KING EDWARD VI
HANDSWORTH GRAMMAR
SCHOOL FOR BOYS

KING EDWARD VI
ACADEMY TRUST
BIRMINGHAM

Year 12

Statistics 1

S1 3 Representations of Data

Booklet

HGS Maths

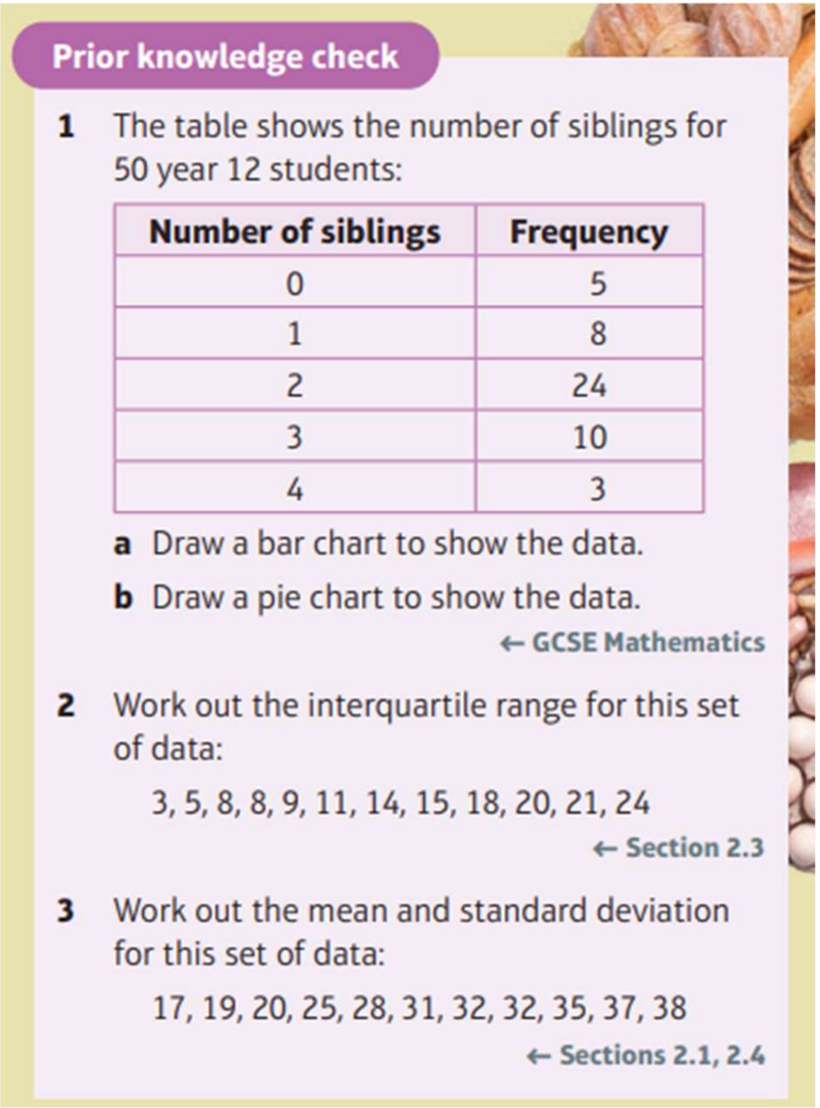
Dr Frost Course

Name: ________________________________

Class: ________________________________

Contents

Prior knowledge check



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:

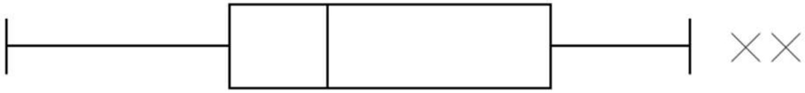
17, 19, 20, 25, 28, 31, 32, 32, 35, 37, 38

← Sections 2.1, 2.4

This Chapter Overview

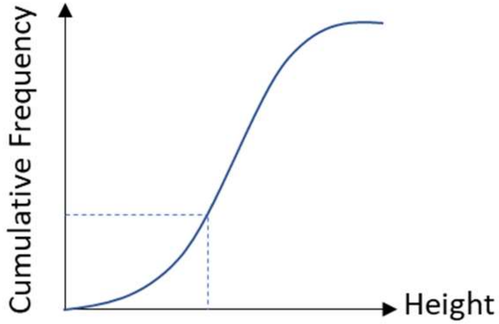
We've seen so far how data is collected and calculations can be made. We now concentrate on how the processed data can be *displayed*.

BOX PLOTS AND OUTLIERS

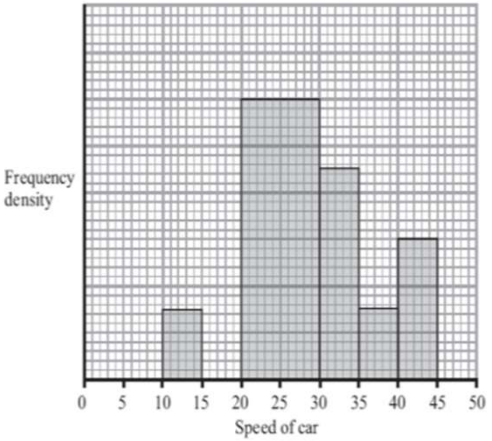


NEW since GCSE! Outliers.

CUMULATIVE FREQ DIAGRAMS



HISTOGRAMS



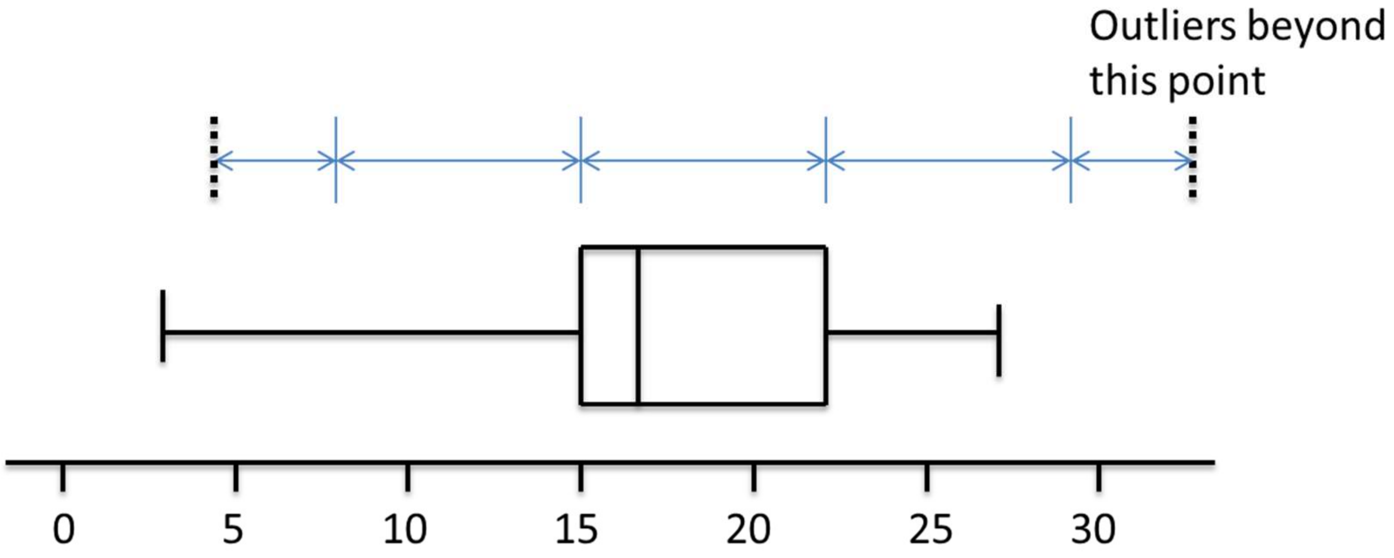
NEW since GCSE! Area is not necessarily equal to frequency.
Forming a frequency polygon by joining midpoints.

Changes since the old 'S1' syllabus:

- Stem and leaf diagrams have been cut. (THANK GOD FOR THAT)
- 'Skew' has been cut.
- Cumulative frequency diagrams have been added.
- Turning histogram into frequency polygon.

3.1 Outliers

An outlier is **an extreme value.**



One common definition of an outlier is when we're **1.5 IQRs** beyond the lower and upper quartiles.

(But you will be told in the exam if the rule differs from this)

Another common definition involves a multiple of the standard deviation from the mean

Notes

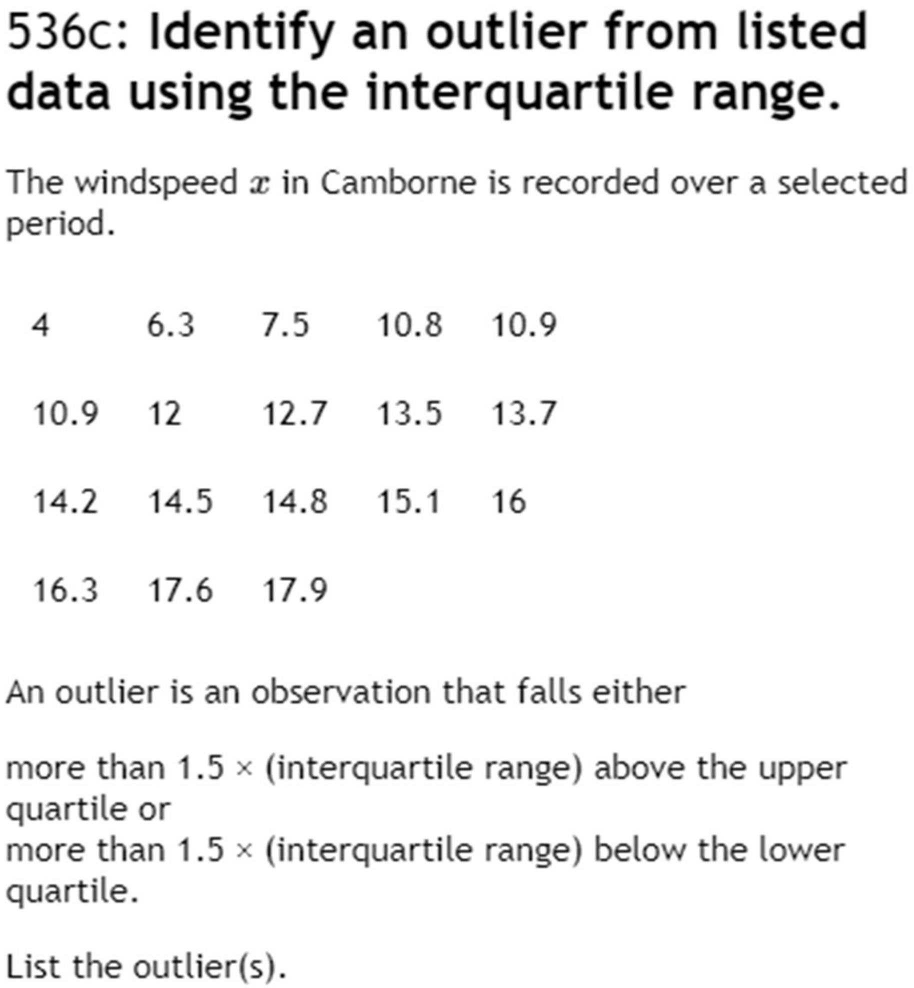
Worked Example

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 x interquartile range above the upper quartile or 1.5 x interquartile range below the lower quartile.

a) Find the quartiles
b) Find any outliers

Worked Example



536c: Identify an outlier from listed data using the interquartile range.

The windspeed x in Camborne is recorded over a selected period.

