

Section 2: Data presentation and interpretation

Notes and Examples – measures of spread

This section deals with

- [The range](#)
- [Quartiles and the inter-quartile range](#)
- [Box and whisker plots](#)
- [Identifying outliers using quartiles](#)
- [Cumulative frequency curves](#)
- [Percentiles](#)
- [Variance and standard deviation](#)
- [The alternative form of the sum of squares](#)
- [Measures of spread using frequency tables](#)
- [Using standard deviation to identify outliers](#)

The range

For a set of data,

$$\text{range} = \text{highest item} - \text{lowest item}$$

This is straightforward to calculate, but is highly sensitive to outliers. For example, consider this set of marks for a maths test:

{45, 50, 43, 49, 52, 58, 48, 10, 50, 82, 56, 40, 47, 39, 51}

The range of the data is $82 - 10 = 72$ marks, but this does not give a good measure of the spread, as most of the marks are in the range 40 – 60. Discounting the ‘10’ and the ‘80’ as outliers gives a range of $58 - 40 = 18$, which is perhaps more representative of the data.

Quartiles and the inter-quartile range

One way of refining the range so that it does not rely completely on the most extreme items of data is to use the interquartile range.

$$\text{Interquartile range} = \text{upper quartile} - \text{lower quartile.}$$

The upper quartile is the median of the upper half of the data, and the lower quartile is the median of the lower half of the data.

For a large data set, 25% of the data lie below the lower quartile, and 75% of the data lie below the upper quartile. The interquartile range measures the range of the middle 50% of the data.

Edexcel AS Maths Data 2 Notes & Examples

For small sets of data, you use a procedure for placing the lower and the upper quartile, similar to that used for placing the median.



Example 1

- (i) Find the interquartile range of the set of marks below from a test taken by 15 students.

50 82 40 51 45 50 48 49 47 10 43 58 56 52 19

- (ii) One student was absent and took the test the following week, scoring 59.
Find the new interquartile range.

Solution

- (i) First arrange the data in order of size:

10 19 40 (43) 45 47 48 (49) 50 50 51 (52) 56 58 82

The lower quartile is the median of the lower 7 marks, which is 43.

There are 15 items of data, so the median is the 8th item, which is 49. Discard this.

The upper quartile is the median of the upper 7 marks, which is 52.

So the interquartile range is $52 - 43 = 9$.

- (ii) The new set of data has 16 items.

10 19 (40 43) 45 47 48 | 49 50 50 (51 52) 56 58 59 82

Median
49.5

The lower quartile is the median of the lower 8 marks, which is 44.

For an even number of data items, the median falls between two items of data, so there is no data item to discard:

The upper quartile is the median of the upper 8 marks, which is 54.

The interquartile range = $54 - 44 = 10$

Note: there are some slightly different ways of finding quartiles, and software may use different methods. However, it makes little practical difference to the result.

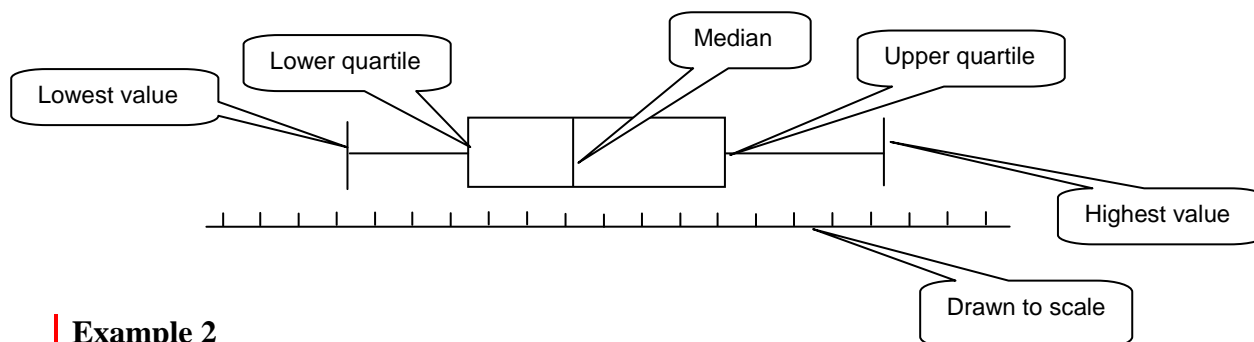
Box-and-whisker plots

The median and quartiles can be displayed graphically by means of a box-and-whisker plot, or boxplot. This gives an extremely useful summary of the data, and can be used to compare sets of data.



