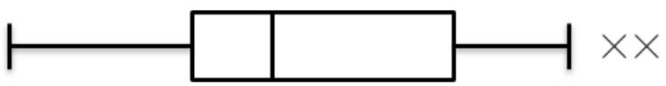
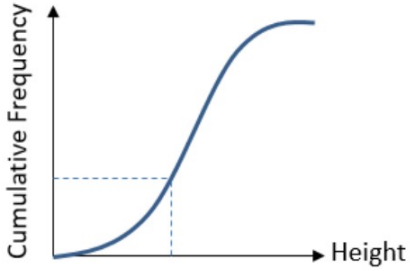
Chapter 3: Representations of Data

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*.

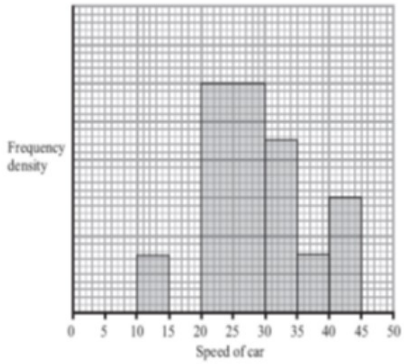
BOX PLOTS AND OUTLIERS



CUMULATIVE FREQ DIAGRAMS



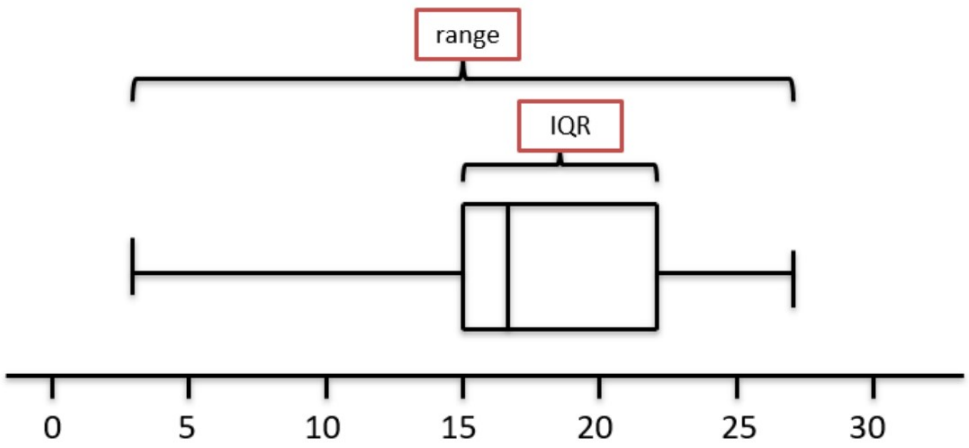
HISTOGRAMS



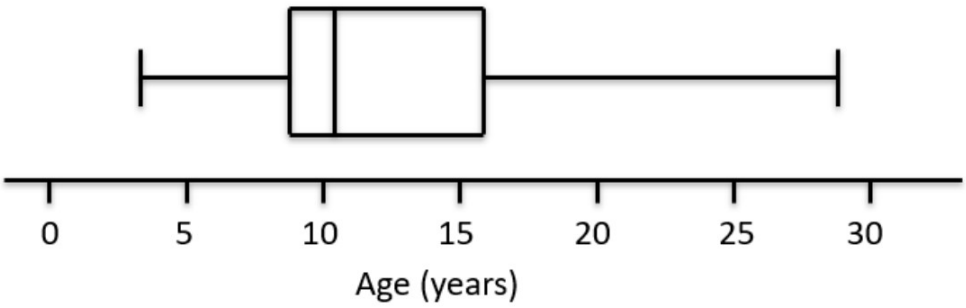
Box Plot recap

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

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



Interpreting a Box Plot



True or false:

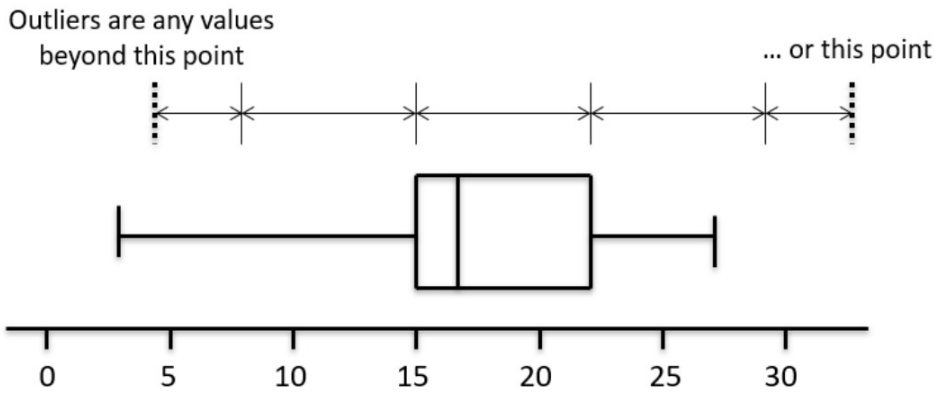
"The right box represents more people than the left box."

"The ages are more spread out above the median."

Outliers

An outlier is **an extreme value.**

One common definition of an outlier is **1.5 IQRs** beyond the lower and upper quartiles
i.e. $Q_3 + 1.5 \times \text{IQR}$ and $Q_1 - 1.5 \times \text{IQR}$



Outliers are marked with a cross

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.

The ages of 15 Lib Dem MPs are given:

11 18 20 27 30 31 32 32 35 36 37 58 63 78 105

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 $\Sigma x = 613$ and $\Sigma x^2 = 33815$

The ages of 9 people at a birthday party are:

12 17 21 33 34 37 42 62 165

An outlier is an observation which lies ± 2 standard deviations from the mean.
Identify any outliers for this data, and clean the data of any anomalies if necessary.
Note, $\bar{x} = 47$ and $\sigma = 44.02$

Outliers which have occurred as a result of an error are called *anomalies*. Removing anomalies is called 'cleaning the data'.

