Edexcel AS Mathematics Collecting and interpreting data

Section 2: Data presentation and interpretation

Notes and Examples – measures of central tendency

This section deals with

- Measures of central tendency
- Comparison of measures of central tendency
- Frequency tables
- Finding measures of central tendency from frequency distributions
- Grouping data
- Estimating the median from grouped data

Measures of central tendency

There are four of these which are commonly used: the *mean*, the *median*, the *mode* and the *midrange*.

You need to be able to calculate these for both discrete and continuous data. You also need to appreciate the different properties of each of these measures.

The mean

When people talk about the average, it is usually the *mean* that they mean! This is the sum of the data divided by the number of items of data.
We can express this using mathematical notation as follows:

For the data set $x_1, x_2, x_3, x_4 ..., x_n$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\bar{x}$ denotes the mean value of x

Σ is the Greek letter sigma and stands for 'the sum of'. The whole expression is saying: 'The mean ($\bar{x}$) is equal to the sum of all the data items (x_i for $i = 1$ to n) divided by the number of data items (n).'

Example 1 shows a very simple calculation set out using this formal notation.

Example 1

Find the mean of the data set {6, 7, 8, 8, 9}.

Solution

$x_1 = 6,\ x_2 = 7,\ x_3 = 8,\ x_4 = 8,\ x_5 = 9,\ n = 5$

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{6+7+8+8+9}{5} = 7.6$$

Edexcel AS Maths Data 2 Notes and Examples

The median

When data is arranged in order, the median is the item of data in the middle. However, when there is an even number of data, the middle one lies between two values, and we use the mean of these two values for the median.

For example, this dataset has 9 items:

1 1 3 4 6 7 7 9 10

There are 4 data items below the 5th and 4 items above; so the middle item is the 5th , which is 6.

If another item of data is added to give 10 items, the middle items are the 5th and 6th:

1 1 3 4 6 7 7 9 10 12

so the median is the mean average of the 5th and 6th items, i.e. $\frac{6+7}{2} = 6.5$.

The mode

The mode is the most common or frequent item of data; in other words the item with the highest frequency.

So for the data set $\{6, 7, 8, 8, 9\}$
the mode is 8 as this appears twice.

There may be more than one mode, if more than one item has the highest frequency.

The midrange

The final measure of average is the midrange. This is halfway between the lowest and highest values, i.e. the mean of the highest and lowest values:

$$\text{midrange} = \frac{\text{highest} + \text{lowest}}{2}$$

So for the data set $\{6, 7, 8, 8, 9\}$

the lowest value is 6 and the highest is 9, so the midrange is $\frac{6+9}{2} = 7.5$

Example 2

For the data displayed in this stem and leaf diagram, find

(i) the median
(ii) the mode
(iii) the midrange.

16	5 5 6 7 8	$n = 20$
17	0 0 1 3 3 7 8 9	17 |3 represents 1.73
18	2 2 2 5 5 8	
19	0	

Solution

(i) Counting from the lowest item (1.65), the 10th is 1.73 and the 11th is 1.77.

The median is therefore $\frac{1.73+1.77}{2}=1.75$.

(ii) No item appears more than twice except for 1.82 which appears three times. The mode is 1.82.

(iii) The midrange $=\frac{1.65+1.90}{2}=1.775$

Comparison of measures of central tendency (averages)

- The mean includes all the data in the average, and takes account of the numerical value of all the data. So exceptionally large or small items of data can have a large effect on the mean – it is *susceptible to outliers*.
- The median is less sensitive to high and low values (outliers), as it is simply the middle value in order of size. If the numerical values of each of the items of data is relevant to the average, then the mean is a better measure; if not, use the median.

- The mode picks out the commonest data item. This is only significant if there are relatively high frequencies involved. It takes no account at all of the numerical values of the data.

- The midrange is calculated solely on the highest and the lowest items of data; this is easy to calculate, but assumes that the data is symmetrical if it is to provide a suitable measure of average.

