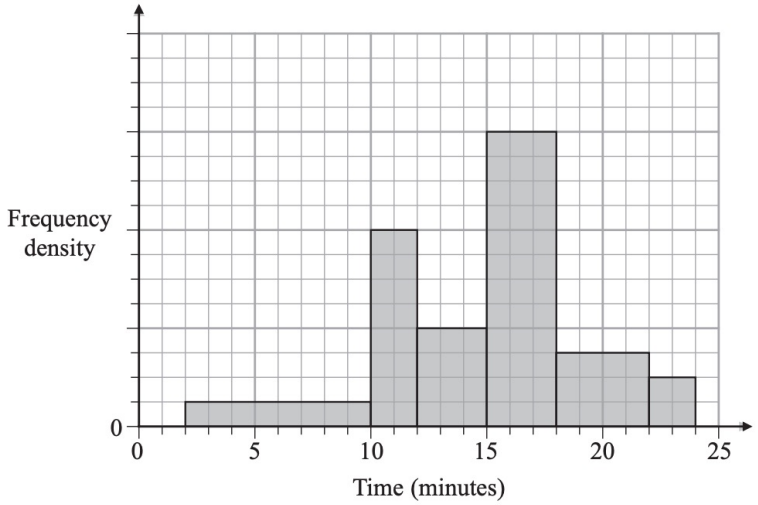


Figure 1

The histogram in Figure 1 shows the times taken to complete a crossword by a random sample of students.

The number of students who completed the crossword in more than 15 minutes is 78

Estimate the percentage of students who took less than 11 minutes to complete the crossword.

(4)

Your Turn

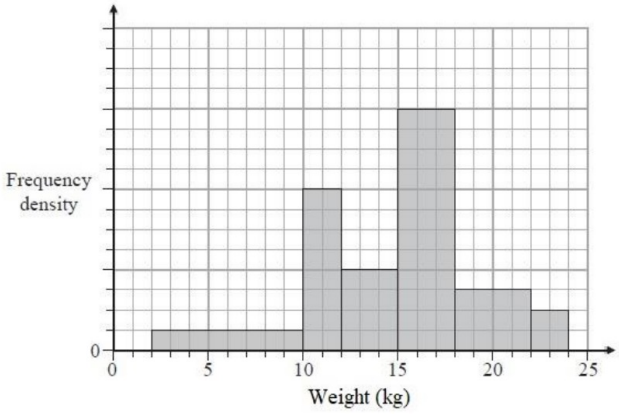


Figure 1

The histogram in Figure 1 shows the weights of dogs at a kennel.

There are 18 dogs who weigh less than 15kg.

Estimate the percentage of dogs who weigh more than 17kg.

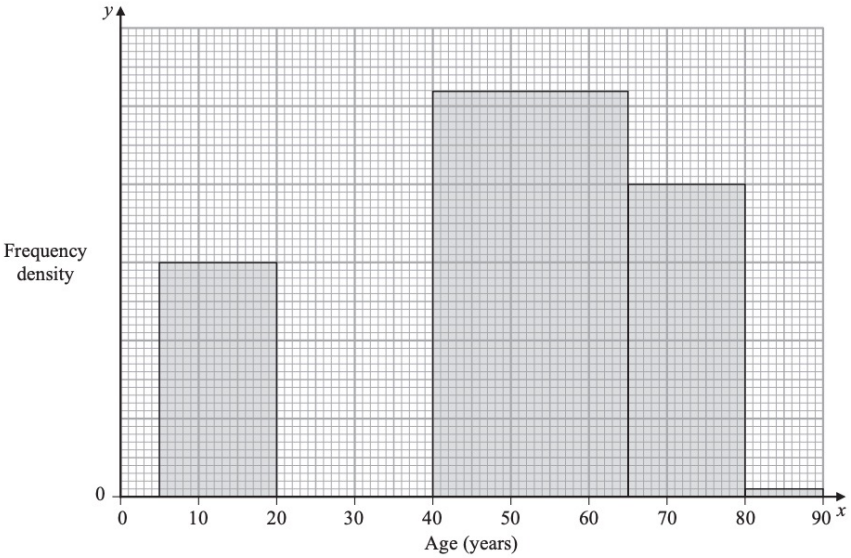
(4)

Worked Example

The partially completed table and partially completed histogram give information about the ages of passengers on an airline.

