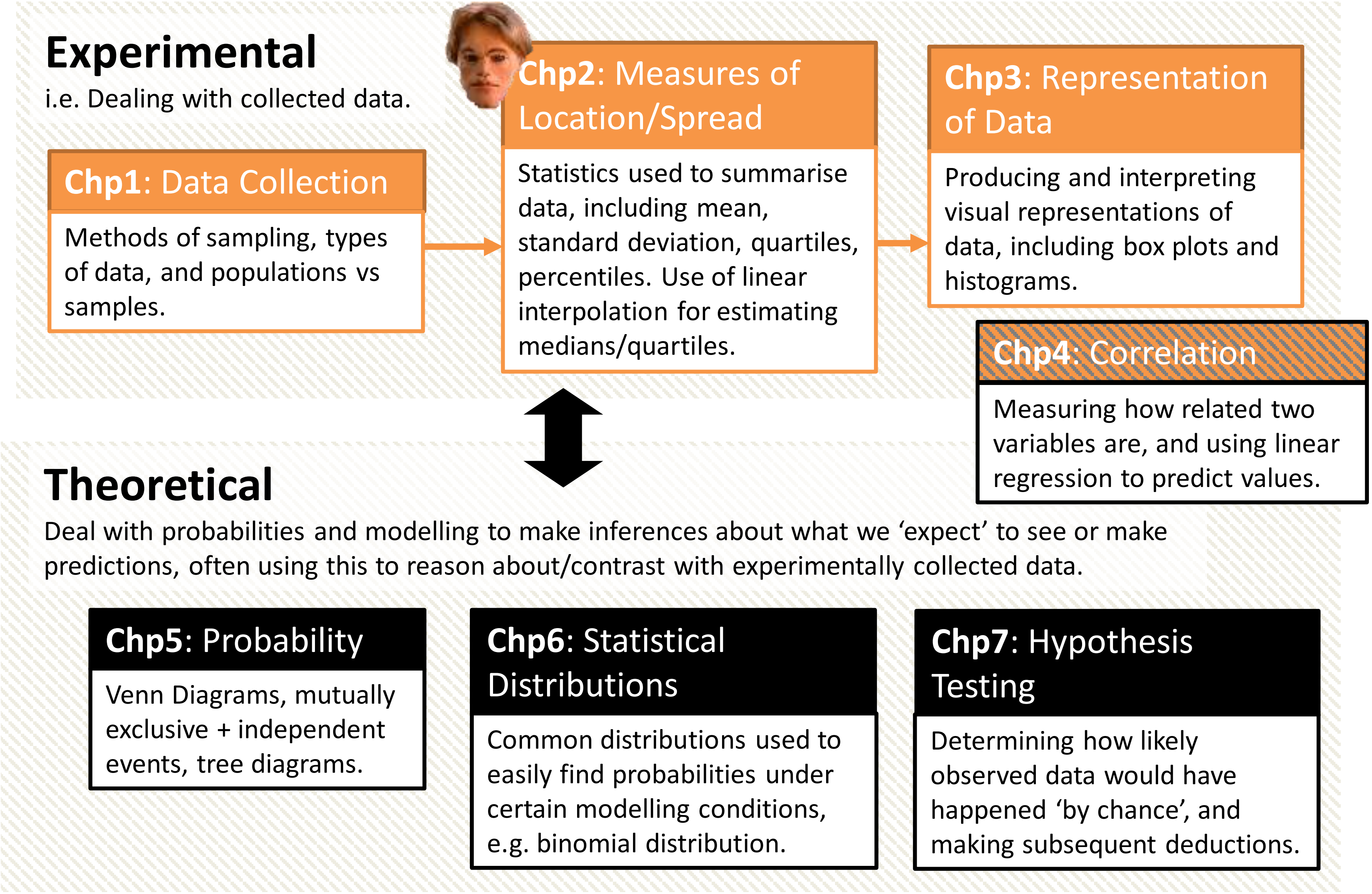
Stats1 Chapter 3 ::

Representations of Data

jfrost@tiffin.kingston.sch.uk

www.drfrostmaths.com
@DrFrostMaths

Last modified: 15th February 2019



Experimental

i.e. Dealing with collected data.

Chp1: Data Collection

Methods of sampling, types of data, and populations vs samples.

Chp2: Measures of Location/Spread

Statistics used to summarise data, including mean, standard deviation, quartiles, percentiles. Use of linear interpolation for estimating medians/quartiles.

Chp3: Representation of Data

Producing and interpreting visual representations of data, including box plots and histograms.

Chp4: Correlation

Measuring how related two variables are, and using linear regression to predict values.

Theoretical

Deal with probabilities and modelling to make inferences about what we ‘expect’ to see or make predictions, often using this to reason about/contrast with experimentally collected data.

Chp5: Probability

Venn Diagrams, mutually exclusive + independent events, tree diagrams.

Chp6: Statistical Distributions

Common distributions used to easily find probabilities under certain modelling conditions, e.g. binomial distribution.

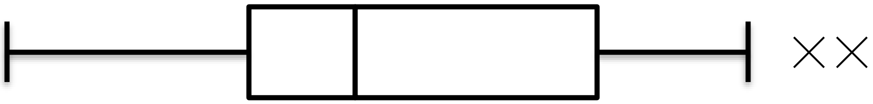
Chp7: Hypothesis Testing

Determining how likely observed data would have happened ‘by chance’, and making subsequent deductions.

This Chapter Overview

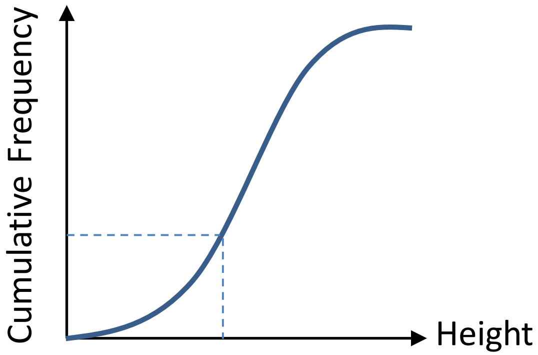
We've seen so far how data is collected and calculations can be made. We now concentrate on how the processed data can be *displayed*.

BOX PLOTS AND OUTLIERS

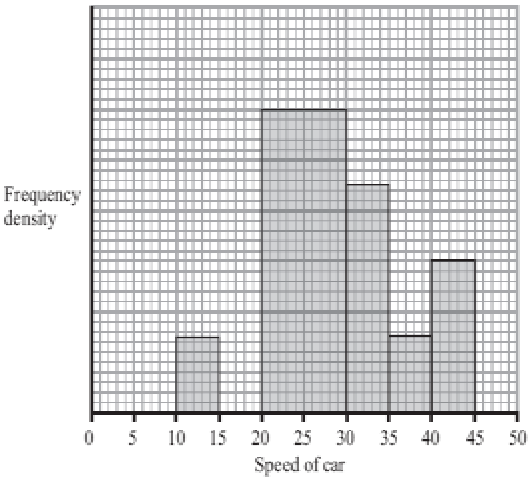


NEW since GCSE! Outliers.

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:

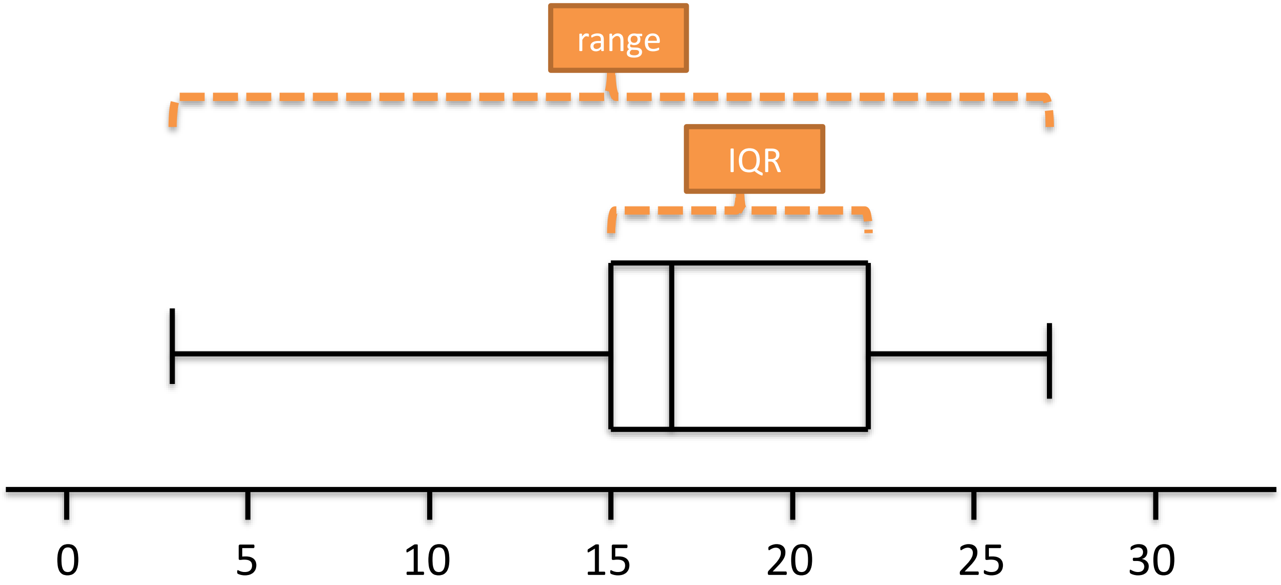
- Stem and leaf diagrams have been cut. (THANK GOD FOR THAT)
- 'Skew' has been cut.
- Cumulative frequency diagrams have been added.
- Turning histogram into frequency polygon.

Box Plot recap

Box Plots allow us to visually represent the distribution of the data.

Minimum	Lower Quartile	Median	Upper Quartile	Maximum
3	15	17	22	27

Sketch Sketch Sketch Sketch Sketch


range

IQR

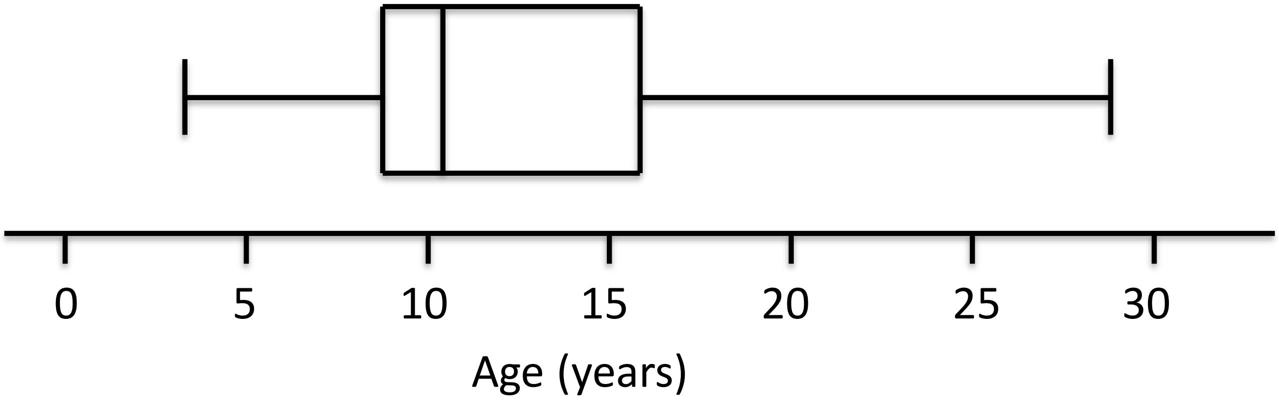
How is the **IQR** represented in this diagram?

Sketch

How is the **range** represented in this diagram?

Sketch

Interpreting a Box Plot



True or false: (click your answer)

“The right box represents more people than the left box.”

False True

Each box represents 25% of people, i.e. the same number of people!

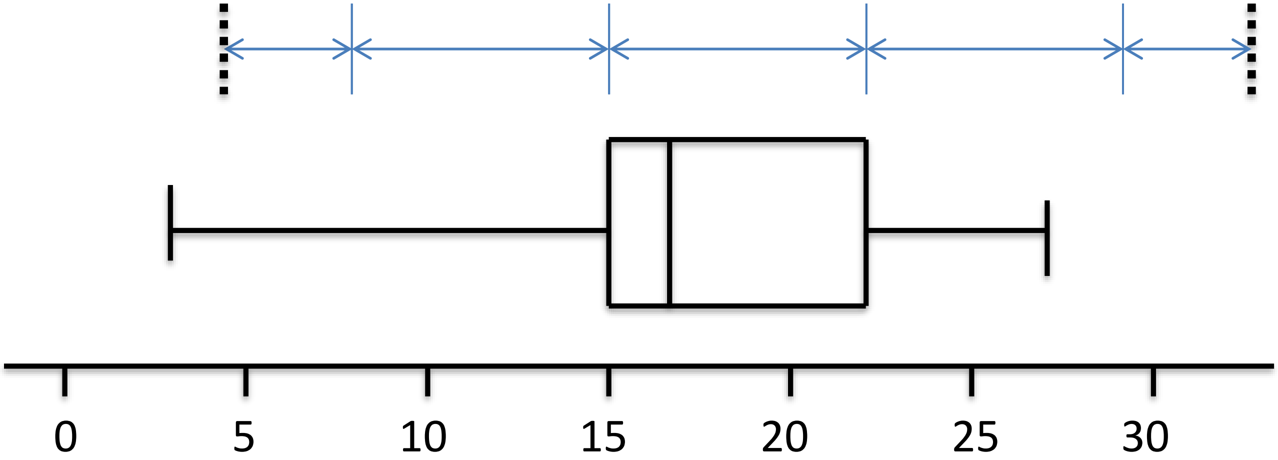
“The ages are more spread out above the median.”

False True

The wider the box or whisker, the more spread out the values are within that 25% of the data. We’d say that the data has “**positive skew**”, but you are not required to know this term.

Outliers

An outlier is: ?