You can use $\ll$ or $\gg$ to denote 'much greater than' or 'much less than'

Your Turn

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

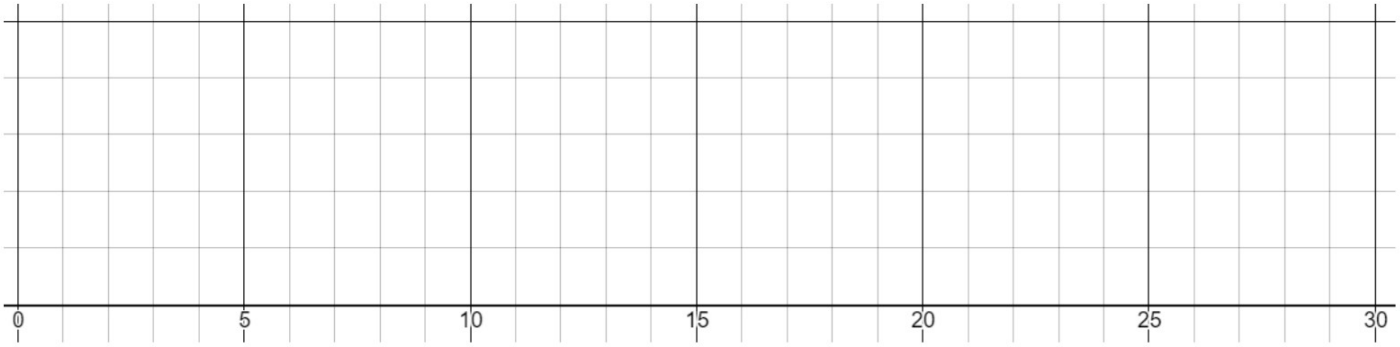
a) Calculate the mean and standard deviation, given that $\Sigma x = 252$ and $\Sigma x^2 = 5468$.
b) An outlier is an observation which lies ± 2 standard deviations from the mean. Identify any outliers for this data.

Box Plot example

Smallest values	Largest values	Lower Quartile	Median	IQR
0, 3	21, 27	8	10	6

Draw a box plot to represent the above data, showing any outliers.
Identify outliers using 1.5 x IQR above or below the quartiles.

When there's an outlier at one end, there's two allowable places to put the end of the 'whisker':
1) The maximum value that is *not* an outlier
2) The outlier boundary



Your Turn

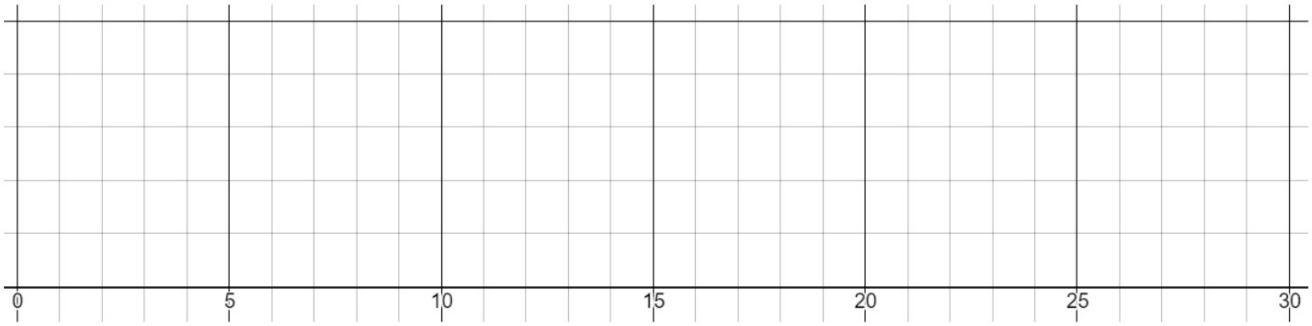
[Jan 2011 Q3] Over a long period of time a small company recorded the amount it received in sales per month. The results are summarised below.

	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)**

(b) 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)**



Comparing Box Plots

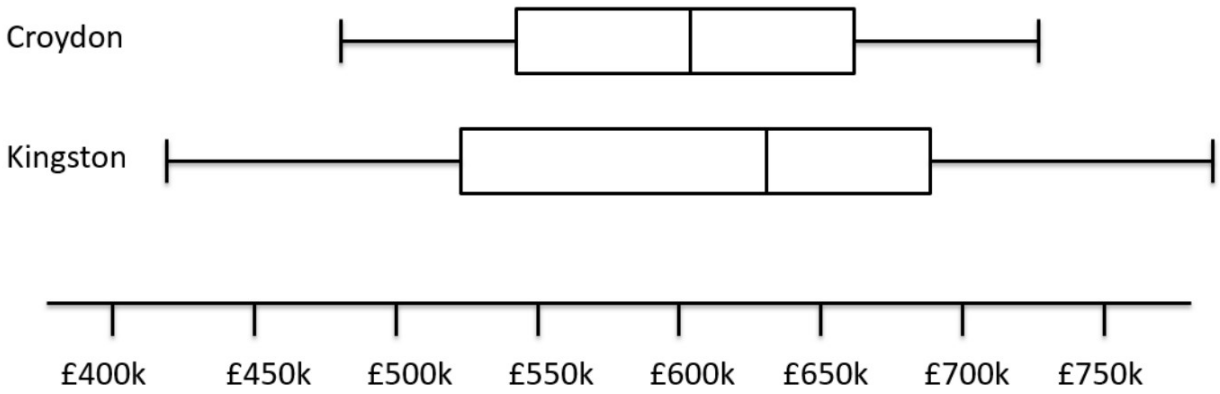
When making comparisons...

...include some measure of **spread**.

...include some measure of **location** (median is best).

...and write about it on context of the question!

Box Plots 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, measure of spread:

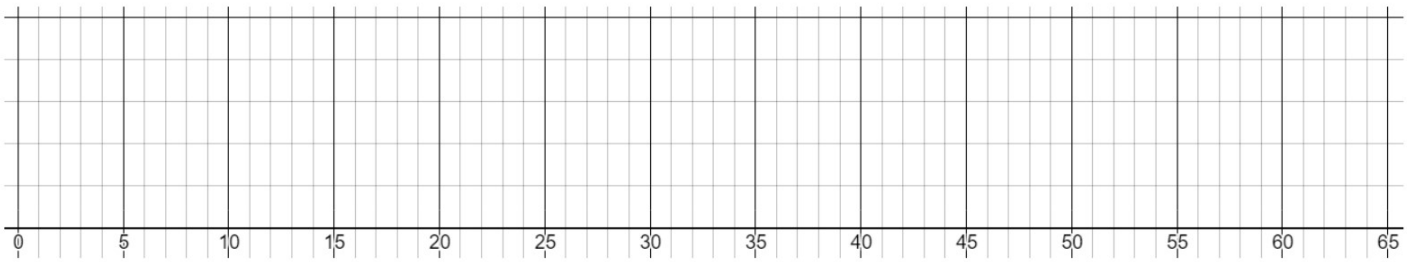
For 1 mark, measure of location:

Ex 3A/B

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.

