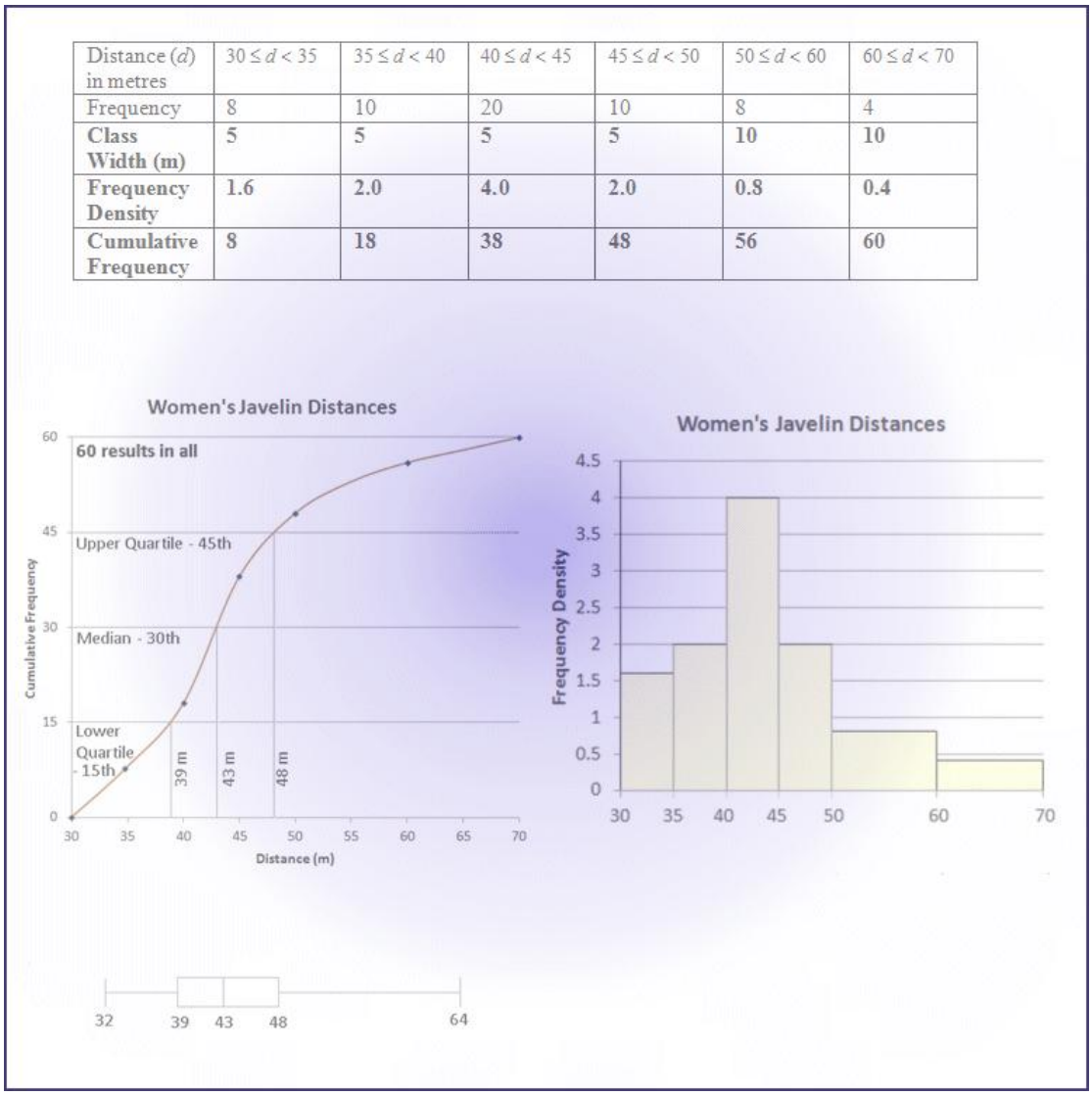
M.K. HOME TUITION

Mathematics Revision Guides
Level: GCSE Higher Tier

HISTOGRAMS, CUMULATIVE FREQUENCY AND BOX PLOTS



Distance (d) in metres	$30 \le d < 35$	$35 \le d < 40$	$40 \le d < 45$	$45 \le d < 50$	$50 \le d < 60$	$60 \le d < 70$
Frequency	8	10	20	10	8	4
Class Width (m)	**5**	**5**	**5**	**5**	**10**	**10**
Frequency Density	**1.6**	**2.0**	**4.0**	**2.0**	**0.8**	**0.4**
Cumulative Frequency	**8**	**18**	**38**	**48**	**56**	**60**

Version: 2.4 Date: 25-10-2015

Histograms.

This section deals with two additional ways of representing data; the histogram and the cumulative frequency diagrams.

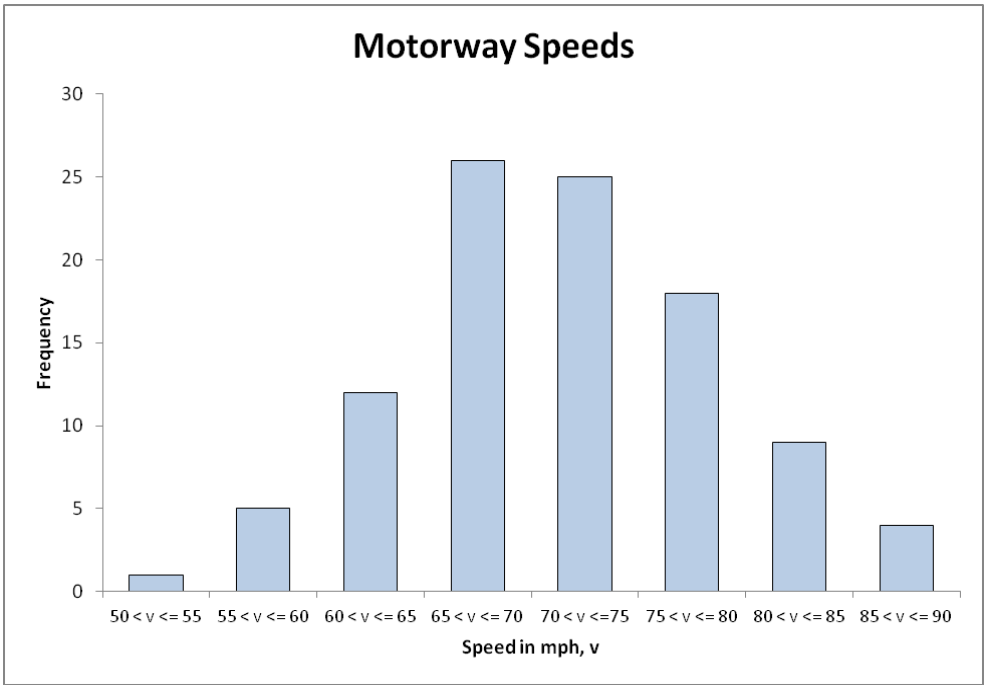
Example (1):

A police officer has measured the speeds of 100 cars on a speed camera on the hard shoulder of a part of the M60 motorway where the speed limit is 70 mph. The table of results is shown below.

Speed (v) in mph	Frequency
$50 \le v < 55$	1
$55 \le v < 60$	5
$60 \le v < 65$	12
$65 \le v < 70$	26
$70 \le v < 75$	25
$75 \le v < 80$	18
$80 \le v < 85$	9
$85 \le v < 90$	4

Two pupils, Amy and Beth, produced bar charts to illustrate the data.

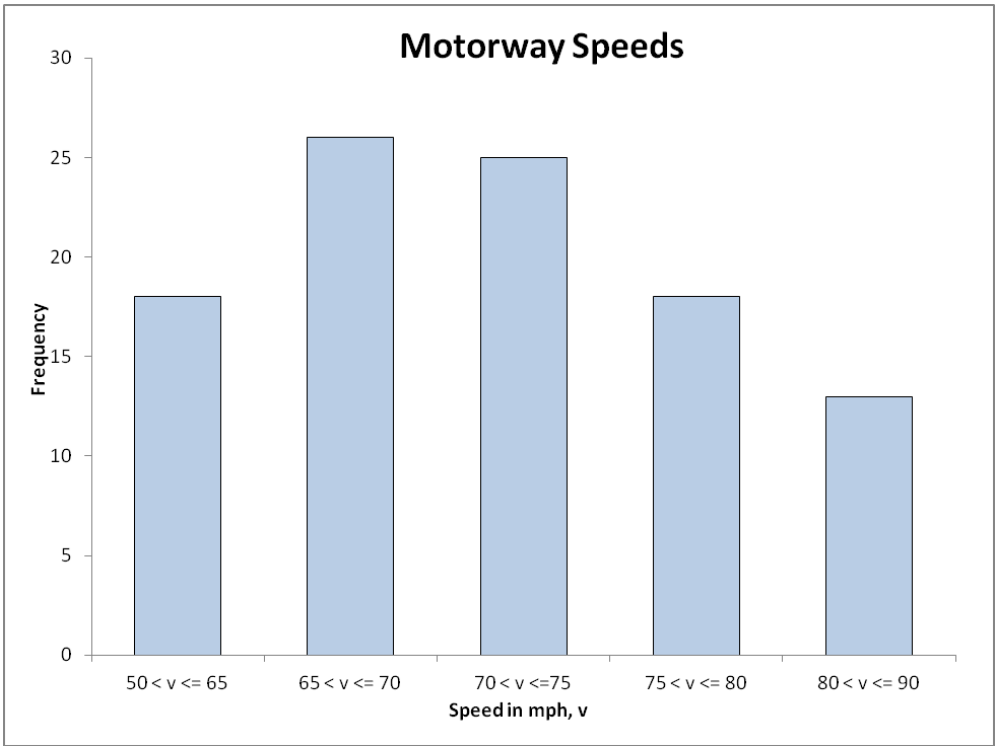
Amy used the police data in its original grouped form, to produce this bar chart:



Notice the strong concentration of cars recorded travelling at 65-75 mph , and the rapid tail-off at the higher, and especially the lower, speeds.

Beth, on the other hand, had regrouped the frequencies for all the cars driven slower than 65 mph, and those driven faster than 80 mph.

Speed (v) in mph	Frequency
$50 \le v < 65$	18
$65 \le v < 70$	26
$70 \le v < 75$	25
$75 \le v < 80$	18
$80 \le v < 90$	13



Beth's bar chart appears to show a more uniform height for the bars, because of the way she had grouped the data.

Histograms versus bar charts.

The bar charts from Example(1) appear to be different in shape, despite the fact that they represent the same data. Amy's bar chart shows a strong concentration of cars recorded at 65-75 mph , with rapid tail-offs towards the higher and lower speeds. By contrast, Beth's bar chart seems to be more rectangular in shape with less concentration towards the middle range of speeds.

Because of this, we have to use **histograms** to represent such data without bias.

A histogram is a special type of bar chart, with the following characteristics:

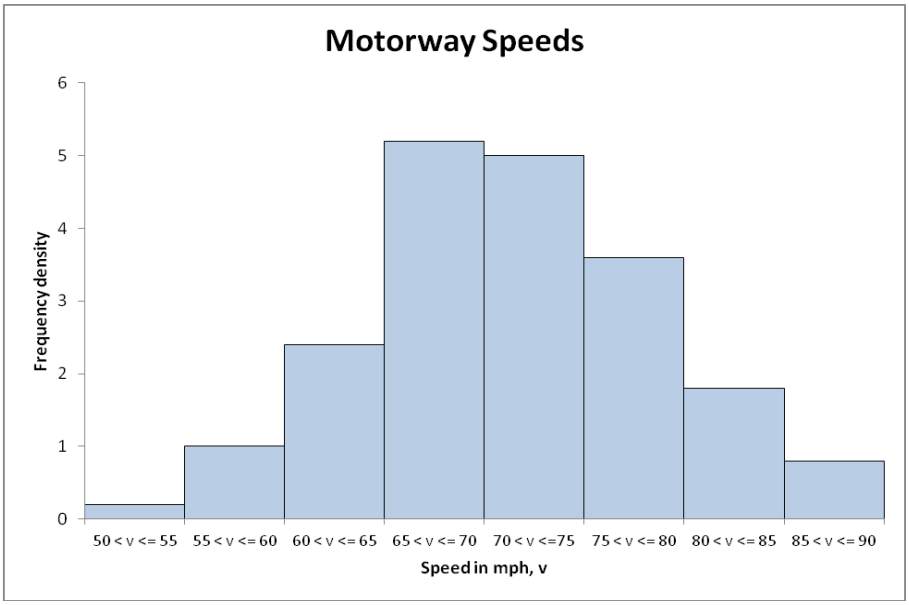
- The frequency of the data is denoted by the **area** of each bar rather than by its height
- The bars are not separated
- The width of the bar corresponds to the width of the class interval
- The data is numeric, but can be continuous or discrete.

A histogram is not merely "a bar chart with the gaps between bars removed".

Because frequency is now represented by area rather than height, we introduce the idea of **frequency density.** This can be obtained by dividing the actual frequency by the class width.

Going back to the original police data, we produce this new table and histogram.