Suppose you are negotiating a salary increase for employees at a small firm. The salaries are currently as follows:

£6000, £12000, £14000, £14000, £15000, £15000, £15000, £15000, £16000, £16000, £18000, £18000, £18000, £20000, £100000

The £6000 is a part-time worker who works only two days a week

The £100000 is the managing director

- The mean salary is £20800
- The median salary is £15000
- The modal salary is also £15000
- The midrange is £53000

Which is the most appropriate measure?

If you were the managing director, and used the midrange, you could argue that the average is £53000 – she would be lucky to get away with this figure! More reasonably, she could point to a mean of £20800, but of the current employees she is the only one who earns more than this amount.

If you were the union representative, you would quote the median or the mode (£15000), as these give the lowest averages. This is certainly more typical of the majority of workers.

There is no ‘right’ answer to the appropriate average to take – it depends on the purpose to which it is put. However, it is clear that:

- The mean takes account of the numerical value of *all* the data, and is higher due to the effect of the £100000 salary, which is an outlier.
- The median and mode are not affected by the outliers (£100000 and £6000)
- The midrange relies entirely on the outliers, and is therefore unreliable and should be discounted.

Example 3

Shanice receives the following marks for her end-of-term exams:

Subject	Mark (%)
Maths	30
English	80
Physics	45
Chemistry	47
French	47
History	50
Biology	46
Religious Education	55

Calculate the mean, median, mode and midrange. Comment on which is the most appropriate measure of average for this data.

Solution

The mean $= \dfrac{30+80+45+47+47+50+46+55}{8} = 50$

In numerical order, the results are: 30, 45, 46, 47, 47, 50, 55, 80

The median is therefore 47.

The mode is 47, as there are two of these and only one each of the other marks.

The midrange is $\dfrac{80+30}{2} = 55$.

The mode is not suitable – there is no significance in getting two scores of 47.

The midrange is based entirely on the outlier results in Maths and English and is not representative.

The median or the mean could be used. The mean is higher since it takes more account of the high English result. The median is perhaps the most representative, and she got 4 scores in the range 45-47; but Shanice would no doubt use the mean to make more of her good English result!

Finding measures of central tendency from frequency distributions

When data are given in the form of a frequency table, the methods for finding measures of central tendency have to be adapted slightly.

The mean

x	f
1	3
2	5
3	2
4	3
5	4
6	3
Total	20

The mean of the data shown in the frequency table above can be written as

$$\bar{x} = \frac{1+1+1+2+2+2+2+2+3+3+4+4+4+5+5+5+5+6+6+6}{20} = \frac{69}{20} = 3.45$$

An alternative way of writing this is

$$\bar{x} = \frac{3\times1+5\times2+2\times3+3\times4+4\times5+3\times6}{3+5+2+3+4+3} = \frac{69}{20} = 3.45$$

This can be expressed more formally as

$$\bar{x} = \frac{\sum_{i=1}^{6} f_i x_i}{\sum_{i=1}^{6} f_i}$$

Each value of x is multiplied by its frequency, and then the results are added together.

The frequencies are added to find the total number of data items

It is helpful to add another column to the frequency table, for the product fx.

x	f	fx
1	3	3
2	5	10
3	2	6
4	3	12
5	4	20
6	3	18
Total	$\sum f = 20$	$\sum fx = 69$

Then you can simply add up the two columns and use the totals to calculate the mean.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{69}{20} = 3.45$$

In general, when the data is given using frequencies, the formula for the mean is:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Example 4

A company collects feedback on its website. Customers give a score on a scale of 1 – 5 to indicate levels of satisfaction.

Here are the data for the first three months of a particular year.

Month	Mean of feedback scores	Number of customers giving feedback
January	4.3	370
February	3.9	280
March	4.2	315

Find the mean satisfaction level over the three month period.

It is tempting to think that you could find the mean satisfaction level by finding the mean of 4.3, 3.9 and 4.2. However, since there were different numbers of customers each month, this would not give a true value for the mean.

Solution

Total score given by customers in January $= 4.3 \times 370 = 1591$

Total score given by customers in February $= 3.9 \times 280 = 1092$

Total score given by customers in March $= 4.2 \times 315 = 1323$

Instead, you need to find the total of the scores given by all the customers, and divide by the total number of customers.