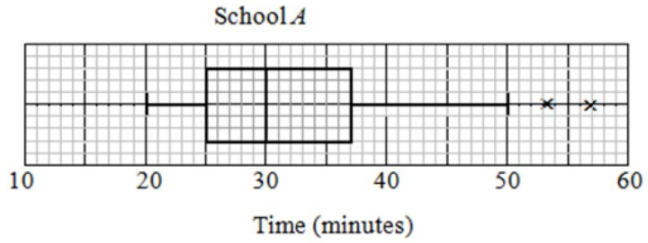
(*a*) On graph paper, draw a box plot to represent these data. **(7)**



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



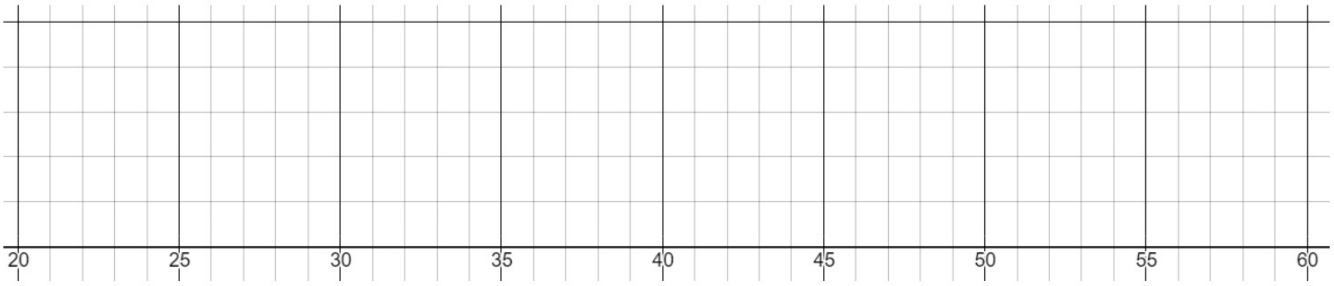
(b) (i) Write down the time by which 75% of the children in school A had completed the run.
(ii) State the name given to this value. **(2)**

(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)**

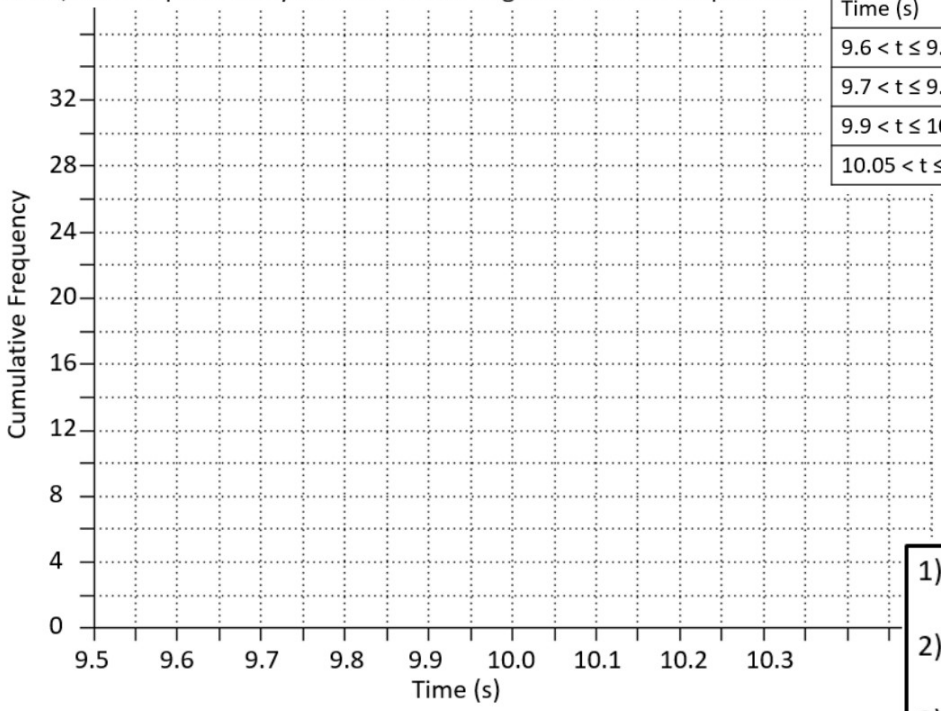
(e) Compare and contrast these two box plots. **(4)**



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.

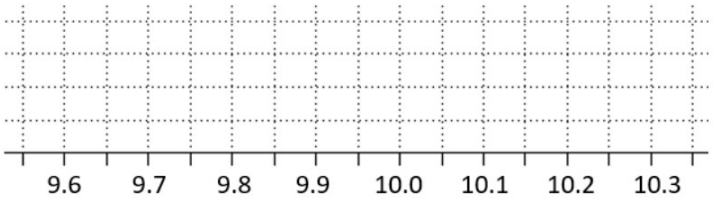
Time (s)	Frequency
$9.6 < t \leq 9.7$	1
$9.7 < t \leq 9.9$	4
$9.9 < t \leq 10.05$	10
$10.05 < t \leq 10.2$	17



1) Estimate the quartiles and IQR.

2) Draw a box plot.

3) Estimate how many runners had a time
a) less than 10.15s
b) more than 9.95s
c) between 9.8s and 10s