Speed (v) in mph	Frequency, f	**Class width, w**	**Frequency density, f/w**
$50 \le v < 55$	1	**5**	**0.2**
$55 \le v < 60$	5	**5**	**1.0**
$60 \le v < 65$	12	**5**	**2.4**
$65 \le v < 70$	26	**5**	**5.2**
$70 \le v < 75$	25	**5**	**5.0**
$75 \le v < 80$	18	**5**	**3.6**
$80 \le v < 85$	9	**5**	**1.8**
$85 \le v < 90$	4	**5**	**0.8**



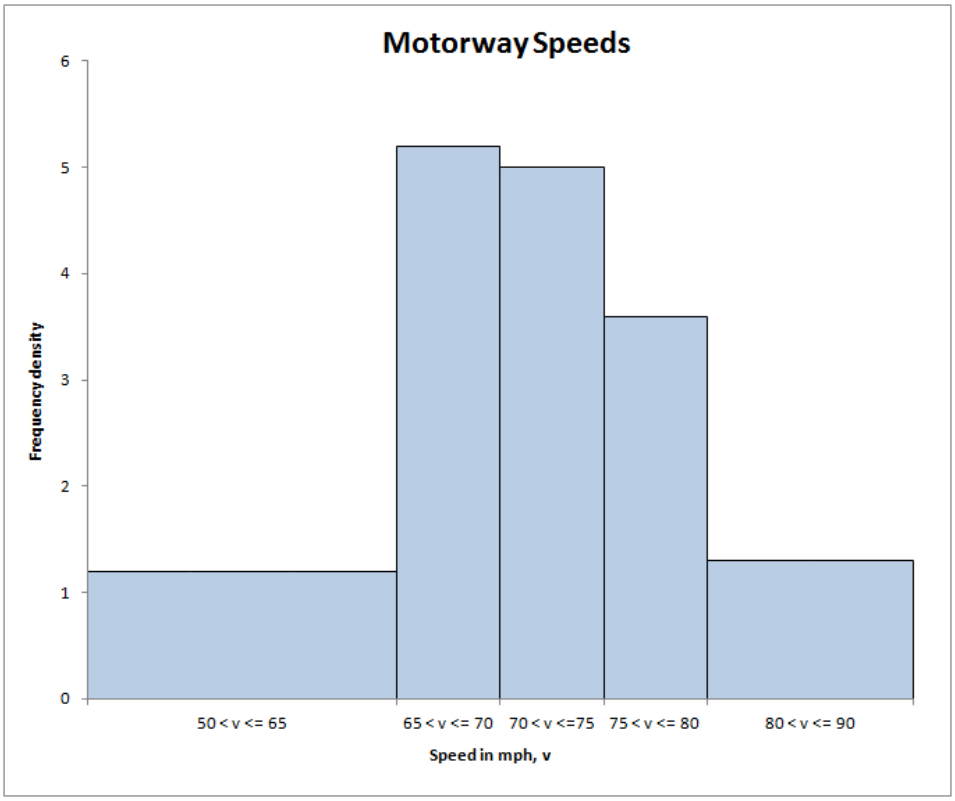
The gaps between the bars have been removed, and the vertical axis now shows frequency density, but the shape of the distribution is no different from that of Amy's original bar chart.

Example (2) : Produce a histogram from Beth's data for motorway speeds.

We calculate frequency densities, remembering that the class intervals are now of varying widths.

Speed (v) in mph	Frequency, f	**Class width, w**	**Frequency density, f/w**
$50 \le v < 65$	18	**15**	**1.2**
$65 \le v < 70$	26	**5**	**5.2**
$70 \le v < 75$	25	**5**	**5**
$75 \le v < 80$	18	**5**	**3.6**
$80 \le v < 90$	13	**10**	**1.3**

The resulting histogram is shown below.



This histogram is less detailed than Amy's, due to the crude grouping of the data, but the overall shape is still the same.

Notice how the "50-65 mph" and "75-80 mph" bars have the same area, given that they each represent 18 cars, even though their linear proportions are different.

Example (3): A garage issued a pledge to its customers whereby any brand new car purchased from them could be returned to have any faults fixed free of charge within the first year, under guarantee.

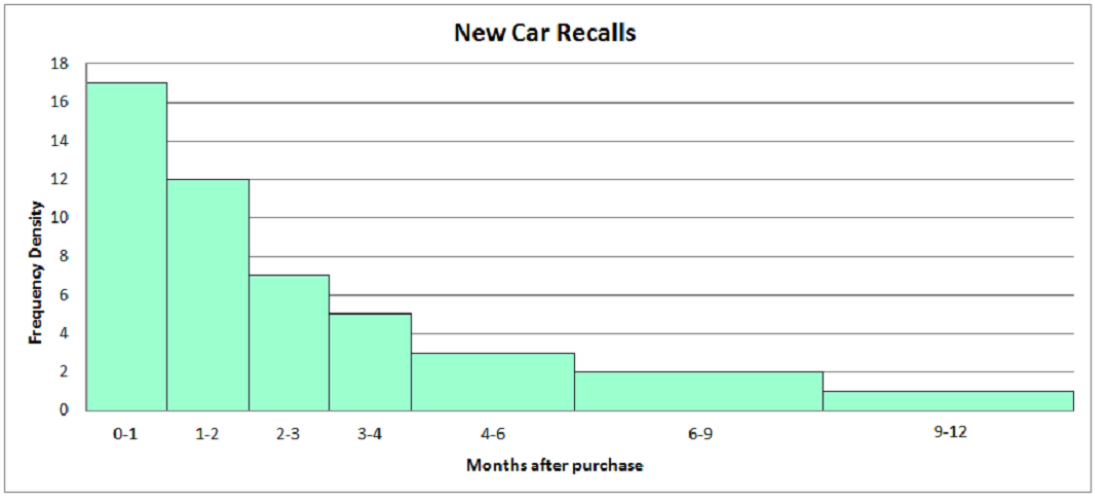
The number of customer returns based on the time after purchasing the car was recorded below.

Months after purchase	0-1	1-2	2-3	3-4	4-6	6-9	9-12
Recalls	17	12	7	5	6	6	3

Draw a histogram to represent the data.

The data is continuous here, so the '0-1' class width is 1, the '4-6' class width is 2, and so on.

Months after purchase	0-1	1-2	2-3	3-4	4-6	6-9	9-12
Recalls (frequency), *f*	17	12	7	5	6	6	3
Class width, *w*	**1**	**1**	**1**	**1**	**2**	**3**	**3**
Frequency density, *f/w*	**17**	**12**	**7**	**5**	**3**	**2**	**1**



Example (4a): The table below shows the distribution of marks (rounded to nearest unit) for 96 pupils taking Maths Paper 1. Plot the results on a histogram.

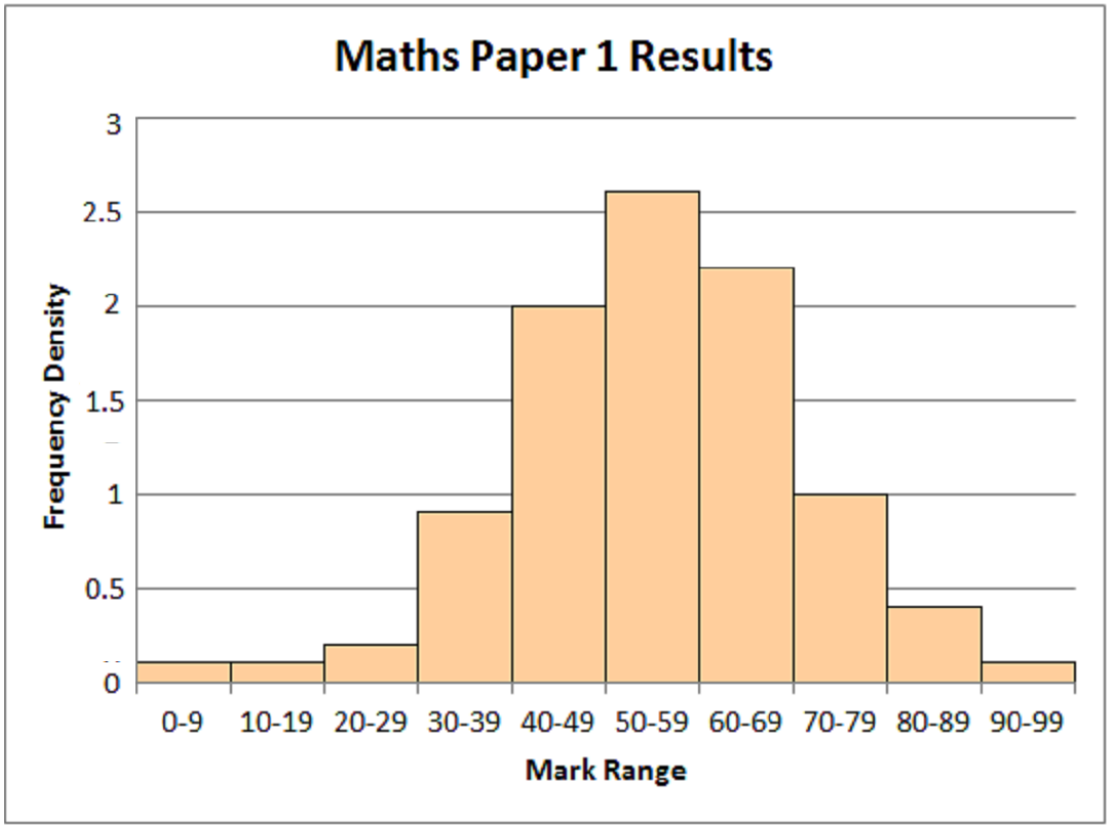
Range of marks	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Frequency	1	1	2	9	20	26	22	10	4	1
Frequency density	**0.1**	**0.1**	**0.2**	**0.9**	**2.0**	**2.6**	**2.2**	**1.0**	**0.4**	**0.1**

Because the marks form a set of rounded discrete data, the true class intervals have a width of 10, such that the "50-59" class is actually $49.5 \leq (\text{mark}) < 59.5$.

We therefore need to divide all the frequencies by 10 to get the frequency densities.

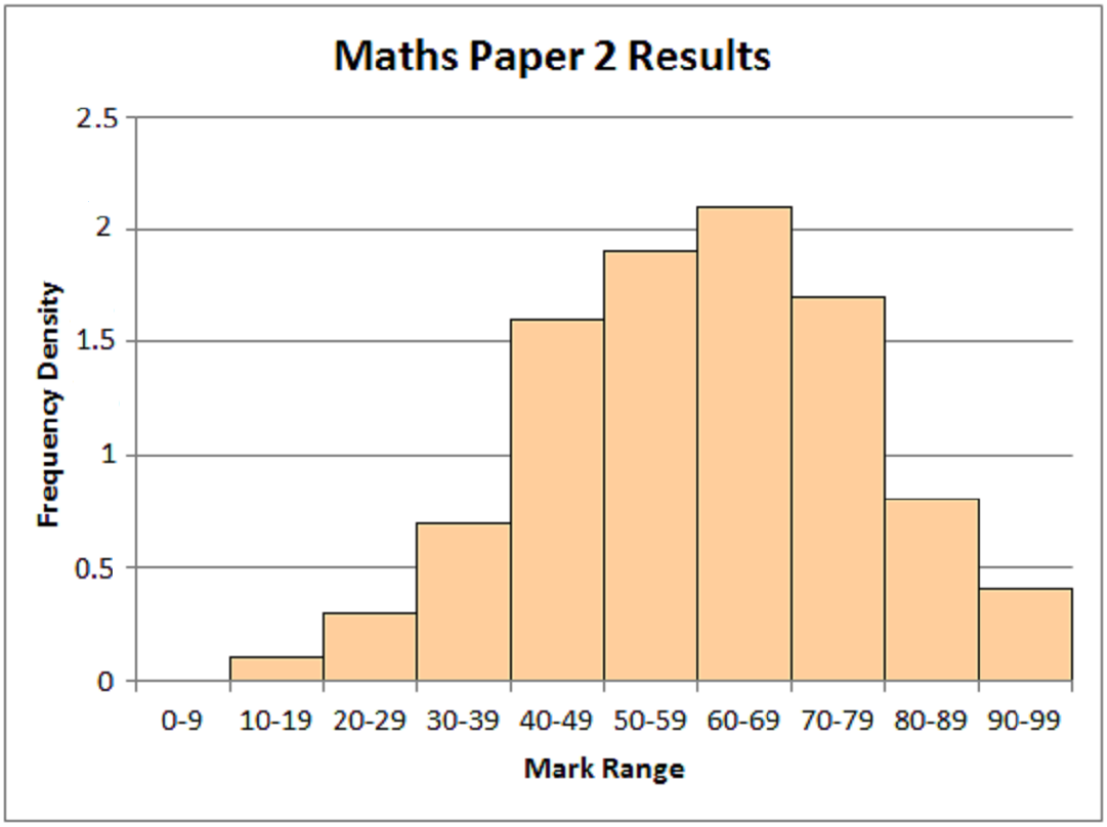
Note how there is a peak at the modal 50-59% class interval and the intervals on either side of it.

Because we are rounding to the nearest unit, the 50-59% class interval is strictly $49.5 \leq (\text{mark}) < 59.5$, but this is not an issue in this particular example.



Example (4b): Plot a similar histogram to that in Example(4) for Maths Paper 2. How does it differ from the previous one ?

Range of marks	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Frequency	0	1	3	7	16	19	21	17	8	4
Frequency density	**0.0**	**0.1**	**0.3**	**0.7**	**1.6**	**1.9**	**2.1**	**1.7**	**0.8**	**0.4**



This time the modal range is now 60-69%, with less of a concentration about the centre than there we had with Paper 1. There also seems to be a skew towards the higher marks.

There are ways of expressing this spread, called variance and standard deviation, but this is outside the scope of GCSE.

Example (5): Sam is a keen gardener and his two most productive apple trees are a Cox and a Gala. The details of his harvest last autumn are as below.

The data values are identical to those in Example (9) of the "Averages" document.

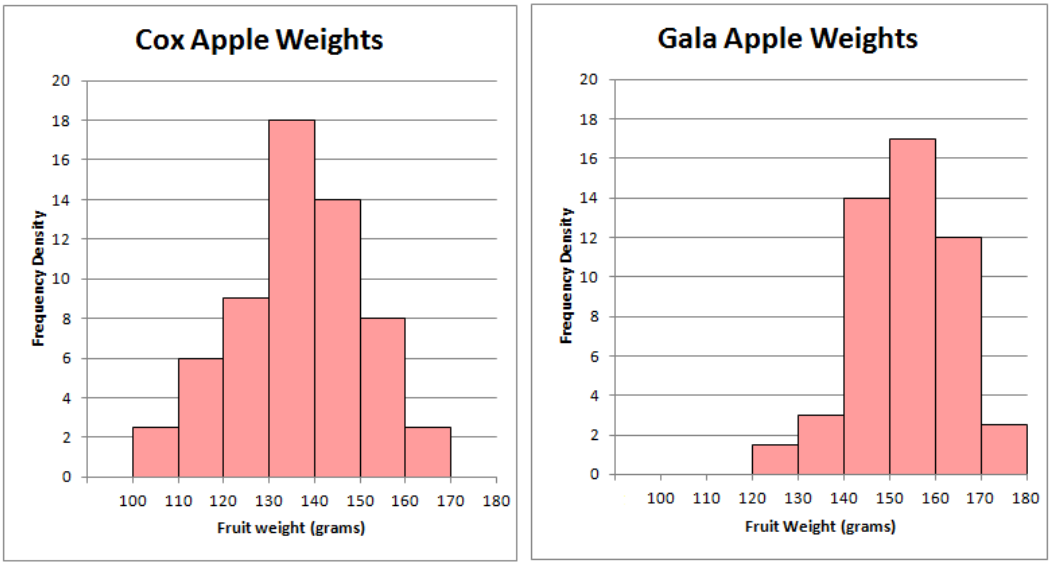
Cox	
Weight of apple, w, (g)	Frequency, f
$100 \le w < 110$	25
$110 \le w < 120$	60
$120 \le w < 130$	90
$130 \le w < 140$	180
$140 \le w < 150$	140
$150 \le w < 160$	80
$160 \le w < 170$	25
TOTAL	**600**

Gala	
Weight of apple, w, (g)	Frequency, f
$120 \le w < 130$	15
$130 \le w < 140$	30
$140 \le w < 150$	140
$150 \le w < 160$	170
$160 \le w < 170$	120
$170 \le w < 180$	25
TOTAL	**500**

i) Represent the data on two separate histograms.
ii) Compare the two sets of data, writing at least two supporting comments.