There were no passengers aged 90 or over.

Age (x years)	$0 \leqslant x < 5$	$5 \leqslant x < 20$	$20 \leqslant x < 40$	$40 \leqslant x < 65$	$65 \leqslant x < 80$	$80 \leqslant x < 90$
Frequency	5	45	90			1



(a) Complete the histogram.

(3)

(b) Use linear interpolation to estimate the median age.

(4)

An outlier is defined as a value greater than $Q_3 + 1.5 \times$ interquartile range.

Given that $Q_1 = 27.3$ and $Q_3 = 58.9$

(c) determine, giving a reason, whether or not the oldest passenger could be considered as an outlier.

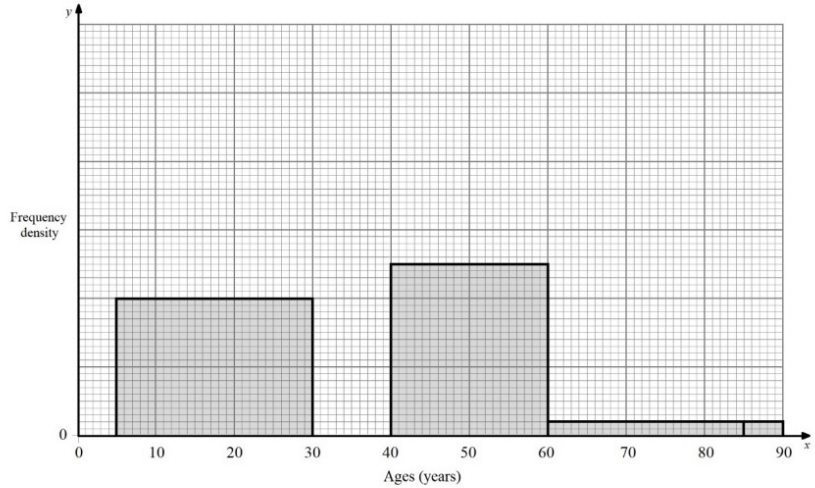
(2)

Your Turn

The partially completed table and partially completed histogram give information about the ages of an audience in the theatre.

There were no audience members aged 90 or over.

Age (x years)	$0 \leqslant x < 5$	$5 \leqslant x < 30$	$30 \leqslant x < 40$	$40 \leqslant x < 60$	$60 \leqslant x < 85$	$85 \leqslant x < 90$
Frequency	5	50	55			1



(a) Complete the histogram.

(3)

(b) Use linear interpolation to estimate the median age.

(4)

An outlier is defined as a value greater than $Q_3 + 1.5 \times$ interquartile range.

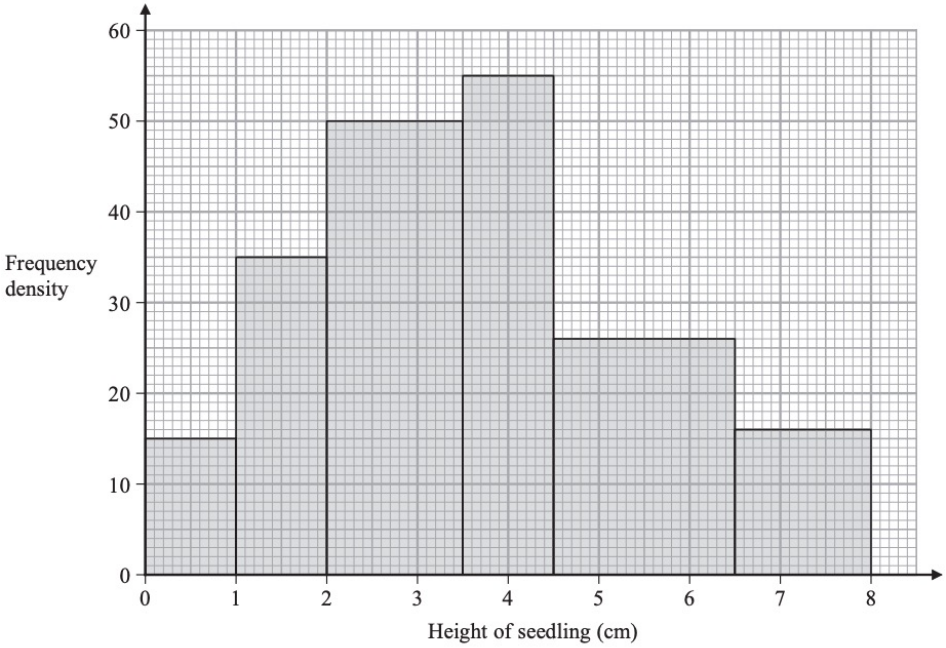
Given that $Q_1 = 23.25$ and $Q_3 = 45.8$