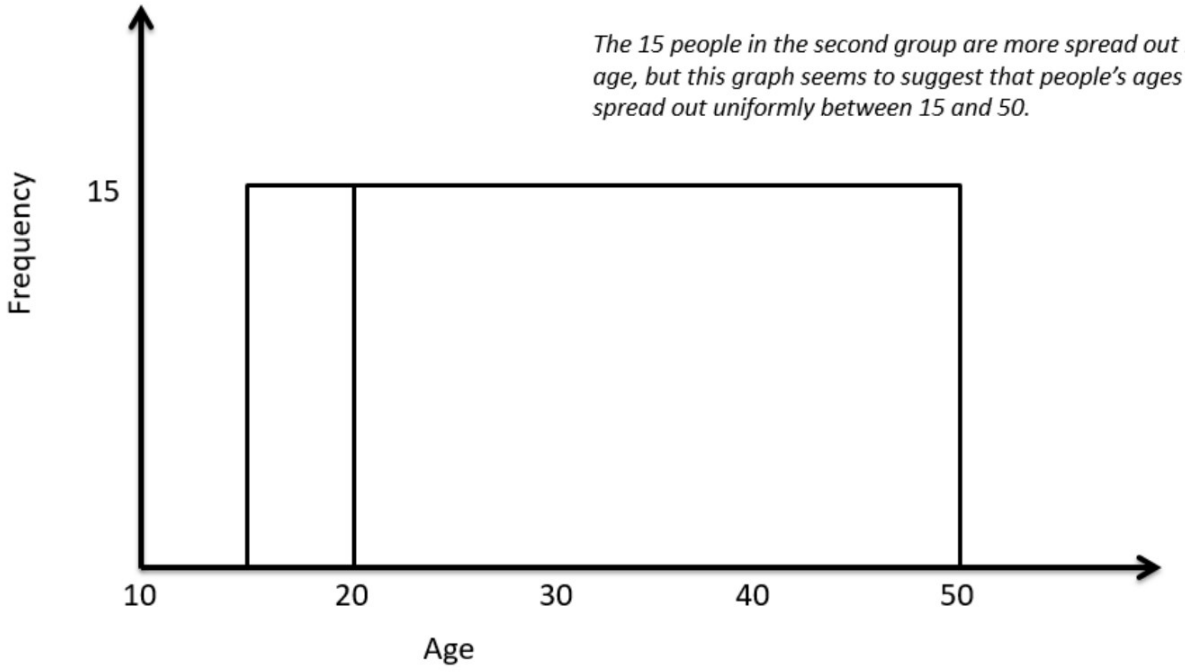
Histograms

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?

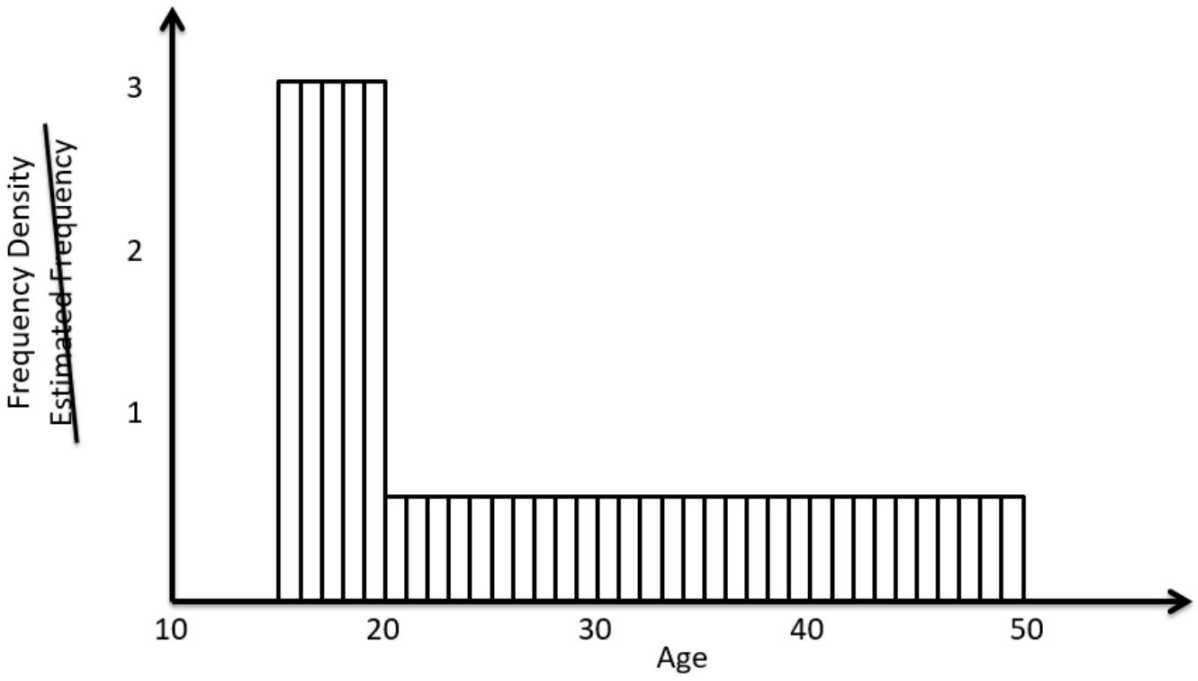
The 15 people in the second group are more spread out in age, but this graph seems to suggest that people's ages are spread out uniformly between 15 and 50.



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.



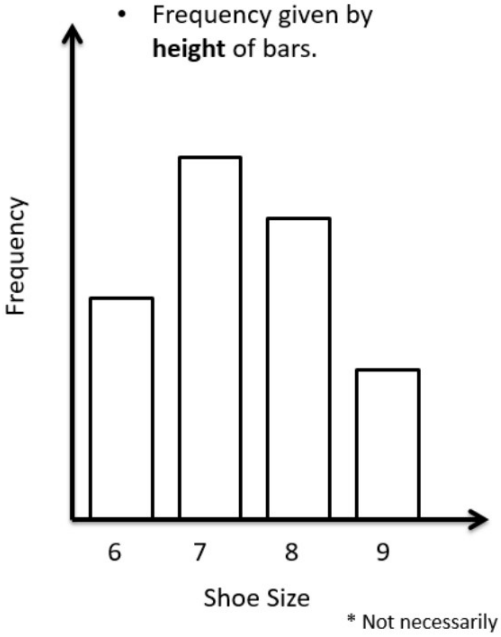
The resulting diagram is known as a **histogram**; it allows us to display the 'concentration' (i.e. density) of people per unit value.
The 'frequency per age' is known as the '**frequency density**'. In general, given the frequency and class width, we can calculate it using:

$$\text{Frequency Density} = \frac{\text{Frequency}}{\text{Class Width}}$$

Bar Charts vs Histograms

Bar Charts

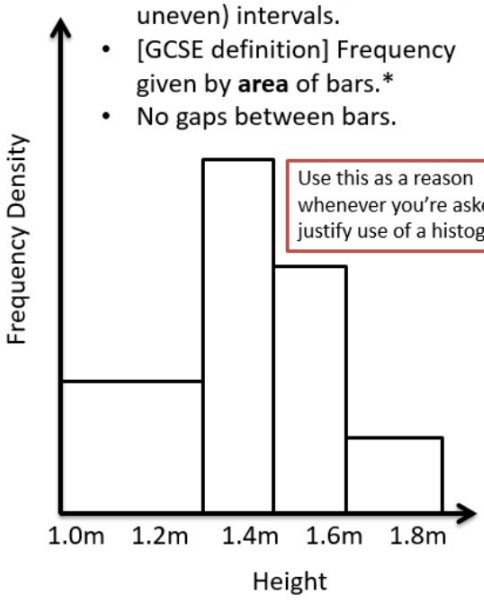
- For **discrete** data.
- Frequency given by **height** of bars.



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.