Outliers beyond this point

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)

Examples

The diameters of 11 different Roman coins are measured in centimetres:

2.2 2.5 2.7 2.7 2.8 3.0 3.1 3.1 3.2 4.0 4.7

Determine the quartiles and hence any outliers.

?

[Textbook] 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.

? a

? b

Context: Recall that the standard deviation is, roughly speaking, the average distance of each value from the mean. So the outlier definition is saying we're at least twice this average distance, which seems like a sensible definition.

In Year 2, you will encounter the **normal distribution**, which can be used to model data which is **clustered about some mean and tails off symmetrical in either direction**. If this data was approximately normally distributed, then there is a 5% chance a random observation would fall outside 2 standard deviations within the mean. You will learn then how to make such probability calculations.

Test Your Understanding

The ages of 15 Lib Dem MPs are given:

11 18 20 27 30 31 32 32 35 36 37 58 63 78 104

a) If an outlier is considered to be 1.5 interquartile ranges below the lower quartile or above the upper quartile, determine any outliers.

b) If instead an outlier is considered to be outside 2 standard deviations within the mean, determine any outliers. Note that $\Sigma x = 612$ and $\Sigma x^2 = 33606$

? a

? b

Box Plot Example

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

Draw a box plot to represent the above data.

?

Exam Tip: You MUST show your outlier boundary calculations.

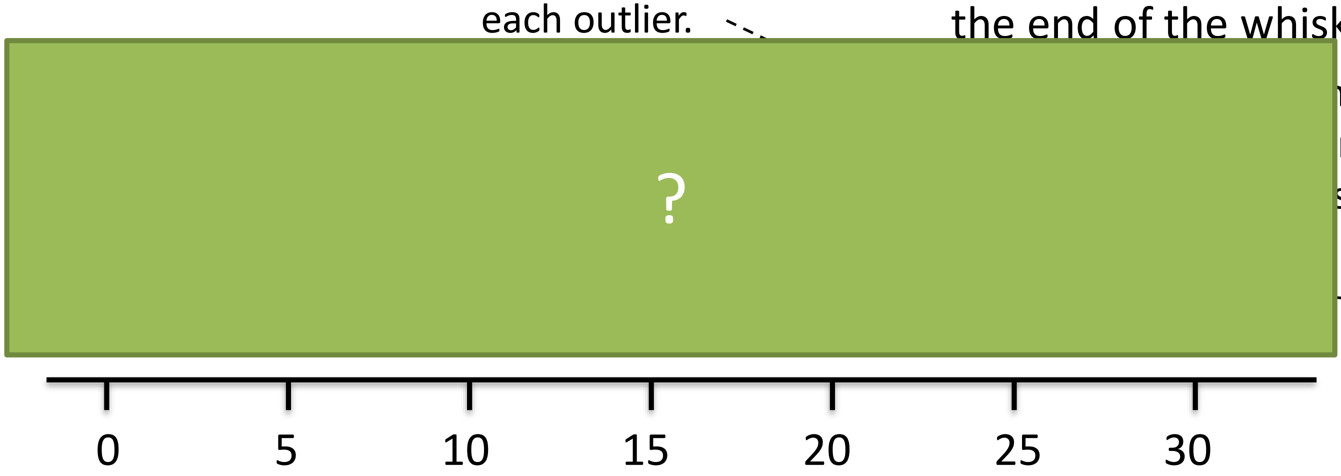
Use a cross for each outlier.

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

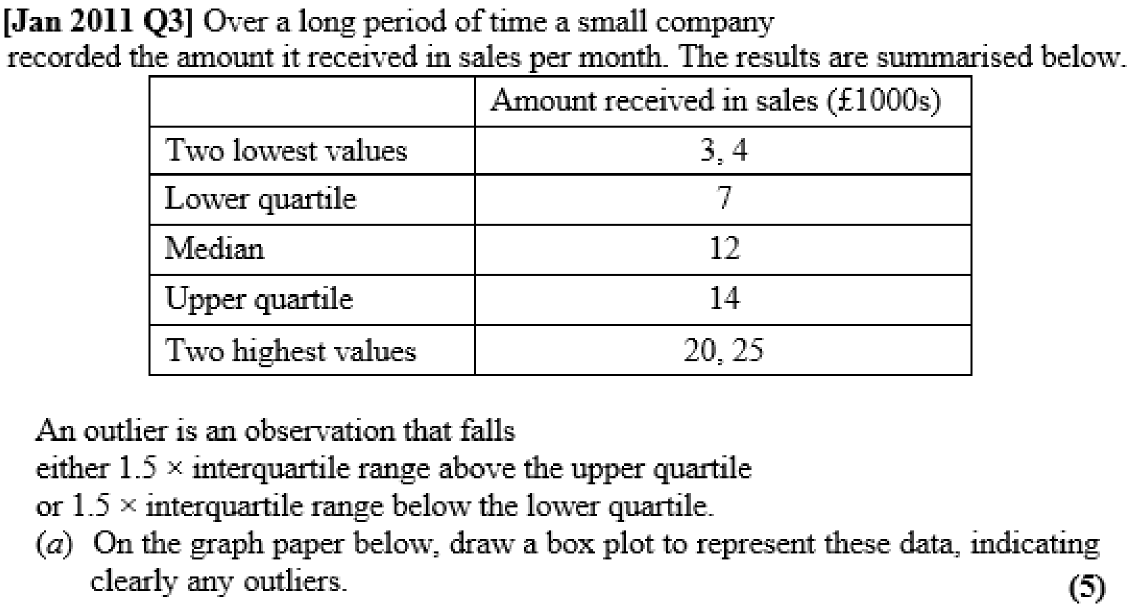
aximum value not an r, 21 (I think this one s most sense).

OR the outlier boundary, 23.

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



Test Your Understanding

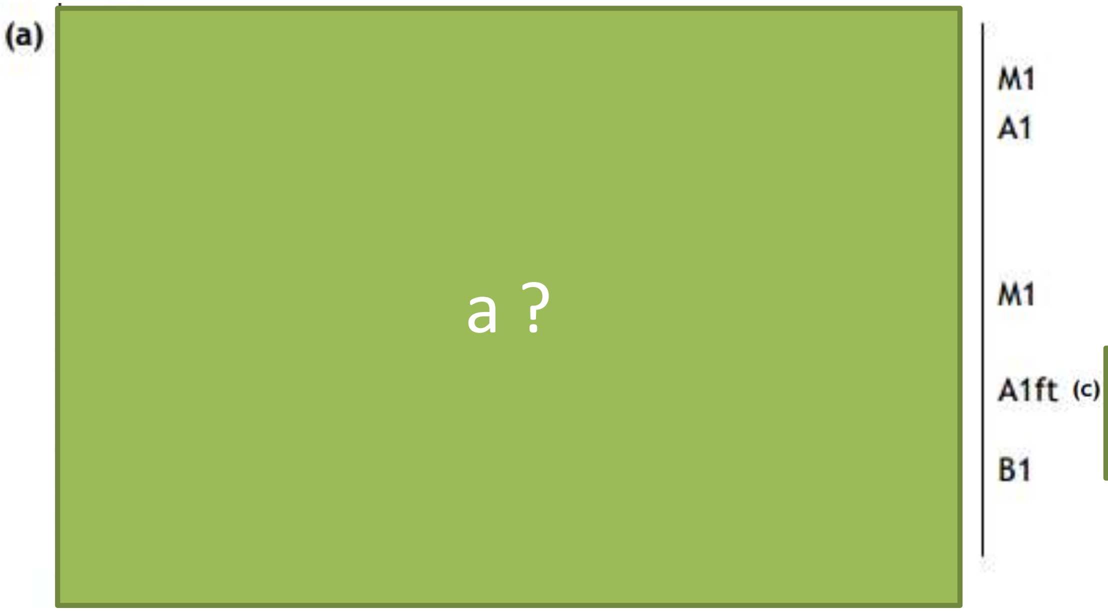


[Jan 2011 Q3] Over a long period of time a small company recorded the amount it received in sales per month. The results are summarised below.

	Amount received in sales (£1000s)
Two lowest values	3, 4
Lower quartile	7
Median	12
Upper quartile	14
Two highest values	20, 25

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 the graph paper below, draw a box plot to represent these data, indicating clearly any outliers. **(5)**



(a)

a ?

M1

A1

M1

A1ft (c)

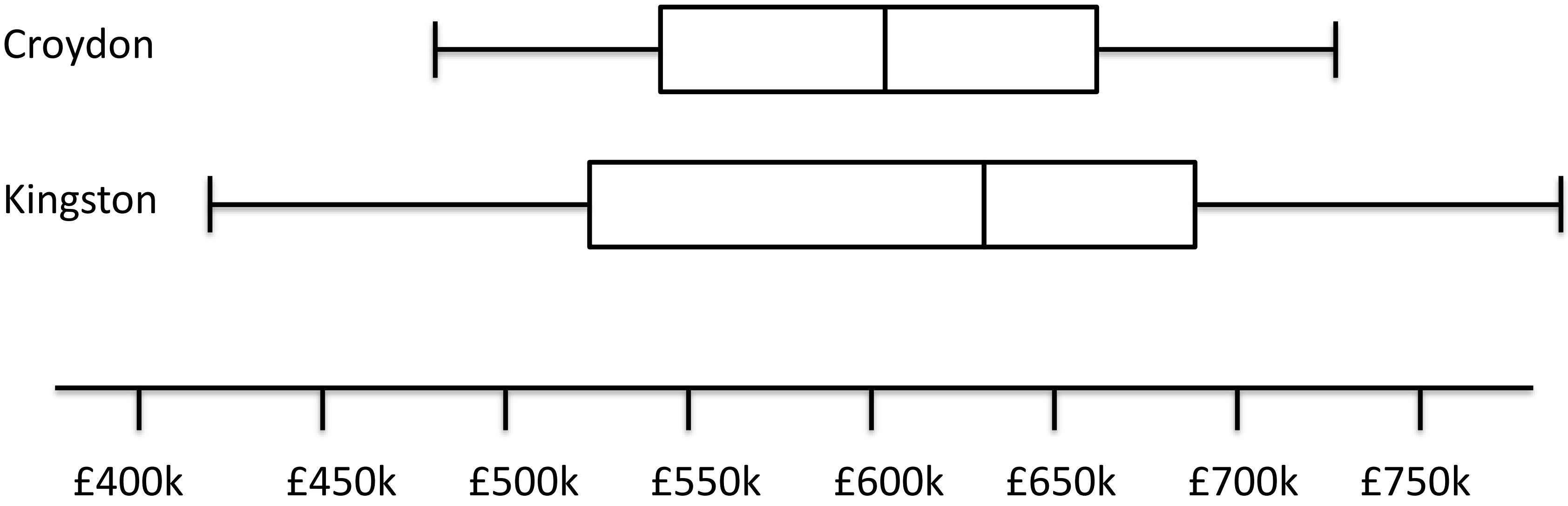
B1

(c) The company claims that for 75% of the months, the amount received per month is greater than £10 000. Comment on this claim, giving a reason for your answer. **(2)**

c ?

Comparing Box Plots

Box Plot comparing house prices of Croydon and Kingston-upon-Thames:



"Compare the prices of houses in Croydon with those in Kingston". (2 marks)

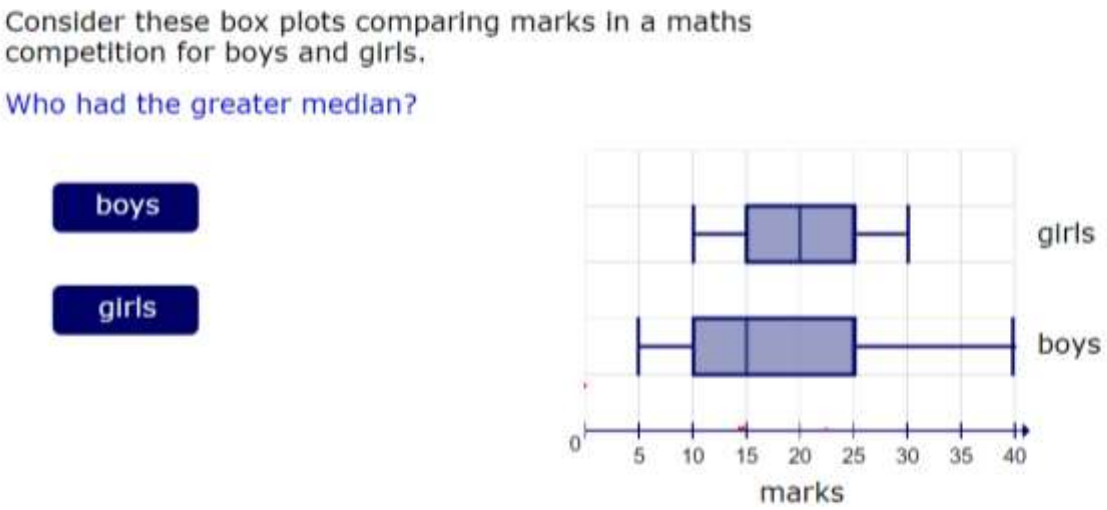
For 1 mark, one of:

?

For 1 mark:

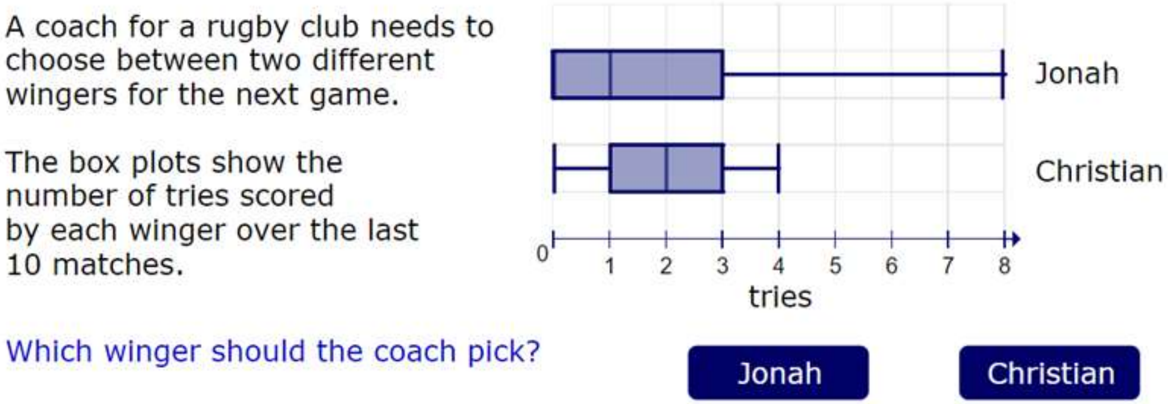
?