(c) Determine, giving a reason, whether or not the oldest person in the audience could be considered as an outlier.

(2)

Worked Example

3. The histogram summarises the heights of 256 seedlings two weeks after they were planted.

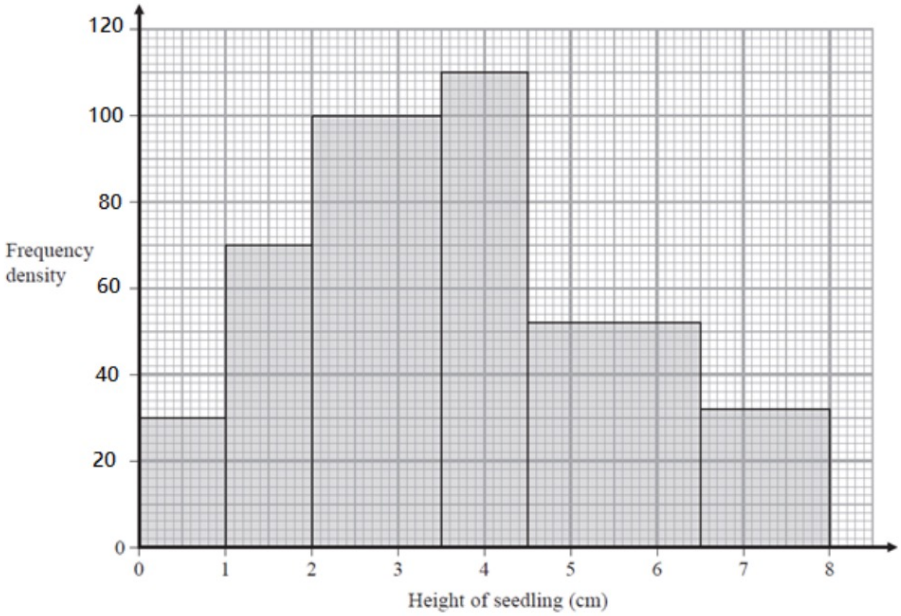


(a) Use linear interpolation to estimate the median height of the seedlings.

(4)

Your Turn

The histogram summarises the heights of 512 seedlings two weeks after they were planted.

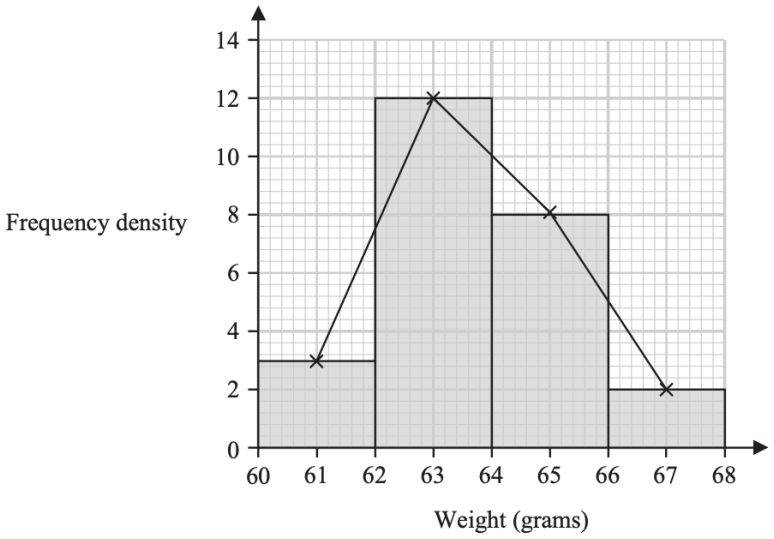


(*a*) Use linear interpolation to estimate the median height of the seedlings.