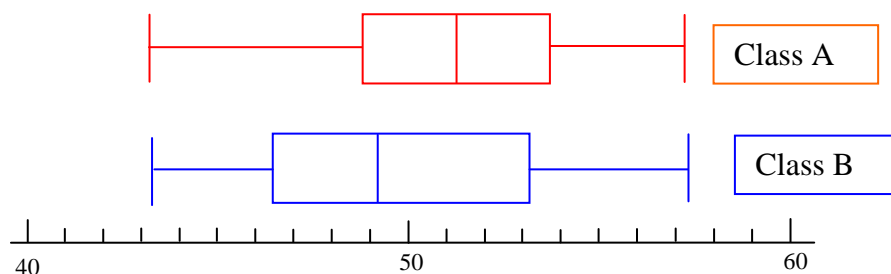
Edexcel AS Maths Data 2 Notes & Examples

In this diagram, a box is drawn from the lower to the upper quartile, and a line drawn in the box showing the position of the median. Whiskers extend from the lowest value to the highest:



Example 2

Compare the following sets of data using their box and whisker plots. They represent marks out of 100 for two classes.



Solution

The ranges of marks are similar, but class A has a lower inter-quartile range than class B, which suggests that the majority of the marks are less spread out for Class A. The median and quartiles for class A are higher than those for class B, so on average class A did slightly better on the test.



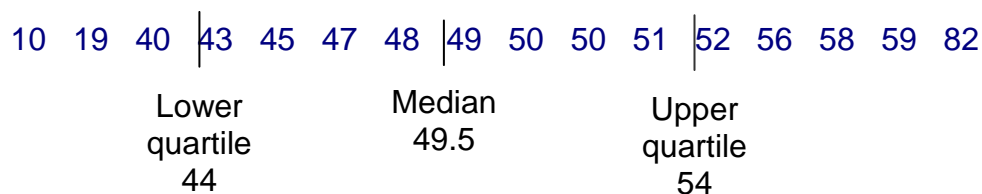
Identifying outliers using quartiles

An *outlier* is an extreme value in a set of data.

One definition of an outlier uses the quartiles and interquartile range. An outlier can be identified as follows (IQR stands for interquartile range):

- any data which are $1.5 \times \text{IQR}$ below the lower quartile;
- any data which are $1.5 \times \text{IQR}$ above the upper quartile.

For example, here is the dataset from Example 1(ii).



Edexcel AS Maths Data 2 Notes & Examples

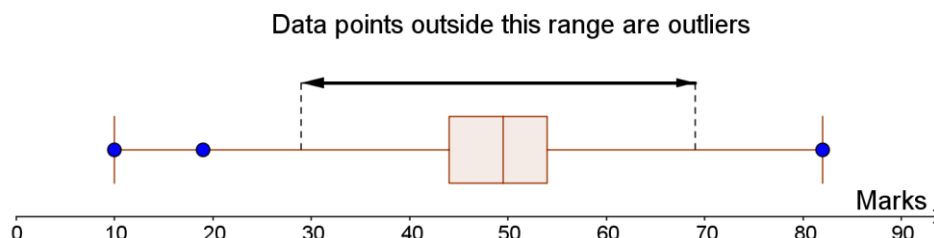
The interquartile range is $54 - 44 = 10$.

$$1.5 \times \text{IQR} = 1.5 \times 10 = 15$$

$1.5 \times \text{IQR}$ below the lower quartile = $44 - 15 = 29$, so 10 and 19 are outliers.

$1.5 \times \text{IQR}$ above the upper quartile = $54 + 15 = 69$, so 82 is an outlier.

The box-and-whisker diagram below shows the outliers.