Total score given by customers over the three months = 4006

Total number of customers giving feedback = 965

Mean $= \dfrac{4006}{965} = 4.15$

Edexcel AS Maths Data 2 Notes and Examples

The median

When you want to find the median of a data set presented in a frequency table, one useful point is that the data is already ordered.

x	f
1	3
2	5
3	2
4	3
5	4
6	3
Total	20

For this data set, there are 20 data items, so the median is the mean of the 10th and 11th items.

For this small set of data, it is easy to see that the 10th data item is 3 and the 11th is 4. The median is therefore 3.5.

However, for a larger set of data it may be more difficult to identify the middle item or items. One way to make this a little easier is to use a **cumulative frequency table**.

x	f	Cum. freq.
1	3	3
2	5	8
3	2	10
4	3	13
5	4	17
6	3	20

The third column gives the **cumulative frequency**. This is the total of the frequencies so far.

You can find each cumulative frequency by adding each frequency to the previous cumulative frequency. E.g., for $x = 4$, the cumulative frequency is 10 + 3 = 13.

The final value of the cumulative frequency (in this case 20) tells you the total of the frequencies. The cumulative frequencies show that the 10th item is 3 and the 11th item is 4. So the median is 3.5.

The mode

Identifying the mode is easy when data are given in a frequency table.

x	f
1	3
2	5
3	2
4	3
5	4
6	3
Total	20

The highest frequency is for $x = 2$. So the mode is 2.

The midrange

Again, it is easy to identify the highest and lowest values from a frequency table. In the table above, the highest value of x is 6 and the lowest is 1.

The midrange is $\frac{6+1}{2}=3.5$

Example 5

For the following set of data

x	f
22	5
23	17
24	23
25	35
26	12
Total	92

find the values of

(i) the mean (ii) the median (iii) the mode (iv) the midrange

Solution

(i) $$\bar{x}=\frac{22\times5+23\times17+24\times23+25\times35+26\times12}{5+17+23+35+12}$$

$$=\frac{2240}{92}=24.3\text{ (3 s.f.)}$$

The mean = 24.3 (3 s.f.)

(ii) Make a cumulative frequency table:

x	f	cf
22	5	5
23	17	22
24	23	45
25	35	80
26	12	92

Since there are 92 data items, the median is the average of the 46th and 47th items.
There are 45 items of 24 or less, and 80 items of 25 or less.
So clearly the 46th and 47th items are both 25.
The median is 25.

(iii) The highest frequency is for x = 25.
The mode is 25.

(iv) The highest value is 26 and the lowest is 22.

The midrange $=\frac{26+22}{2}=24$.

Estimating the mean from grouped data

When the data is grouped into classes, you can still estimate the mean by using the midpoint of the classes (the mid-interval value). This means that you assume that all the values in each class interval are equally spaced about the mid-point.

You can show most of the calculations in a table, as shown in the following example.

Example 6
Estimate the mean weight for the following data:

Weight, *w*, (kg)	Frequency
$50 \le w < 60$	3
$60 \le w < 70$	5
$70 \le w < 80$	7
$80 \le w < 90$	3
$90 \le w < 100$	2
Total	20

Solution

The mid-interval value is the mean of the upper and lower bound of the weight.

Weight, *w*, (kg)	Mid-interval value, *x*	Frequency, *f*	*fx*
$50 \le w < 60$	55	3	165
$60 \le w < 70$	65	5	325
$70 \le w < 80$	75	7	525
$80 \le w < 90$	85	3	255
$90 \le w < 100$	95	2	190
		$\sum f = 20$	$\sum fx = 1460$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{1460}{20} = 73$$

The mean weight is estimated to be 73 kg.