The histograms for the above data are shown below.

(All the class widths are 10 here, so all frequency densities are one-tenth of the actual frequencies)



The modal classes are different for the two varieties of apple; for the Cox apples, the modal weight class is 130-140 grams, but for the Gala apples, the 150-160 gram class is the mode.

The range of weight values for the Cox apples is potentially 100 to 170 grams, i.e. a 70 gram difference, but for the Gala apples, this range is 60 grams, from 120 to 180 grams.
The data includes no actual extreme values, so these ranges are not accurate.

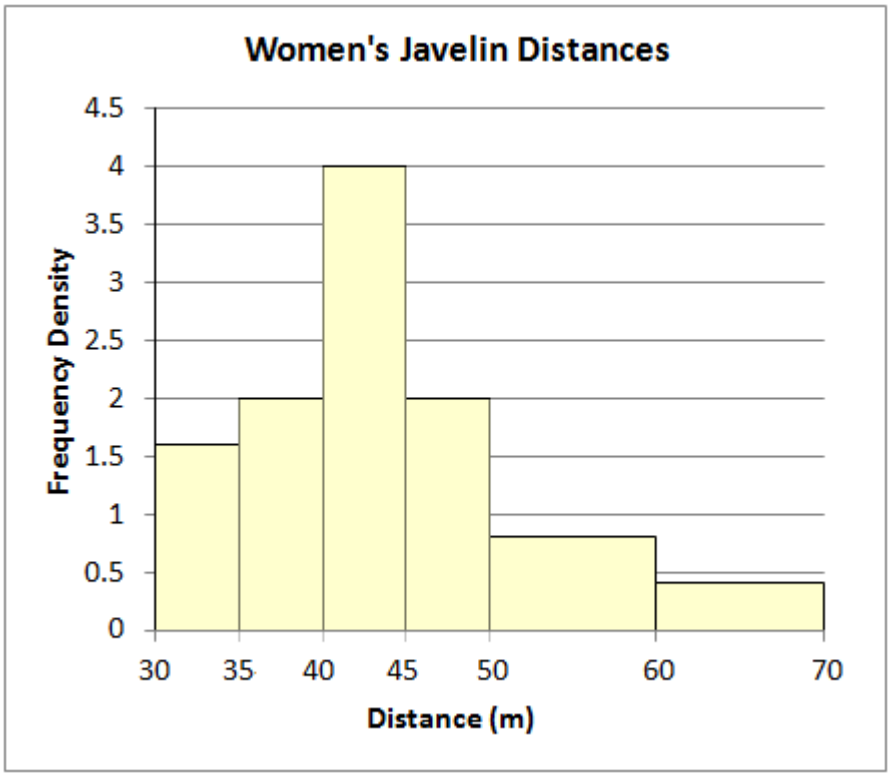
Example(6): An athletics club has analysed the results of the women's javelin throwing events over a season, and the following results obtained:

Draw and describe the histogram corresponding to the results below:

Distance (d) in metres	$30 \le d < 35$	$35 \le d < 40$	$40 \le d < 45$	$45 \le d < 50$	$50 \le d < 60$	$60 \le d < 70$
Frequency	8	10	20	10	8	4

Note that the data is continuous here, and also that the class intervals are of differing widths. We therefore need to work out frequency densities as follows:

Distance (d) in metres	$30 \le d < 35$	$35 \le d < 40$	$40 \le d < 45$	$45 \le d < 50$	$50 \le d < 60$	$60 \le d < 70$
Frequency	8	10	20	10	8	4
Class width	5	5	5	5	10	10
Frequency density	**1.6**	**2.0**	**4.0**	**2.0**	**0.8**	**0.4**



Cumulative frequency.

Another method of displaying grouped data is the cumulative frequency diagram, which keeps a running total of frequencies.

Example (7a): Plot a cumulative frequency graph for the results of Maths Paper 1.

Range of marks	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Frequency	1	1	2	9	20	26	22	10	4	1

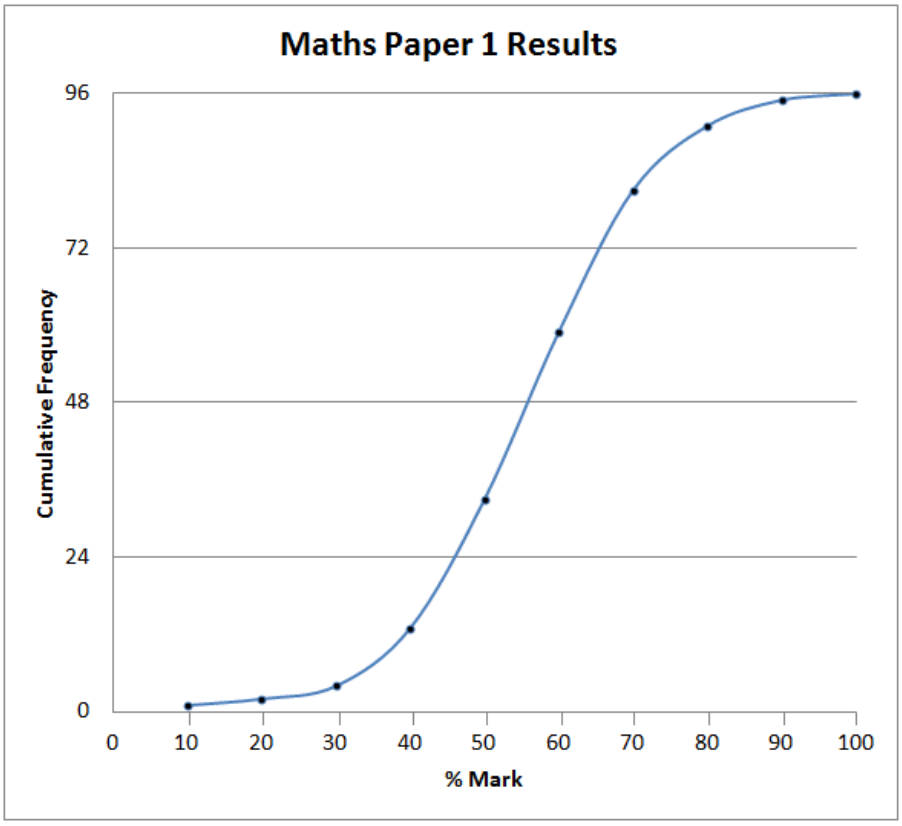
When we look at the frequency table, we can see that 1 pupil scored up to 9%, 1 + 1 or 2 pupils scored up to 19%, 1 + 1 + 2 or 4 pupils scored up to 29%, and so on until 96 pupils scored up to 99%.

We therefore put the running totals in a **cumulative frequency** row as follows:

Marks	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Frequency	1	1	2	9	20	26	22	10	4	1

Cumulative marks	0-9	0-19	0-29	0-39	0-49	0-59	0-69	0-79	0-89	0-99
Cumulative frequency	**1**	**2**	**4**	**13**	**33**	**59**	**81**	**91**	**95**	**96**

The results can then be put into a graph as shown below, joining points with a smooth curve.



Note how cumulative totals are lined up at 9.5, 19.5, 29.5 and so forth on the data axis.
This is because the class interval 40-49 (say) can take in all values from 39.5 to 49.5 (when rounded).

Example (7b): Plot a corresponding cumulative frequency graph for the results of Maths Paper 2.

Compare and contrast it with the previous graph.

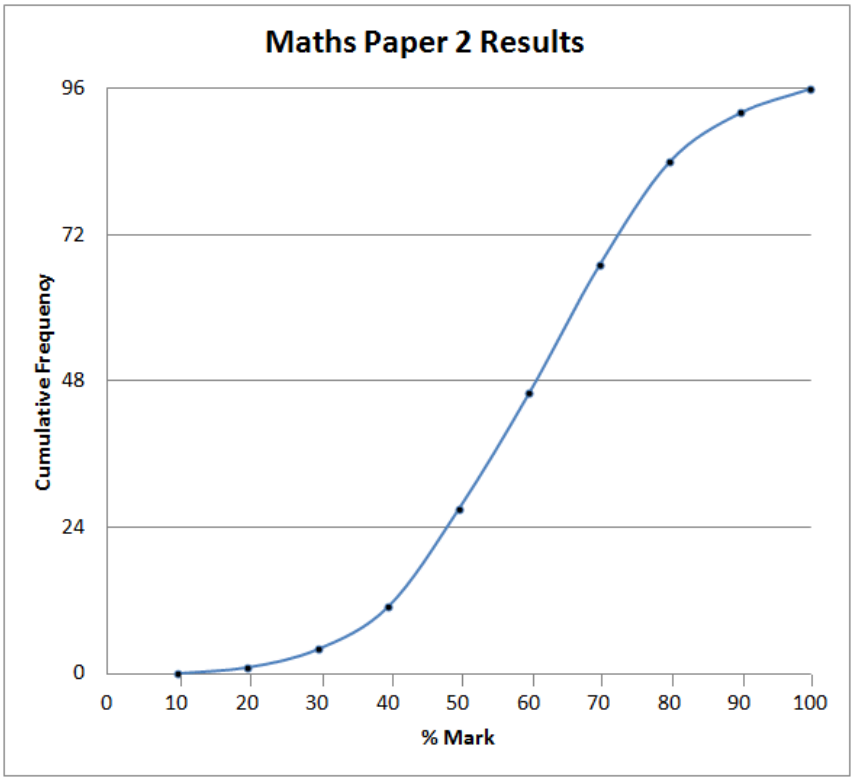
Marks	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Frequency	0	1	3	7	16	19	21	17	8	4

Firstly we work out the cumulative frequency at the end of each class interval :

Marks	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Frequency	0	1	3	7	16	19	21	17	8	4

Cumulative marks	0-9	0-19	0-29	0-39	0-49	0-59	0-69	0-79	0-89	0-99
Cumulative frequency	**0**	**1**	**4**	**11**	**27**	**46**	**67**	**84**	**92**	**96**

Finally, plot the cumulative frequencies (2 for 0-9, 7 for 10-19, 15 for 20-29....) against 9.5, 19.5, 29.5... on the data axis.

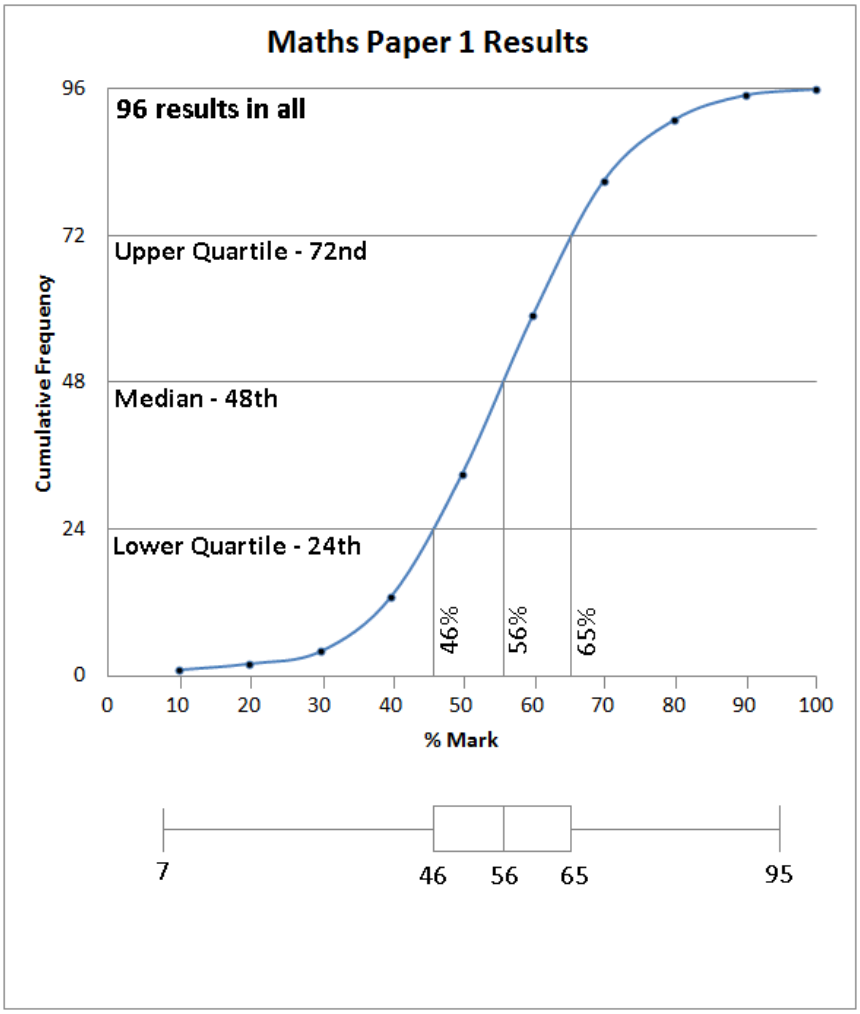


The cumulative frequency graph has a somewhat gentler slope overall, particularly around the median, suggesting a lesser concentration of marks around it.

Interpreting cumulative frequency graphs – finding the median and quartiles.

We will now look at how to read off information from a cumulative frequency graph, in this case the results of Maths Paper 1 from Example (7a).

Additionally, we are told that the lowest mark was 7 and the highest mark was 95.



There are marks for 96 pupils in all, so the median will correspond to the 48th mark out of the 96.