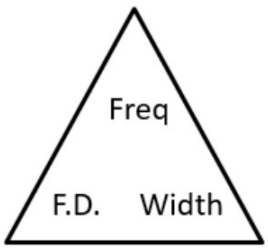


* Not necessarily true. We'll correct this in a sec.

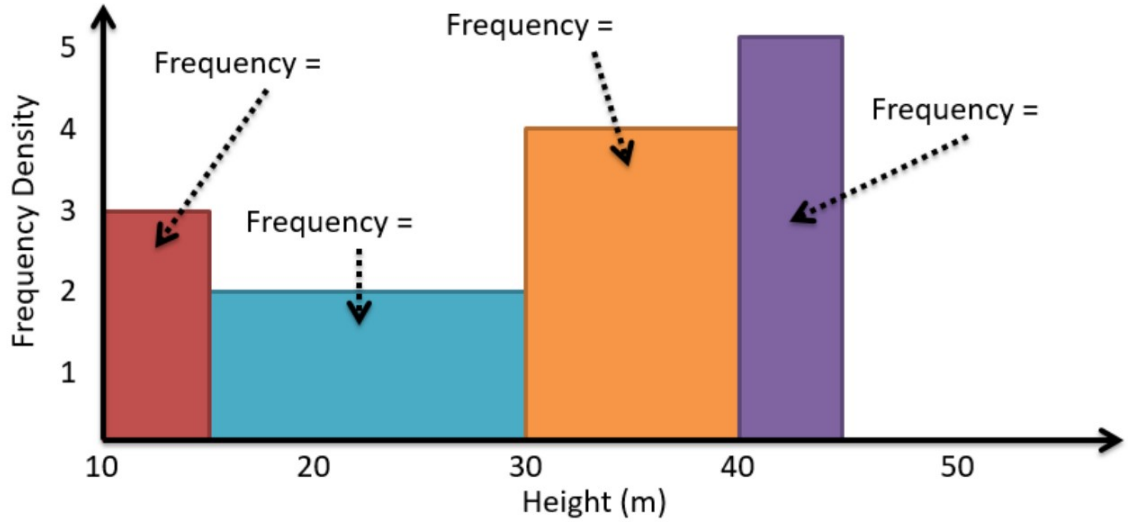
Q1

Weight (w kg)	Frequency	Frequency Density
$0 < w \leq 10$	40	
$10 < w \leq 15$	6	
$15 < w \leq 35$		2.6
$35 < w \leq 45$		1

Still using the **incorrect** GCSE formula:



Q2



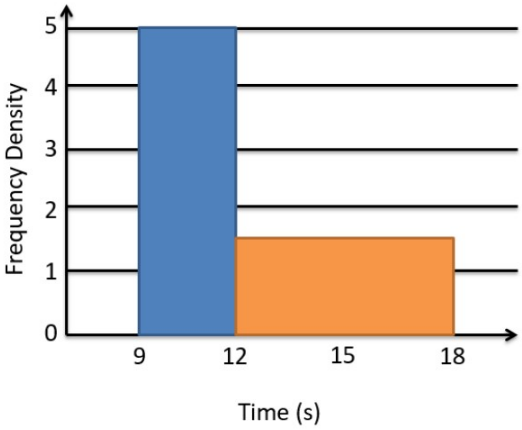
SKILL #1 Area = frequency x k

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

Identify the scaling using a known area with a known frequency (either total area/frequency or just one bar)

$$area = freq \times k$$ where k is a scaling constant

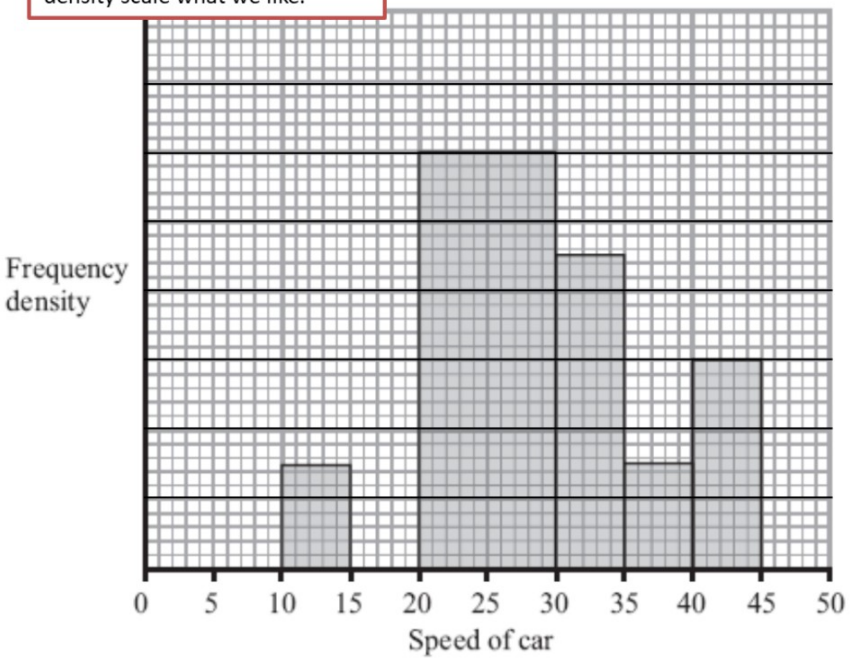
There were 60 runners in a 100m race. The following histogram represents their times. Determine the number of runners with times above 14s.



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.

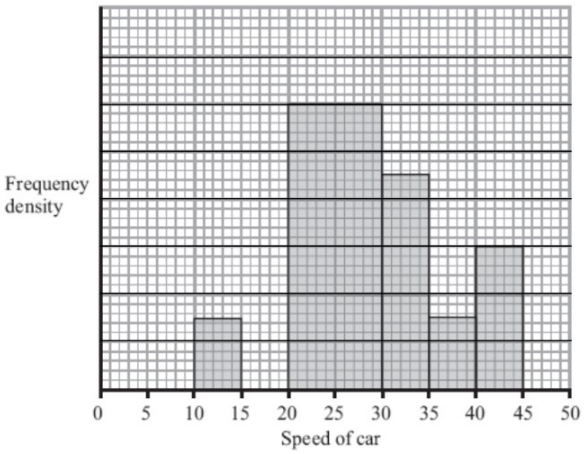
(*a*) Calculate the number of cars that were exceeding the speed limit by at least 5 mph in the sample.

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



(b) Estimate the value of the mean and median speed of the cars in the sample.

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.