(4)

Worked Example

The histogram and its frequency polygon below give information about the weights, in grams, of 50 plums.



(a) Show that an estimate for the mean weight of the 50 plums is 63.72 grams.

(2)

(b) Calculate an estimate for the standard deviation of the 50 plums.

(2)

Later it was discovered that the scales used to weigh the plums were broken.

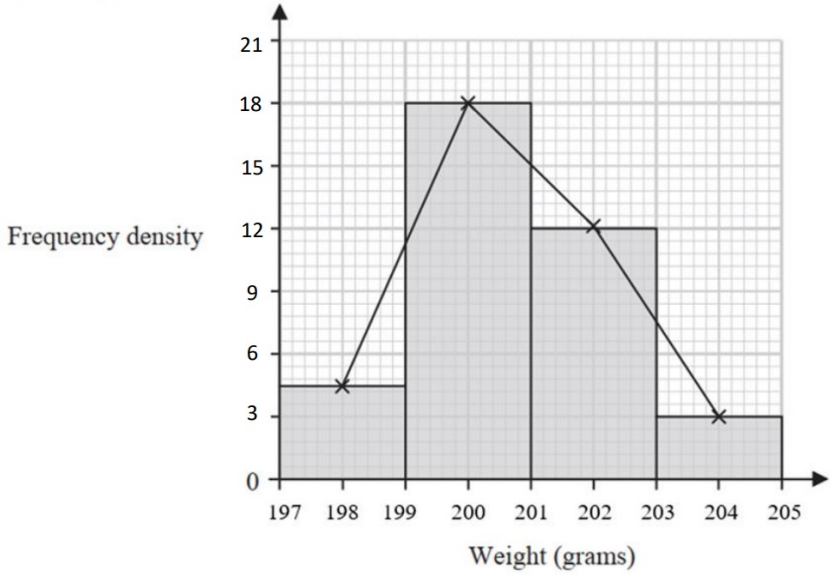
Each plum actually weighs 5 grams less than originally thought.

(c) State the effect this will have on the estimate of the standard deviation in part (b). Give a reason for your answer.

(1)

Your Turn

The histogram and its frequency polygon below give information about the weights, in grams, of 75 apples.



(a) Show that an estimate for the mean weight of the 75 apples is 200.72 grams.

(2)

(b) Calculate an estimate for the standard deviation of the 75 apples.

(2)

Later it was discovered that the scales used to weigh the apples were broken.

Each apple actually weighs 6 grams more than originally thought.

(c) State the effect this will have on the estimate of the standard deviation in part (b). Give a reason for your answer.

(1)

Worked Example

A coach recorded the heights of some adult rugby players and found the following summary statistics.

Median = 1.85 m

Range = 0.28 m

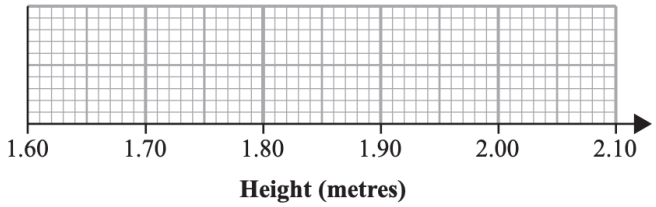
Interquartile range = 0.11 m

The coach also noticed that

- the height of the shortest player is 1.72 m
- 25% of the players' heights are below the height of a player whose height is 1.81 m

Draw a box and whisker plot to represent this information on the grid below.

(4)



Your Turn

A farmer recorded the heights of sunflowers growing in their field and found the following summary statistics.