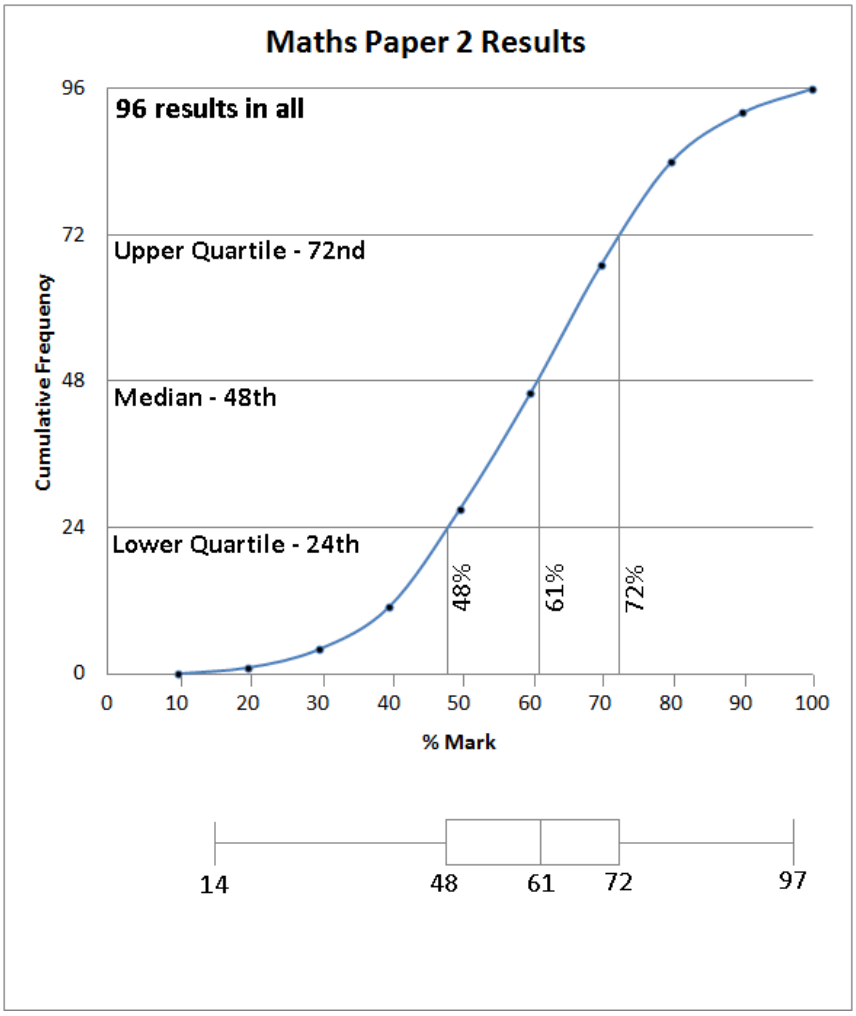
When using cumulative frequency graphs, we are not pedantically concerned with there being 'two middle numbers', the 48th and 49th . This is because we can, in any case, only estimate the median as we do not have the original data items.

We can find the median by drawing a horizontal line along the halfway point (48) of the cumulative frequency axis to meet the curve, and then drawing a vertical from the curve to the data axis, meeting it at about 56 – the estimated median.

Other statistical points are the upper and lower quartiles. Those points are three quarters and one quarter of the way up the cumulative frequency axis, and can be found in a similar way to the median.

The upper quartile corresponds to the 72nd result, and meets the data axis at 65%.
The lower quartile corresponds to the 24th result, and meets the data axis at 46%.

The **inter-quartile range** is the difference between the upper and lower quartiles, namely 19%.
Here are the corresponding results from the Maths Paper 2 results of Example (7b), where we have also been told that the lowest and highest marks are 14 and 97.



The inter-quartile range is rather higher than with Maths Paper 1 - 24% compared to 19%.
Also, the median and upper quartile are considerably higher than those for Maths Paper 1.

The diagrams at the bottom of the cumulative frequency plots are known as **box-and whisker** plots, and will be discussed in greater detail in the next section.

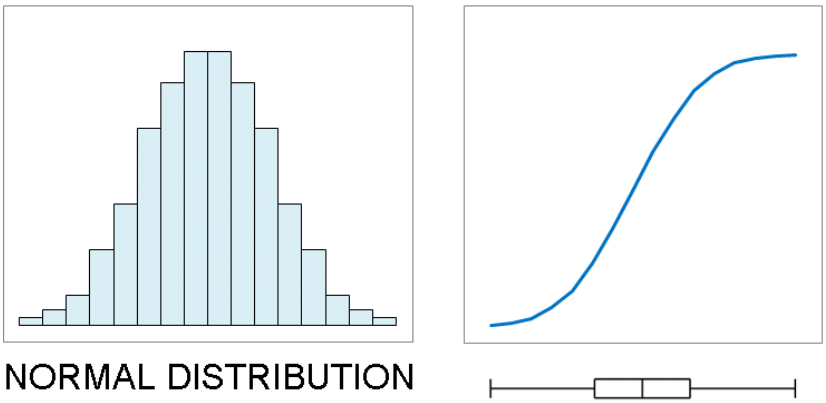
The Box and Whisker Plot.

The box-and-whisker plots shown at the bottom of the last two cumulative frequency plots are another 'shorthand' way of summarising distributions. In a box-and-whisker pot, the quartiles and median are displayed in a 'box' and the two 'extreme' values are shown as 'whiskers'.

The vertical lines in the box-and-whisker plot coincide with the crucial vertical lines in the graph..

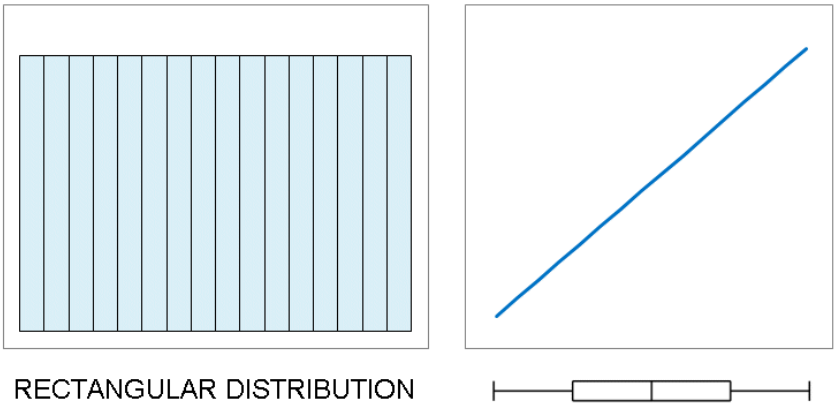
Box-and-whisker plots can show irregularities in a distribution, namely **positive and negative skew**.

There are four main distribution patterns to recognise at GCSE; normal, rectangular, positive skew and negative skew. The histograms, cumulative frequency graphs and box-and-whisker plots are shown on the next two pages for comparison.



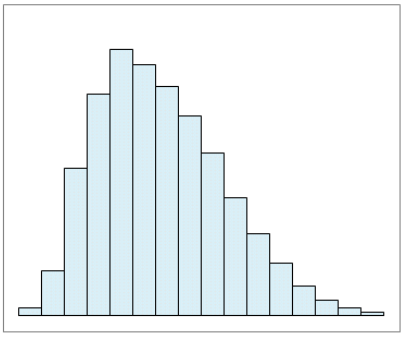
NORMAL DISTRIBUTION

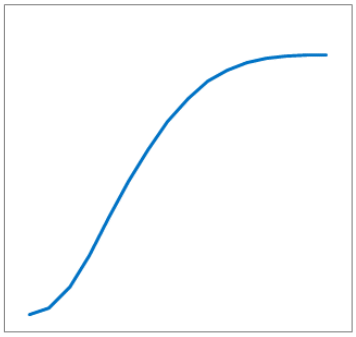
Most histograms at GCSE display a symmetrical or 'normal' distribution, characterised by a bell-shaped histogram and a balanced cumulative frequency diagram and box plot, where the quartiles are closer to the median than they are to the extreme values. In addition, there is little difference between the mean and the median.



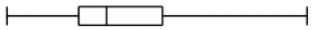
RECTANGULAR DISTRIBUTION

In a rectangular distribution, there is no concentration of marks about the centre, all the columns of the histogram are equal in height, the cumulative frequency curve is a straight line, and the quartiles are equally distant from the median and the extremes. Again, the mean and median are practically coincident.

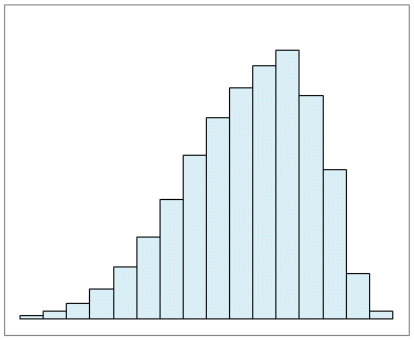


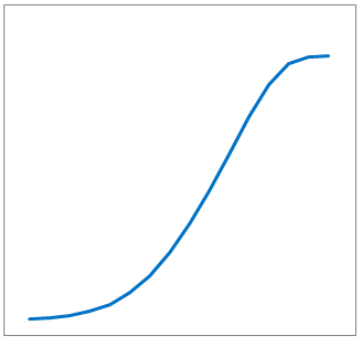


POSITIVE SKEW



This time, we have an asymmetrical distribution where the median is less than the mean, and is closer to the lower quartile. This is evident in the histogram where the highest frequencies occur to the left of the central value. In addition, the 'whiskers' are unequal in length.





NEGATIVE SKEW

This is the mirror case to the previous one – here, the median is greater than the mean, and is closer to the upper quartile, with the 'whiskers' again unequal in length.

The length of the 'box' compared with that of the 'whiskers' can also give information about the spread of the data.

Example (8a): Repeat Example (5) but plot two cumulative frequency diagrams for the weights of the apples in Sam's garden, including the box and whisker plots.

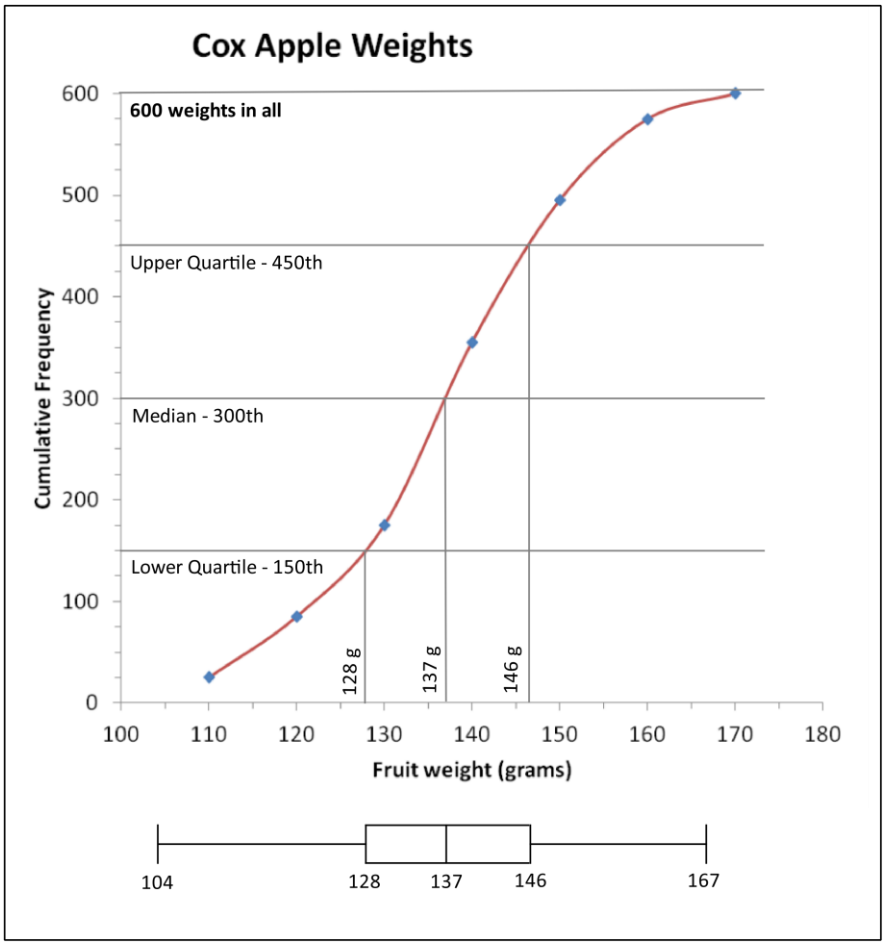
Cox apples :

Weight (w) in grams	$100 \le w < 110$	$110 \le w < 120$	$120 \le w < 130$	$130 \le w < 140$	$140 \le w < 150$	$150 \le w < 160$	$160 \le w < 170$
Frequency	25	60	90	180	140	80	25

Cumulative weight (w)	$100 \le w < 110$	$100 \le w < 120$	$100 \le w < 130$	$100 \le w < 140$	$100 \le w < 150$	$100 \le w < 160$	$100 \le w < 170$
Cumulative Frequency	**25**	**85**	**175**	**355**	**495**	**575**	**600**

We are also given that the lightest Cox apple weighed 104 g and the heaviest, 167 g.

Because the data is assumed to be continuous, we line up the points at 110, 120, 130 ... on the data axis. the completed cumulative frequency graph is shown below, complete with box-and-whisker plot.



Gala apples :

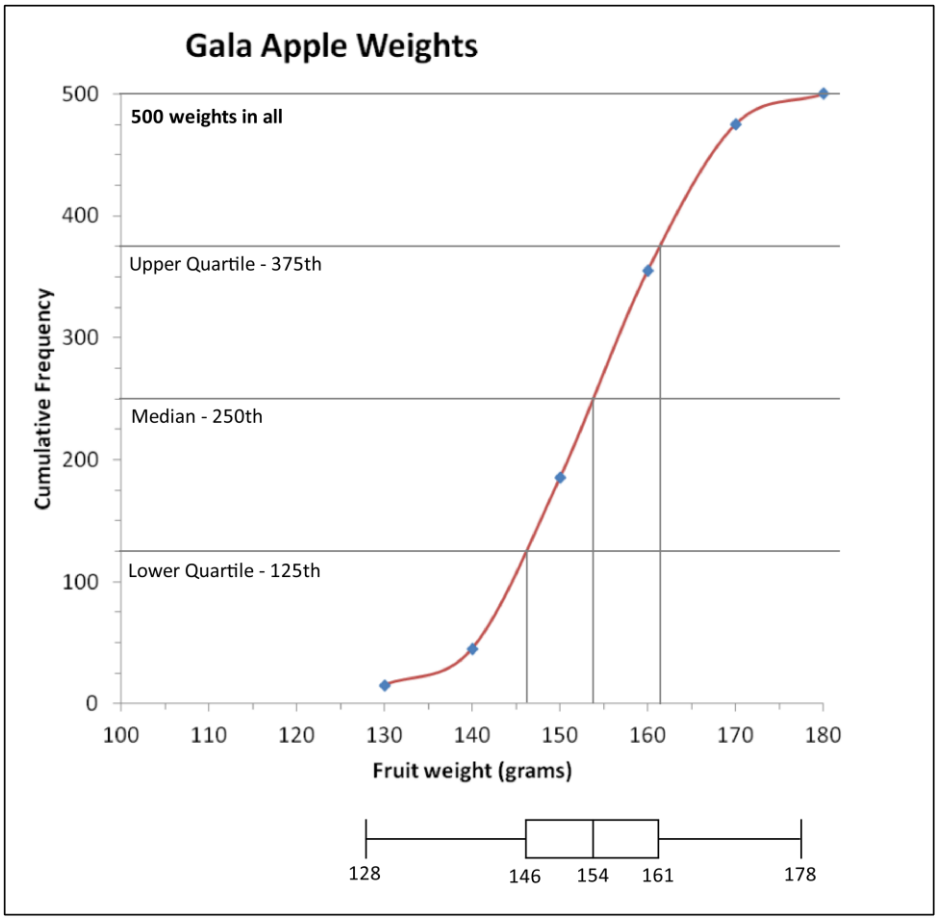
Weight (w) in grams	$120 \le w < 130$	$130 \le w < 140$	$140 \le w < 150$	$150 \le w < 160$	$160 \le w < 170$	$170 \le w < 180$
Frequency	15	30	140	170	120	25

Cumulative weight	$100 \le w < 130$	$100 \le w < 140$	$100 \le w < 150$	$100 \le w < 160$	$100 \le w < 170$	$100 \le w < 180$
Cumulative Frequency	**15**	**45**	**185**	**355**	**475**	**500**

The corresponding cumulative frequency graph for the Gala apples is shown below, with the box-and-whisker plot. (Continuous data - we line up the points at 130, 140, 150 ... on the data axis).