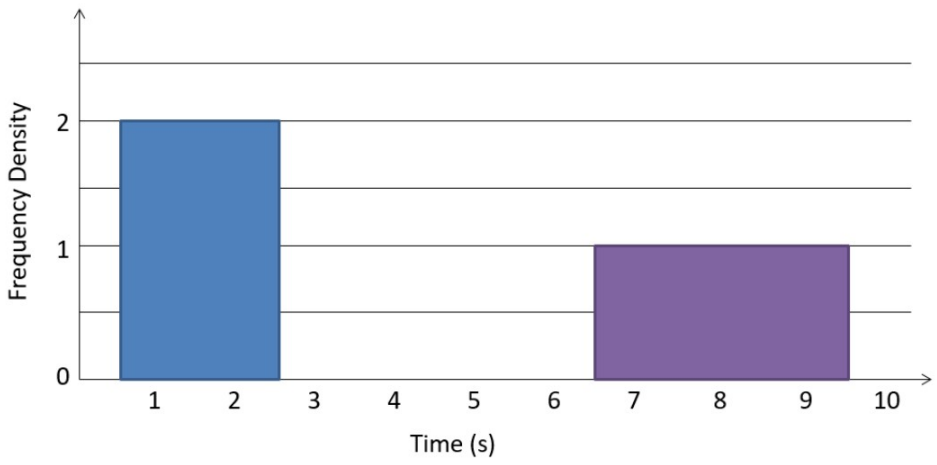


SKILL #2 Gaps

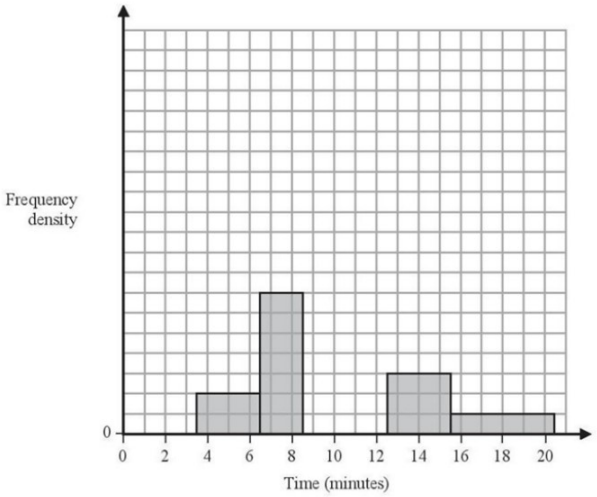
Weight (to nearest kg)	Frequency	F.D.
1-2	4	
3-6	3	
7-9		

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

1) Note the gaps affects class width
2) 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.



2. The partially completed histogram and the partially completed table show the time, to the nearest minute, that a random sample of motorists were delayed by roadworks on a stretch of motorway.



Delay (minutes)	Number of motorists
4 – 6	6
7 – 8	
9	17
10 – 12	45
13 – 15	9
16 – 20	

Estimate the percentage of these motorists who were delayed by the roadworks for between 8.5 and 13.5 minutes.

(5)

SKILL #3 Width and height on a diagram

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

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: Find the scaling for
class width → *drawn width* or
frequency density → *drawn height* or
area → *frequency*

[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.

1. The number of hours of sunshine each day, y, for the month of July at Heathrow are summarised in the table below.

Hours	$0 \leqslant y < 5$	$5 \leqslant y < 8$	$8 \leqslant y < 11$	$11 \leqslant y < 12$	$12 \leqslant y < 14$
Frequency	12	6	8	3	2

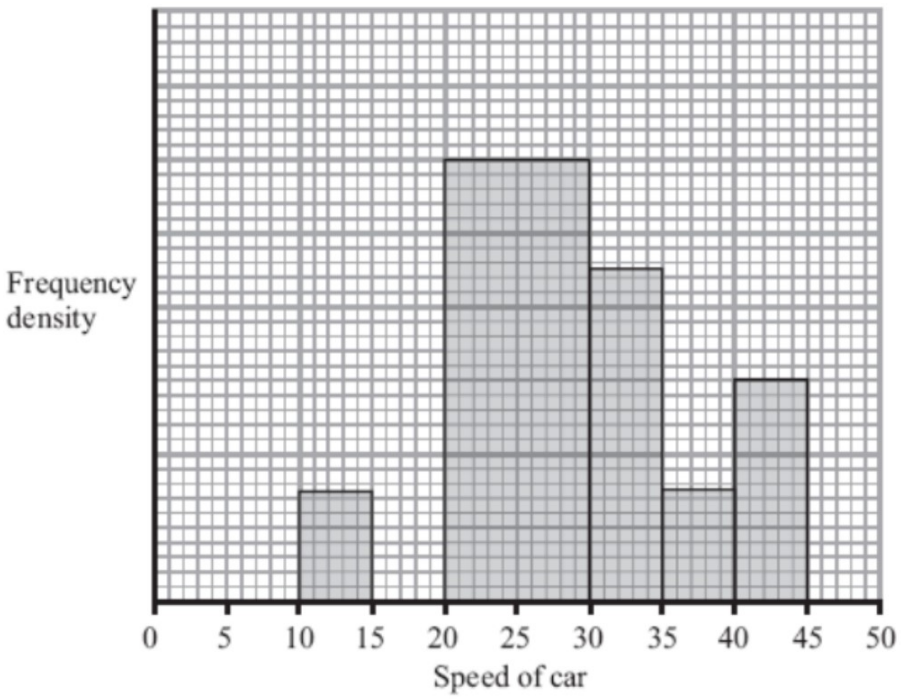
A histogram was drawn to represent these data. The $8 \leqslant y < 11$ group was represented by a bar of width 1.5 cm and height 8 cm.

(a) Find the width and the height of the $0 \leqslant y < 5$ group.

(3)

SKILL #3 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.