Median = 2.15 m

Range = 0.42 m

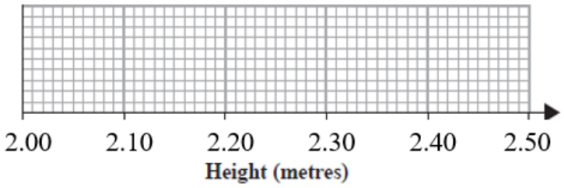
Interquartile range = 0.16 m

The farmer also noticed that

- the height of the tallest sunflower is 2.48 m
- 25% of the sunflowers are shorter than 2.11 m

Draw a box and whisker plot to represent this information on the grid below.

(4)



Worked Example

Customers in a shop have to queue to pay.

The partially completed table below and partially completed histogram opposite, give information about the time, x minutes, spent in the queue by each of 112 customers one day.

Time in queue (x minutes)	Frequency
1–2	64
2–3	
3–4	13
4–6	
6–8	3

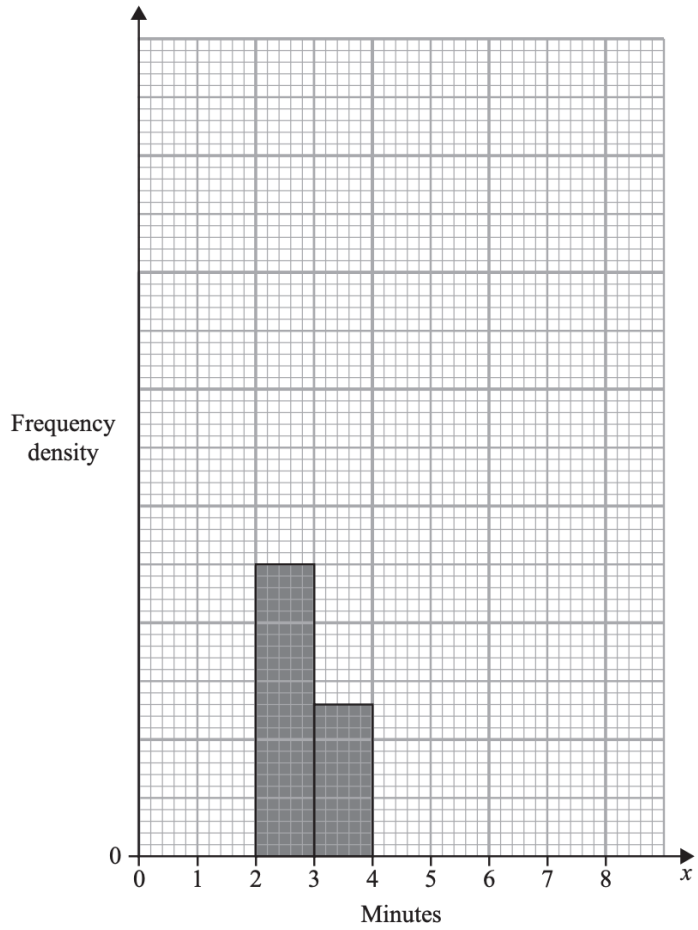
No customer spent less than 1 minute or longer than 8 minutes in the queue.

(a) Complete the table.

(2)

(b) Complete the histogram.

(2)