4 6.3 7.5 10.8 10.9

10.9 12 12.7 13.5 13.7

14.2 14.5 14.8 15.1 16

16.3 17.6 17.9

An outlier is an observation that falls either

more than 1.5 × (interquartile range) above the upper quartile or
more than 1.5 × (interquartile range) below the lower quartile.

List the outlier(s).

Worked Example

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 $\sum x = 252$ and $\sum x^2 = 5468$

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

Worked Example

Clean this data on ages of people in a group:
$12, 13, 14, 12, 13, 156$

3.2 Box Plots

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

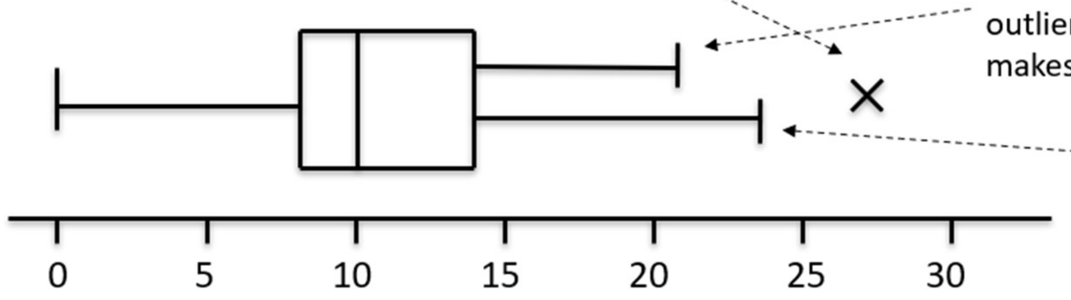
$$IQR = 14 - 8 = 6$$

Outlier boundaries:

$$\mathbf{14 + (1.5 \times 6) = 23}$$
$$\mathbf{8 - (1.5 \times 6) = -1}$$

When there’s an outlier at one end, there’s two allowable places to put the end of the whisker:

Use a cross for each outlier.



The maximum value not an outlier, 21 (I think this one makes most sense).

OR the outlier boundary, 23.

Use one or the other **(not both)**.

Notes

Worked Example

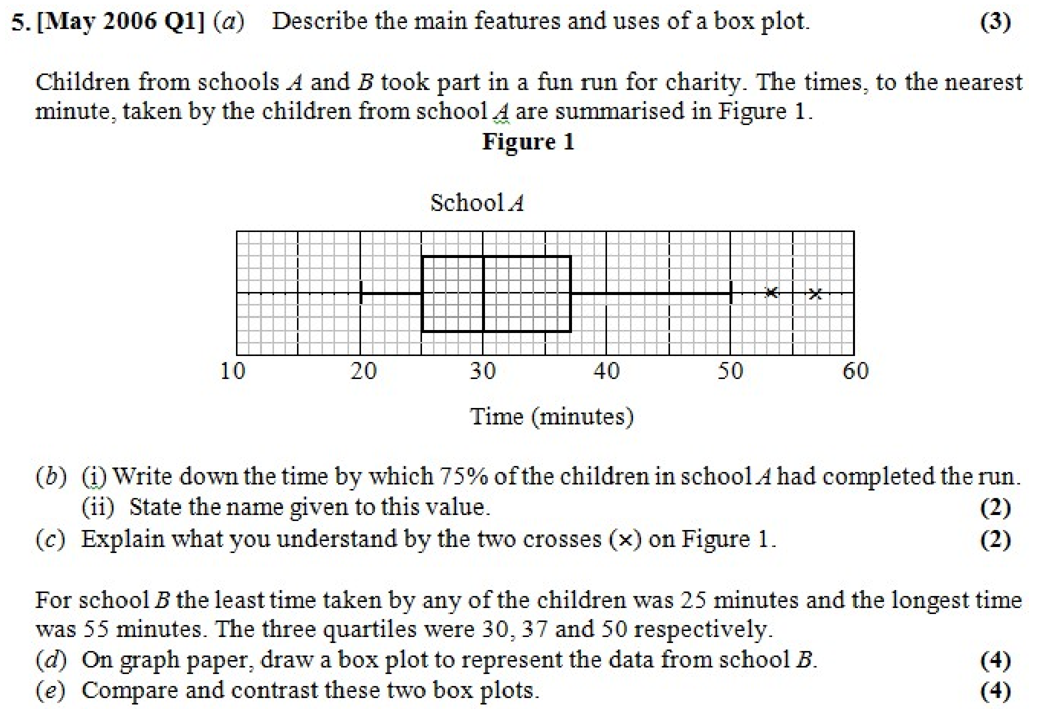
a) Draw a box plot for the data on blood glucose levels of females from Worked Example 1.
The blood glucose level of 30 males is recorded. The results, in mmol/litre, are summarised below:
Lower quartile = 3.6
Upper quartile = 4.7
Median = 4.0
Lowest value = 1.4
Highest value = 5.2

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

Supplementary Questions:



5. [May 2006 Q1] (*a*) Describe the main features and uses of a box plot. **(3)**

Children from schools *A* and *B* took part in a fun run for charity. The times, to the nearest minute, taken by the children from school *A* are summarised in Figure 1.

Figure 1

School *A*
10
20
30
40
50
60
Time (minutes)

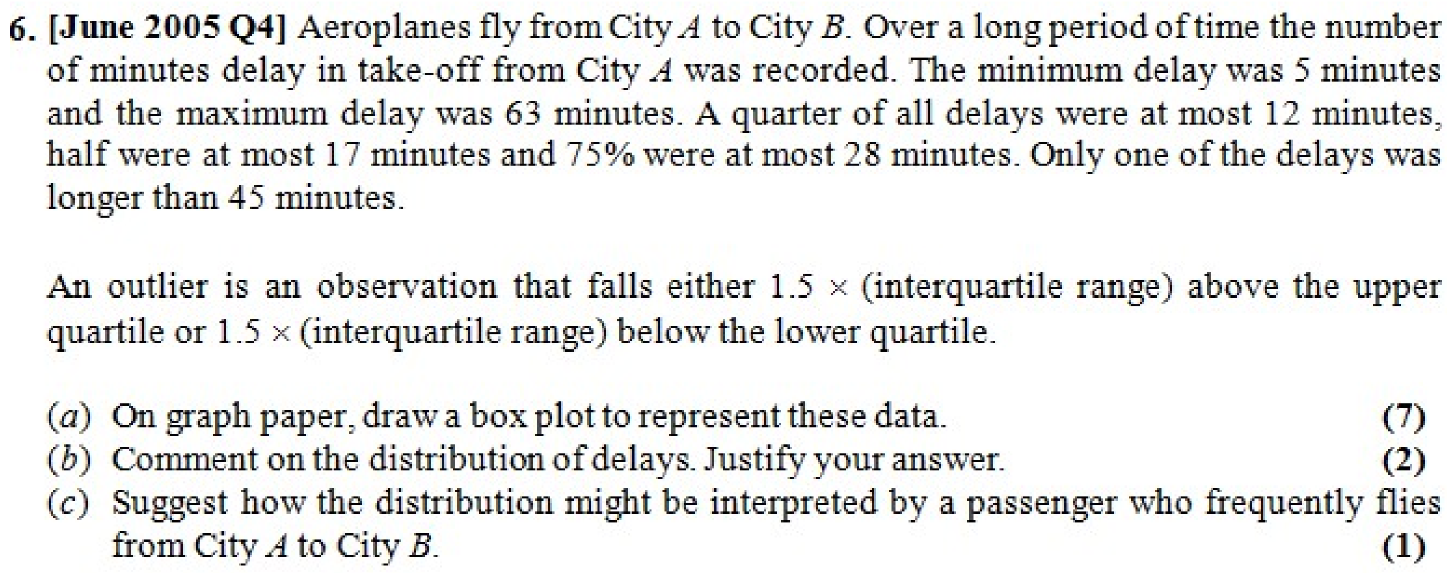
(*b*) (i) Write down the time by which 75% of the children in school *A* had completed the run.
(ii) State the name given to this value. **(2)**

(*c*) Explain what you understand by the two crosses (×) on Figure 1. **(2)**

For school *B* the least time taken by any of the children was 25 minutes and the longest time was 55 minutes. The three quartiles were 30, 37 and 50 respectively.

(*d*) On graph paper, draw a box plot to represent the data from school *B*. **(4)**

(*e*) Compare and contrast these two box plots. **(4)**



6. [June 2005 Q4] Aeroplanes fly from City *A* to City *B*. Over a long period of time the number of minutes delay in take-off from City *A* was recorded. The minimum delay was 5 minutes and the maximum delay was 63 minutes. A quarter of all delays were at most 12 minutes, half were at most 17 minutes and 75% were at most 28 minutes. Only one of the delays was longer than 45 minutes.

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*) On graph paper, draw a box plot to represent these data. **(7)**

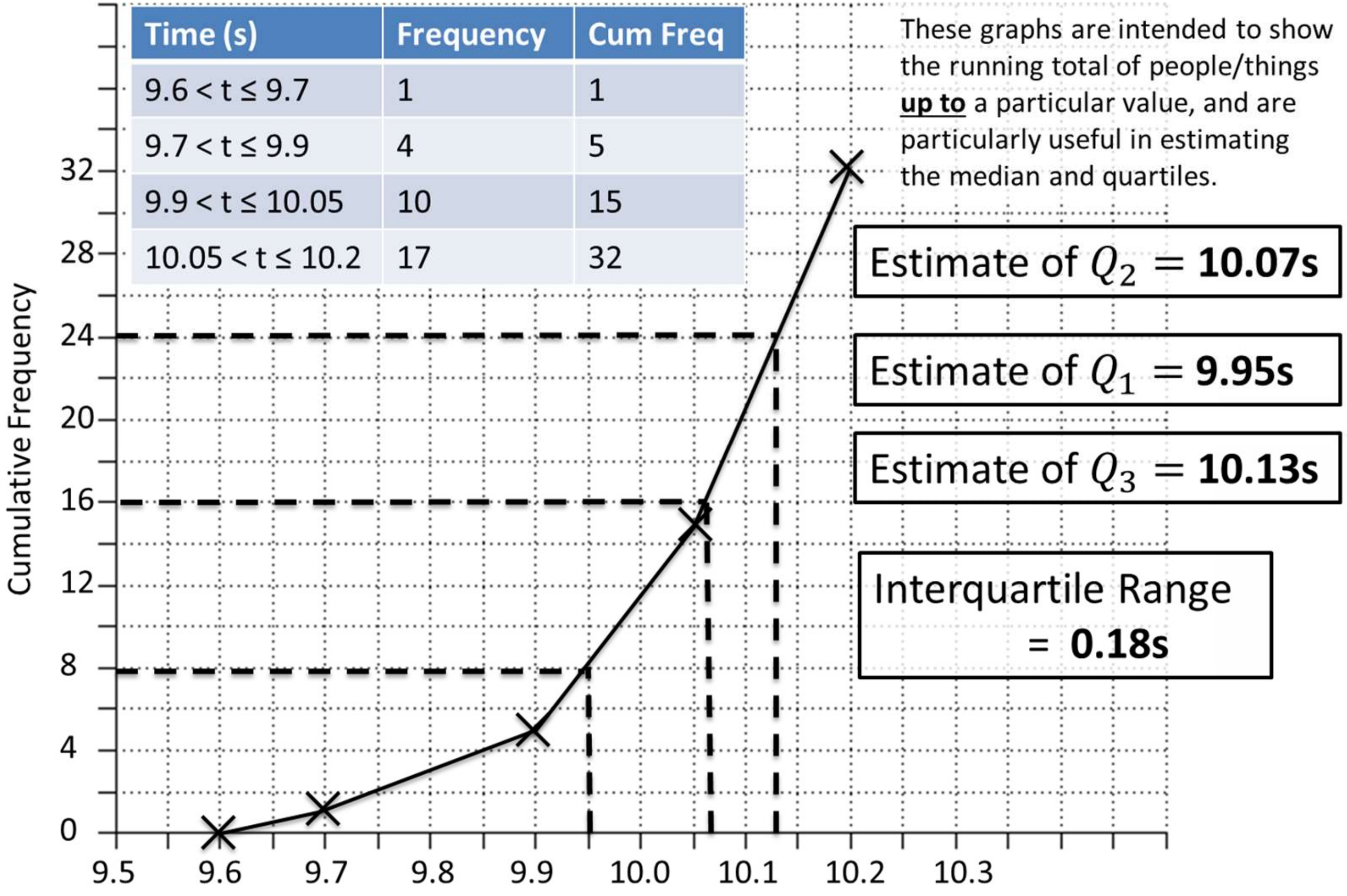
(*b*) Comment on the distribution of delays. Justify your answer. **(2)**