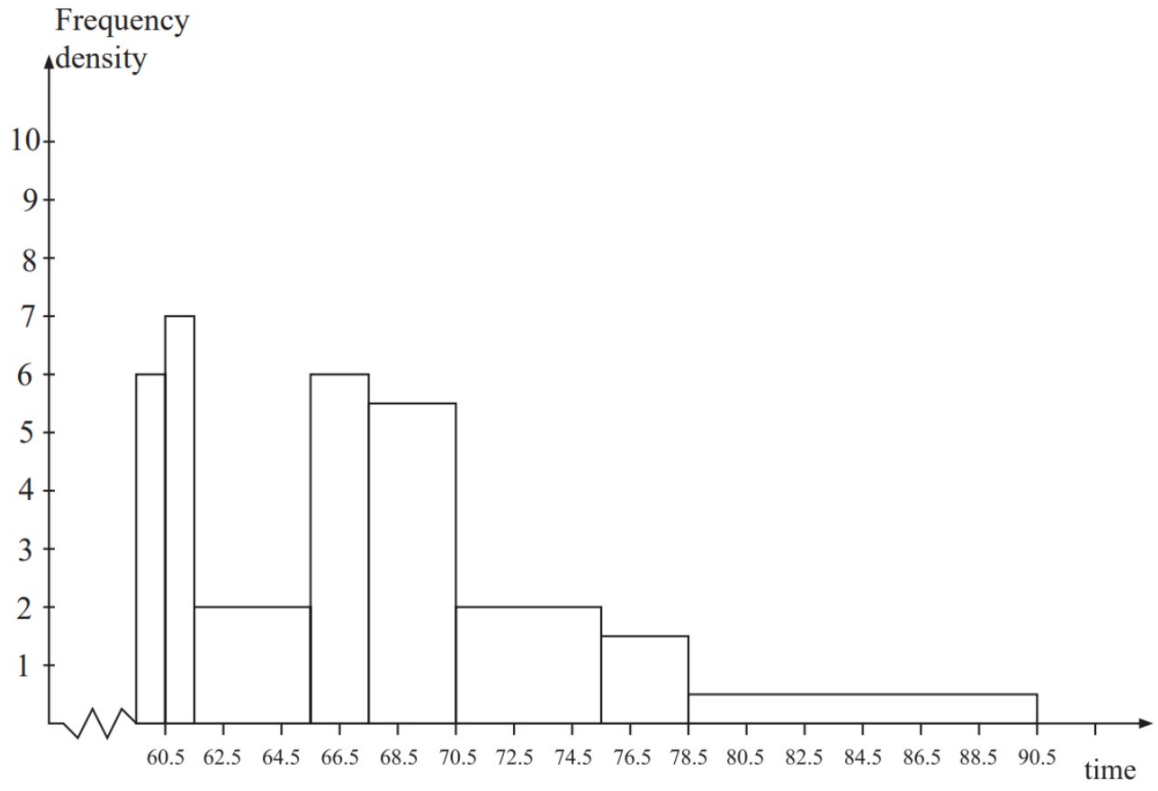


Ex 3D + Extra Questions
Note - Ex 3E has no new content, and is just comparing data

Extra Questions

1) 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.



2) 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)

3)

[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)**

4)

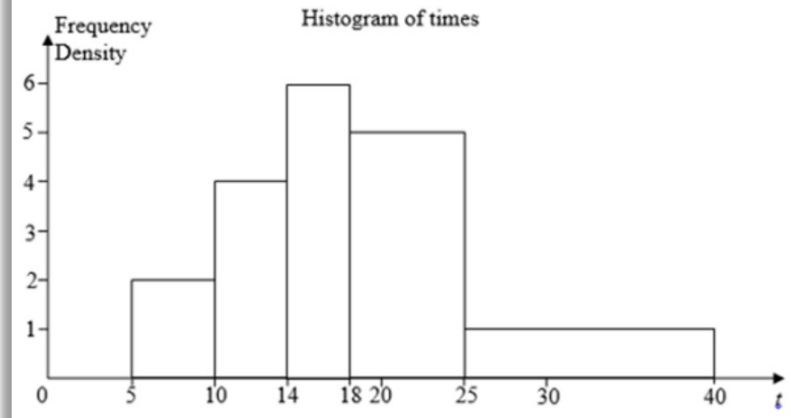


Figure 2 shows a histogram for the variable t which represents the time taken, in minutes, by a group of people to swim 500 m.

(*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)**

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

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

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

A histogram was drawn and the class $200 \leq y < 240$ 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 $320 \leq y < 400$ **(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)**

6) **[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)**

7) **[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)**

1)

Width	1	1	4	2	3	5	3	12
Freq. Density	6	7	2	6	5.5	2	1.5	0.5

M1

0.5×12 or 6 A1

Total area is $(1\times6)+(1\times7)+(4\times2)+\ldots.,=70$

$(90.5-78.5)\times\frac{1}{2}\times\frac{140}{\text{their } 70}$ M1

"70 seen anywhere" B1

Number of runners is 12 A1

2)

Answer: Distance is continuous

Note that gaps in the class intervals!
4 / 5 = 0.8
19 / 5 = 3.8
53 / 10 = 5.3
...

3)

(a) Width = 2×1.5 = **3 (cm)** B1

Area = $8\times1.5=12$ cm^2 Frequency = 24 so 1 cm^2 = 2 plants (o.e.) M1

Frequency of 12 corresponds to area of 6 so height = **2 (cm)** A1 (3)

(b) $[Q_2=]\ (5+)\ \frac{19}{24}\times5$ **or** (use of $(n+1)$) $(5+)\frac{19.5}{24}\times5$ M1

$=8.9583\ldots$ **awrt 8.96** **or** $9.0625\ldots$ awrt 9.06 A1 (2)

(c) $[\bar{x}=]\ \frac{755}{70}$ or **awrt 10.8** B1

$[\sigma_x=]\ \sqrt{\frac{12037.5}{70}-\bar{x}^2}\ =\sqrt{55.6326\ldots}$ M1A1ft

$=$ **awrt 7.46** (Accept s = awrt 7.51) A1 (4)

4)

5(a) 18-25 group, area=7x5=35
25-40 group, area=15x1=15

(b) (25-20)x5+(40-25)x1=40

(c) Mid points are 7.5, 12, 16, 21.5, 32.5
$\sum f=100$
$\frac{\sum ft}{\sum f}=\frac{1891}{100}=18.91$

(d) $\sigma_t=\sqrt{\frac{41033}{100}-\bar{t}^2}$ $\sqrt{\frac{n}{n-1}\left(\frac{41033}{100}-\bar{t}^2\right)}$ alternative OK

$\sigma_t=\sqrt{52.74\ldots}=7.26$

(e) $Q_2=18$ or 18.1 if (n+1) used

$Q_1=10+\frac{15}{16}\times4=13.75$ or 15.25 numerator gives 13.8125

$Q_3=18+\frac{25}{35}\times7=23$ or 25.75 numerator gives 23.15

5)

Width = 4 (cm) B1

Area of 14 cm^2 represents frequency 28 and area of $4h$ represents 18 M1

Or $\frac{4h}{18}=\frac{14}{28}$ (o.e.) h = **2.25** (cm) A1

(b) $m=(240)+\frac{10}{22}\times80$ (o.e.) M1

$=276.36\ldots$ $(\frac{3040}{11})$ $((£)\mathbf{276}\le m<(£)276.5)$ A1

(c) $\sum fy=31600$ leading to $\bar{y}=316$ M1A1

$\sigma_y=\sqrt{\frac{12452800}{100}-(\bar{y})^2}$ $=157.07\ldots$ (awrt **157**) Allow $s=157.86\ldots$ M1A1