Comparing Box Plots



Consider these box plots comparing marks in a maths competition for boys and girls.

Who had the greater median?



A coach for a rugby club needs to choose between two different wingers for the next game.

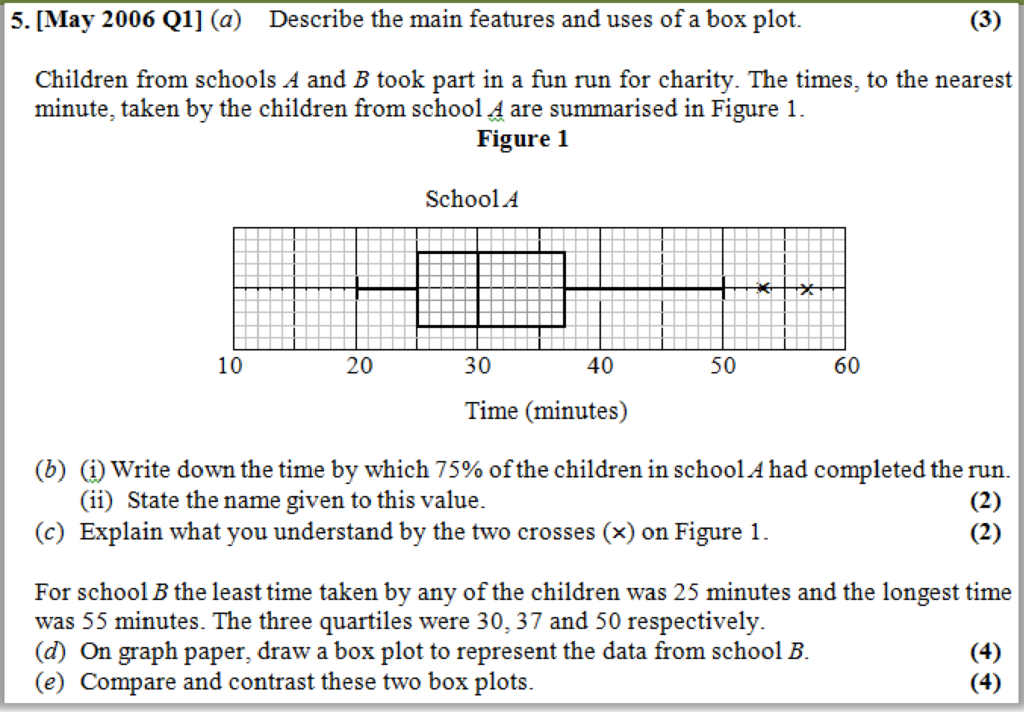
The box plots show the number of tries scored by each winger over the last 10 matches.

Which winger should the coach pick?

Exercise 3A/3B

Pearson Pure Mathematics Year 1/AS
Pages 42-43, 45

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)**

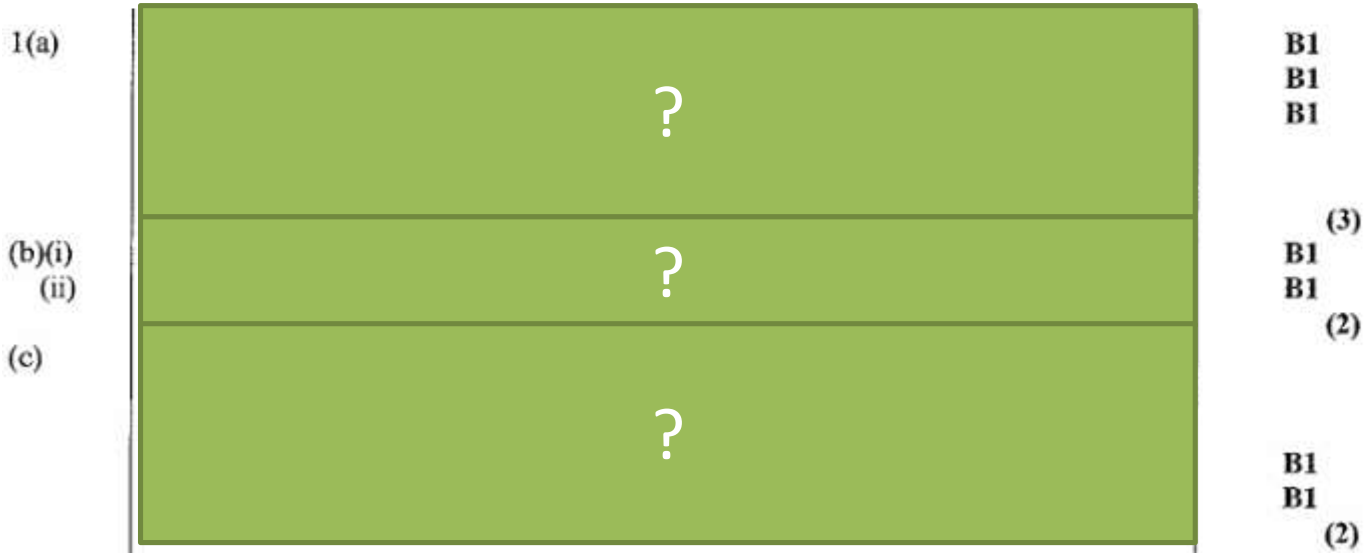
(*c*) Explain what you understand by the two crosses (×) on Figure 1. **(2)**

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)**

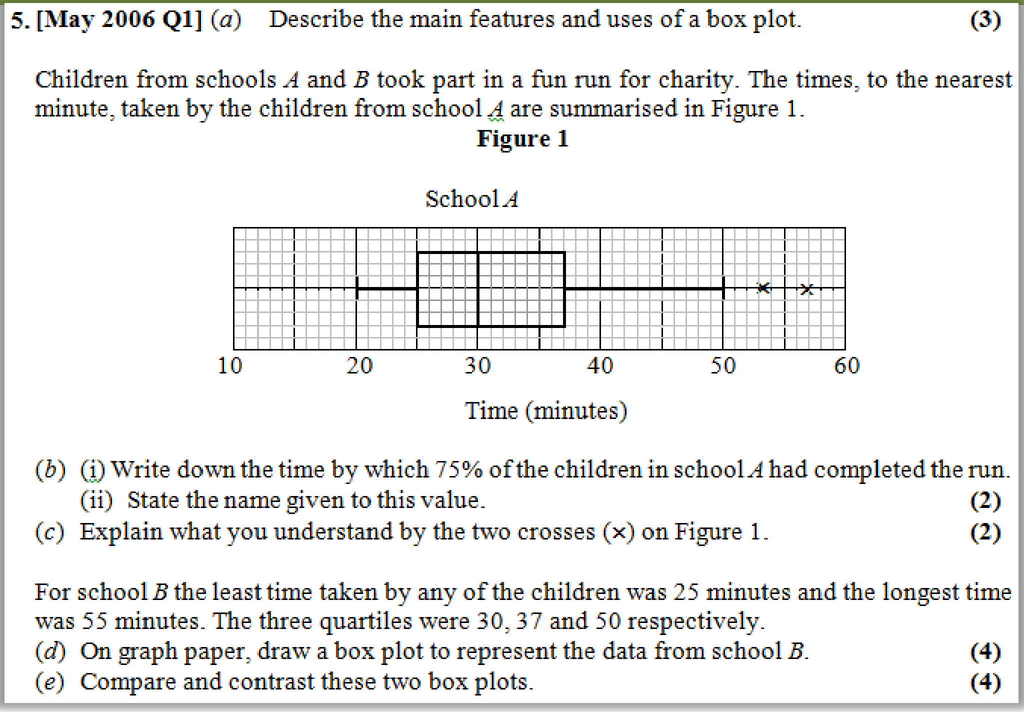
(*e*) Compare and contrast these two box plots. **(4)**

(Solutions to (d) and (e) on next slide)



1(a)	?	**B1** **B1** **B1** **(3)**
(b)(i) (ii)	?	**B1** **B1** **(2)**
(c)	?	**B1** **B1** **(2)**

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)**

(d)

?

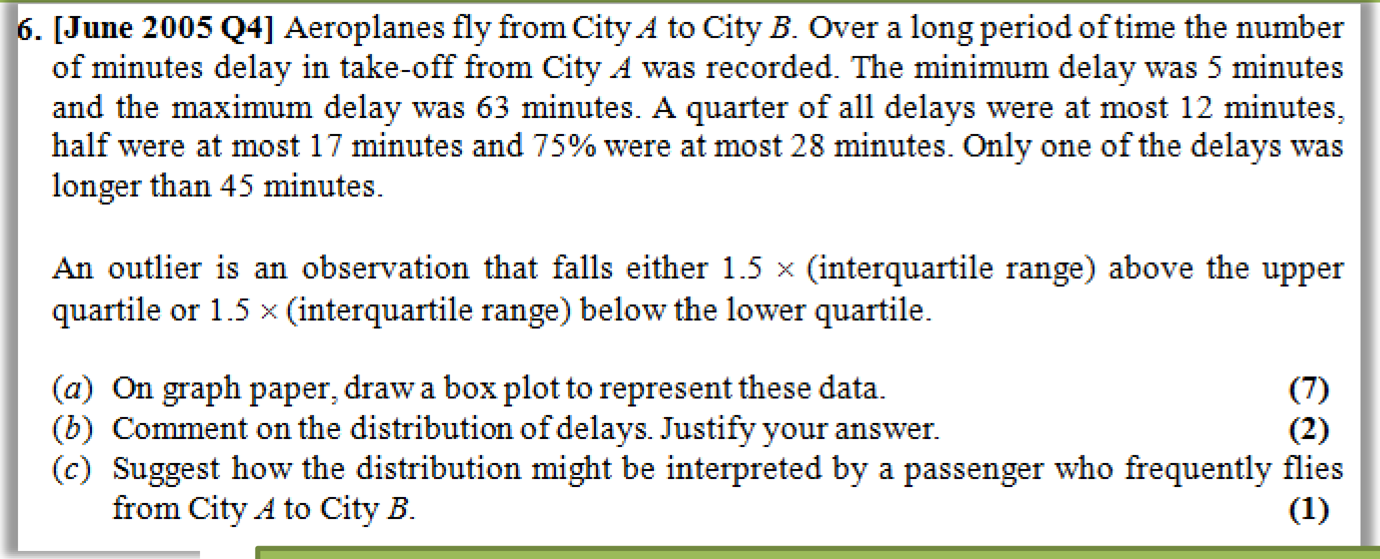
(e)

?

B1
B1
B1
B1

Supplementary Questions:

(on your printed sheet)



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)**

(a) ?

(b) ? B1; B1 (2)

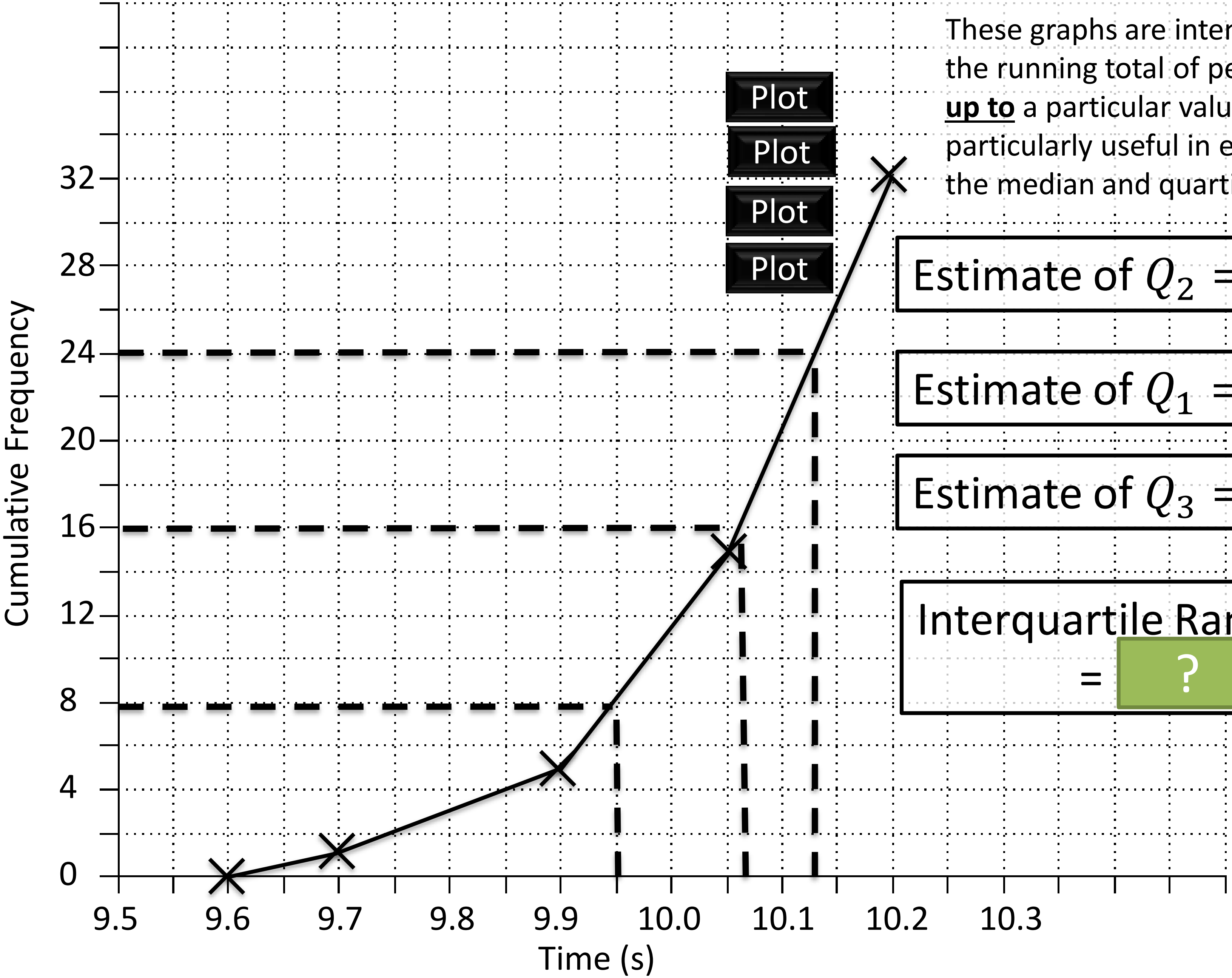
(c) ? B1 (1)

Cumulative Frequency Diagrams

Example: The table below shows the time taken for a group of runners to run 50m. Draw a Cumulative Frequency curve for the data. Use your graph to estimate the median, LQ, UQ and IQR

Time (s)	Frequency	C. Freq
$9.6 < t \leq 9.7$	1	1
$9.7 < t \leq 9.9$	4	5
$9.9 < t \leq 10.05$	10	15
$10.05 < t \leq 10.2$	17	32

Cumulative Frequency Diagrams



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.