(*c*) Suggest how the distribution might be interpreted by a passenger who frequently flies from City *A* to City *B*. **(1)**

3.3 Cumulative Frequency

Key points:

- Plot first lower bound (x) with y=0
- Plot the **upper class bound (x)** with the **cumulative frequency (y)**
- Join using **straight lines** at A level
- E.g.:



Time (s)	Frequency	Cum Freq
9.6 < t ≤ 9.7	1	1
9.7 < t ≤ 9.9	4	5
9.9 < t ≤ 10.05	10	15
10.05 < t ≤ 10.2	17	32

These graphs are intended to show the running total of people/things **up to** a particular value, and are particularly useful in estimating the median and quartiles.

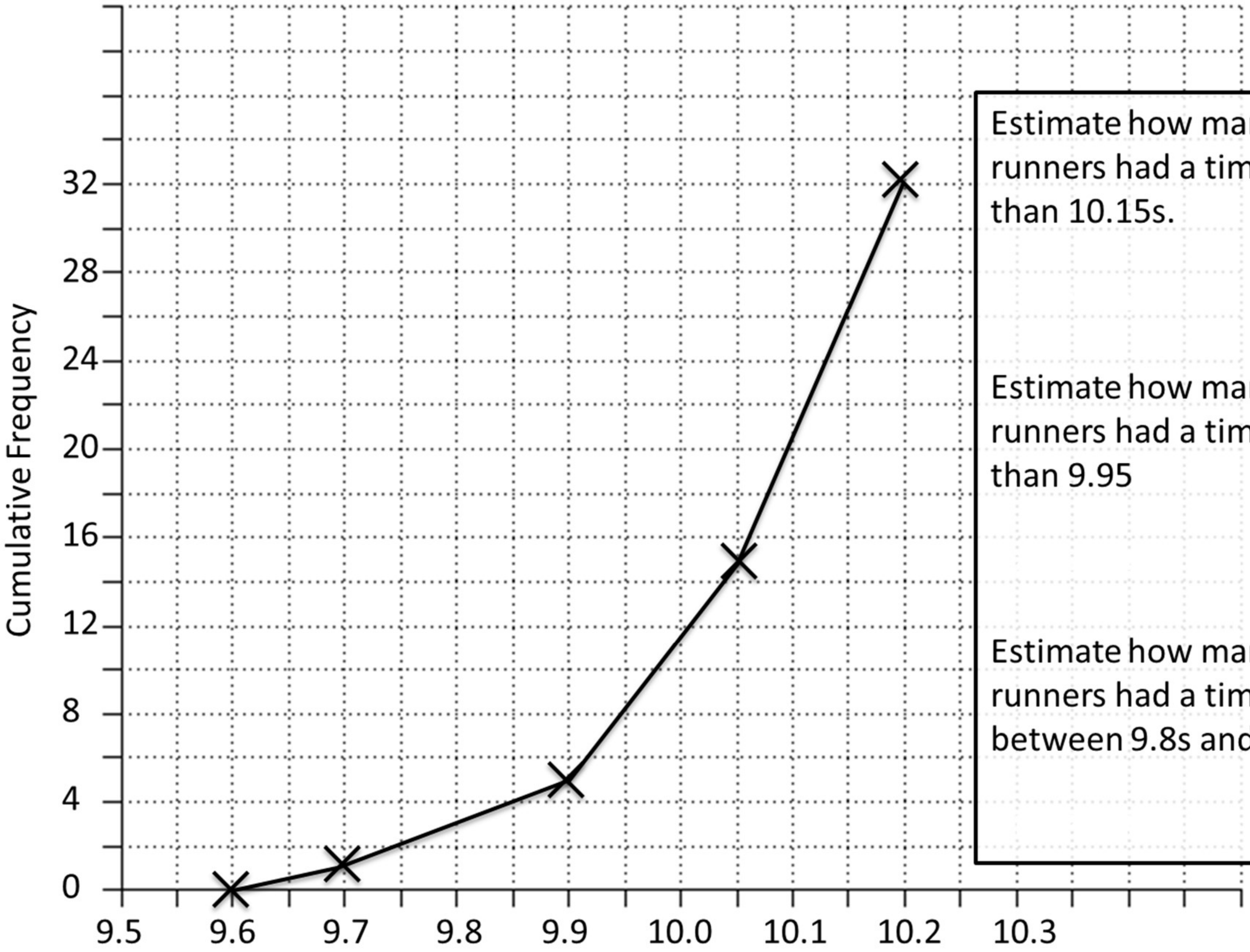
Estimate of Q_2 = **10.07s**

Estimate of Q_1 = **9.95s**

Estimate of Q_3 = **10.13s**

Interquartile Range = **0.18s**

Notes



Estimate how many runners had a time less than 10.15s.

Estimate how many runners had a time more than 9.95

Estimate how many runners had a time between 9.8s and 10s

Notes

Worked Example

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

a) Draw a cumulative frequency diagram

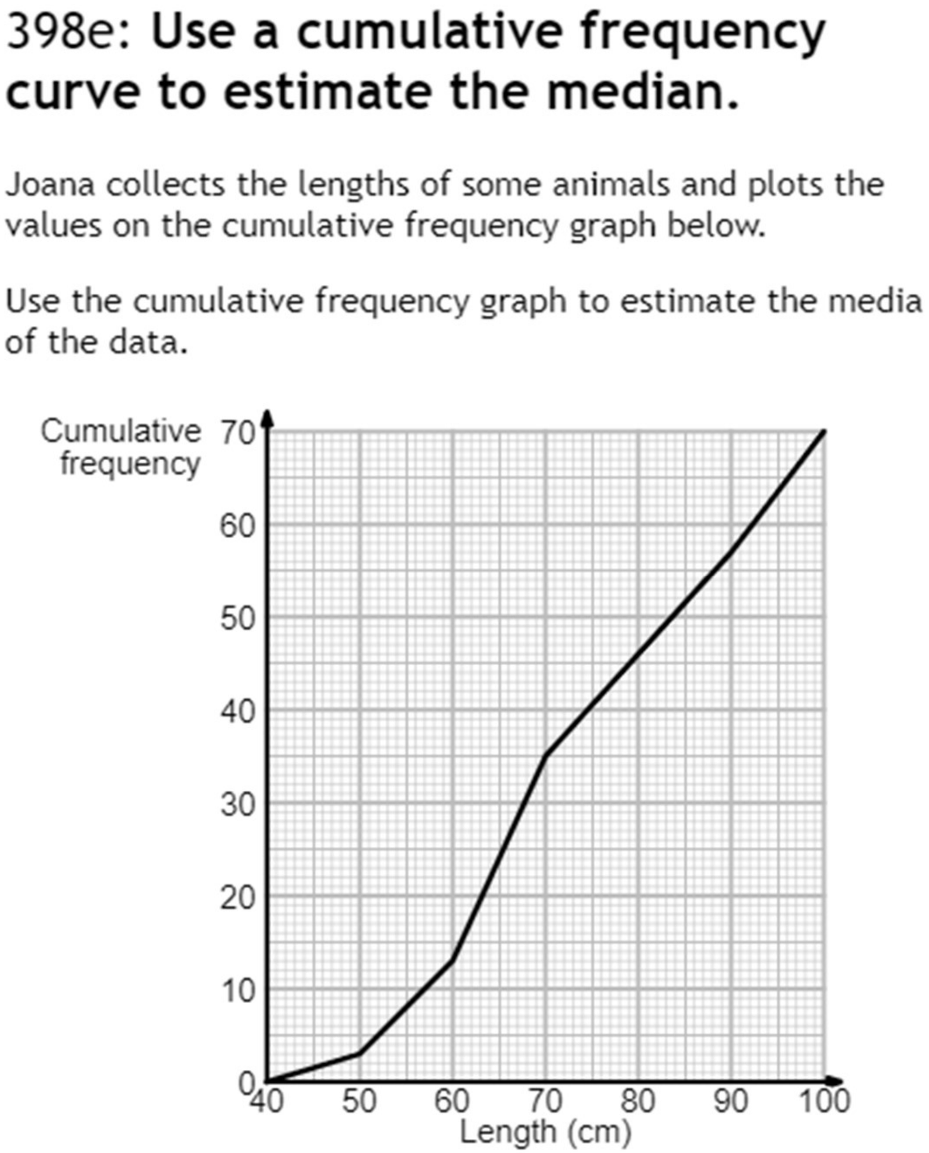
b) Estimate the median height of the giraffes

c) Estimate the lower quartile and the 90th percentile

d) Draw a box plot to represent this data

Height, $\mathbf{h}$ (cm)	**Frequency**
$4.6 \leq h < 4.8$	4
$4.8 \leq h < 5.0$	7
$5.0 \leq h < 5.2$	15
$5.2 \leq g < 5.4$	33
$5.4 \leq h < 5.6$	17
$5.6 \leq h < 5.8$	4

Worked Example

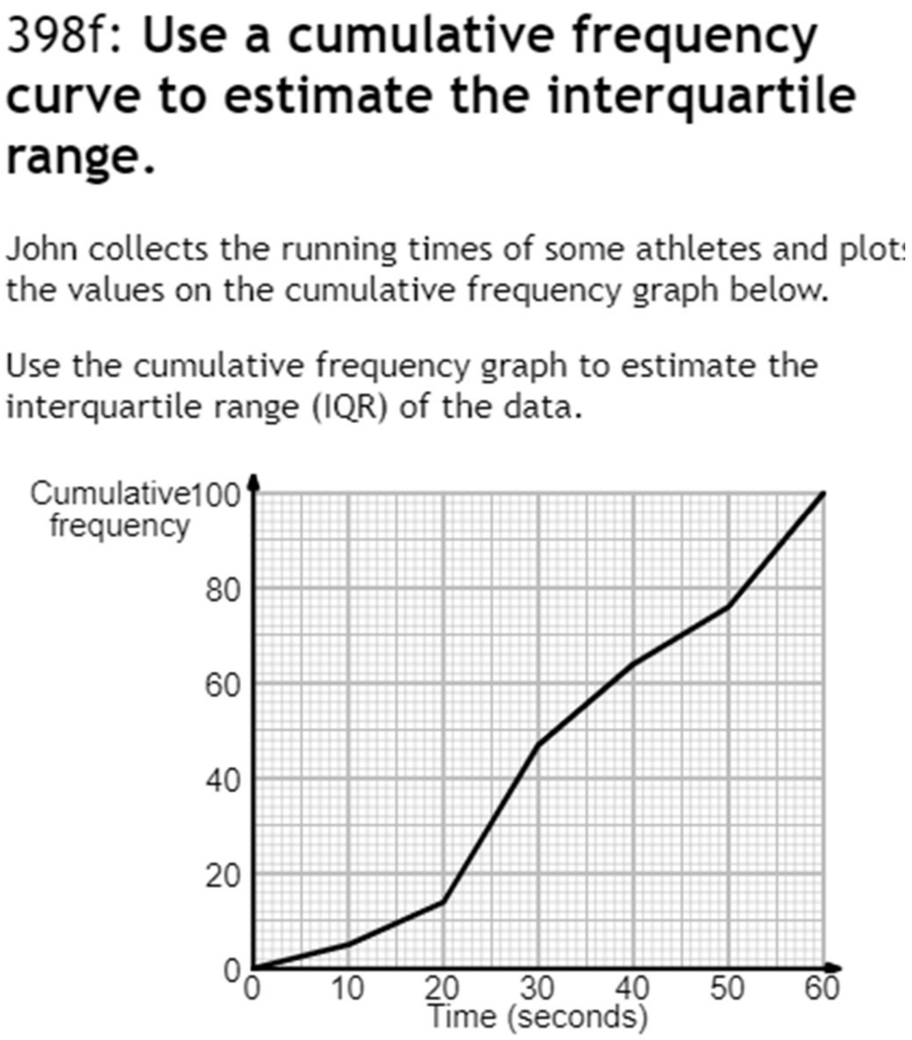


398e: **Use a cumulative frequency curve to estimate the median.**

Joana collects the lengths of some animals and plots the values on the cumulative frequency graph below.