To find mid-interval values, you need to think carefully about the upper and lower bounds of each interval. In the example above, it is clear what these bounds are. However, if the intervals had been expressed as 50 – 59, 60 – 69 and so on, then it is clear that the original weights had been rounded to the nearest kilogram, and the intervals were actually $49.5 \le w < 59.5$, $59.5 \le w < 69.5$, etc. So in that case the mid-interval values would be 54.5, 64.5 and so on.
Note: If your data involves ages, and you used the groups 20-29, 30-39 etc, the group for 20-29 would include everyone from their 20th birthday to the day

before their 30th birthday. So in that case the mid-interval values would be 25, 35 etc.

Estimating the median from grouped data

If data are only available as a grouped frequency distribution, then it is not possible to find an exact value for the median.

However, it is possible to estimate the value of the median. The working below shows three methods for doing so. You may find it helpful to work with a sketch diagram in the working you do.
Suppose, for example, that you want to estimate the median of the data set shown below.

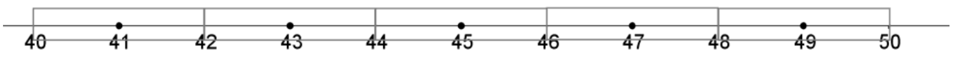
Weight w, (kg)	Frequency
$30 \le w < 35$	2
$35 \le w < 40$	3
$40 \le w < 50$	5
$50 \le w < 60$	2
$60 \le w < 65$	1
over 65	0

There are 13 data items, so you need to find the value of the 7th data item.

It is helpful to add the cumulative frequencies to this table. This shows clearly that the 7th data item lies in the class interval $40 \le w < 50$, and it is the 2nd item in this class interval (which contains 5 items).

Weight (kg)	Frequency	Weight	Cumulative frequency
		$w < 30$	0
$30 \le w < 35$	2	$w < 35$	2
$35 \le w < 40$	3	$w < 40$	5
$40 \le w < 50$	5	$w < 50$	10
$50 \le w < 60$	2	$w < 60$	12
$60 \le w < 65$	1	$w < 65$	13

To find the median, assumptions need to be made about where the data values lie. When estimating the mean, you assumed that the data values were symmetrically distributed around the midpoint of each class interval. A similar assumption for the class interval $40 \le w < 50$ gives the locations of the data values as shown in the following diagram.

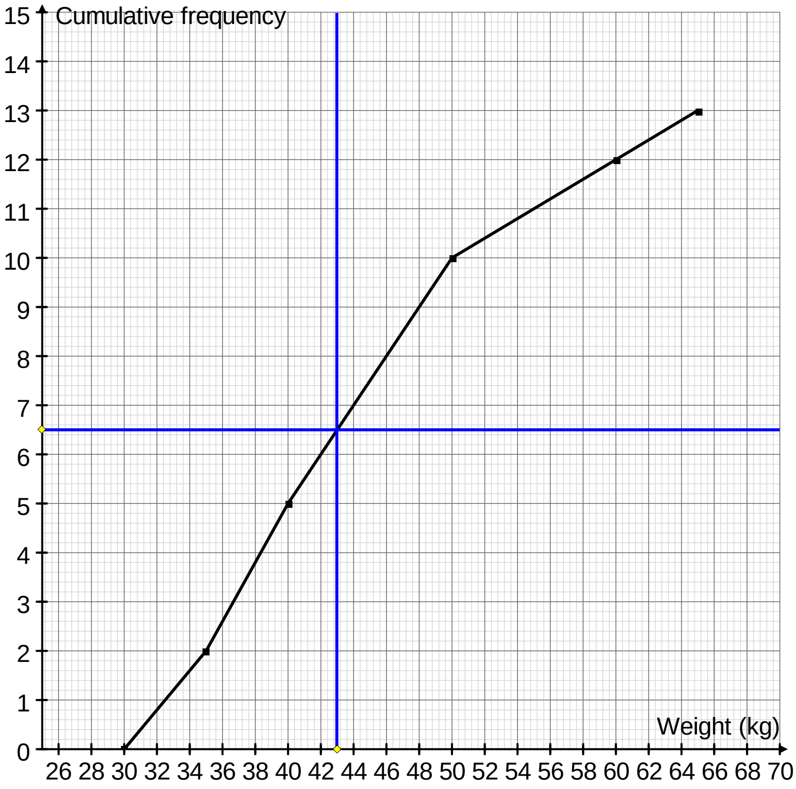


It is easiest to find the locations by splitting the section of the number line into 5 equal intervals and then finding the midpoint of each interval. The median is the second data value in the interval and so it is 43. The estimate of the median is 43 kg.