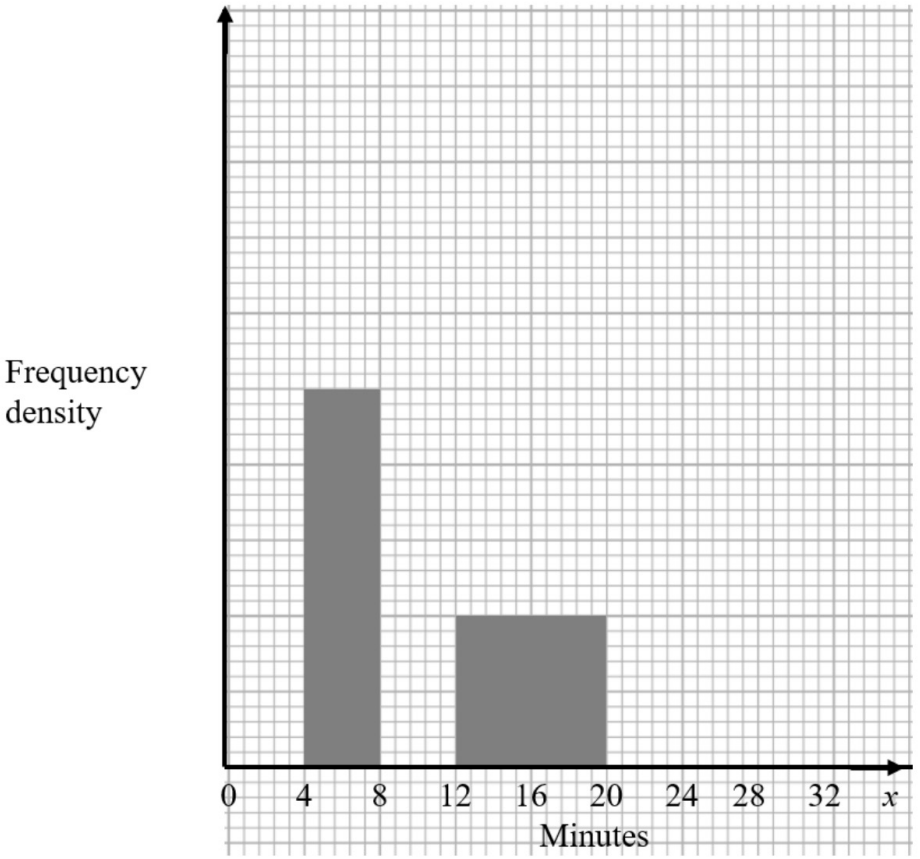


Your Turn

Visitors at a conference were asked to solve a puzzle during their lunch break.

The partially completed table below and partially completed histogram opposite, give information about the time, x minutes, spent solving the puzzle by each of 67 visitors.

Time to solve puzzle (x minutes)	Frequency
4–8	20
8–12	12
12–20	
20–24	
24–32	12



No visitor spent less than 4 minutes or longer than 32 minutes to solve the puzzle.

(a) Complete the table.

(2)

(b) Complete the histogram on the next page.

(2)

Joanne decides to model the frequency density for these 67 customers by a curve with equation

$$y = \frac{k}{x^2} \qquad 4 \leq x \leq 32$$

where k is a constant.

(c) Find the value of k

(3)

Worked Example

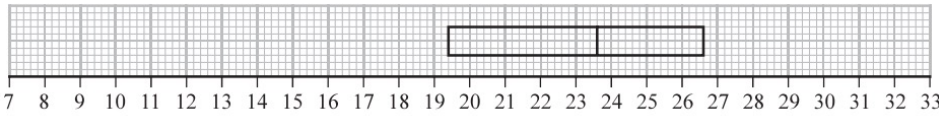


Figure 1

The partially completed box plot in Figure 1 shows the distribution of daily mean air temperatures using the data from the large data set for Beijing in 2015

An outlier is defined as a value
more than 1.5 × IQR below Q_1 or
more than 1.5 × IQR above Q_3

The three lowest air temperatures in the data set are 7.6 °C, 8.1 °C and 9.1 °C
The highest air temperature in the data set is 32.5 °C

(a) Complete the box plot in Figure 1 showing clearly any outliers.

(4)

(b) Using your knowledge of the large data set, suggest from which month the two outliers are likely to have come.

(1)

Using the data from the large data set, Simon produced the following summary statistics for the daily mean air temperature, x °C, for Beijing in 2015

$$n = 184 \qquad \sum x = 4153.6 \qquad S_{xx} = 4952.906$$

(c) Show that, to 3 significant figures, the standard deviation is 5.19 °C

(1)

Your Turn

Figure 1

The partially complete box plot in Figure 1 above shows the distribution of daily mean air temperatures using the data from a large data set for Beijing in 2015

An outlier is defined as a value
more than 1.5 × IQR below Q_1 or
more than 1.5 × IQR above Q_3

The three lowest air temperatures in the data set are 10.9 °C, 7.6 °C and 9.5 °C.
The highest air temperature in the data set is 31.6 °C.