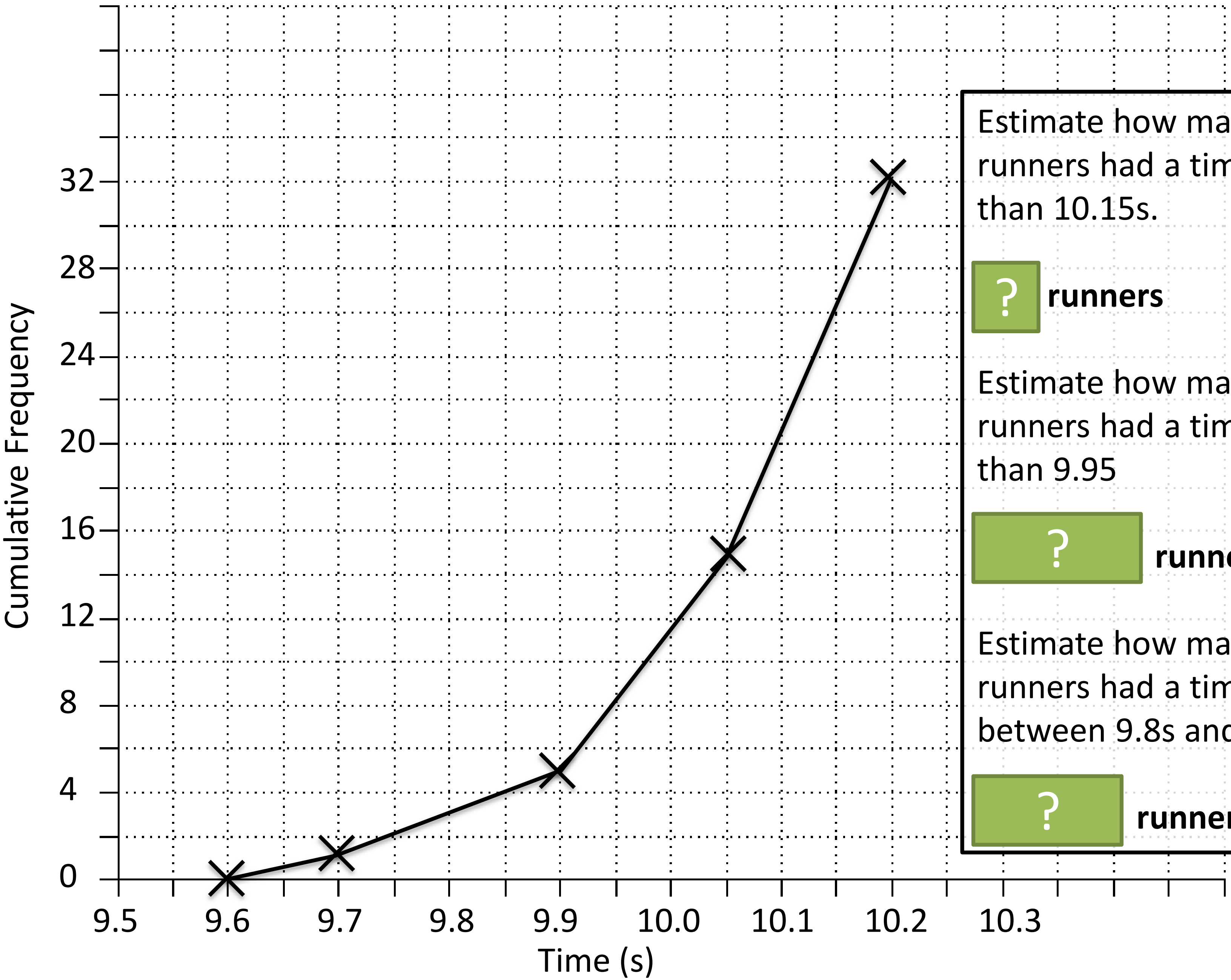
Estimate of Q_2 = ?

Estimate of Q_1 = ?

Estimate of Q_3 = ?

Interquartile Range = ?

Cumulative Frequency Diagrams



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

? **runners**

Estimate how many runners had a time more than 9.95

? **runners**

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

? **runners**

Exercise 3C

Pearson Pure Mathematics Year 1/AS
Pages 47-48

(Students already confident with cumulative frequency graphs may want to skip this exercise)

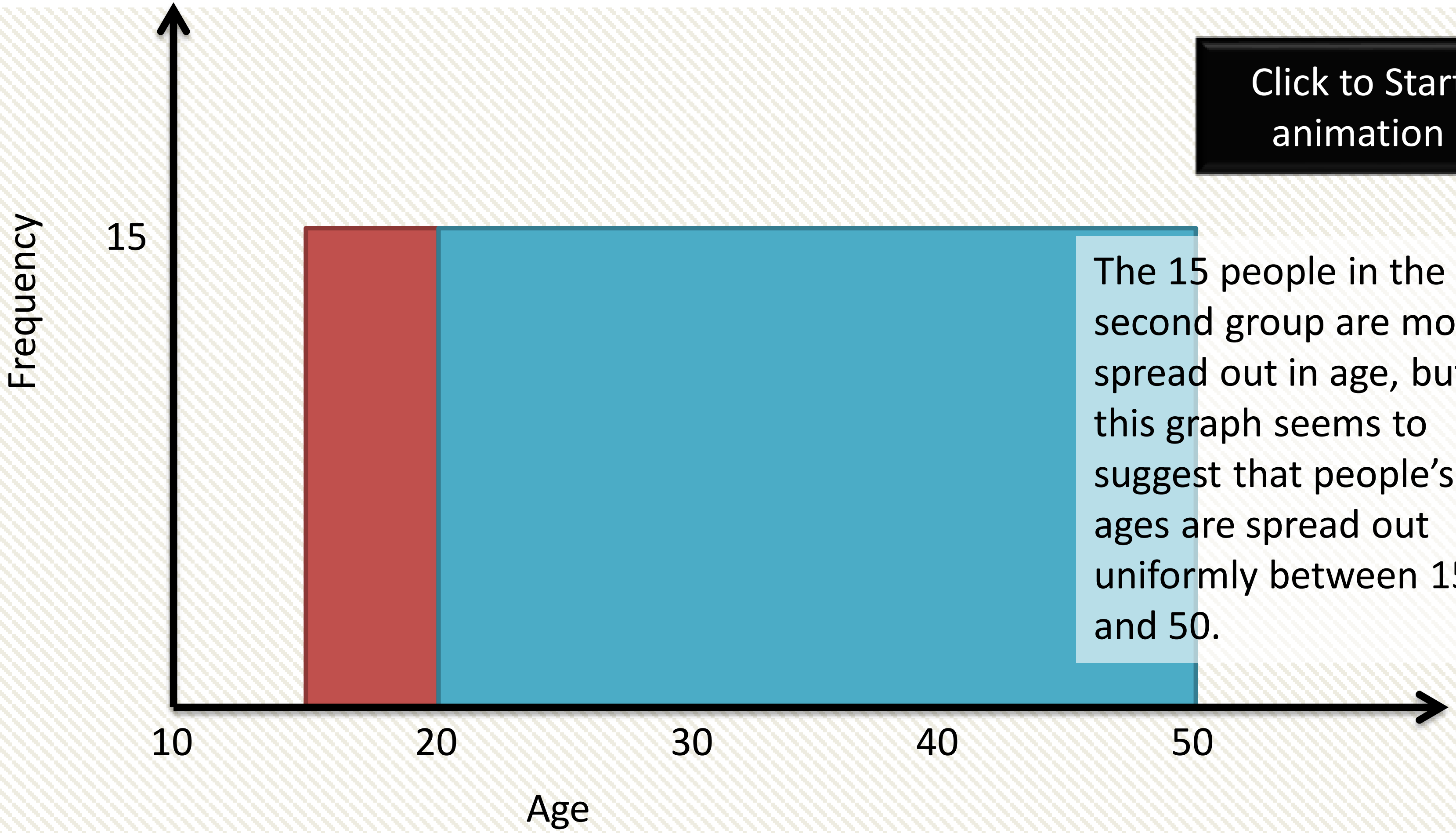
Histograms

Age (years)	Frequency
$15 \leq a < 20$	15
$20 \leq a < 50$	15

Pablo is hosting a party. He counts how many people are between 15 and 20, and 20 and 50.

Why is below graph somewhat unhelpful. How could we fix it?

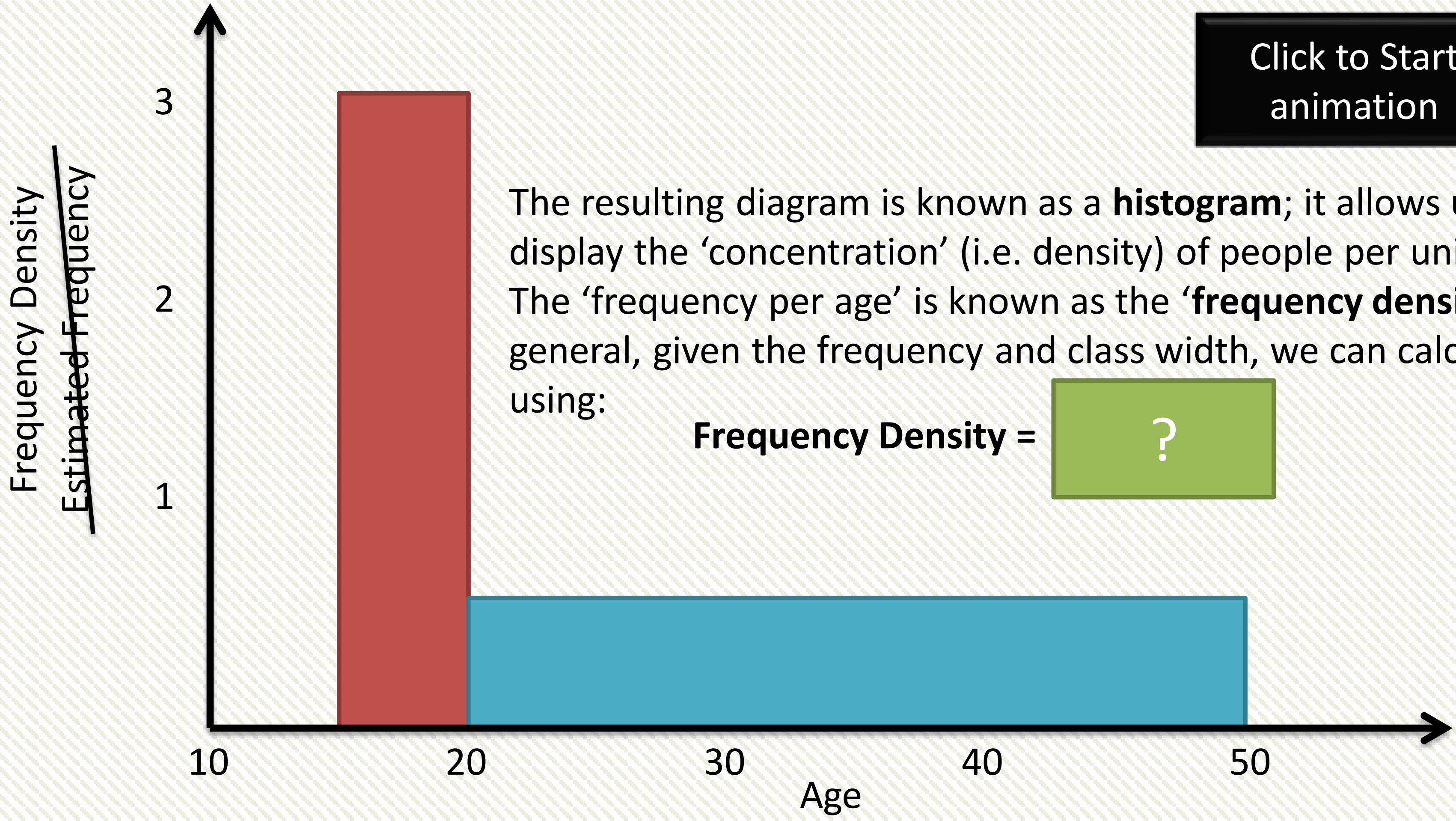
Click to Start animation



The 15 people in the second group are more spread out in age, but this graph seems to suggest that people's ages are spread out uniformly between 15 and 50.

Age (years)	Frequency
15 ≤ a < 20	15
20 ≤ a < 50	15

Let’s presume that within each age group, the ages are evenly spread.
Then there would ? people of each age in the 15-20 group, and ? people of each age in the 20-50 group.