Edexcel AS Maths Data 2 Notes and Examples

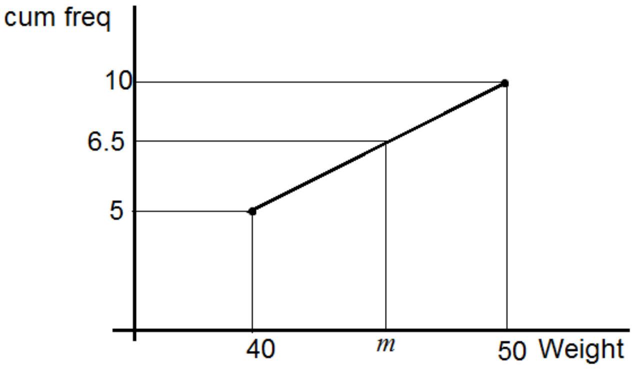
Drawing this kind of diagram to help find the median can be difficult if there are a lot of data values so it is usually easier to use one of the methods below.

The median can be found by drawing a cumulative frequency diagram.



We don't really think about individual data items for a cumulative frequency graph. We think of it as a mapping from the data values to the numbers 0 to *n* so the centre point (median) is at value n/2 on the vertical axis. In this case, 13/2 = 6.5.

Instead of drawing the whole cumulative frequency graph, you can draw a sketch of the relevant section instead.