(a) Complete the box plot in Figure 1. Write down any outliers.

(4)

(*b*) Using your knowledge of the large data set, suggest from which month the two outliers are likely to have come.

(1)

Using the data from the same large set, Craig produced the following summary statistics for the daily mean air temperature, x °C, for Beijing in 2015.

$$n = 166 \quad \sum x = 4222.8 \quad S_{xx} = 4877.585$$

(*c*) Show that, to 3 significant figures, the standard deviation is 5.42 °C

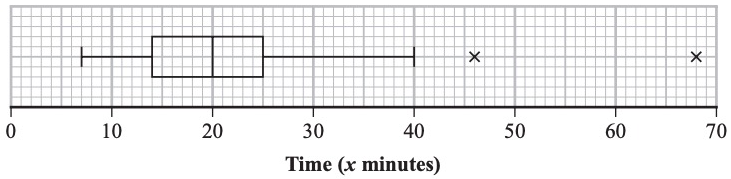
(1)

Worked Example

Each member of a group of 27 people was timed when completing a puzzle.

The time taken, x minutes, for each member of the group was recorded.

These times are summarised in the following box and whisker plot.



(a) Find the range of the times.

(1)

(b) Find the interquartile range of the times.

(1)

For these 27 people $\sum x = 607.5$ and $\sum x^2 = 17\,623.25$

(c) calculate the mean time taken to complete the puzzle,

(1)

(d) calculate the standard deviation of the times taken to complete the puzzle.

(2)

Taruni defines an outlier as a value more than 3 standard deviations above the mean.

(e) State how many outliers Taruni would say there are in these data, giving a reason for your answer.

(1)

Adam and Beth also completed the puzzle in a minutes and b minutes respectively, where $a > b$.
When their times are included with the data of the other 27 people

- the median time increases
- the mean time does not change

(f) Suggest a possible value for a and a possible value for b, explaining how your values satisfy the above conditions.

(3)

(g) Without carrying out any further calculations, explain why the standard deviation of all 29 times will be lower than your answer to part (d).

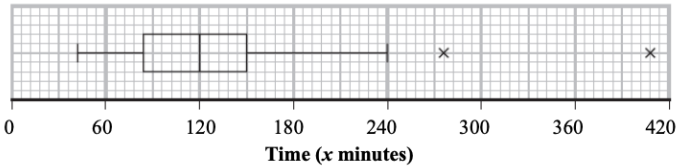
(1)

Your Turn

Each member of a group of 35 people went walking on a weekend.

The time taken for the walk, x minutes, for each member of the group was recorded.

These times are summarised in the following box and whisker plot.



(a) Find the range of the times.

(1)

(b) Find the interquartile range of the times.

(1)

For these 35 people $\sum x = 3850$ and $\sum x^2 = 575960$

(c) calculate the mean time taken on the walk,

(1)

(d) calculate the standard deviation of the times taken on the walk.

(2)

Louise defines an outlier as a value more than 3 standard deviations above the mean.

(e) State how many outliers Louise would say there are in these data, giving a reason for your answer.

(1)

Alana and Buda also went walking for a minutes and b minutes respectively, where $a > b$.

When their times are included with the data of the other 35 people

- the median time decreases
- the mean time does not change

(f) Suggest a possible value for a and a possible value for b, explaining how your values satisfy the above conditions.

(3)

(g) Without carrying out any further calculations, explain why the standard deviation of all 37 times will be lower than your answer to part (d).