Click to Start animation

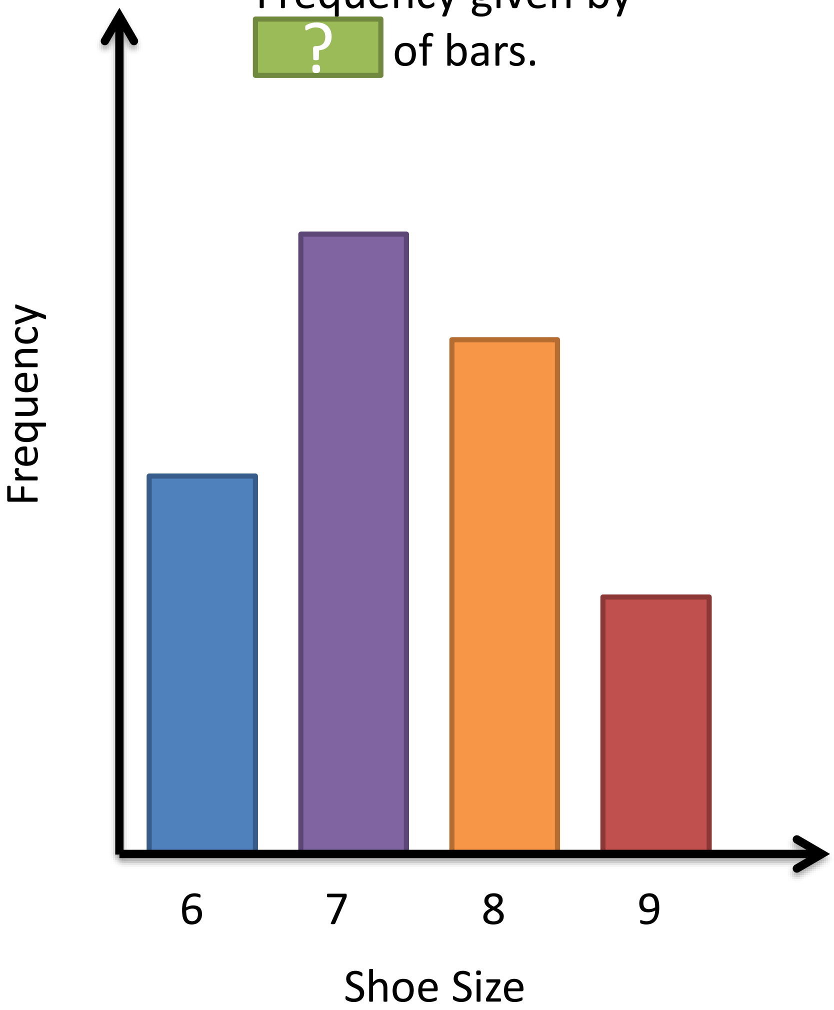
The resulting diagram is known as a **histogram**; it allows us to display the ‘concentration’ (i.e. density) of people per unit value. The ‘frequency per age’ is known as the ‘**frequency density**’. In general, given the frequency and class width, we can calculate it using:

Frequency Density = ?

Bar Charts vs Histograms

Bar Charts

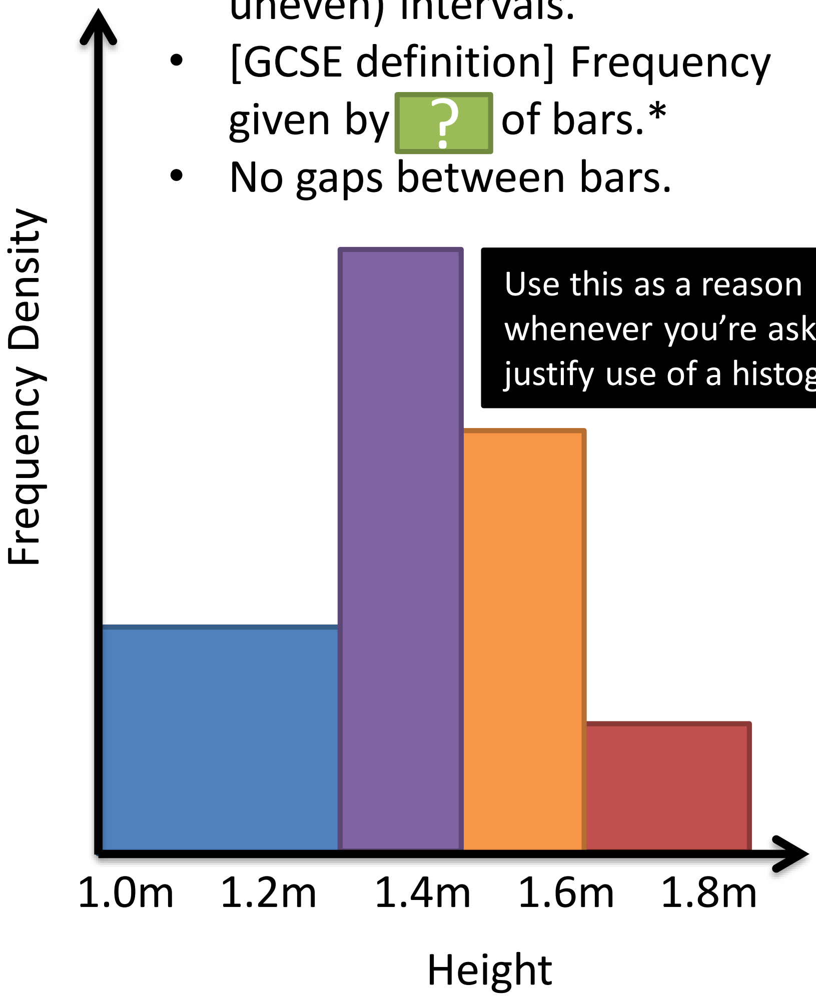
- For ? data.
- Frequency given by ? of bars.



Histograms

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

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



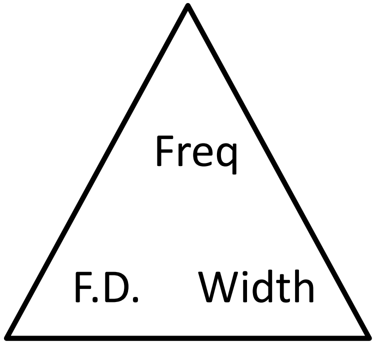
* Not necessarily true. We'll correct this in a sec.

Bar Charts vs Histograms

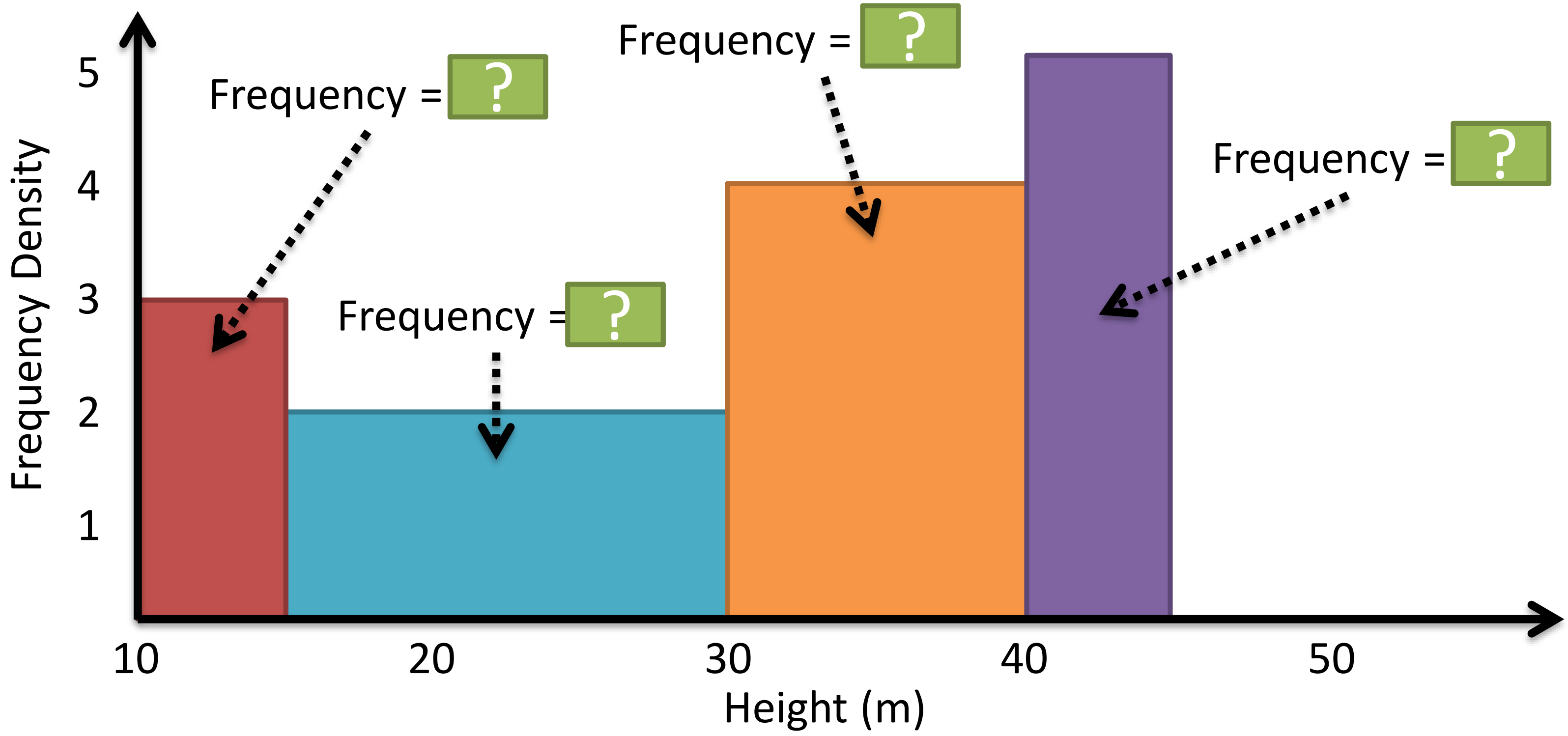
Q1

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

Still using the **incorrect** GCSE formula:



Q2



Frequency = ?

Frequency = ?

Frequency = ?

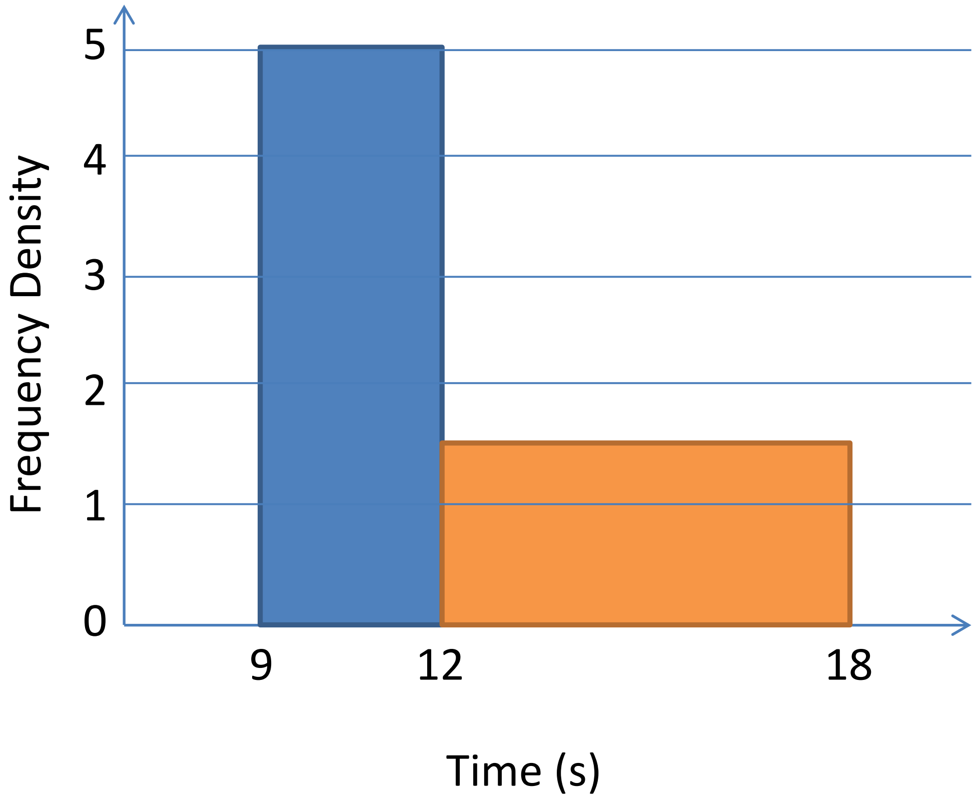
Frequency = ?

SKILL #1 :: Area = frequency?

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'.

?

Then use this scaling along with the desired area.

?