Using the gradient of the line

$$\frac{10-5}{50-40}=\frac{6.5-5}{m-40}$$

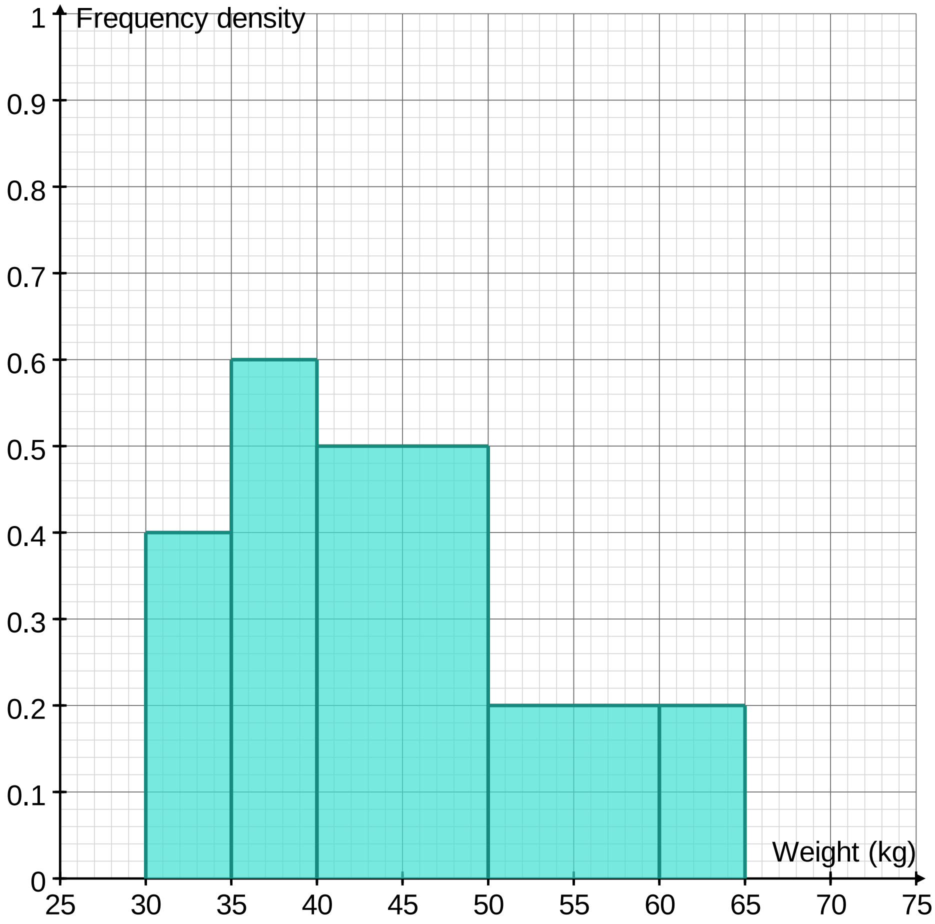
$$\frac{5}{10}=\frac{1.5}{m-40}$$

$$\frac{m-40}{1.5}=2$$

$$m=40+2\times1.5=43$$

The estimate of the median is 43 kg.

A vertical line at the median cuts the area of the histogram into two equal halves.

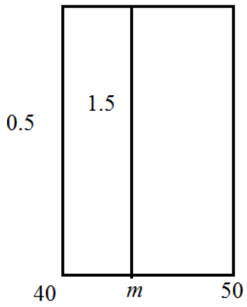


You know the median lies between 40 and 50. This bar represents a frequency of 5. Area in a histogram represents frequency so the area of the bar is 5. Its width is 10 so its height is 0.5.

The total area of the histogram is the total frequency; this is 13. The area of the first two bars together is 5. Half the total area is 13/2 = 6.5.

6.5 – 5 = 1.5

The vertical line at the median has an area of 1.5 in front of it in the third bar.



$$0.5(m-40)=1.5$$
$$m-40=3$$
$$m=43$$

The estimate of the median is 43 kg.

Example 7

Estimate the median of the following dataset, which gives the mass of 200 eggs.

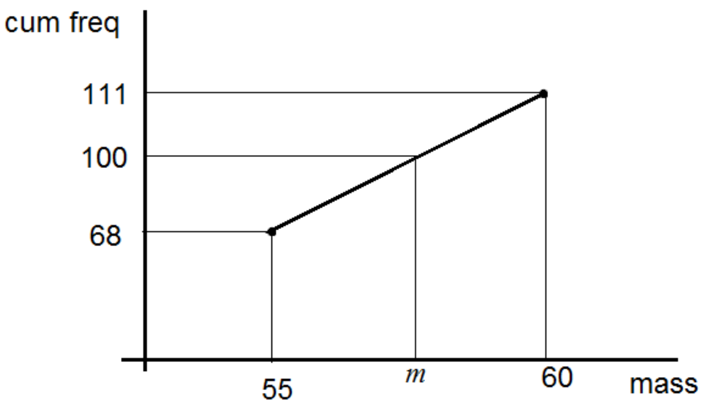
Mass, m (g)	Frequency
$40 \le m < 45$	8
$45\le\ m < 50$	29
$50 \le m < 55$	31
$55 \le m < 60$	43
$60 \le m < 65$	35
$65 \le m < 70$	33
$70 \le m < 75$	21
$75 \le m < 80$	0

Solution

Mass, m (g)	Frequency	Mass	Cumulative frequency
		$m < 40$	0
$40 \le m < 45$	8	$m < 45$	8
$45\le\ m < 50$	29	$m < 50$	37
$50 \le m < 55$	31	$m < 55$	68
$55 \le m < 60$	43	$m < 60$	111
$60 \le m < 65$	35	$m < 65$	146
$65 \le m < 70$	33	$m < 70$	179
$70 \le m < 75$	21	$m < 75$	200

There are 200 data values numbered 1 to 200. The median is the 100.5th data item. This lies in the $55 \le m < 60$ class interval.

The cumulative frequency diagram has a vertical axis from 0 to n, the middle of this is at $\frac{n}{2}$ i.e. at 100. The relevant section of the cumulative frequency graph is shown below.



Using the gradient of the line

$$\frac{111-68}{60-55}=\frac{100-68}{m-55}$$

$$\frac{43}{5}=\frac{32}{m-55}$$

$$\frac{m-55}{32}=\frac{5}{43}$$

$$m=55+32\times\frac{5}{43}\approx 58.7$$

The estimate of the median is 58.7 g.