The lightest Gala apple weighed 128 g, and of the heaviest one, 178 g.

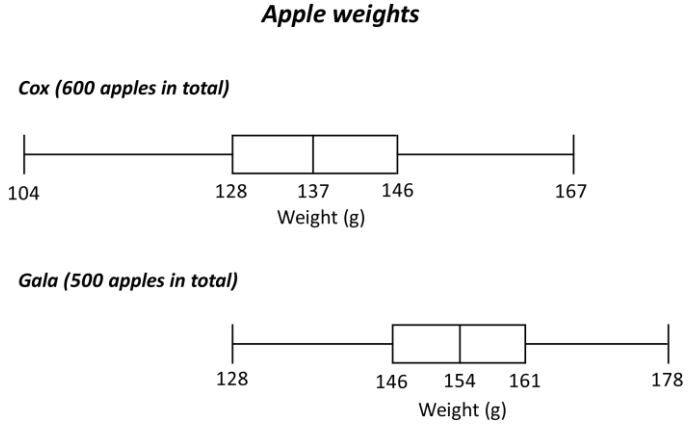
The horizontal axis begins on 100 grams to provide a direct comparison with the cumulative frequency graph for the Cox apples.



The box-and-whisker plots can be shown by themselves to provide an instant comparison of data sets. .

In an examination question, such plots will not generally have actual values stated, but will be shown on a graph background, with matching horizontal scales.

For a box-and-whisker example without a supporting cumulative frequency diagram, see Example (15) later in the document.



Example (8b): Rose has been looking at the box plots above for Sam's apple weights and has written down the following conclusions :

- A quarter of the Cox apples were lighter than the lightest Gala apple.
- The number of Gala apples weighing *below* 146 grams was the same as the number of Cox apples weighing *above* 146 grams.
- The Gala apples were heavier overall than the Cox apples.
- The Gala apples were slightly more consistent in weight than the Cox apples.

Comment on her results.

The lower quartile weight for the Cox apples was 128 grams, meaning that a quarter of them weighed less than that. By coincidence, the lightest of the Gala apples weighed 128 grams.

All the 'key' weights for the Gala apples – median, quartiles and extremes – were higher than those for the Cox apples, implying that the Gala apples were heavier overall than the Cox apples. .

The total weight range for the Gala apples was 178-128, or 50 grams, and for the Cox apples, it was 167-104, or 63 grams. The inter-quartile ranges were 146-128, or 18 grams, for the Cox apples, and 161-146, or 15 grams, for the Gala apples. This implies that the Gala apples were slightly more consistent in weight than the Cox apples.

Rose's first, third and fourth statements were therefore correct, but the second one was subtly wrong. The 146 gram cut-off weight was the *lower* quartile for the Gala apples, implying that a quarter of the Gala apples were lighter, and that same weight was the *upper* quartile for the Cox apples, implying that a quarter of the Cox apples were heavier.

The *proportions* of apples in those two categories were equal, but the *numbers* were not, as a quarter of 600 Cox apples is not the same as a quarter of 500 Gala apples.

(If Rose had used the word "proportion" or "fraction" instead of "number" in her second statement, she would have been correct.)

Example (8c): Sam thought about selling his apples at the farmers' market. The weight rules were that the Cox apples should weigh between 4½ and 5½ ounces (128 – 156 grams) each, and that the Gala apples should weigh between 5 and 6 ounces (142 – 170 grams) each.

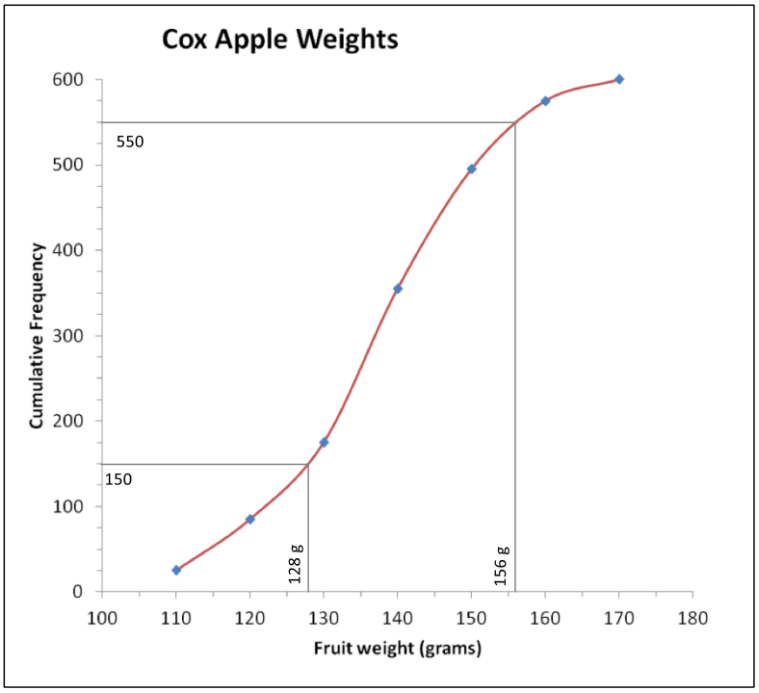
Use the cumulative frequency graphs to show that Sam would have been able to sell about two-thirds of his Cox apples and four-fifths of his Gala apples using the weight criteria above.

It can be estimated from the cumulative frequency graph that 150 out of 600 Cox apples weighed up to 128 grams. (This value of 128 grams corresponds to the lower quartile in the earlier part of the question).

Reading off the graph in the same way also gives us the estimated number of Cox apples weighing up to 156 grams. On the graph, this reads as 550.

Hence an estimated (550-150), or 400 out of 600 Cox apples fell within the 128-156 g weight range.

Now $\frac{400}{600} = \frac{2}{3}$, so Sam would have been able to sell two-thirds of his Cox apples at the market.

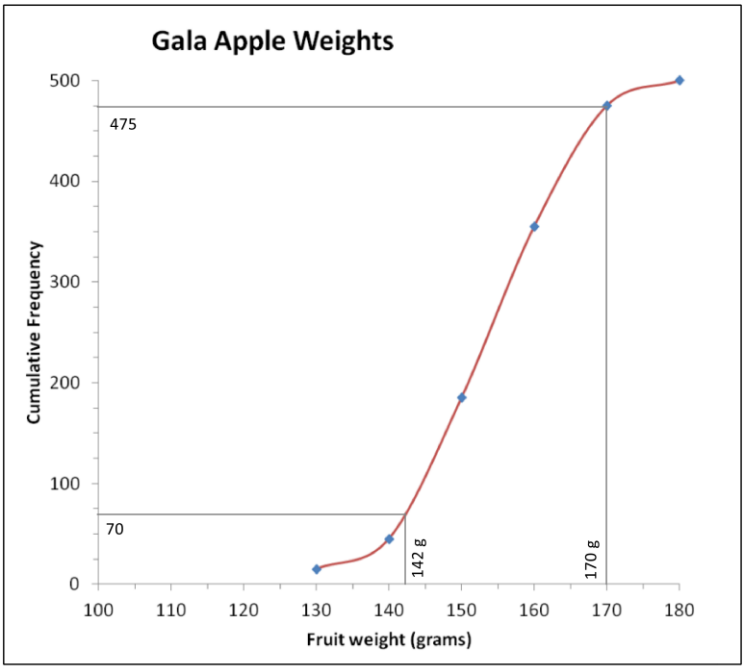


It can be estimated from the cumulative frequency graph that 475 out of 500 Gala apples weighed up to 170 grams. (This value of 475 could also have been taken from the original data in the earlier part of the question).

Similarly, the estimated number of Gala apples weighing up to 142 grams can be read off as 70

Hence an estimated (475-70), or 405, out of 500 Gala apples fell within the 142-170 g weight range.

Now $\frac{405}{500} = 0.81$, so Sam would have been able to sell 81%, or just over four fifths, of his Gala apples at the market.



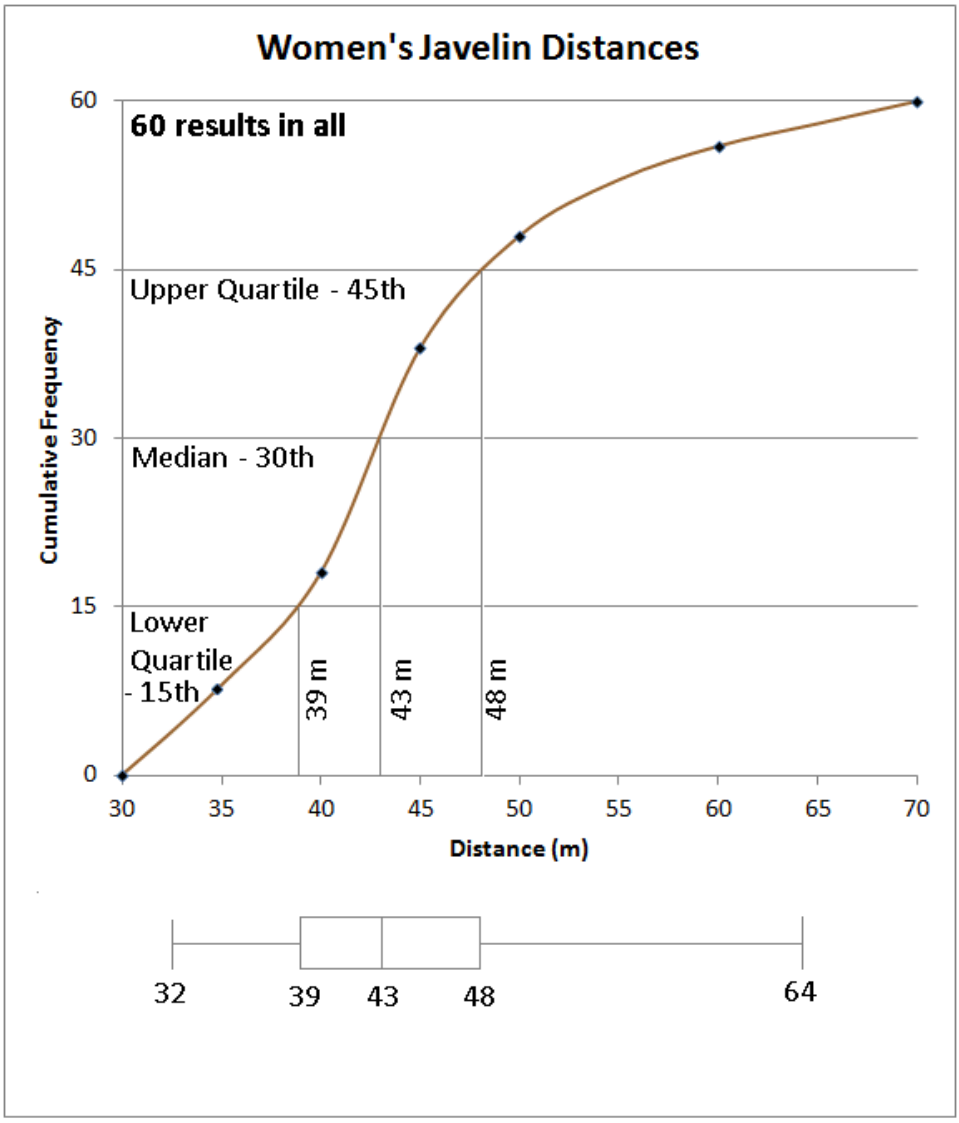
Example (9a):
Plot a cumulative frequency graph for the women's javelin results, using the data from Example (6).

In addition, draw a box-and-whisker plot for the same data, given that the shortest distance thrown is 32 m and the longest 64 m. Describe the distribution.

Distance (d) in metres	$30 \le d < 35$	$35 \le d < 40$	$40 \le d < 45$	$45 \le d < 50$	$50 \le d < 60$	$60 \le d < 70$
Frequency	8	10	20	10	8	4

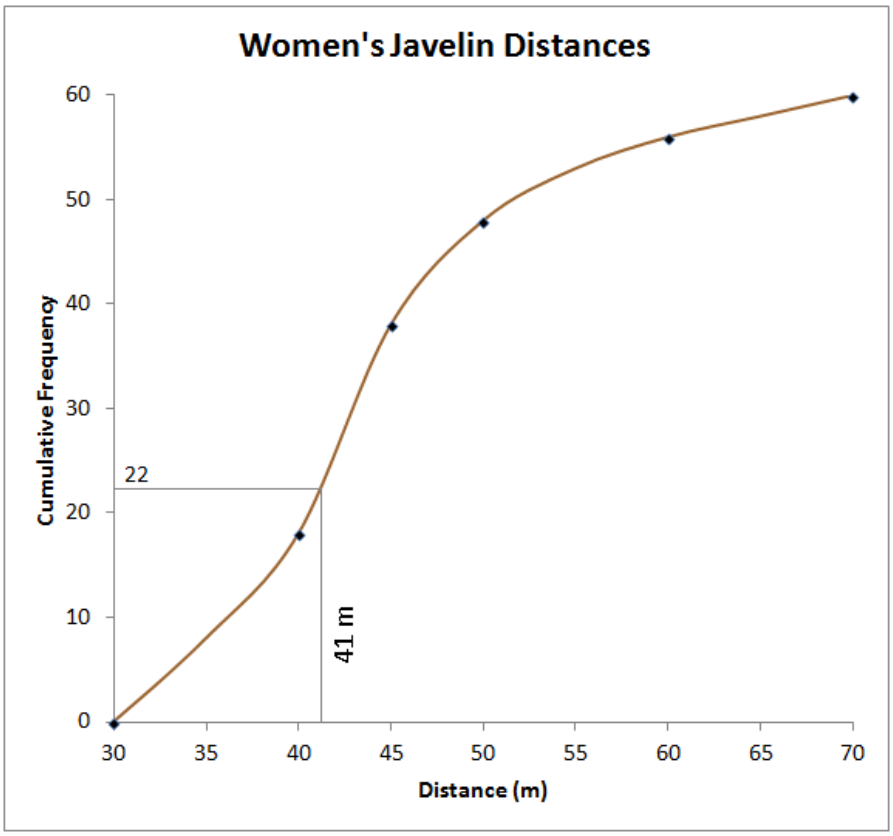
Cumulative distance (d)	$30 \le d < 35$	$30 \le d < 40$	$30 \le d < 45$	$30 \le d < 50$	$30 \le d < 60$	$30 \le d < 70$
Cumulative Frequency	**8**	**18**	**38**	**48**	**56**	**60**

(Continuous data - we line up the points at 30, 35, 40 ... on the data axis)



The distribution appears to have appreciable positive skew.