Use the cumulative frequency graph to estimate the media of the data.

Worked Example

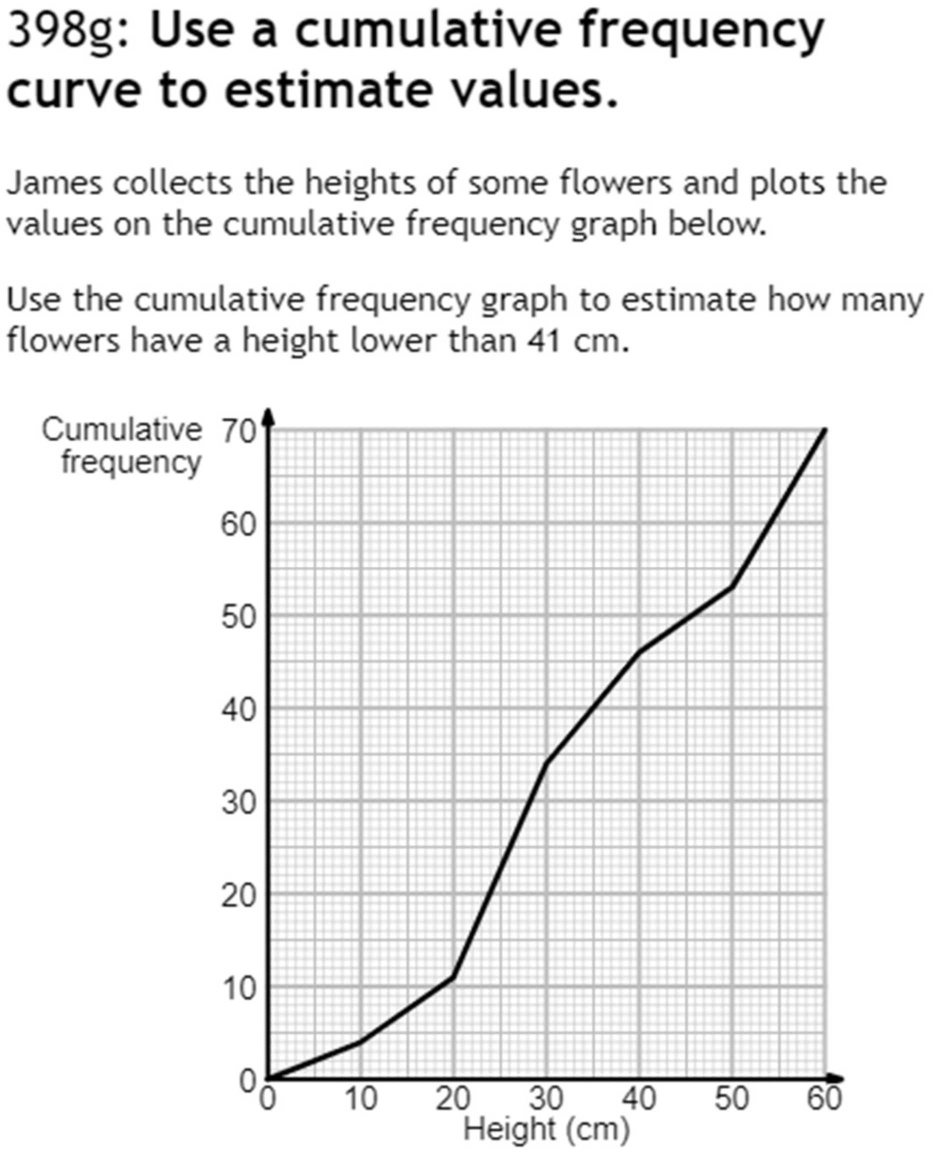


398f: Use a cumulative frequency curve to estimate the interquartile range.

John collects the running times of some athletes and plot: the values on the cumulative frequency graph below.

Use the cumulative frequency graph to estimate the interquartile range (IQR) of the data.

Worked Example



398g: Use a cumulative frequency curve to estimate values.

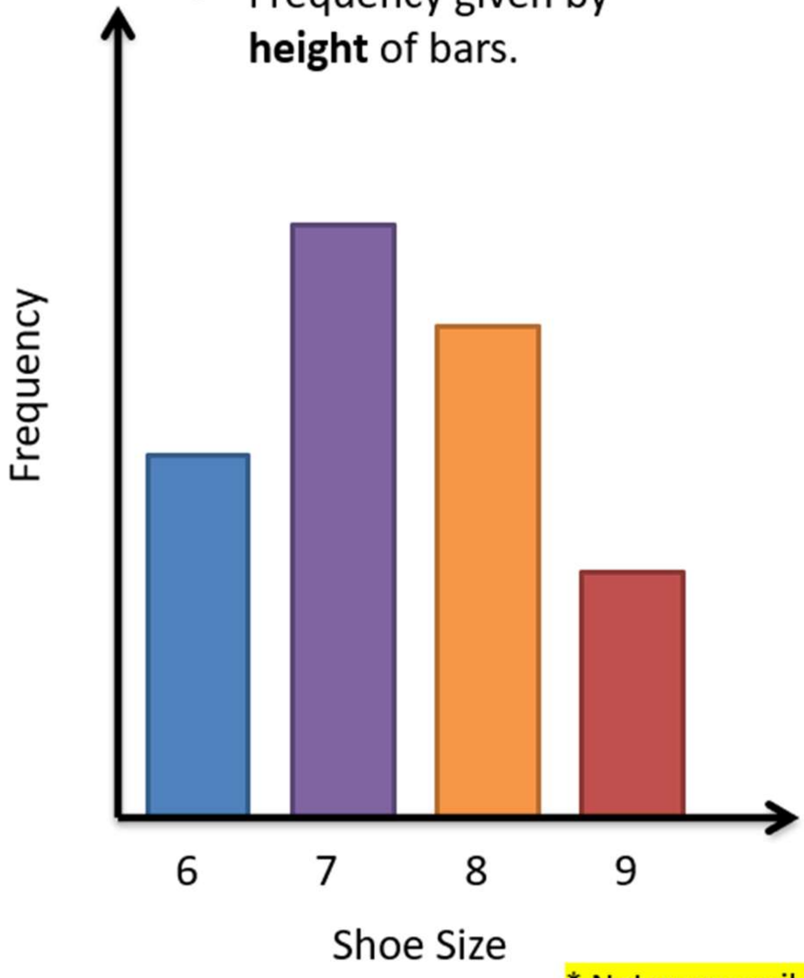
James collects the heights of some flowers and plots the values on the cumulative frequency graph below.

Use the cumulative frequency graph to estimate how many flowers have a height lower than 41 cm.

3.4 Histograms

Bar Charts

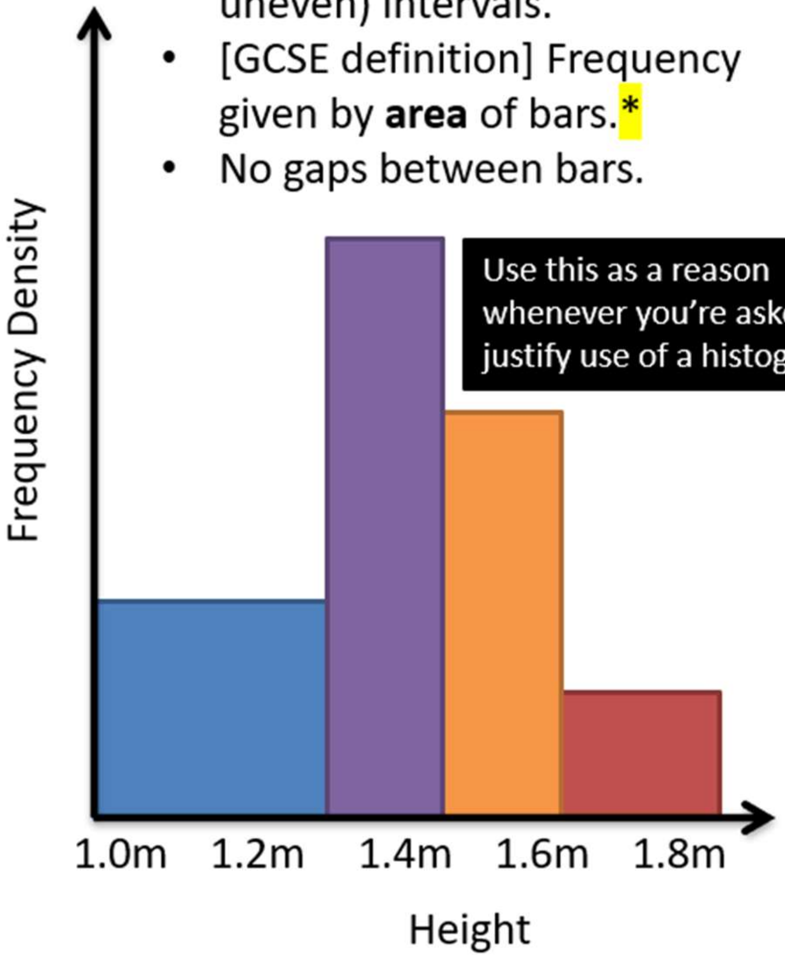
- For **discrete** data.
- Frequency given by **height** of bars.



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.