(1)

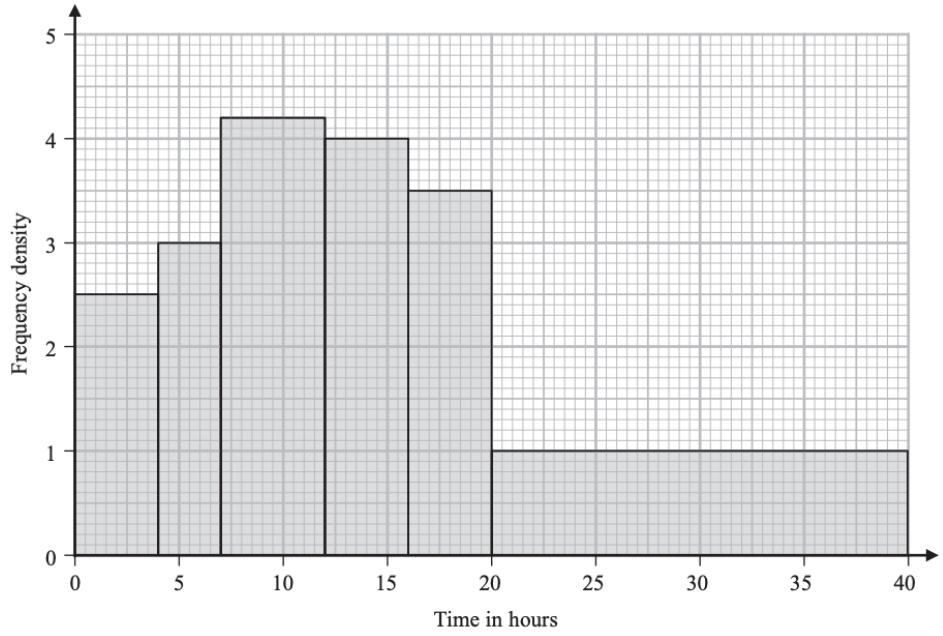
Worked Example

A medical researcher is studying the number of hours, T, a patient stays in hospital following a particular operation.

The histogram on the page opposite summarises the results for a random sample of 90 patients.

(a) Use the histogram to estimate $P(10 < T < 30)$

(2)



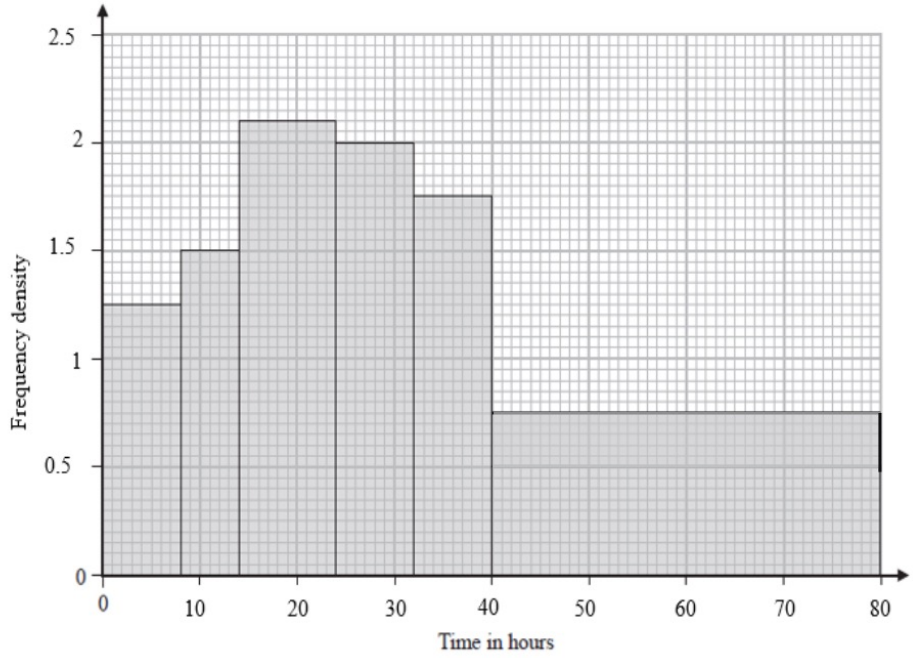
Your Turn

A medical researcher is studying the number of hours, T, a patient stays in hospital following a particular operation.

The histogram on the next page summarises the results for a random sample of 100 patients.

(a) Use the histogram to estimate $P(20 < T < 40)$

(2)



Summary

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 The vertical scale on a histogram shows the frequency density:

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

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