Example (9b) : A throw in the women's javelin is considered by the AAA to be of Grade 1 standard if the length of throw is greater than 41 metres. How many of the throws in Examples (6) and (9a) qualified for that standard, and what percentage of the total is that ?



It can be estimated from the cumulative frequency diagram that 22 out of the 60 throws recorded were less than or equal to 41 metres in distance.

Hence an estimated (60 – 22), or 38 , throws were of AAA Grade 1 standard.
That equates to 63% of the total.

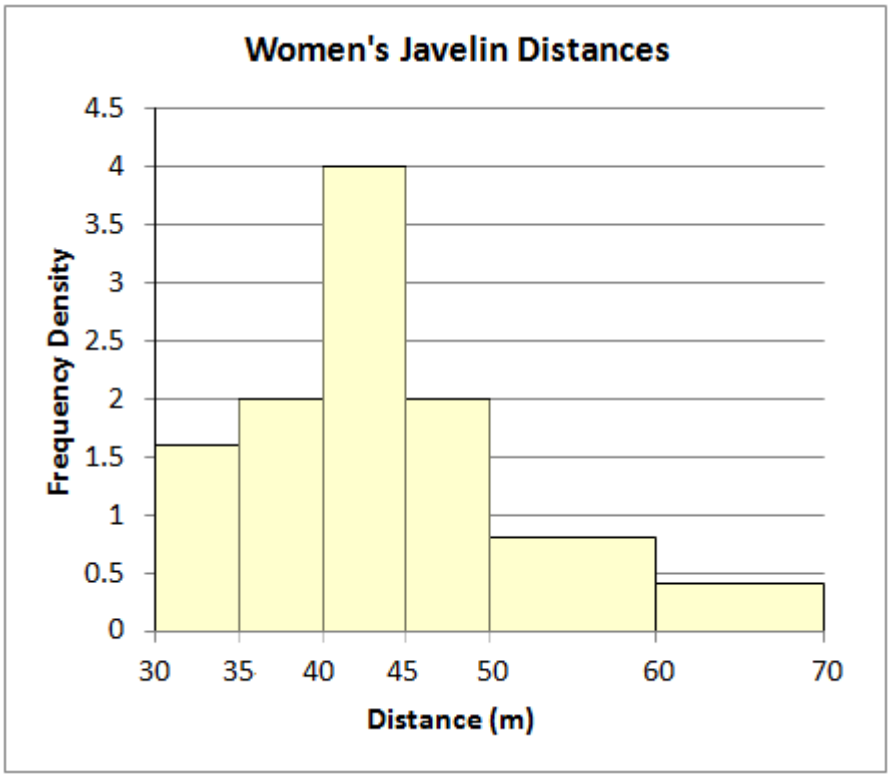
Estimating the median and quartiles from a histogram, without a cumulative frequency plot.

Sometimes we might be asked to estimate the median and/or quartiles of a set of values directly from a histogram, without using a cumulative frequency plot.

Example (10) (Recall Examples (6) and (9)) Once again, we have the table of women's javelin results, along with the histogram.

Distance (d) in metres	$30 \le d < 35$	$35 \le d < 40$	$40 \le d < 45$	$45 \le d < 50$	$50 \le d < 60$	$60 \le d < 70$
Frequency	8	10	20	10	8	4
Freq.Density	1.6	2.0	4.0	2.0	0.8	0.4
Cumulative Frequency	**8**	**18**	**38**	**48**	**56**	**60**

It is not vital to complete the cumulative frequency table, but it is helpful when looking for the median and the quartiles.



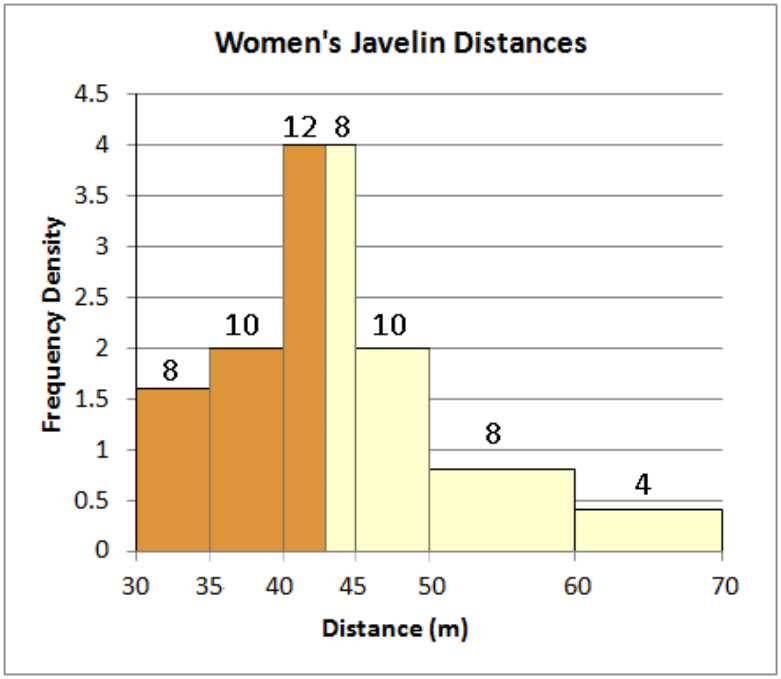
Estimate the median and quartiles for the resulting histogram.

To visualise the median on a histogram, we imagine a vertical line dividing the histogram into two equal areas. Because there are 60 values in all, the dividing line would occur at the point that the cumulative frequency reaches the halfway stage, i.e. on the 30th result.

The cumulative frequency reaches 18 at the end of the "35≤ d < 40" class interval , i.e. at 40 metres, and 38 at the end of the "40 ≤ d < 45" interval, or at 45 metres. The 30th result would therefore lie in the "40 ≤ d < 45" interval, which contains the 20 data items from the 19th to the 38th.

In other words, the 30th result is $\frac{12}{20}$, or $\frac{3}{5}$, of the way, along the "40 ≤ d < 45" interval, by interpolation.

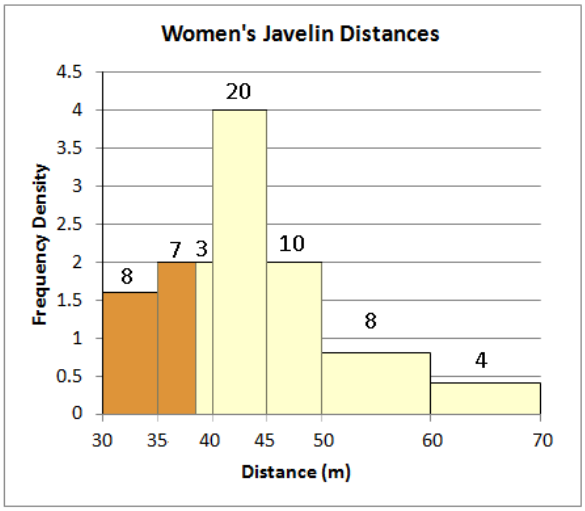
Now $\frac{3}{5}$ of 5 (the class width) = 3, so the median is approximately 40 + 3, or 43 metres.

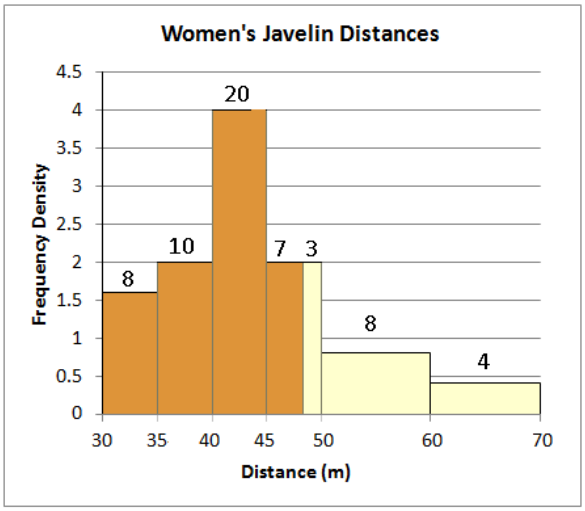


(Actual frequencies are shown at the top of the columns, along with the 'split' column containing the median.)

The quartiles can be estimated in a similar way. Thus, the lower quartile or 15th result will be $\frac{7}{10}$ of the way along the "35 ≤ d < 40" class interval, or at 35 + $\frac{7}{10}$ of 5 , or 38.5 metres (see lower left).

The upper quartile or 45th result will be $\frac{7}{10}$ of the way along the "45 ≤ d < 50 " interval, i.e. at 45 + $\frac{7}{10}$ of 5 , or 48.5 metres (see lower right).

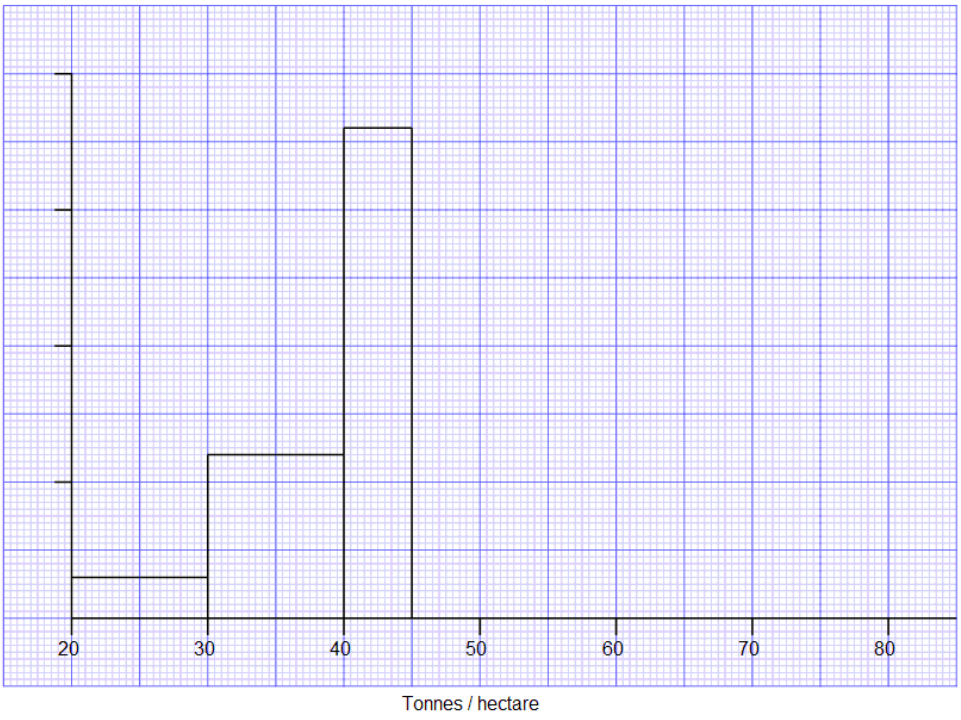




Sometimes an exam question on histograms will provide only 'skeleton' data.

Example (11): The following data apply to potato yields per hectare on 60 farms in Cheshire in 2013.

Yield per hectare, y, in tonnes	Frequency
$20 \le y < 30$	
$30 \le y < 40$	12
$40 \le y < 45$	
$45 \le y < 50$	14
$50 \le y < 60$	9
$60 \le y < 80$	



i) Use the data in the histogram to complete the table.
ii) Use the data in the table to complete the histogram.
iii) Estimate the number of farms whose potato yields were 35 tonnes per hectare or less.

i) Firstly, we need to find an item of data common to both the histogram and the table to fix the vertical frequency density scale. The class interval $30 \le y < 40$ is common to both, and as its class width is 10 and its frequency is 12, its frequency density is therefore 1.2.

We can now enter the values 0.0, 1.0, 2.0, 3.0 and 4.0 on the vertical frequency density tick marks.

Next, we can read off the frequency density for the $20 \le y < 30$ class. Its value of 0.3 coupled with the class interval width of 10 corresponds to a frequency of 3.

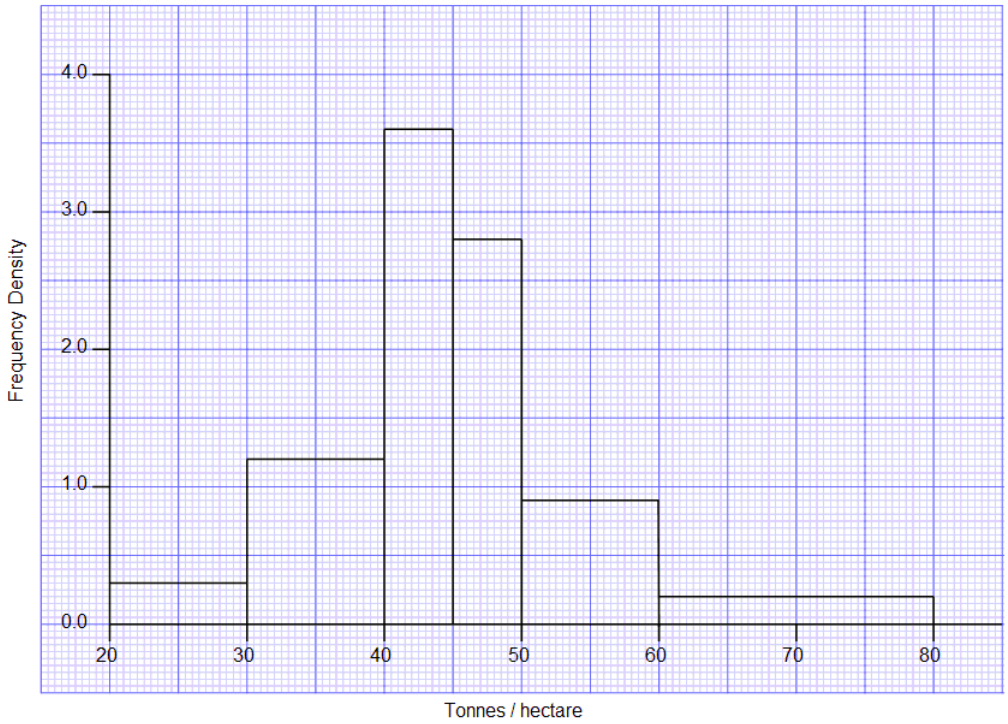
Similarly, for the $40 \le y < 45$ class, the frequency density reads as 3.6, which equates to an actual frequency of 18 given the class width of 5.