Recap of GCSE

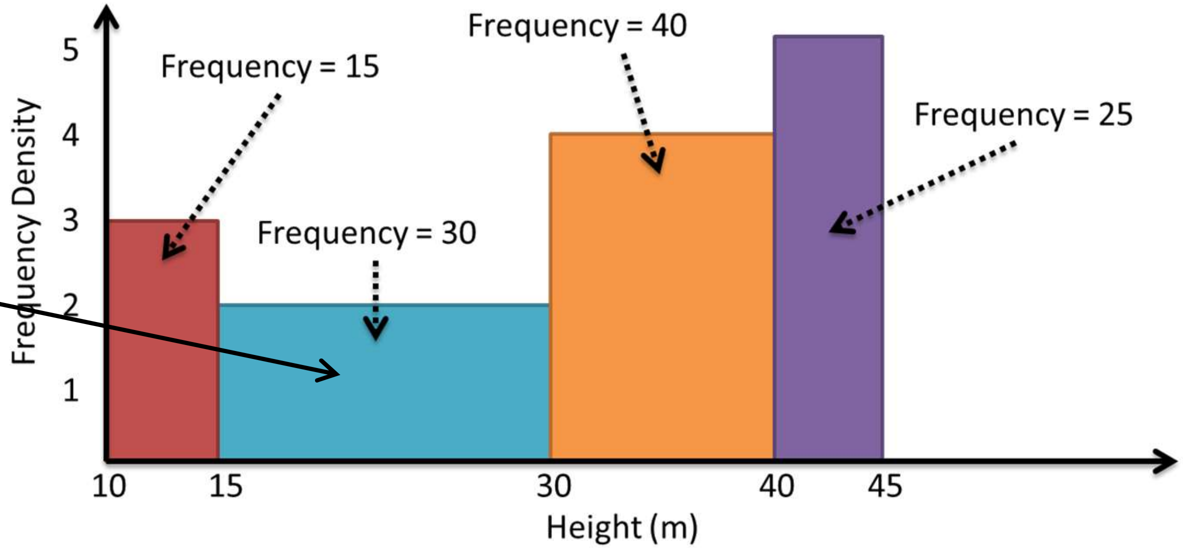
Weight (w kg)	Frequency	Frequency Density
0 < w ≤ 10	40	4
10 < w ≤ 15	6	1.2
15 < w ≤ 35	52	2.6
35 < w ≤ 45	10	1

$$\text{f.d.} = \frac{\text{freq.}}{\text{c.w.}}$$

e.g. $\frac{6}{15-10} = 1.2$

because

$$\underset{\text{freq.}}{30} = \underset{\text{c.w.}}{(30-15)} \times \underset{\text{f.d.}}{2}$$

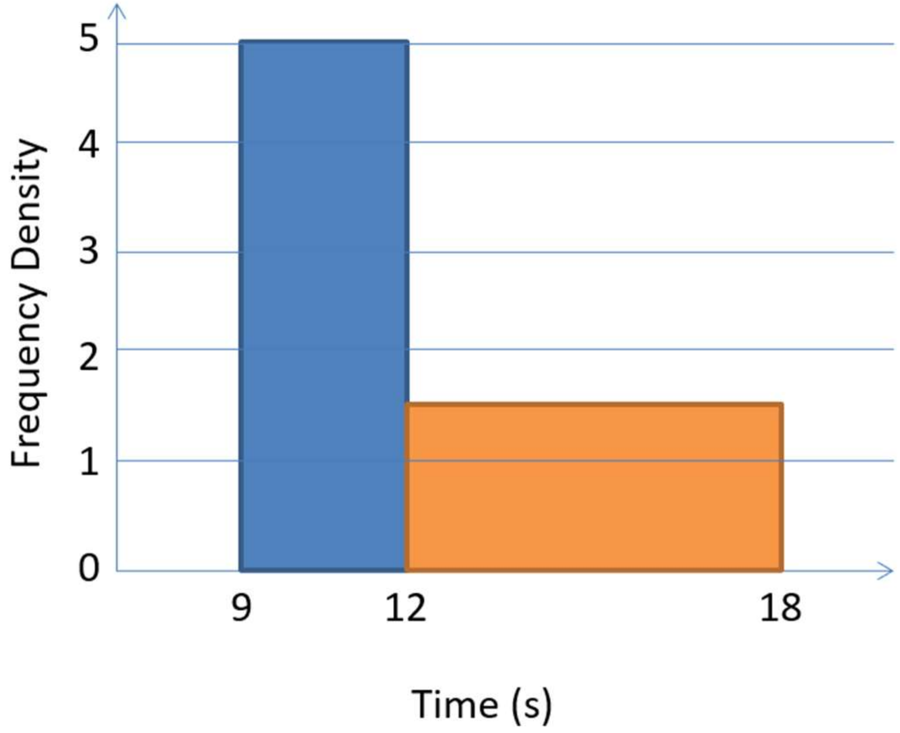


Frequency NOT ALWAYS = Area at A level

Unlike at GCSE, the area of a bar is not necessarily equal to the frequency; there are just **proportional**.

Identify the scaling $area \xrightarrow{\times k} frequency$ using a known area with known frequency (which may be total area/frequency or just one bar)

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



Total frequency is known; therefore find total area and hence the 'scaling'.

Total area = 15 + 9 = 24

Area		Freq	
24	$\xrightarrow{\times k}$	60	$k = \frac{60}{24} = 2.5$

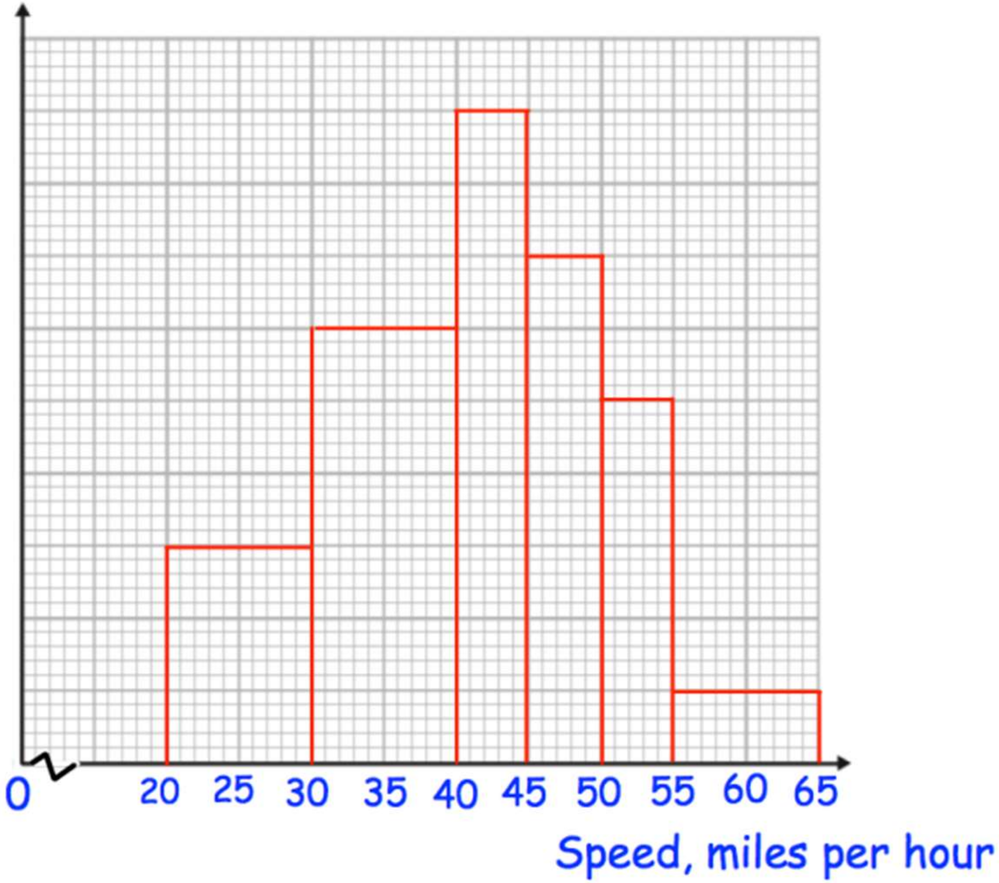
Then use this scaling along with the desired area.

$$\text{Area} = 4 \times 1.5 = 6$$
$$\text{Frequency} = 6 \times 2.5 = 15$$

Notes

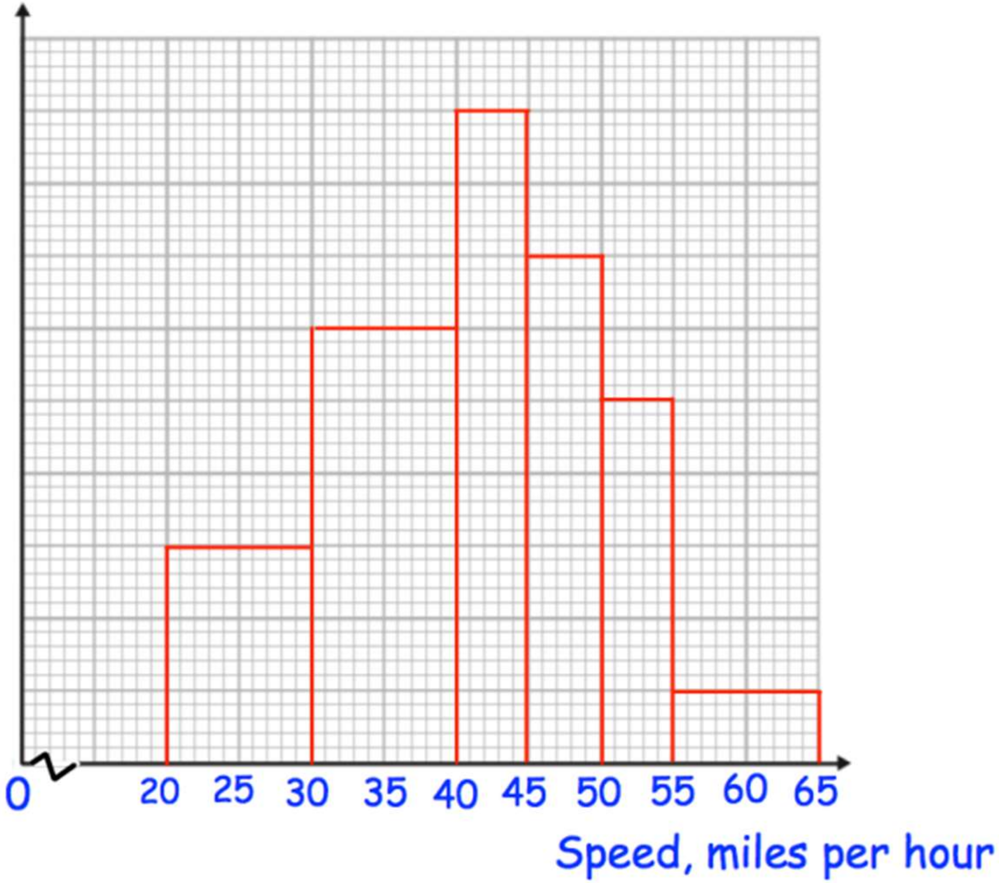
Worked Example

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

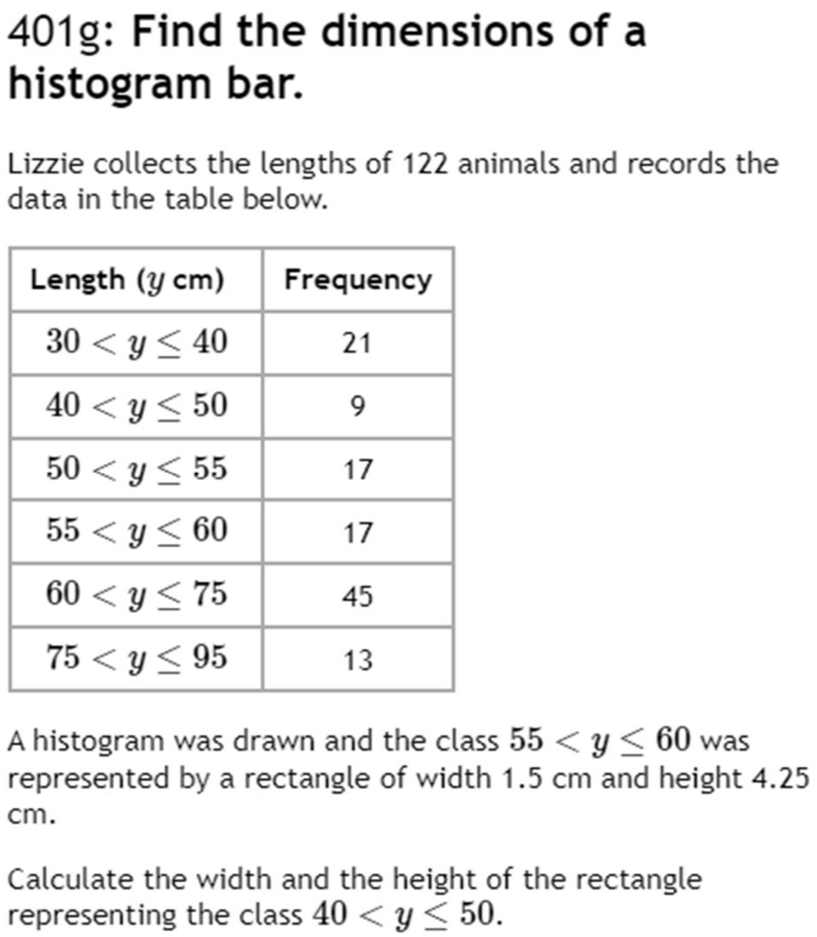


Worked Example

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



Worked Example



401g: **Find the dimensions of a histogram bar.**

Lizzie collects the lengths of 122 animals and records the data in the table below.