Test Your Understanding

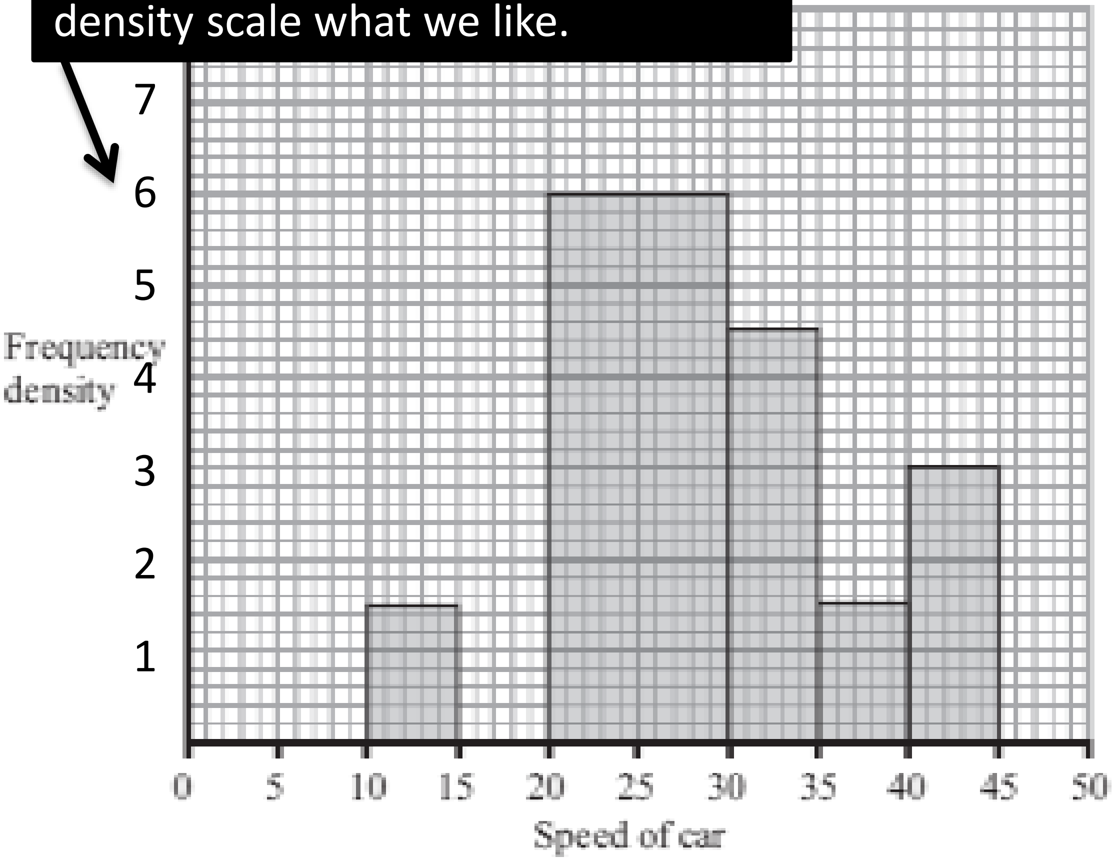
(on your printed sheet)

Edexcel S1 May 2012 Q5

A policeman records the speed of the traffic on a busy road with a 30 mph speed limit. He records the speeds of a sample of 450 cars. The histogram in Figure 2 represents the results.

(*a*) Calculate the number of cars that were exceeding the speed limit by at least 5 mph in the sample. ***(4 marks)***

Tip: We can make the frequency density scale what we like.



M1 A1: Determine what one small square or one large square is worth.

(i.e. work out $area \to freq$ scaling)

?

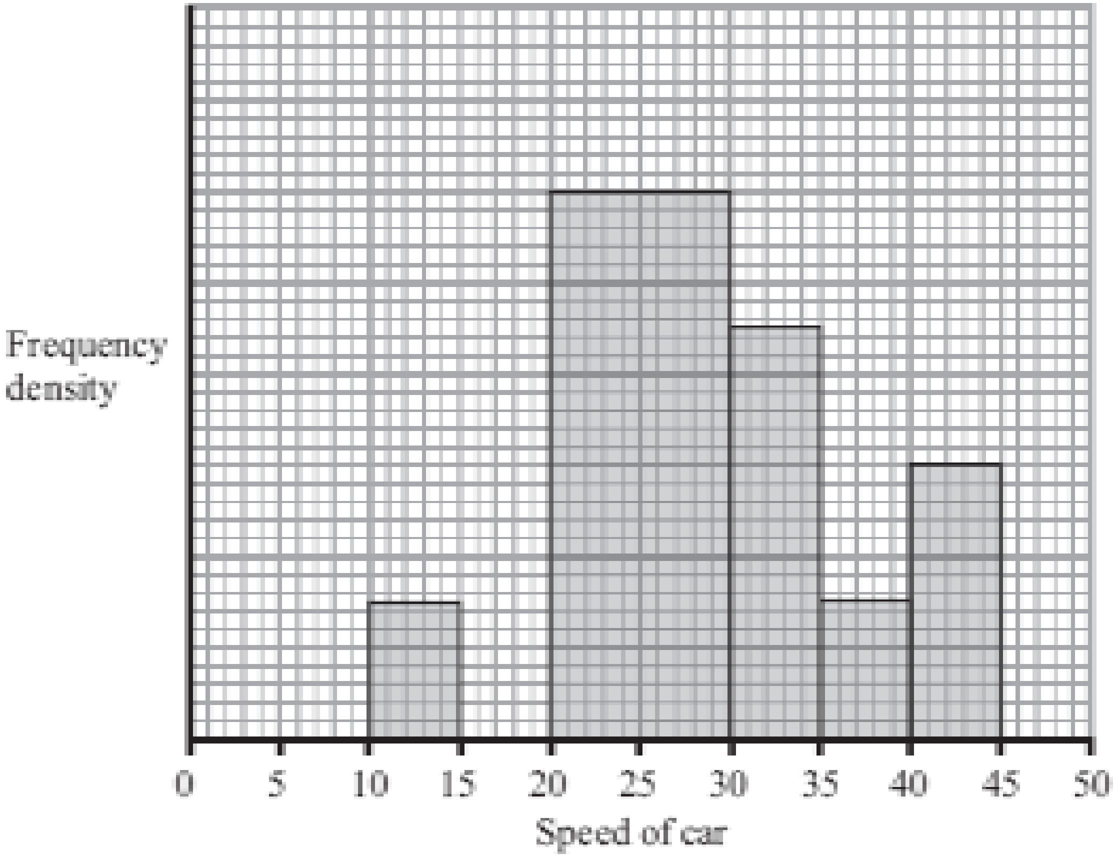
M1 A1: Use this to find number of cars travelling >35mph.

?

Test Your Understanding

(on your printed sheet)

(b) Estimate the value of the mean speed of the cars in the sample. ***(3 marks)***



M1 M1: Use histogram to construct sum of speeds.

?

A1 Correct value

?

Tip: Whenever you are asked to calculate mean, median or quartiles from a histogram, form a grouped frequency table. Use your scaling factor to work out the frequency of each bar.

Test Your Understanding

(on your printed sheet)

(*c*) Estimate, to 1 decimal place, the value of the median speed of the cars in the sample. **(2)**

~~(*d*) Comment on the shape of the distribution. Give a reason for your answer.~~ ~~**(2)**~~

~~(*e*) State, with a reason, whether the estimate of the mean or the median is a better representation of the average speed of the traffic on the road.~~ ~~**(2)**~~

(crossed out questions would not appear in new syllabus)

(c)	?	M1 A1	 (2)
(d)	?	B1ft dB1ft	 (2)
(e)	?	B1 dB1	 (2)

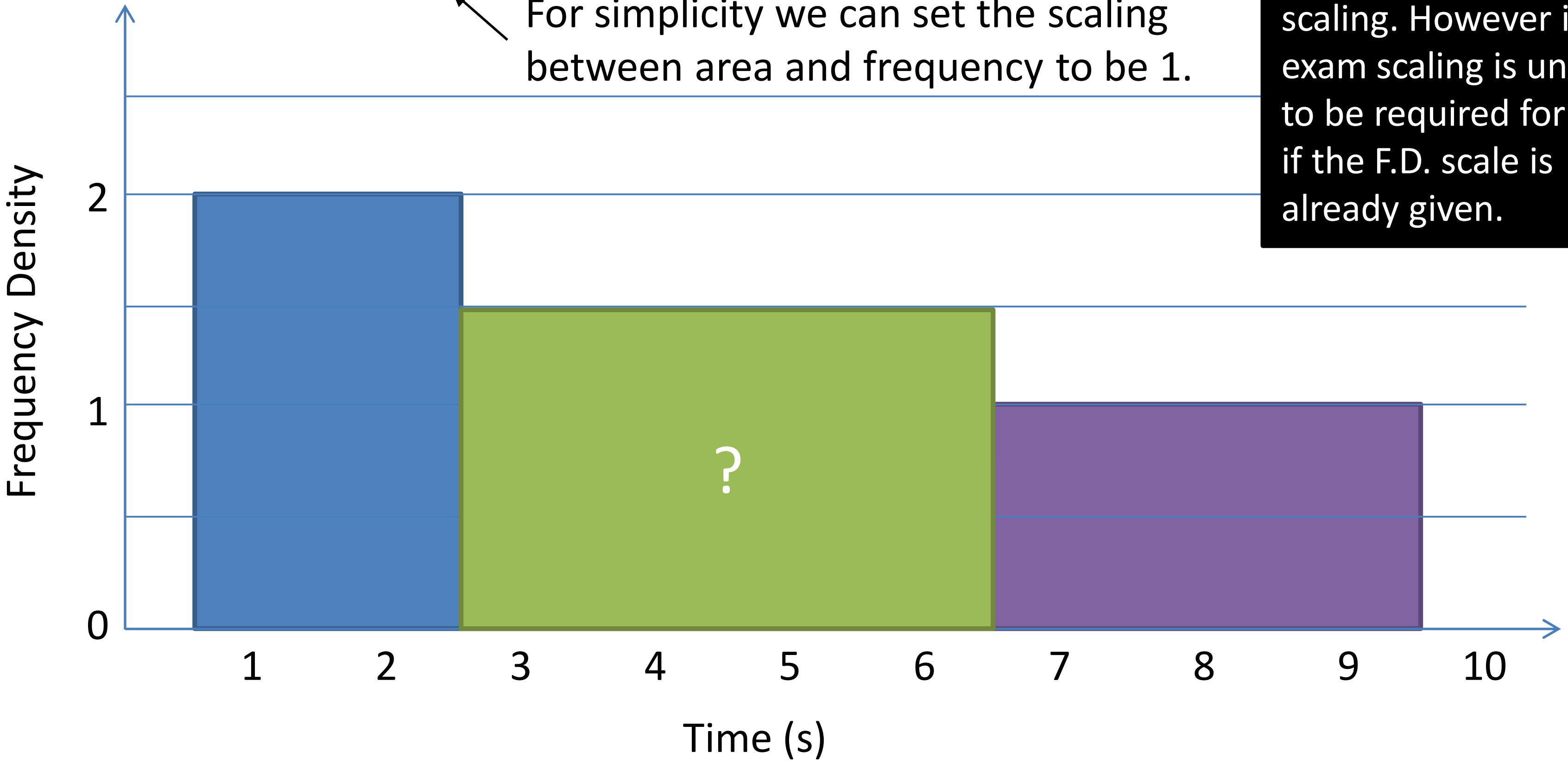
SKILL #2 :: Gaps between classes

Weight (to nearest kg)	Frequency	F.D.
1-2	4	?
3-6	3	?
7-9	?	?

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

Note the gaps affects class width!

Remember the frequency density axis is only correct to scale, so there may be some scaling. However in an exam scaling is unlikely to be required for F.D. if the F.D. scale is already given.

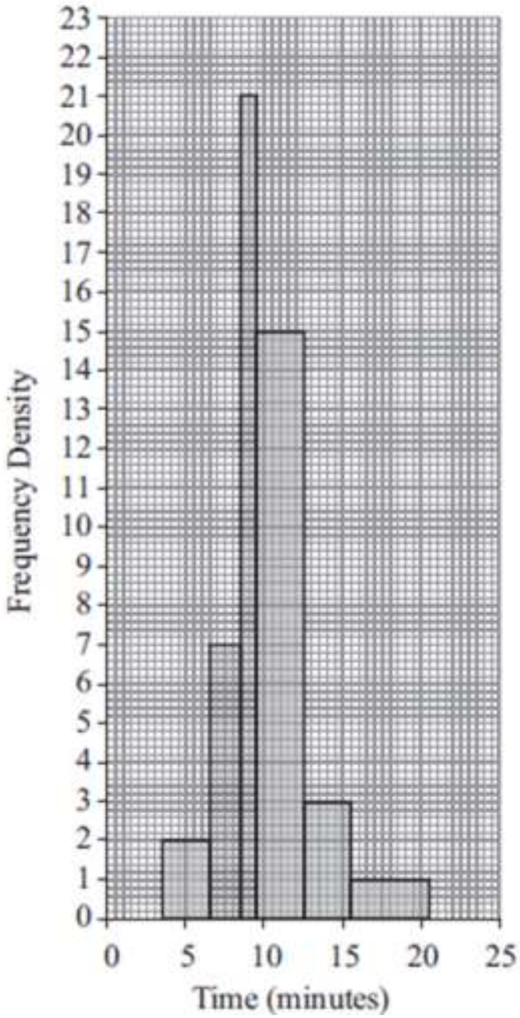


Test Your Understanding

(on your printed sheet)

Jan 2012 Q1

The histogram in Figure 1 shows the time, to the nearest minute, that a random sample of 100 motorists were delayed by roadworks on a stretch of motorway.



Tip: Be careful that you use the correct class widths!

(a) Complete the table.

Delay (minutes)	Number of motorists
4 – 6	6
7 – 8	?
9	21
10 – 12	45
13 – 15	9
16 – 20	?

(2)

(*b*) Estimate the number of motorists who were delayed between 8.5 and 13.5 minutes by the roadworks.

(2)

?

SKILL #3 :: Width and height on diagram

An exam favourite is to ask what width and height we'd draw a bar in a drawn histogram.

Q: The frequency table shows some running times. On a histogram the bar for 0-4 seconds is drawn with width 6cm and height 8cm. 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

Tip: Strategy ?

Solution ?

Solution ?