The only missing entry for frequency for the table is that for the $60 \le y \le 80$ interval. All the other frequencies sum to 56, and there are 60 farms in all, so the missing frequency must therefore be 4.

The completed data table now looks like this (calculated values in bold):

Yield per hectare, y, in tonnes	Frequency	Class width	Frequency density
$20 \le y < 30$	**3**	10	0.3
$30 \le y < 40$	12	10	1.2
$40 \le y < 45$	**18**	5	3.6
$45 \le y < 50$	14	5	**2.8**
$50 \le y < 60$	9	10	**0.9**
$60 \le y < 80$	**4**	20	**0.2**

ii) We now have all the required data to complete the histogram.



iii) To estimate how many potato farms yielded 35 tonnes or less per hectare, we have to include:

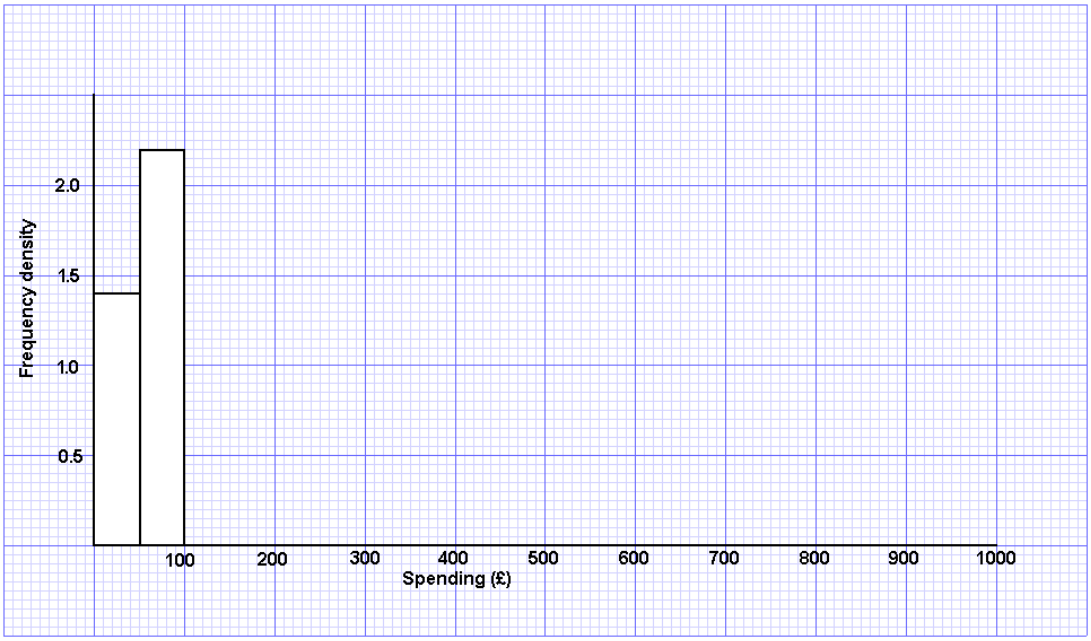
a) the whole of the $20 \le y < 30$ class, or 3 farms, plus
b) half of the $30 \le y < 40$ class, or an estimate of 6 farms, since 35 is midway between 30 and 40.

This gives an estimate of 9 farms in total yielding 35 tonnes of potatoes or less per hectare.

The next example is an omnibus exam-style question taking in a variety of points.

Example(12): (Same data as Example (8) in "Averages")
An electrical retailer has analysed the money spent by customers in the first 500 transactions in the January sales, excluding purchases greater than £1000.

Spending, x, in £	Frequency	Frequency density
$x < 50$	70	1.4
$50 \le x < 100$		2.2
$100 \le x < 200$	100	
$200 \le x < 400$	80	
$400 \le x < 600$	60	
$600 \le x \le 1000$		



(i) Complete the data table and partial histogram above.
(ii) Use the histogram produced in (i) to estimate the median and the inter-quartile range.

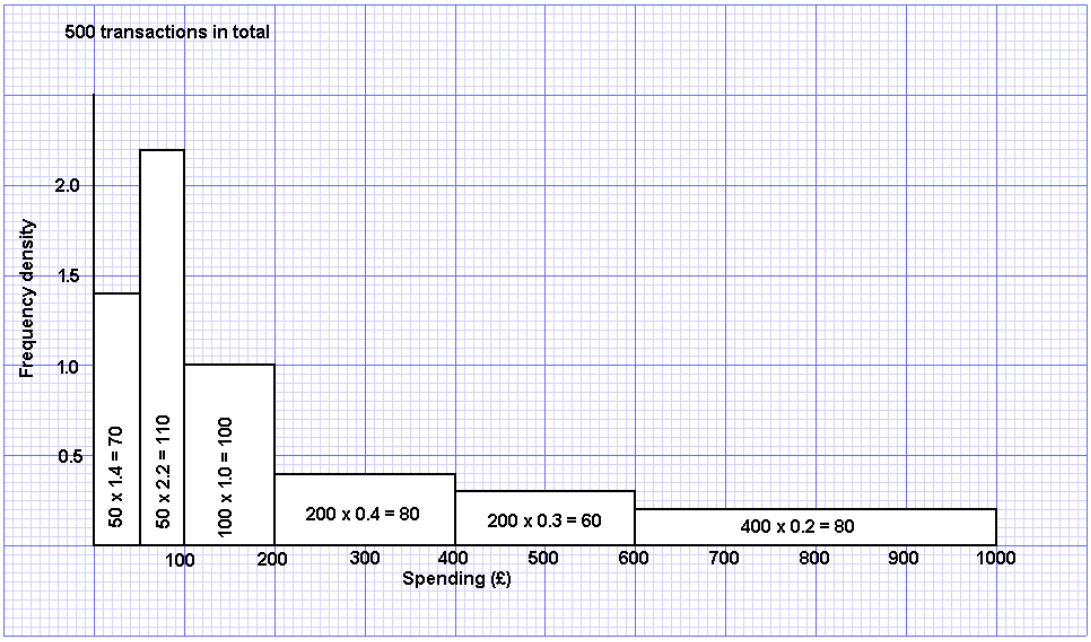
(i) First, we find the ratio between the frequency density and the actual frequency.
It can be seen from the values in the row from the 0-50 class interval that the frequency density is given by dividing the actual frequency (70) by the class interval width (50).

From there, we can find the actual frequency for the 50-100 class interval, namely $2.2 \times 50 = 110$.
The 100-200 class interval has a width of 100, so its frequency density is $\frac{100}{100}$ or 1.0.
The 200-400 class interval has a width of 200, so its frequency density is $\frac{80}{200}$ or 0.4.
Similar working gives a frequency density of 0.3 for the 400-600 class interval.

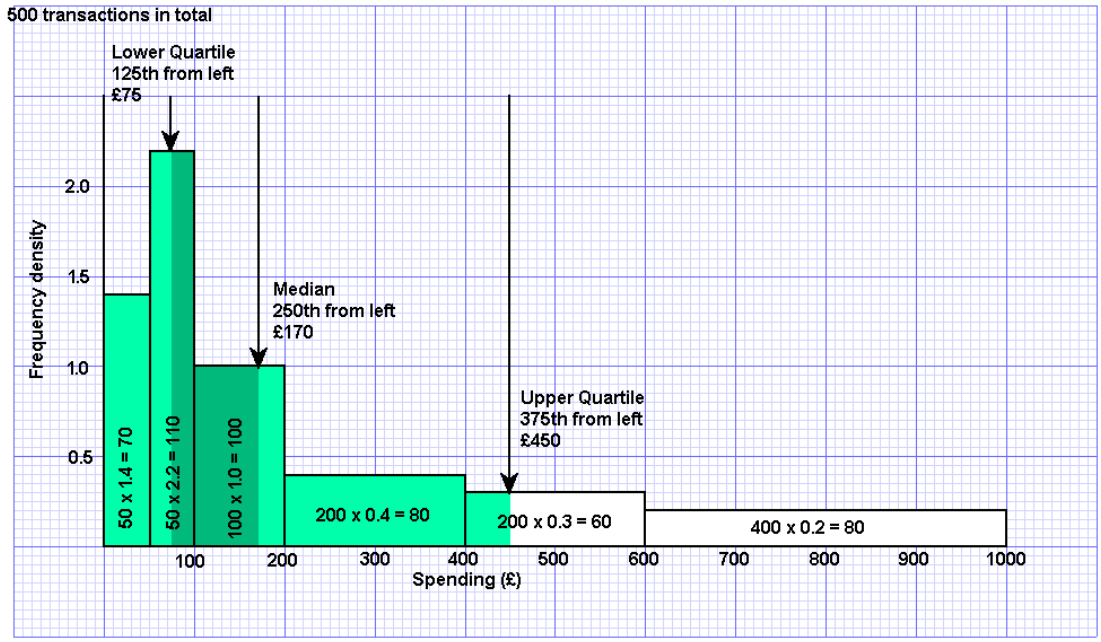
We are told that 500 transactions have been analysed, and therefore the frequency for the 600-1000 class interval is the sum of all the previous frequencies subtracted from 500, or 80.

The completed frequency table and histogram therefore look like this, with a class width column added for information:

Spending, x, in £	Frequency	Class Width (£)	Frequency density
$x < 50$	70	50	1.4
$50 \le x < 100$	**110**	50	2.2
$100 \le x < 200$	100	100	**1.0**
$200 \le x < 400$	80	200	**0.4**
$400 \le x < 600$	60	200	**0.3**
$600 \le x \le 1000$	**80**	400	**0.2**



To find the median and quartiles, and hence the inter-quartile range, we reckon up the appropriate areas on the histogram., using linear interpolation as shown in the diagram below



Because there are 500 transactions in total, the median occurs on the 250th transaction and the lower and upper quartiles on the 125th and 375th transactions.

Scanning the histogram from left to right, the 70th transaction occurs at the end of the 0-50 class interval and the 180th at the end of the 50-100 interval. The 125th is exactly halfway along the 50-100 interval, so the lower quartile is £75.

Continuing the scan, the 100-200 interval includes the 181st to the 280th transactions – the 250th one is seven-tenths of the way along, so the median is £170.

Finally, the 400-600 interval includes the 361st to the 420th transactions – the 375th is one-quarter of the way along, and so the upper quartile is £450.

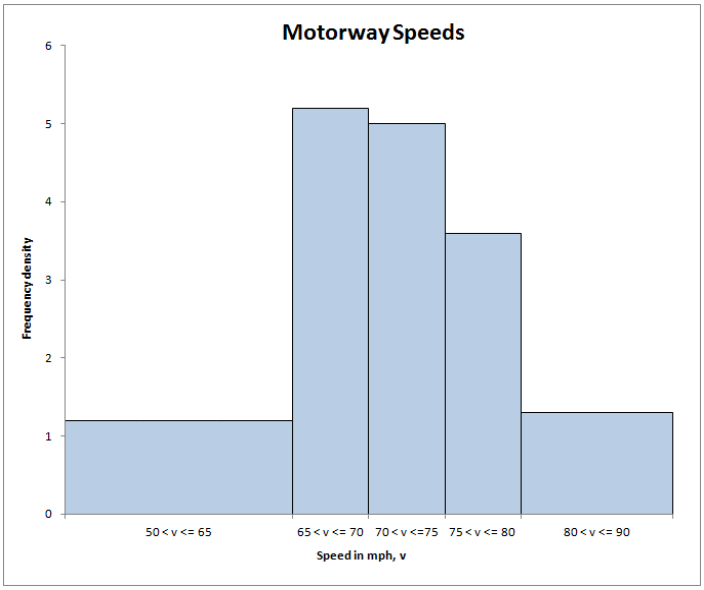
Therefore the median value of the transactions is £170, and the inter-quartile range is £(450-75) or £375.

Note the positive skew of the data – the median of £175 is less than the mean of £286 calculated in Example 7 of the section "Averages".

Example (13): A police officer had measured the speeds of 100 cars on a mobile speed camera on the hard shoulder of a part of the M60 motorway where the speed limit is 70 mph.

The data and histogram are shown below, and are the same as in Example (1).

Speed (v) in mph	Frequency density
$50 \le v < 65$	1.2
$65 \le v < 70$	5.2
$70 \le v < 75$	5
$75 \le v < 80$	3.6
$80 \le v < 90$	1.3



The police made the following comment after looking at the data.

"There are too many drivers breaking the speed limit, and in addition, I reckon that a third of the drivers doing under 65 are hogging the middle lane."

The police therefore examined the possibility of a crackdown during off-peak hours, whereby drivers exceeding 77 mph, as well as middle lane "hogs" driving at under 65 mph, would be fined £100.