Length (y cm)	**Frequency**
$30 < y \leq 40$	21
$40 < y \leq 50$	9
$50 < y \leq 55$	17
$55 < y \leq 60$	17
$60 < y \leq 75$	45
$75 < y \leq 95$	13

A histogram was drawn and the class $55 < y \leq 60$ was represented by a rectangle of width 1.5 cm and height 4.25 cm.

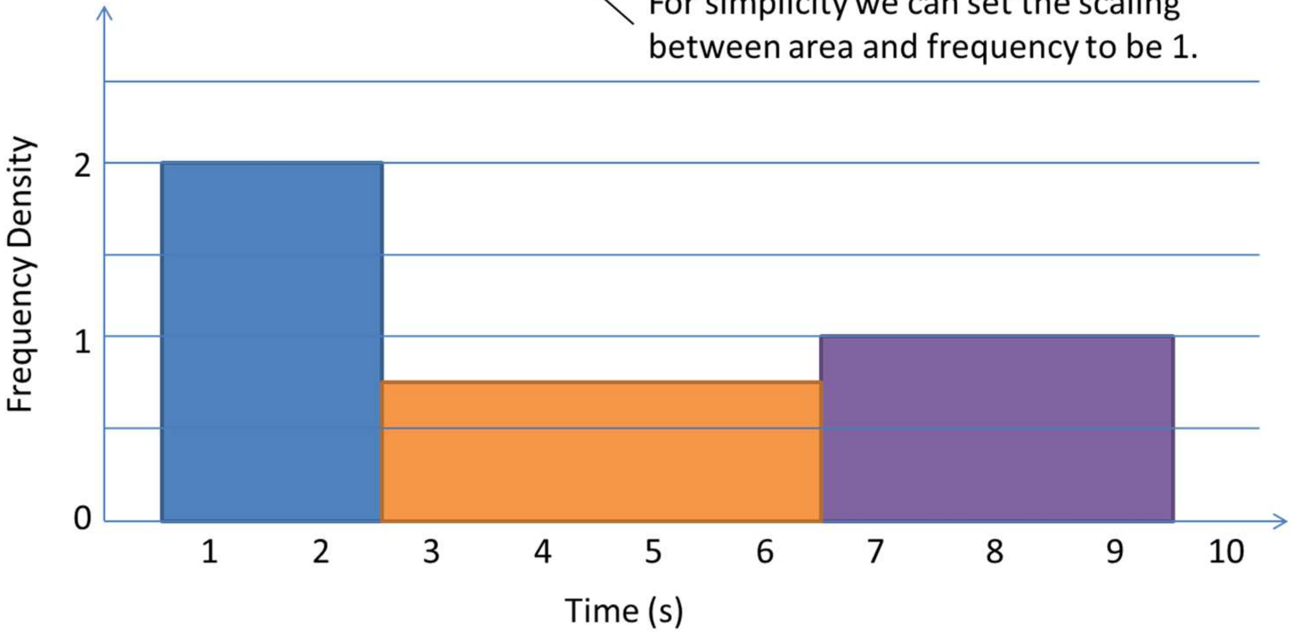
Calculate the width and the height of the rectangle representing the class $40 < y \leq 50$.

Gaps – error intervals

Weight (to nearest kg)	Frequency	F.D.
1-2	4	$\mathbf{4 \div 2 = 2}$
3-6	3	$\mathbf{3 \div 4 = 0.75}$
7-9	$3 \times 1 = 3$	$\mathbf{1}$

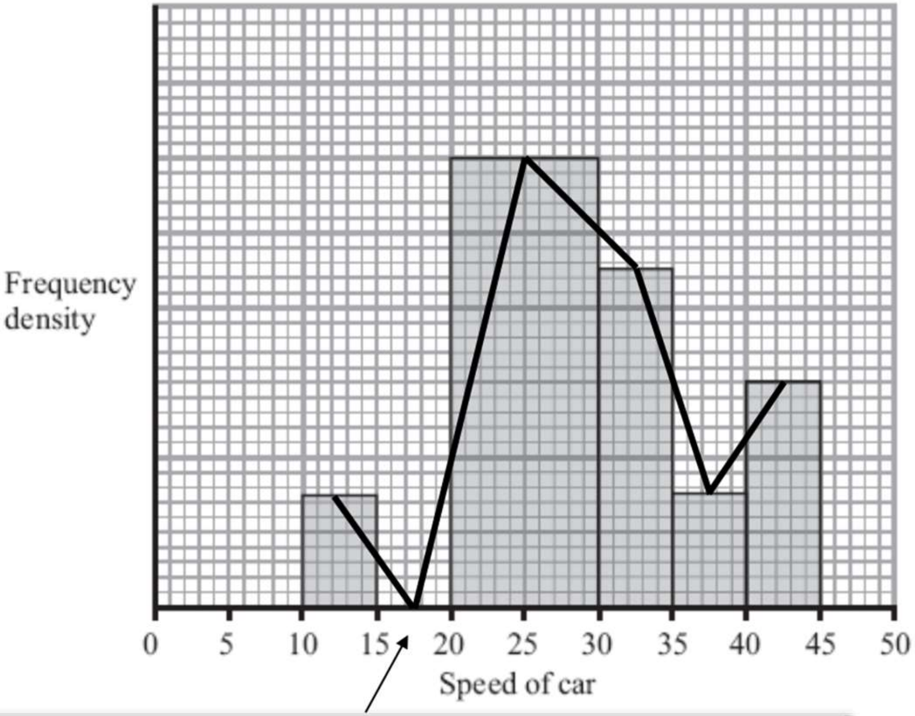
$1-2 \equiv \underbrace{0.5 \leq w < 1.5}_{c.w = 2}$

For simplicity we can set the scaling between area and frequency to be 1.



Forming a frequency polygon

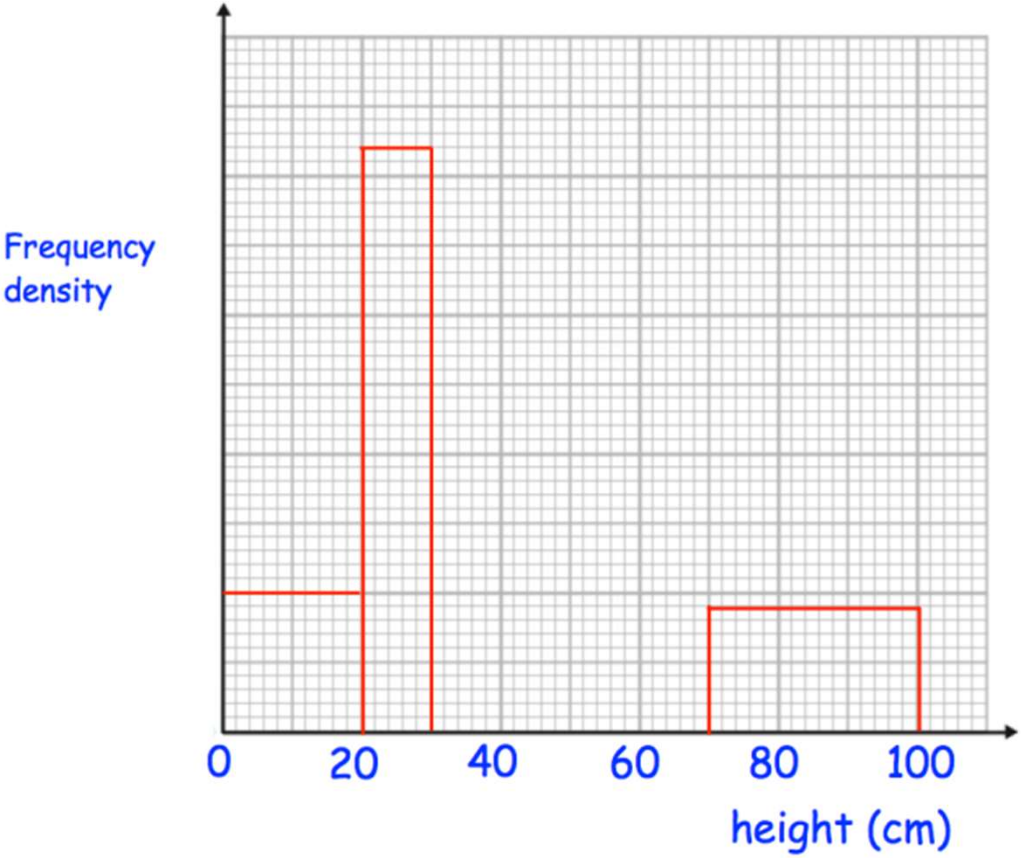
Recall that a frequency polygon can be drawn by using the midpoint of each interval. This corresponds to the midpoint of the top of each bar in a histogram.



Note that the frequency in this interval is 0. That needs to be reflected in the frequency polygon.

Worked Example

Draw a frequency polygon.



Worked Example

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.

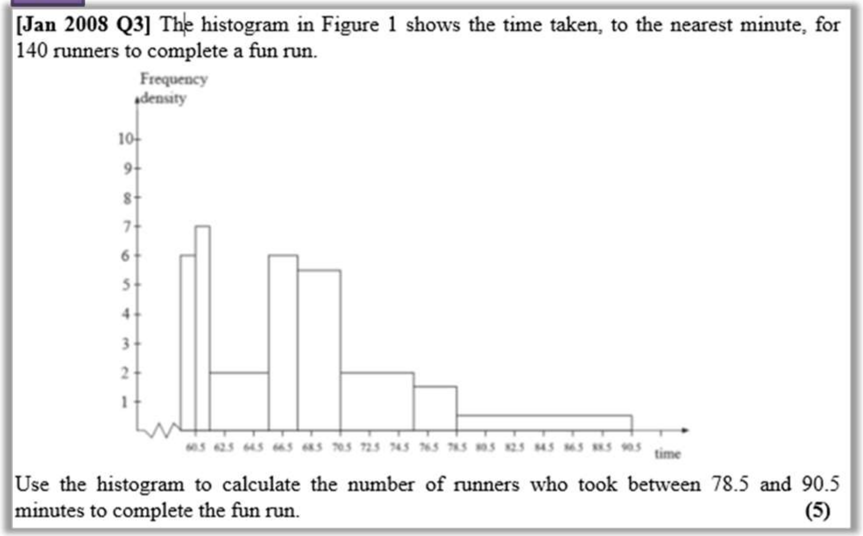
On the histogram the rectangle representing the $16 \leq T < 18$ class was 3.2 cm high and 2 cm wide. The frequency for this class was 8.

c) Show that each day is represented by an area of 0.8 cm^2.

d) Given that the total area under the histogram was 48 cm^2, find the total number of days in the sample.