Scaling for width =

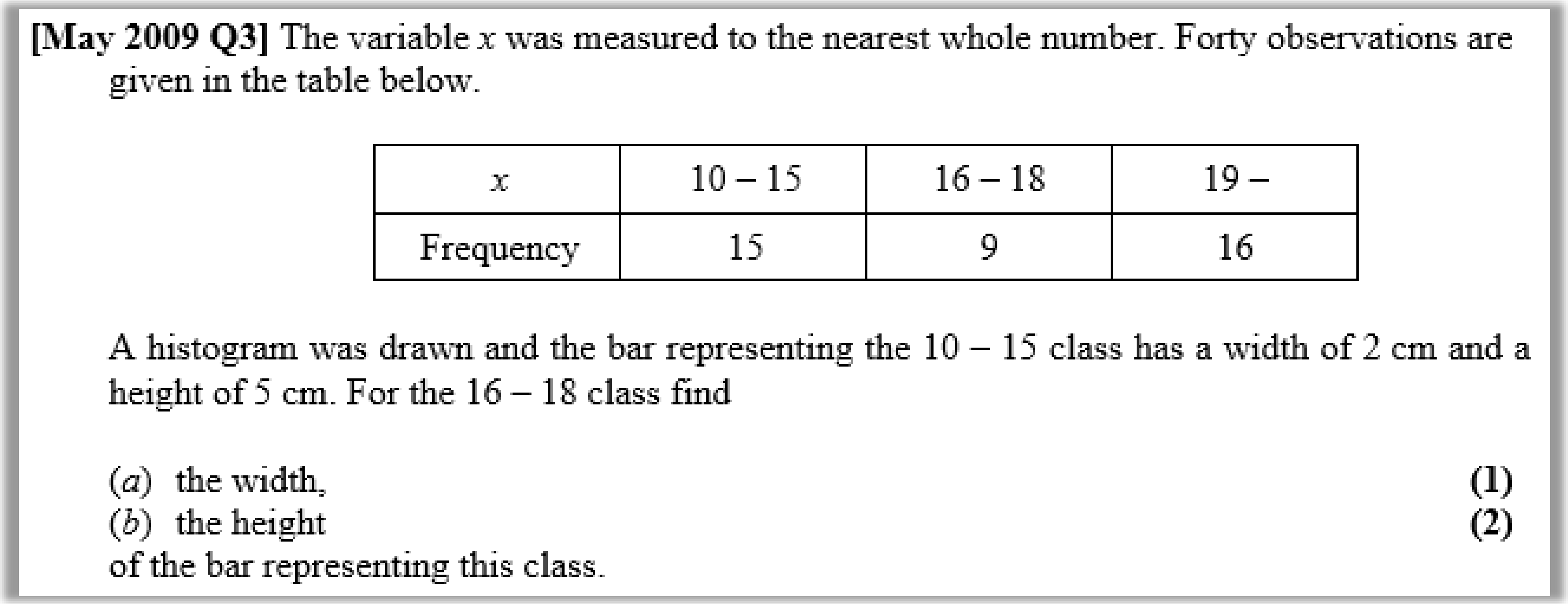
Scaling for height:

4-6 bar:

Solution ?

Test Your Understanding

(on your printed sheet)



[May 2009 Q3] The variable x was measured to the nearest whole number. Forty observations are given in the table below.

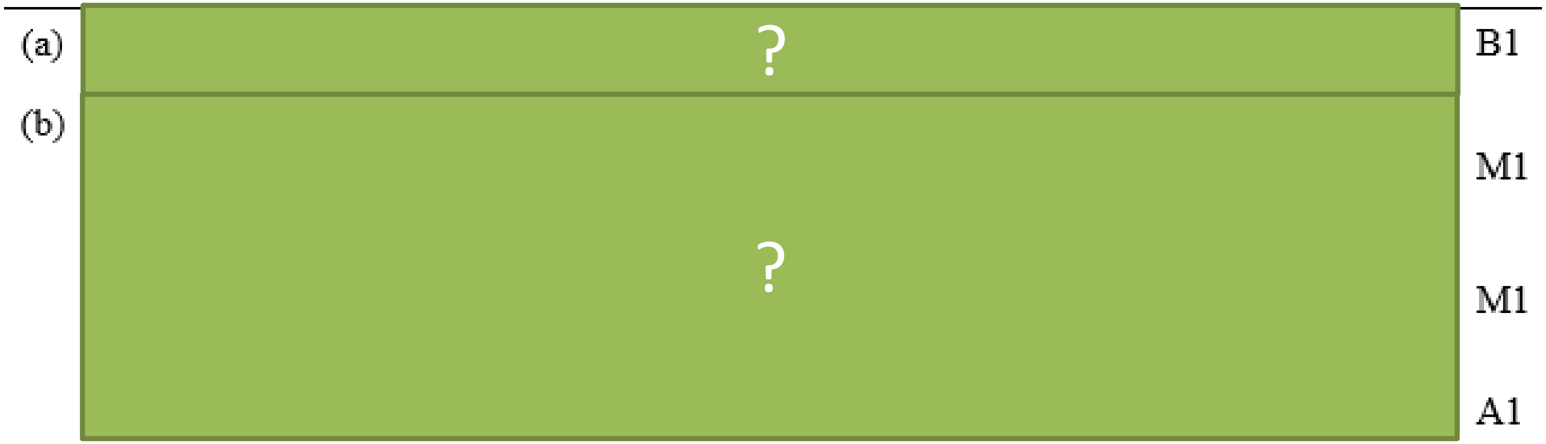
x	10 – 15	16 – 18	19 –
Frequency	15	9	16

A histogram was drawn and the bar representing the 10 – 15 class has a width of 2 cm and a height of 5 cm. For the 16 – 18 class find

(*a*) the width, **(1)**

(*b*) the height **(2)**

of the bar representing this class.

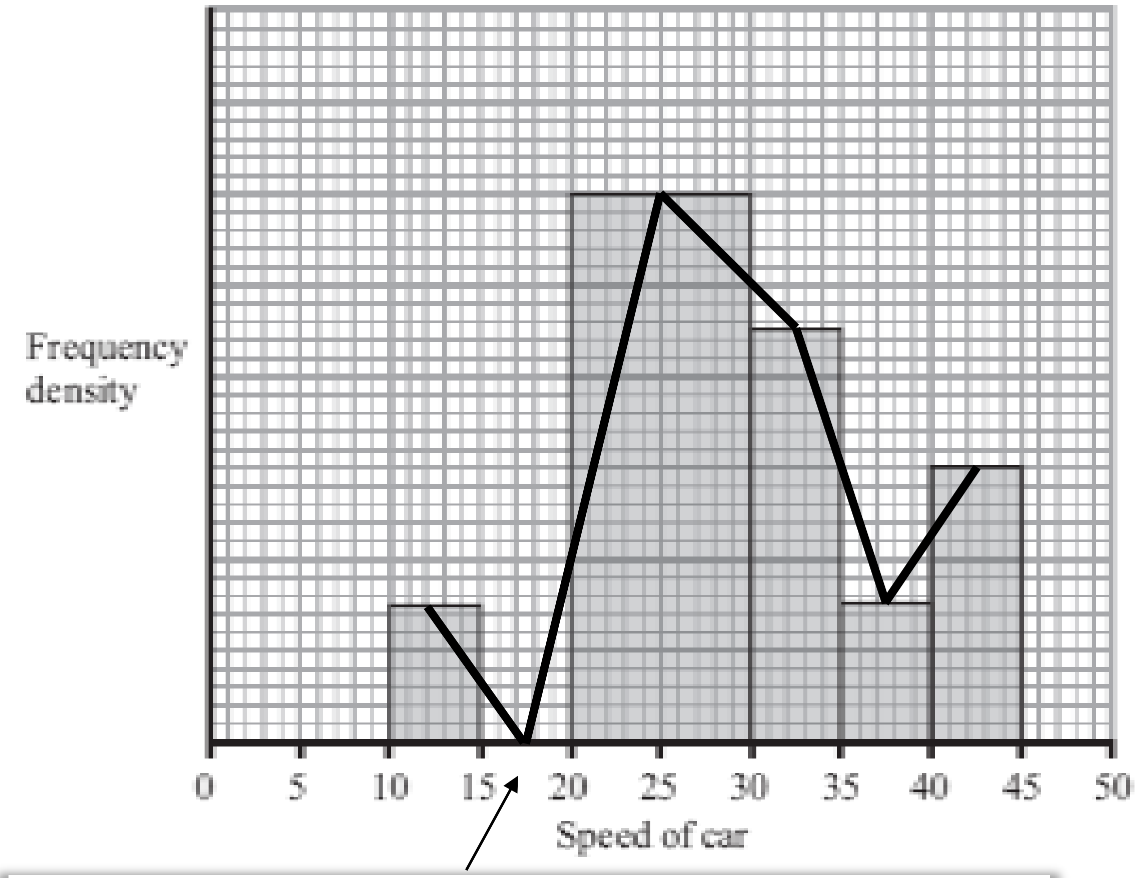


(a) ? B1

(b) ? M1 M1 A1

SKILL #4 :: 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.

Click to Sketch

Exercise 3D

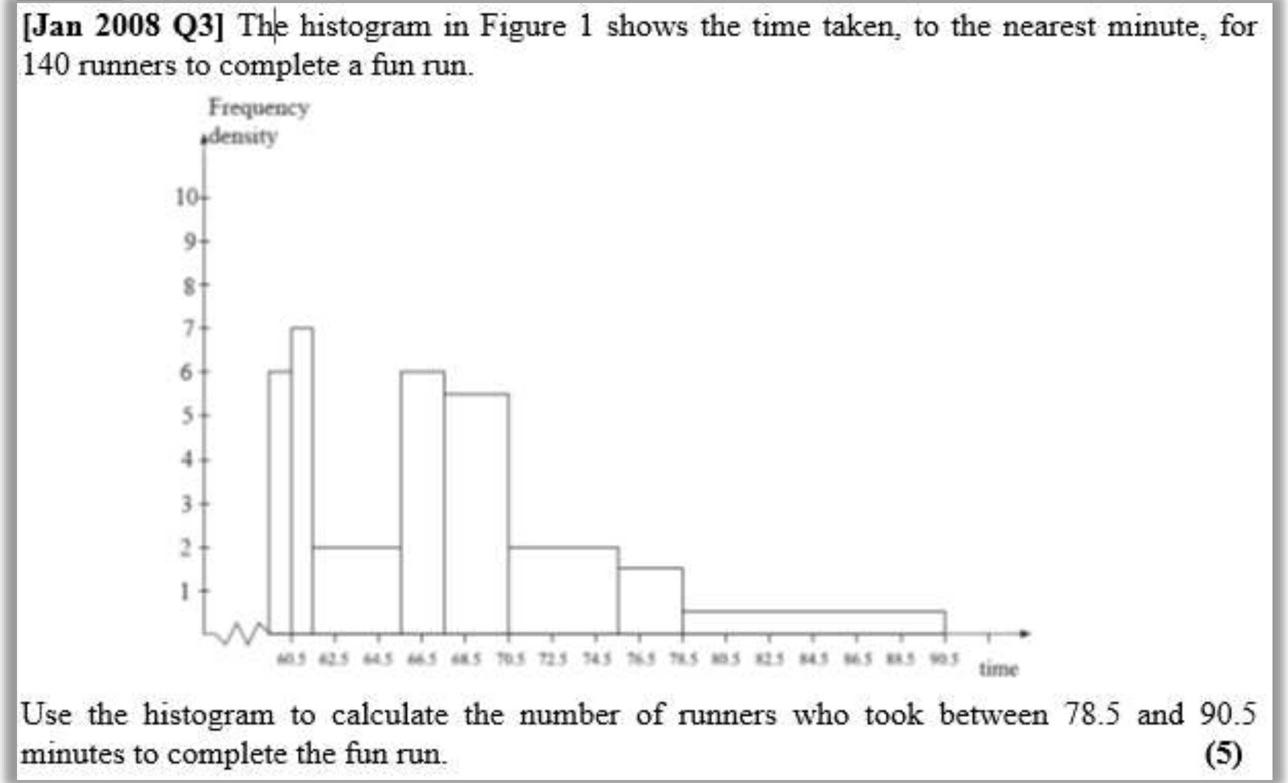
Pearson Pure Mathematics Year 1/AS
Pages 47-48

There is a supplementary exercise (available as a separate file for printing) with solutions on the next slides…

Supplementary Exercise

(on your printed sheet)

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)**

?

Supplementary Exercise

(on your printed sheet)

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.

(DO NOT DRAW THE HISTOGRAM)

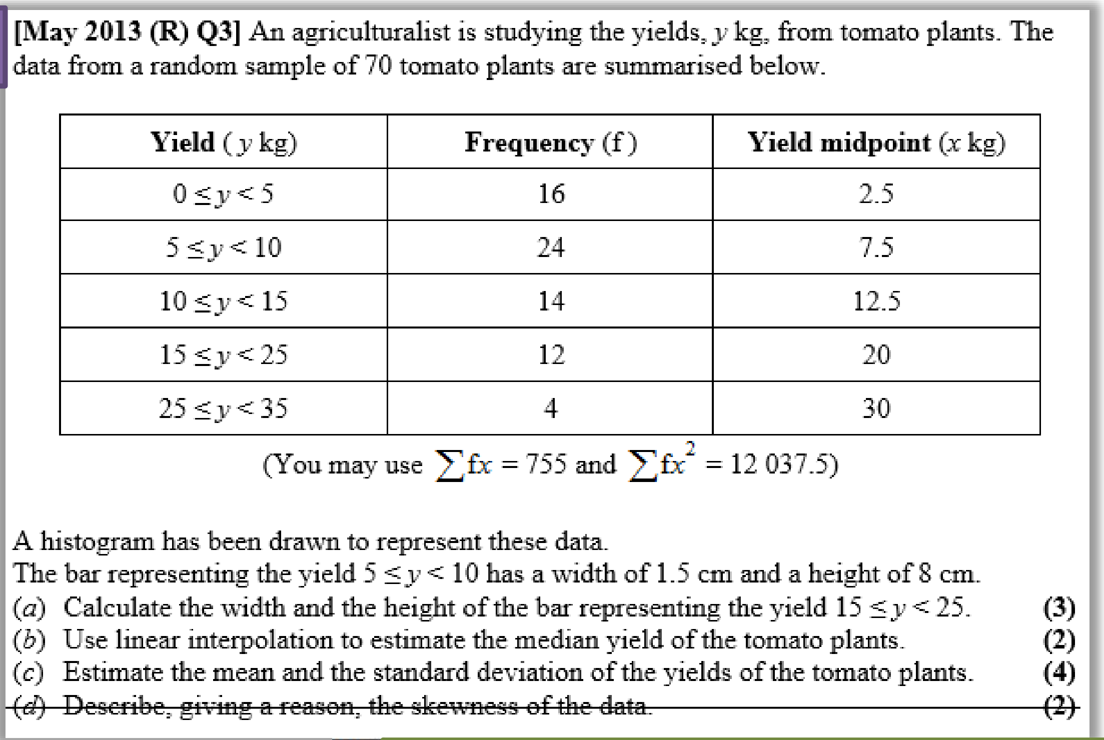
(2)

?