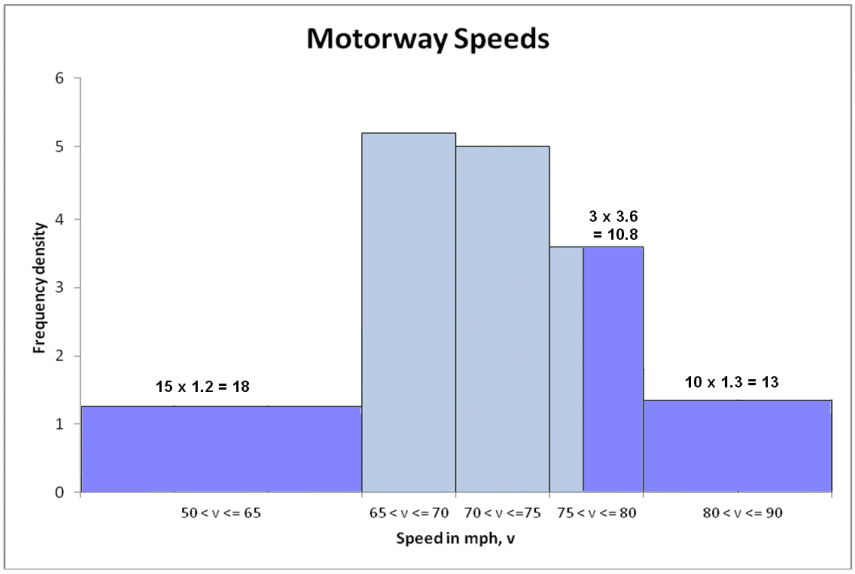
The police chief said: "We get about 8,000 cars using this stretch of motorway on any day during off-peak hours. Even if we were to stop only one offender in every ten, this crackdown is going to end up with us issuing over £20,000 per day in fines."

i) Use the histogram to estimate how many drivers out of that sample of 100 exceeded 77 mph.

ii) Explain why six drivers were likely to have been middle lane "hogs".

iii) Hence calculate the expected revenue in fines for one day of the crackdown. Was the police chief's claim justified ?

The areas crucial to the question are highlighted in the histogram below.



i) The speeding drivers are represented by two bars of the histogram.

All of the $80 \le v < 90$ class is included, or $10 \times 1.3 = 13$ drivers.
(Class width is 10, frequency density = 1.3)

The cut-off limit for speeding fines is at 77 mph, which is two-fifths of the way from 75 to 80. This means that we have to include three-fifths of the $75 \le v < 80$ class to account for the drivers who had been doing between 77 and 80 mph. This split class interval has a width of 3 and frequency density of 3.6, so it represents 3×3.6, or 11 drivers, to the nearest whole.

The estimated number of speeders is therefore 24 out of the sample of 100.

ii) The slow drivers are in the $50 \le v < 65$ class, representing 15×1.2 or 18 drivers. The police reckon that a third of those slow drivers are middle lane "hogs", and one third of 18 is 6.

iii) If the police crackdown were applied to the 100 drivers in the sample, then there would be 24 speeders and 6 middle lane "hogs" issued with fines – a total of 30 drivers, or 30% of the sample.

If we were to project this same ratio to the 8000 cars using the motorway on an off-peak day, then 30% of 8000, or 2400 drivers, would be performing an offence.

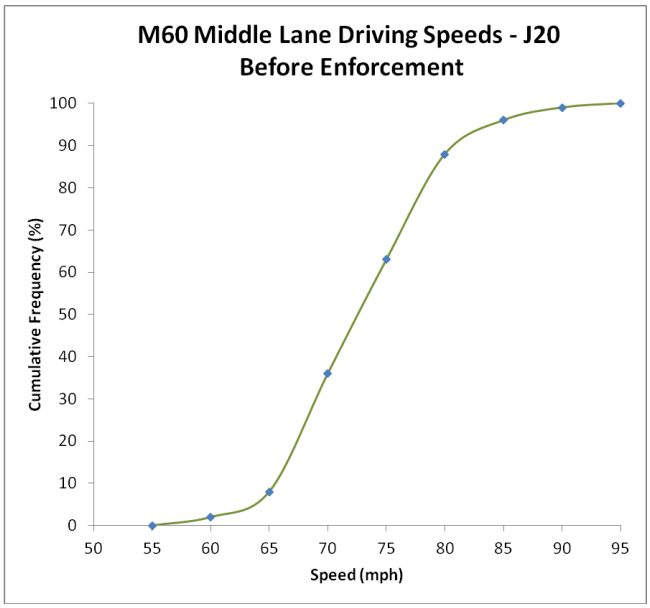
Of those 2400 offenders, one-tenth of them, or 240, would be issued with a fine of £100.

This would bring the total in fines to $240 \times$ £100, or £24,000.

Hence the police chief's claim was justifiable.

Example 14(b): The police then decided to refine the model by monitoring the speeds of over 1000 cars in the middle lane only of the M60 at Junction 20 a few days before the crackdown.

The distribution of speeds is shown in the cumulative frequency graph on the right. Note that the vertical axis shows percentages of the total rather than actual frequencies.



i) Estimate the percentage of offending drivers, namely those travelling at either over 77 mph or under 65 mph.

ii) Hence estimate the daily number of offending drivers, if 8000 people a day used this stretch of motorway during off-peak times.

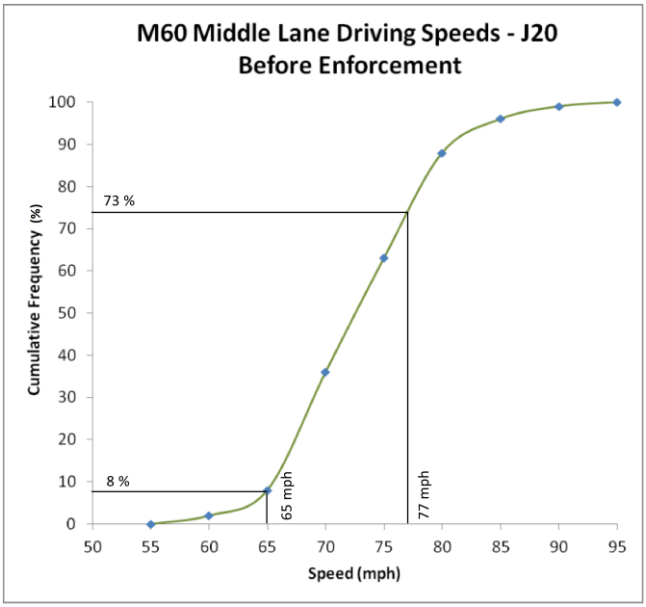
iii) If 10% of the offending drivers in part ii) ended up paying a £100 fine, how much money would be expected to be raised in fines in one day ?

i) From the cumulative frequency diagram, it can be seen that 8% of the drivers travelled at under 65 mph, and 73% travelled at under 77 mph. Hence (100-73)% or 27% were exceeding 77 mph.

Therefore (8+27)%, or **35%**, of the drivers were travelling at either over 77 mph or under 65 mph.

ii) The estimated number of daily offending drivers is 35% of 8000, or **2800**.

iii) 10% of 2800 is 280, so the expected daily revenue in fines would be 280 × £100, or **£28,000**.

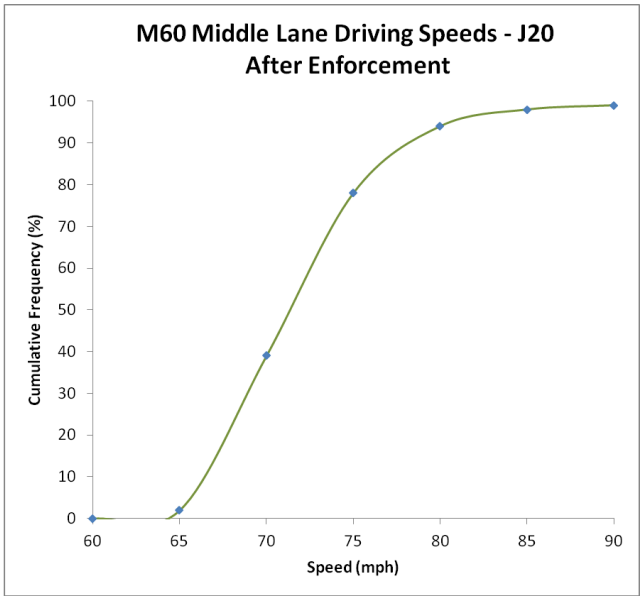


The police then decided to carry out their crackdown on speeding and middle lane hogging, and several weeks later, monitors recorded the speeds of over 1000 cars and plotted the results on the cumulative frequency diagram on the right.

(Again, the vertical axis shows percentages of the total.)

The police chief commented: **"**We've cut right down on middle lane hogging, and halved the percentage of speeders in the middle lane, so our deterrent fines seem to be working quite well."

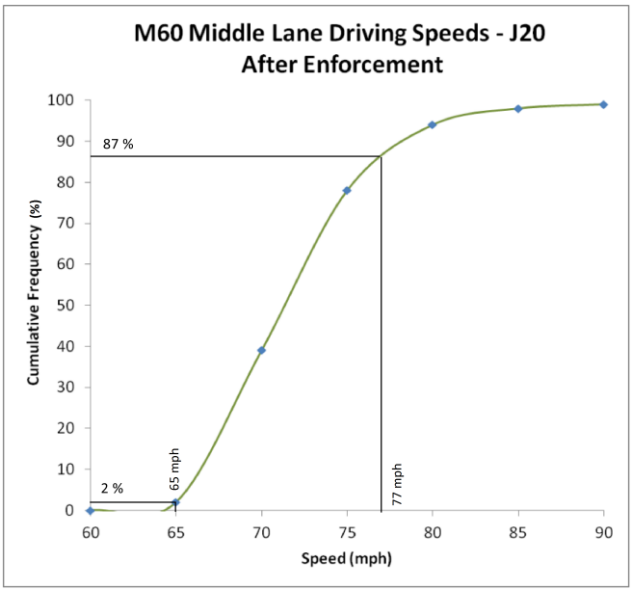
iv) Use the cumulative frequency diagram to support or contradict the police chief's claims.



iv) The percentage of drivers travelling at under 65 mph in the middle lane had declined from 8% of the total to 2%, and so the police chief was correct in claiming that middle lane hogging had been much reduced.

The percentage of speeding drivers in the middle lane had declined to (100-87)%, or 13% of the total, compared to the 27% before the crackdown.

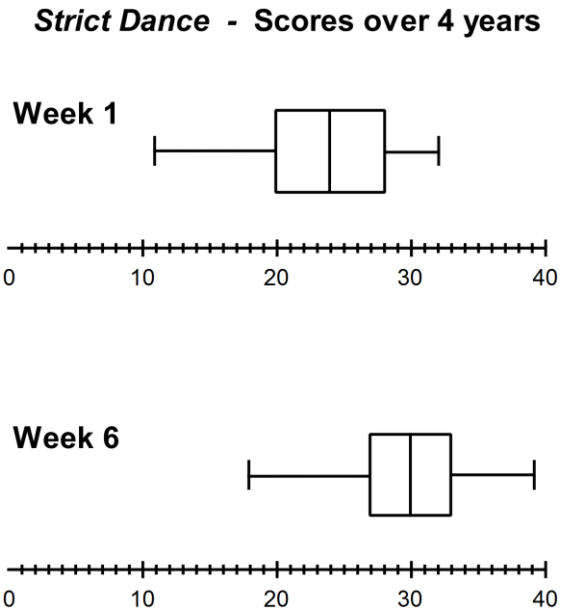
The police chief was therefore also correct in claiming that the percentage of speeding drivers in the middle lane had been halved.



More on Box Plots.

Many examination questions concern comparative box plots without a cumulative frequency diagram.

Example (15): The box plots below show the judges' scoring trends for two weeks of the *Strict Dance* TV show over the last four years, with 60 results analysed in total for the first week of the show, and 40 results for the sixth week of the show. The totals of the judges' scores are all whole numbers.



i) Compare and contrast the box plots for the two weeks, with at least two supporting statements.

ii) Jay assumes that 15 people scored more than 28 points in Week 1, and that 10 people scored 27 or less in Week 6. Is he correct ?

i) Here are a few possible answers:

All the key values (extremes, quartiles and median) were higher in Week 6 than in Week 1, suggesting an overall score improvement across the board. For example, the median was up by 6 points.

The inter-quartile range was lower in week 6 (6 points) than in week 1 (8), which suggests greater consistency in the middle half of the table.

The overall range had not changed in the intervening five weeks. This could be due to outliers, but there is no actual data to confirm or deny that.

ii) The upper quartile score for Week 1 was 28, suggesting that a quarter of 60, or 15, scores were higher than that. By similar reasoning, the lower quartile score for Week 6 was 27, suggesting that a quarter of 40, or 10 scores, were equal or lower.

The only misgiving about Jay's assumption is the possibility of tied scores in the neighbourhood of the quartiles.

Further Note on Quartiles.

Just as the median can be found from discrete numeric data, so can the quartiles. One method is to find the median, split the data into two halves to either side of it, and then find the median of each half.

If there are an even number of data items in the list, there will be two 'middle numbers' instead of an actual median. In this case, we must exclude the computed median from the two halves of the data list.

The lower quartile is the median of the lower half of the data; the upper quartile is the median of the upper half of the data.

Example (16): Find the lower quartile, median and upper quartile from this list;
14, 16, 19, 22, 25, 29, 32, 34, 37, 41.

There are 10 numbers in the list, i.e. $n = 10$; the position of the median is $\frac{1}{2}(n+1)$ or 5½.
There are two "middle numbers", 25 and 29, so the median is halfway between 25 and 29, or 27.

After chopping the data in two, the lower half contains 14, 16, 19, 22, 25 and the upper half contains 29, 32, 34, 37, 41. Note that the calculated median of 27 is not included in either half.

The lower quartile is the median of 14, 16, 19, 22, 25, or 19.
The upper quartile is the median of 29, 32, 34, 37, 41, or 34.

If there are an odd number of data items in the list, then there is an actual median.

There seems to be an inconsistency in such cases. Some schools of thought exclude the median when 'halving' the data, whereas others include it. Still others may use a different method, such as reckoning the positions of the quartiles as $\frac{1}{4}(n+1)$ and $\frac{3}{4}(n+1)$.[*1] **Check with your exam board !**

Example (15): Find the lower quartile, median and upper quartile from this sorted set of numbers;
10, 12, 15, 19, 22, 26, 28, 31, 34. (Exclude the actual median from the two data halves)

The median is the middle number in the list, namely 22.

The lower half of the data contains 10, 12, 15 and 19. The lower quartile is the median of *that* set, and is halfway between 12 and 15, or 13½.

The upper half of the data contains 26, 28, 31 and 34. The upper quartile is the median of *that* set, and is halfway between 28 and 31, or 29½.

[*1] This topic had appeared on recent past paper questions from AQA and Edexcel. One question included a list of 19 items in a stem-and-leaf diagram; another included a list of 31. In the first case the median was the 10^{th} item and the quartiles 5^{th} and 15^{th} ; in the second, the median was the 16^{th} item and the quartiles the 8^{th} and 24^{th} . This implies *excluding* the median when dividing the data into two halves; also, the numbers of items in the lists became exact multiples of 4 on adding 1, making the quartile positions easy to find; thus, 19+1 = 20 and the quartile positions of 5, 10 and 15.