6)

(a) 23, 35.5 (may be in the table)

(b) Width of 10 units is 4 cm so width of 5 units is **2 cm**
Height = $2.6\times4=$**10.4 cm**

(c) $\sum fx=1316.5\Rightarrow\bar{x}=\frac{1316.5}{56}=$ awrt **23.5**

$\sum fx^2=37378.25$ can be implied

So $\sigma=\sqrt{\frac{37378.25}{56}-\bar{x}^2}=$ awrt**10.7** allow $s=10.8$

(d) $Q_2=(20.5)+\frac{(28-21)}{11}\times5=23.68\ldots$ **awrt 23.7 or 23.9**

7)

(a) 8-10 hours: width = 10.5 - 7.5 = 3 represented by 1.5cm
16-25 hours: width = 25.5 - 15.5 = 10 so represented by 5 cm B1
8- 10 hours: height = fd = 18/3 = 6 represented by 3 cm M1
16-25 hours: height = fd = 15/10 = 1.5 represented by 0.75 cm A1 (3)

(b) $Q_2=7.5+\frac{(52-36)}{18}\times3=10.2$ M1 A1

$Q_1=5.5+\frac{(26-20)}{16}\times2[=6.25\ \text{ or }\ 6.3]$ or $5.5+\frac{(26.25-20)}{16}\times2$ [=6.3] A1

$Q_3=10.5+\frac{(78-54)}{25}\times5[=15.3]$ or $10.5+\frac{(78.75-54)}{25}\times5$ [=15.45 \15.5] A1

IQR = (15.3 - 6.3) = 9 A1ft (5)

(c) $\sum fx=1333.5\Rightarrow\bar{x}=\frac{1333.5}{104}=$ AWRT 12.8 M1 A1

(d) $\sum fx^2=27254\Rightarrow\sigma_x=\sqrt{\frac{27254}{104}-\bar{x}^2}=\sqrt{262.05-\bar{x}^2}$ AWRT 9.88 M1 A1

Exam Questions

4. Helen is studying the daily mean wind speed for Camborne using the large data set from 1987. The data for one month are summarised in Table 1 below.

Windspeed	n/a	6	7	8	9	11	12	13	14	16
Frequency	13	2	3	2	2	3	1	2	1	2

Table 1

(a) Calculate the mean for these data.

(1)

(b) Calculate the standard deviation for these data and state the units.

(2)

The means and standard deviations of the daily mean wind speed for the other months from the large data set for Camborne in 1987 are given in Table 2 below. The data are not in month order.

Month	*A*	*B*	*C*	*D*	*E*
Mean	7.58	8.26	8.57	8.57	11.57
Standard Deviation	2.93	3.89	3.46	3.87	4.64

Table 2

(c) Using your knowledge of the large data set, suggest, giving a reason, which month had a mean of 11.57

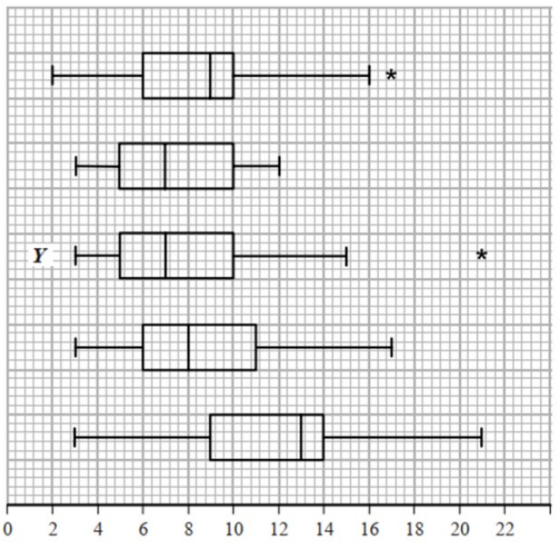
(2)

The data for these months are summarised in the box plots on the opposite page.
They are not in month order or the same order as in Table 2.

(d) (i) State the meaning of the * symbol on some of the box plots.

(ii) Suggest, giving your reasons, which of the months in Table 2 is most likely to be summarised in the box plot marked *Y*.

(3)



4. Charlie is studying the time it takes members of his company to travel to the office. He stands by the door to the office from 08 40 to 08 50 one morning and asks workers, as they arrive, how long their journey was.

(a) State the sampling method Charlie used.

(1)

(b) State and briefly describe an alternative method of non-random sampling Charlie could have used to obtain a sample of 40 workers.

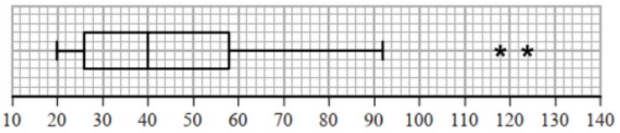
(2)

Taruni decided to ask every member of the company the time, x minutes, it takes them to travel to the office.

(c) State the data selection process Taruni used.

(1)

Taruni's results are summarised by the box plot and summary statistics below.



Journey time (minutes)

$n = 95 \quad \sum x = 4133 \quad \sum x^2 = 202\,294$

(d) Write down the interquartile range for these data.

(1)

(e) Calculate the mean and the standard deviation for these data.

(3)

(f) State, giving a reason, whether you would recommend using the mean and standard deviation or the median and interquartile range to describe these data.

(2)

Rana and David both work for the company and have both moved house since Taruni collected her data.

Rana's journey to work has changed from 75 minutes to 35 minutes and David's journey to work has changed from 60 minutes to 33 minutes.

Taruni drew her box plot again and only had to change two values.

(g) Explain which two values Taruni must have changed and whether each of these values has increased or decreased.

(3)

2.

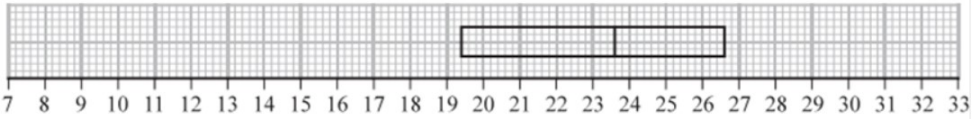


Figure 1

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)

Simon decides to model the air temperatures with the random variable

$$T \sim N(22.6, 5.19^2)$$

(d) Using Simon's model, calculate the 10th to 90th interpercentile range.

(3)