Supplementary Exercise

(on your printed sheet)

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)
$0 \le y < 5$	16	2.5
$5 \le y < 10$	24	7.5
$10 \le y < 15$	14	12.5
$15 \le y < 25$	12	20
$25 \le y < 35$	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 $5 \le y < 10$ 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 $15 \le y < 25$. **(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)**

~~(*d*) Describe, giving a reason, the skewness of the data. **(2)**~~

(a) ?

(3)

(b) ?

(2)

(c) ?

(4)

(d) ?

(2)

Supplementary Exercise

(on your printed sheet)

Q4 [June 2007 Q5]

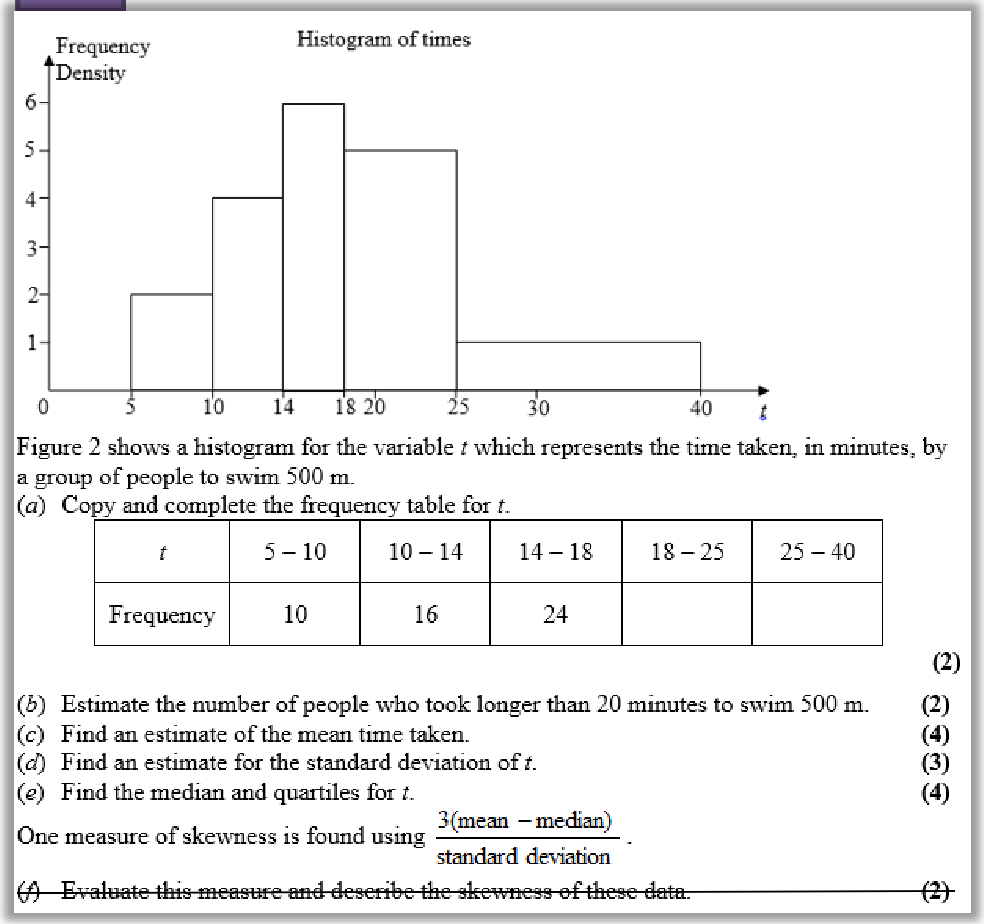

Frequency Density
Histogram of times
6
5
4
3
2
1
0
5
10
14
18
20
25
30
40
t

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)**

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

~~(*f*) Evaluate this measure and describe the skewness of these data. **(2)**~~

5(a) ?

(b) ?

(c) ?

(d) ?

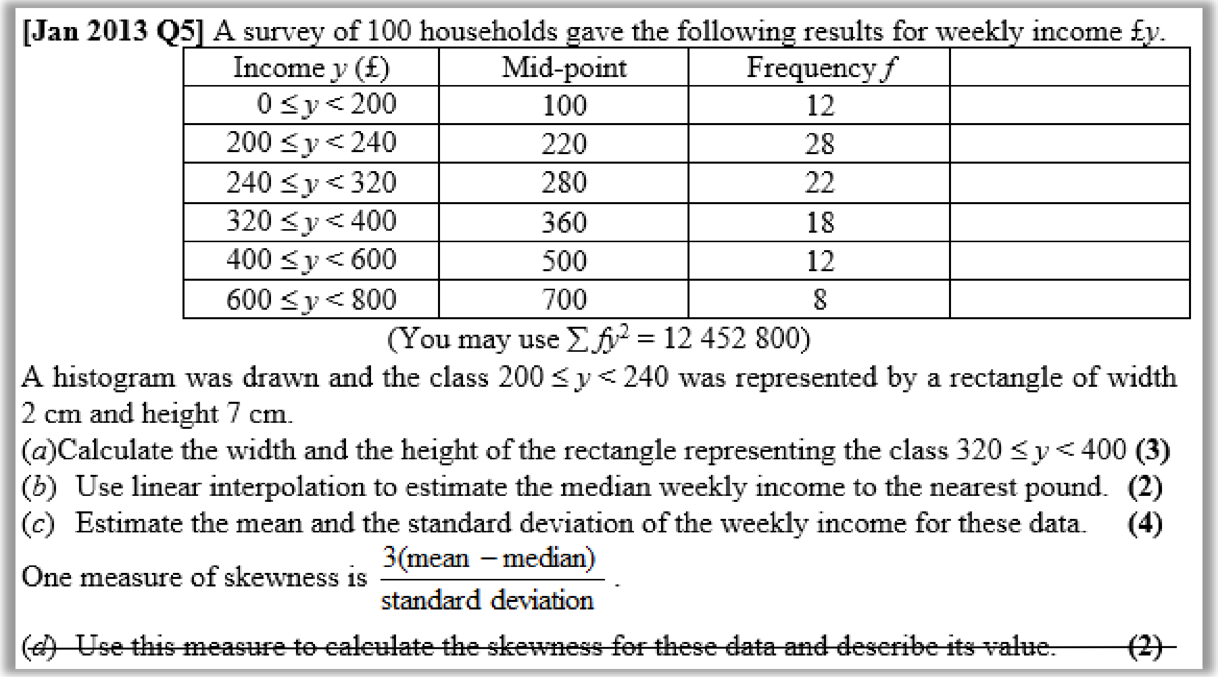
(e) ?

(f) ?

Supplementary Exercise

(on your printed sheet)

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 $\dfrac{3(\text{mean} - \text{median})}{\text{standard deviation}}$.

~~(*d*) Use this measure to calculate the skewness for these data and describe its value. **(2)**~~

(a)

?

(b)

?

(c)

?

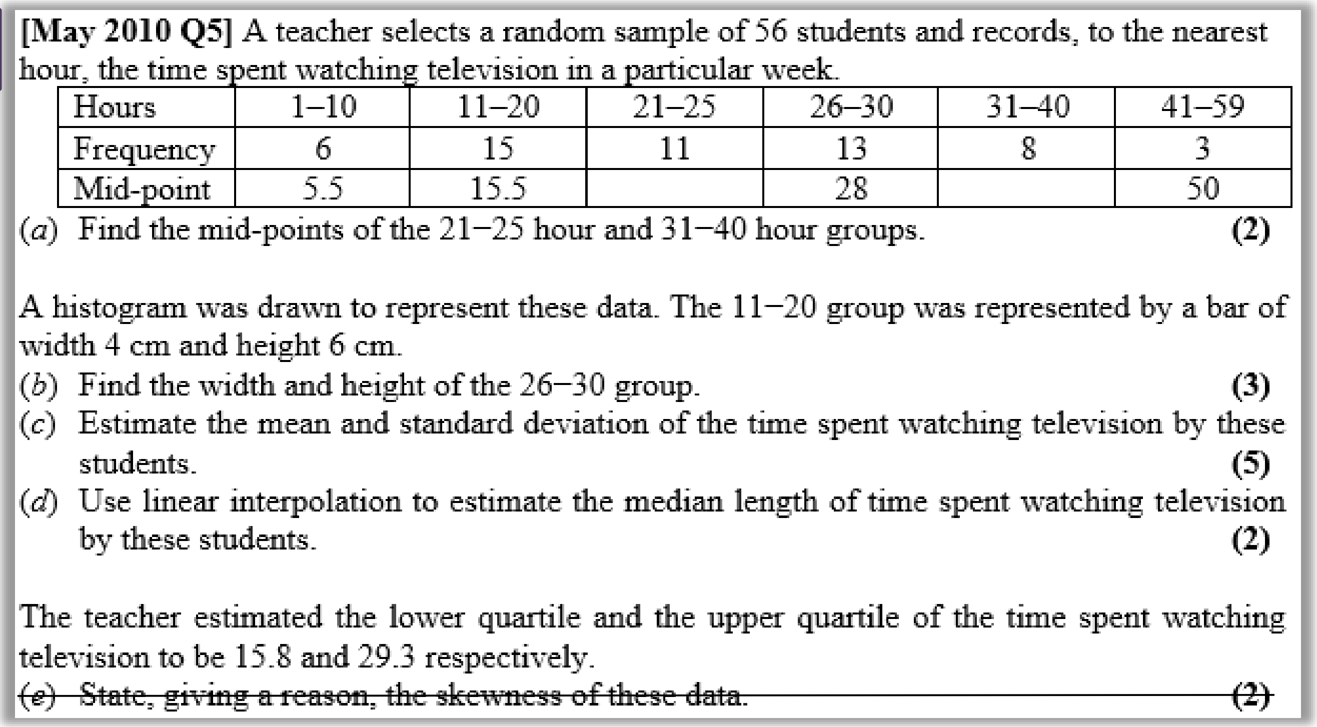
(d)

?

Supplementary Exercise

(on your printed sheet)

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.

~~(*e*) State, giving a reason, the skewness of these data. **(2)**~~

(a) ?

(b) ?

(c) ?

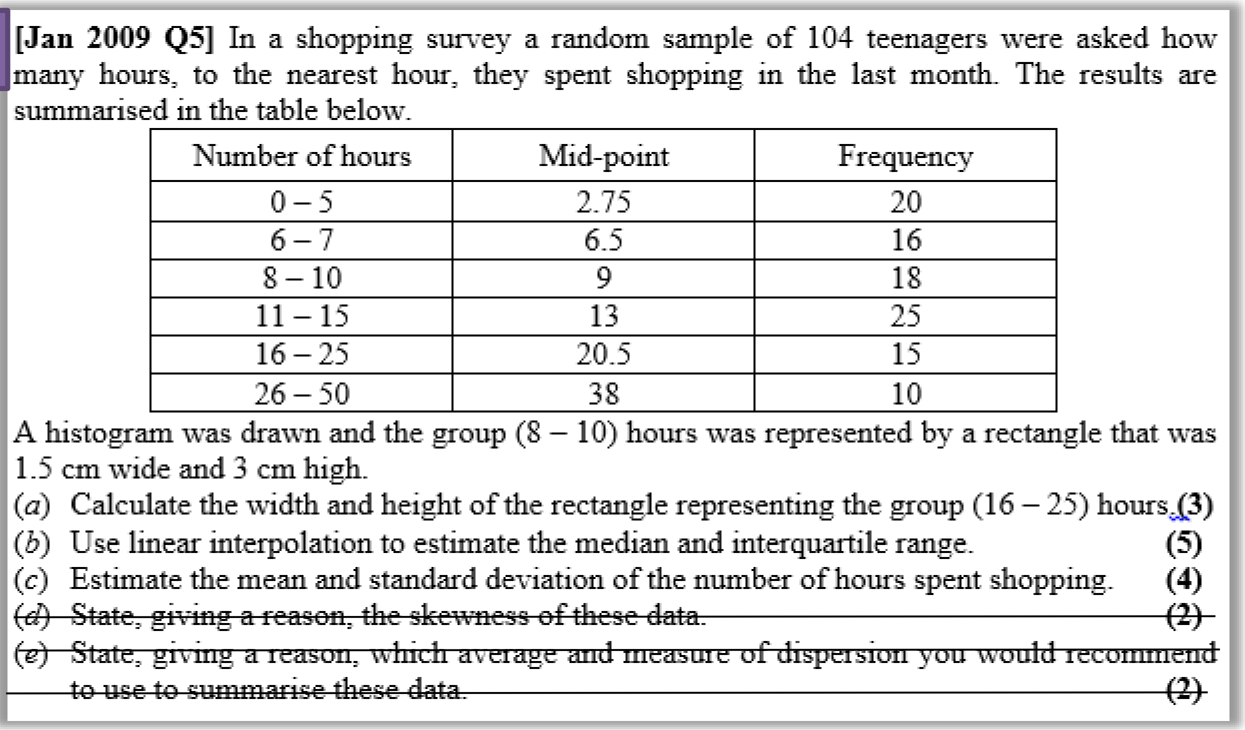
(d) ?

(e) ?

Supplementary Exercise

(on your printed sheet)

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)**

~~(*d*) State, giving a reason, the skewness of these data. **(2)**~~

~~(*e*) State, giving a reason, which average and measure of dispersion you would recommend to use to summarise these data. **(2)**~~

(a) a ?

(b) b ?

(c) c ?

(d) d ?

(e) e ?

(5)