Supplementary Exercise

Q1



[Jan 2008 Q3] The histogram in Figure 1 shows the time taken, to the nearest minute, for 140 runners to complete a fun run.

Use the histogram to calculate the number of runners who took between 78.5 and 90.5 minutes to complete the fun run. **(5)**

Q2

The following table summarises the distances, to the nearest km, that 134 examiners travelled to attend a meeting in London.

Distance (km)	Number of examiners
41–45	4
46–50	19
51–60	53
61–70	37
71–90	15
91–150	6

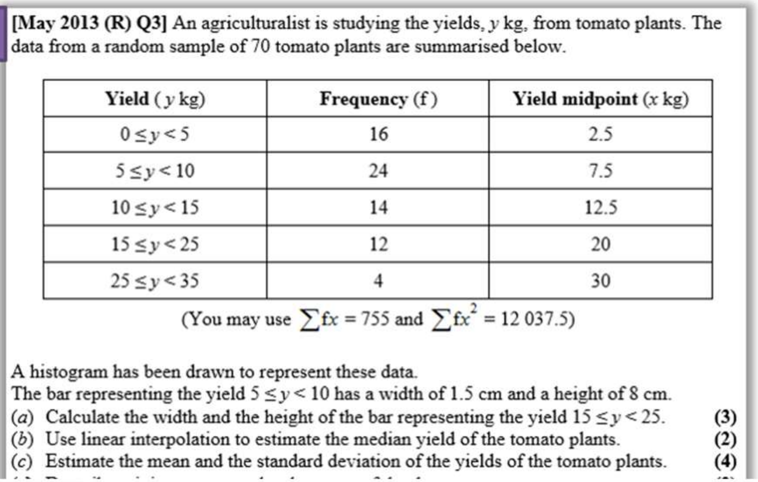
(*a*) Give a reason to justify the use of a histogram to represent these data. **(1)**

?

(*b*) Calculate the frequency densities needed to draw a histogram for these data. **(2)**

(DO NOT DRAW THE HISTOGRAM)

Q3



[May 2013 (R) Q3] An agriculturalist is studying the yields, y kg, from tomato plants. The data from a random sample of 70 tomato plants are summarised below.

Yield (y kg)	Frequency (f)	Yield midpoint (x kg)
$0 \leq y < 5$	16	2.5
$5 \leq y < 10$	24	7.5
$10 \leq y < 15$	14	12.5
$15 \leq y < 25$	12	20
$25 \leq y < 35$	4	30

(You may use $\sum fx = 755$ and $\sum fx^2 = 12\,037.5$)

A histogram has been drawn to represent these data.
The bar representing the yield $5 \leq y < 10$ has a width of 1.5 cm and a height of 8 cm.

(*a*) Calculate the width and the height of the bar representing the yield $15 \leq y < 25$. **(3)**

(*b*) Use linear interpolation to estimate the median yield of the tomato plants. **(2)**

(*c*) Estimate the mean and the standard deviation of the yields of the tomato plants. **(4)**

Q4 **[June 2007 Q5]**

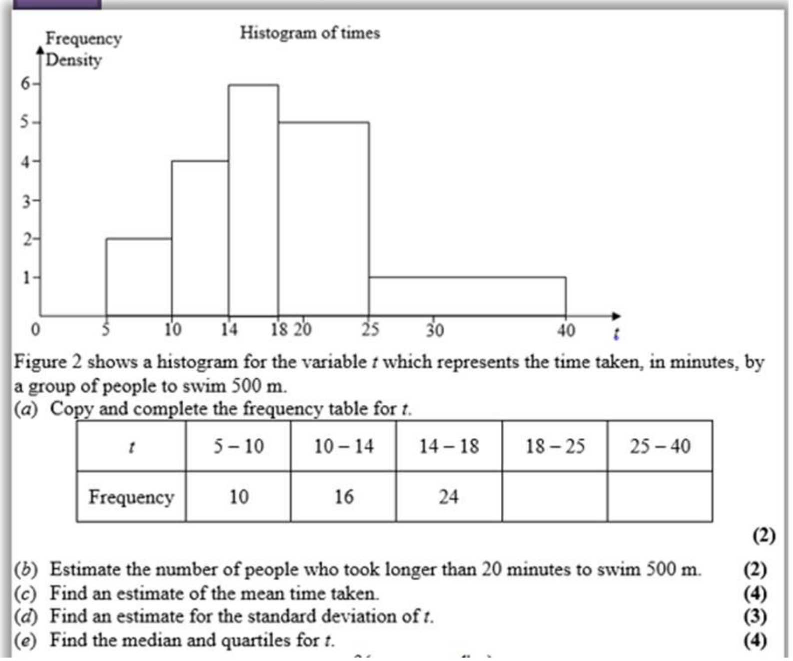


Figure 2 shows a histogram for the variable t which represents the time taken, in minutes, by a group of people to swim 500 m.

(*a*) Copy and complete the frequency table for t.

t	5 – 10	10 – 14	14 – 18	18 – 25	25 – 40
Frequency	10	16	24		

(2)

(*b*) Estimate the number of people who took longer than 20 minutes to swim 500 m. **(2)**

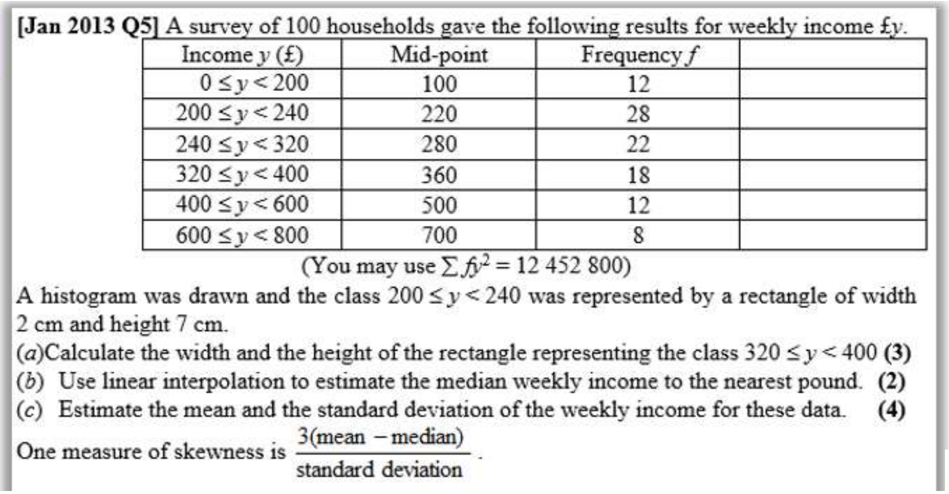
(*c*) Find an estimate of the mean time taken. **(4)**

(*d*) Find an estimate for the standard deviation of t. **(3)**

(*e*) Find the median and quartiles for t. **(4)**

Supplementary Exercise

Q5



[Jan 2013 Q5] A survey of 100 households gave the following results for weekly income £y.

Income y (£)	Mid-point	Frequency f	
$0 \le y < 200$	100	12	
$200 \le y < 240$	220	28	
$240 \le y < 320$	280	22	
$320 \le y < 400$	360	18	
$400 \le y < 600$	500	12	
$600 \le y < 800$	700	8	

(You may use $\sum fy^2 = 12\,452\,800$)

A histogram was drawn and the class $200 \le y < 240$ was represented by a rectangle of width 2 cm and height 7 cm.

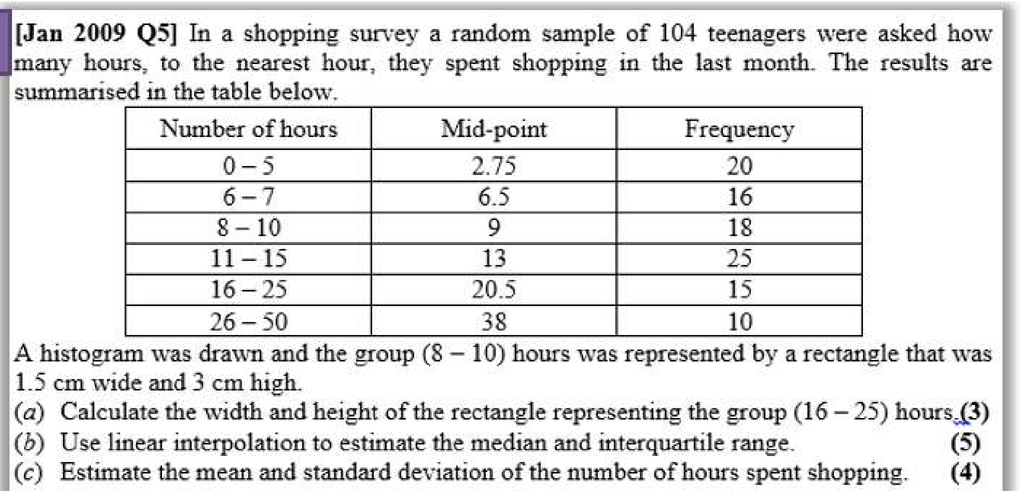
(*a*)Calculate the width and the height of the rectangle representing the class $320 \le y < 400$ **(3)**

(*b*) Use linear interpolation to estimate the median weekly income to the nearest pound. **(2)**

(*c*) Estimate the mean and the standard deviation of the weekly income for these data. **(4)**

One measure of skewness is $\dfrac{3(\text{mean} - \text{median})}{\text{standard deviation}}$.

Q7



[Jan 2009 Q5] In a shopping survey a random sample of 104 teenagers were asked how many hours, to the nearest hour, they spent shopping in the last month. The results are summarised in the table below.

Number of hours	Mid-point	Frequency
0 – 5	2.75	20
6 – 7	6.5	16
8 – 10	9	18
11 – 15	13	25
16 – 25	20.5	15
26 – 50	38	10

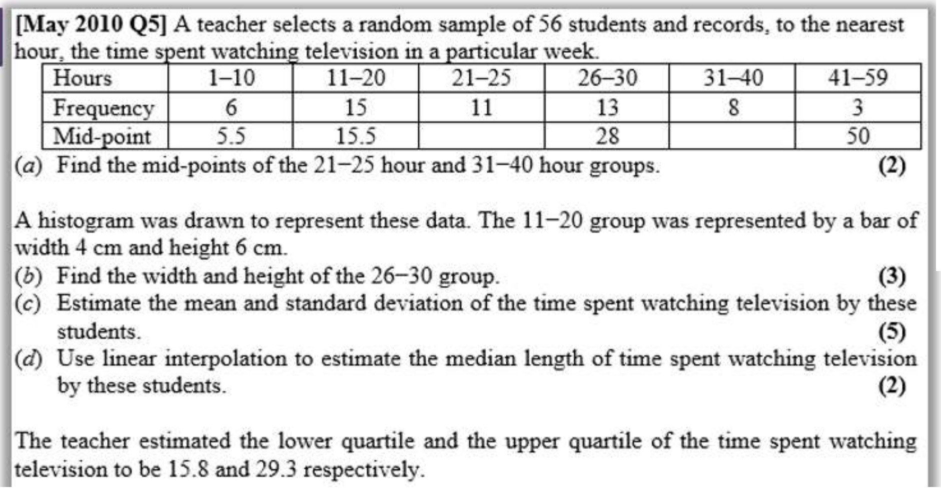
A histogram was drawn and the group (8 – 10) hours was represented by a rectangle that was 1.5 cm wide and 3 cm high.

(*a*) Calculate the width and height of the rectangle representing the group (16 – 25) hours.**(3)**

(*b*) Use linear interpolation to estimate the median and interquartile range. **(5)**

(*c*) Estimate the mean and standard deviation of the number of hours spent shopping. **(4)**

Q6



[May 2010 Q5] A teacher selects a random sample of 56 students and records, to the nearest hour, the time spent watching television in a particular week.

Hours	1–10	11–20	21–25	26–30	31–40	41–59
Frequency	6	15	11	13	8	3
Mid-point	5.5	15.5		28		50

(*a*) Find the mid-points of the 21–25 hour and 31–40 hour groups. **(2)**

A histogram was drawn to represent these data. The 11–20 group was represented by a bar of width 4 cm and height 6 cm.

(*b*) Find the width and height of the 26–30 group. **(3)**

(*c*) Estimate the mean and standard deviation of the time spent watching television by these students. **(5)**

(*d*) Use linear interpolation to estimate the median length of time spent watching television by these students. **(2)**

The teacher estimated the lower quartile and the upper quartile of the time spent watching television to be 15.8 and 29.3 respectively.

3.5 Comparing Data

Worked Example

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 $\sum x^2 = 10301.2$

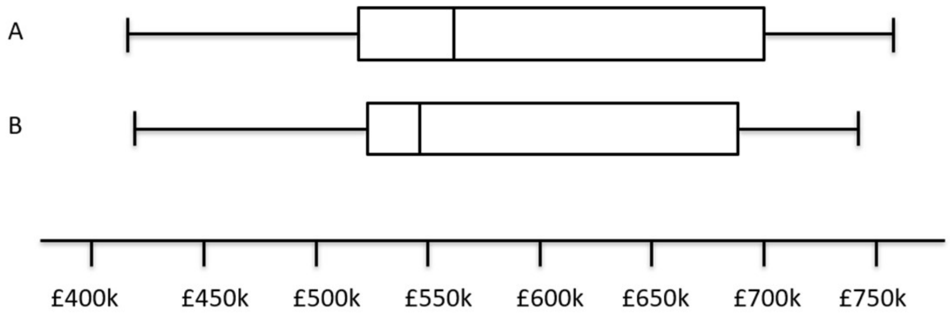
a) Calculate the mean and standard deviation for Heathrow

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

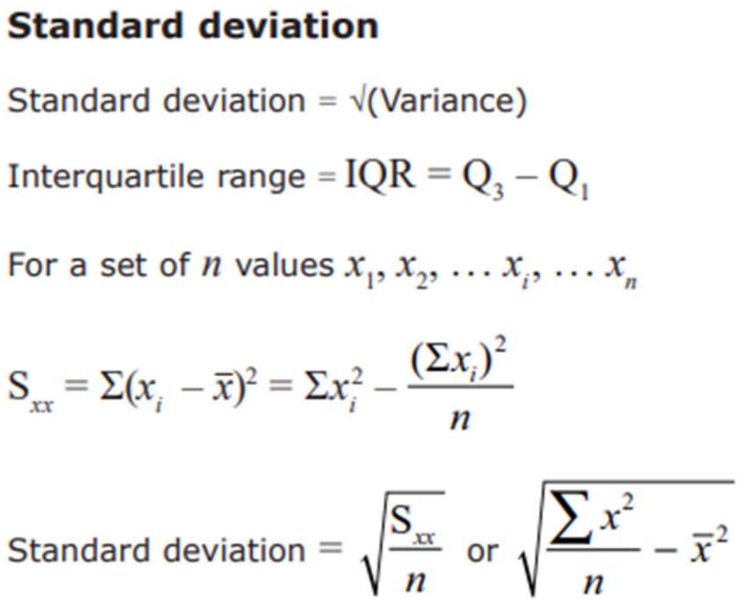
b) Compare the data for the two locations using the information given

Worked Example

Compare the house prices of locations A and B



Extract from Formulae book



Standard deviation

Standard deviation = √(Variance)

Interquartile range = $\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$

For a set of n values $x_1, x_2, \ldots x_i, \ldots x_n$

$$\mathrm{S}_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$$

Standard deviation = $\sqrt{\frac{\mathrm{S}_{xx}}{n}}$ or $\sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$

1.

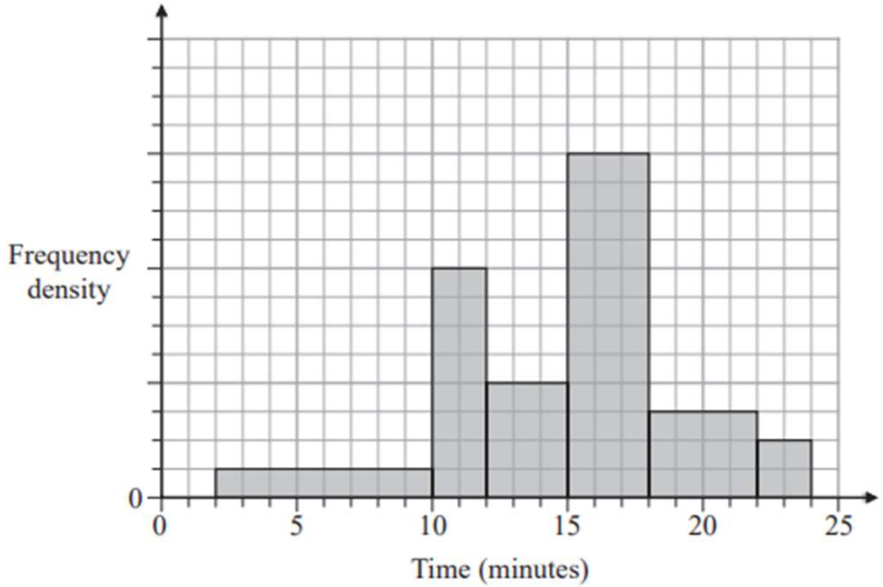


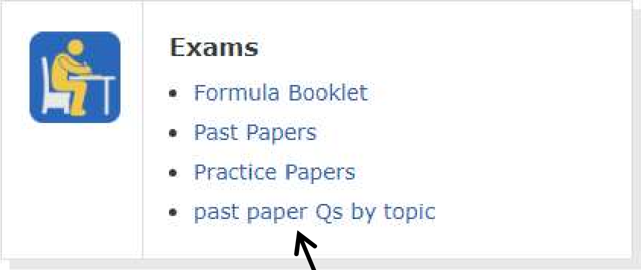
Figure 1

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)



Exams

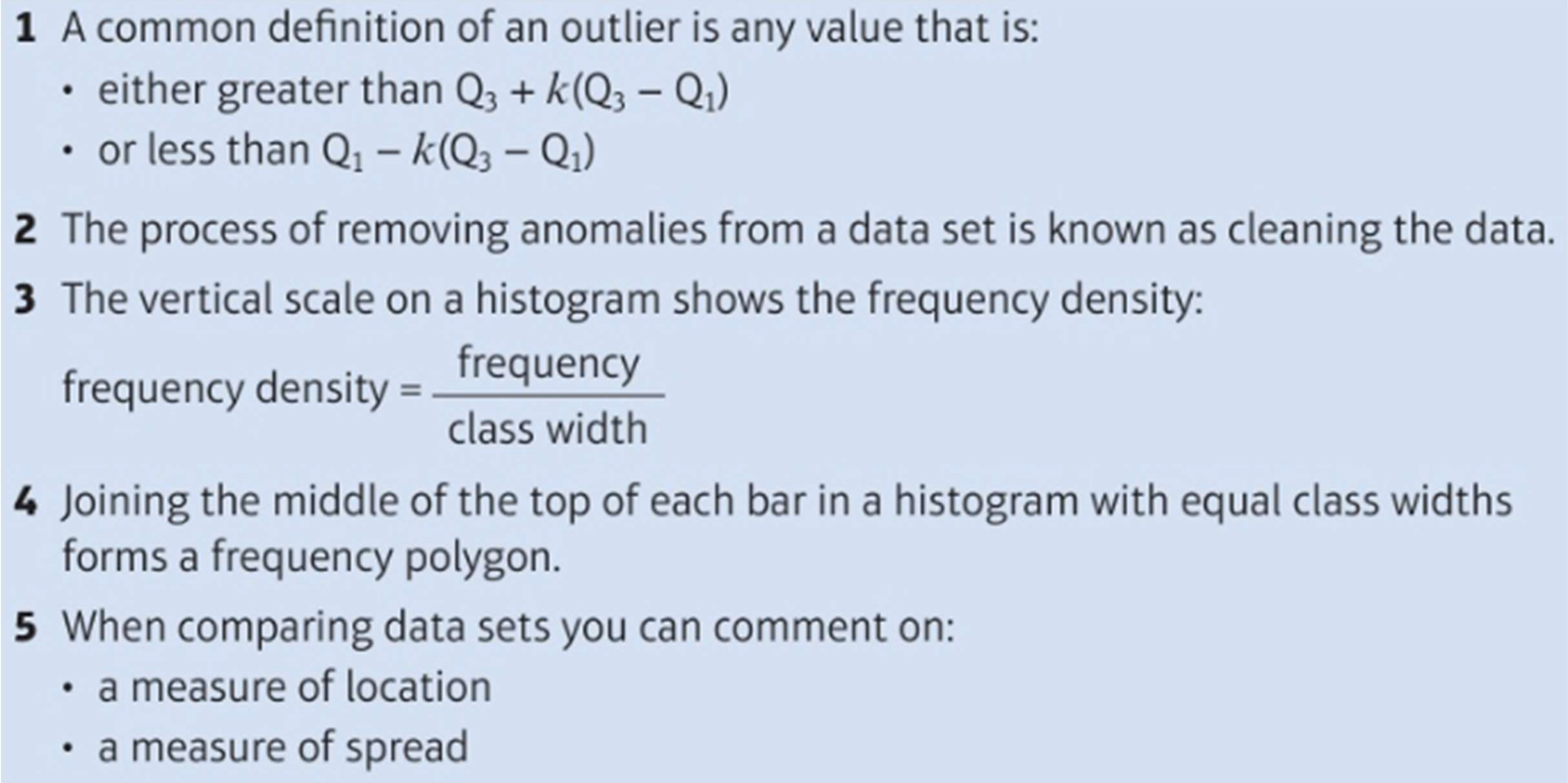
- Formula Booklet
- Past Papers
- Practice Papers
- past paper Qs by topic

Past paper practice by topic. Both new and old specification can be found via this link on hgsmaths.com

1 square is $\dfrac{78}{12\times3+3\times4+2\times2}=\left[\dfrac{78}{52}=1.5\right]$ **and** $(8\times1+1\times8)\times$"1.5"	M1	3.1a
24 students took less than 11 minutes	A1	1.1b
Percentage of students = $\dfrac{\text{"24"}}{78+\text{"24"}+1\times8\times\text{"1.5"}+3\times4\times\text{"1.5"}}\times100$	M1	3.1b
= 18.18... awrt 18%	A1	1.1b
	(4)	

Total 4

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 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