Cumulative frequency tables and curves

Cumulative frequency curves enable us to estimate how many of the items of data fall below any particular value. For large data sets, they are also used to estimate medians, quartiles and percentiles for the data.

For grouped data, cumulative frequencies must be plotted against the upper class boundaries. Here is some data on the weights of eggs

Mass m (g)	Frequency
$40 \leq m < 45$	4
$45 \leq m < 50$	15
$50 \leq m < 55$	15
$55 \leq m < 60$	22
$60 \leq m < 65$	17
$65 \leq m < 70$	16
$70 \leq m < 75$	11
$75 \leq m < 80$	0

The cumulative frequency table is shown below:

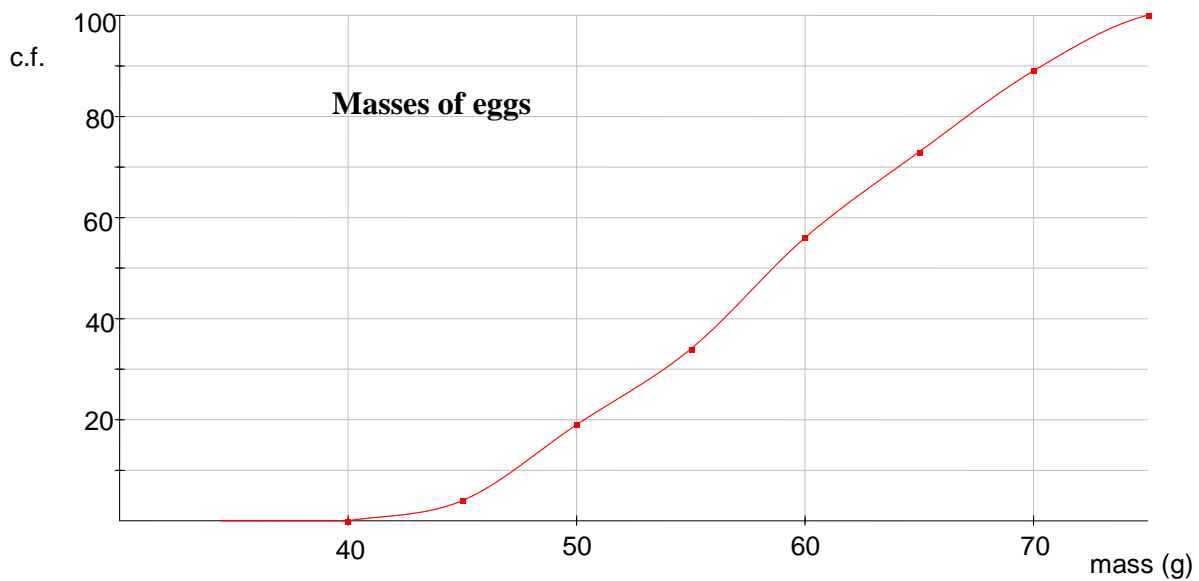
Mass m (g)	Frequency	Mass	Cumulative frequency
		$m < 40$	0
$40 \leq m < 45$	4	$m < 45$	4
$45 \leq m < 50$	15	$m < 50$	19
$50 \leq m < 55$	15	$m < 55$	34
$55 \leq m < 60$	22	$m < 60$	56
$60 \leq m < 65$	17	$m < 65$	73
$65 \leq m < 70$	16	$m < 70$	89
$70 \leq m < 75$	11	$m < 75$	100

Check that you understand the relationship between the frequency column and the cumulative frequency column.

This row shows the endpoint of the graph, in this case (40, 0)

This row tells you that 73 of the eggs have a mass of less than 65 grams

Edexcel AS Maths Data 2 Notes & Examples



IMPORTANT

Note: for the data above, however, the cumulative frequencies are given as the frequencies for $m < 40$, $m < 50$ and so on. Since the data is continuous, there is no distinction between $m < 40$ and $m \leq 40$, so there is no problem with this. However, when you are dealing with discrete data, you must ensure that cumulative frequencies relate to “less than or equal to” a value.



Example 3

Draw a cumulative frequency curve for the following data giving weights of passengers on a bus, and use it to estimate how many passengers weigh over 55 kg.

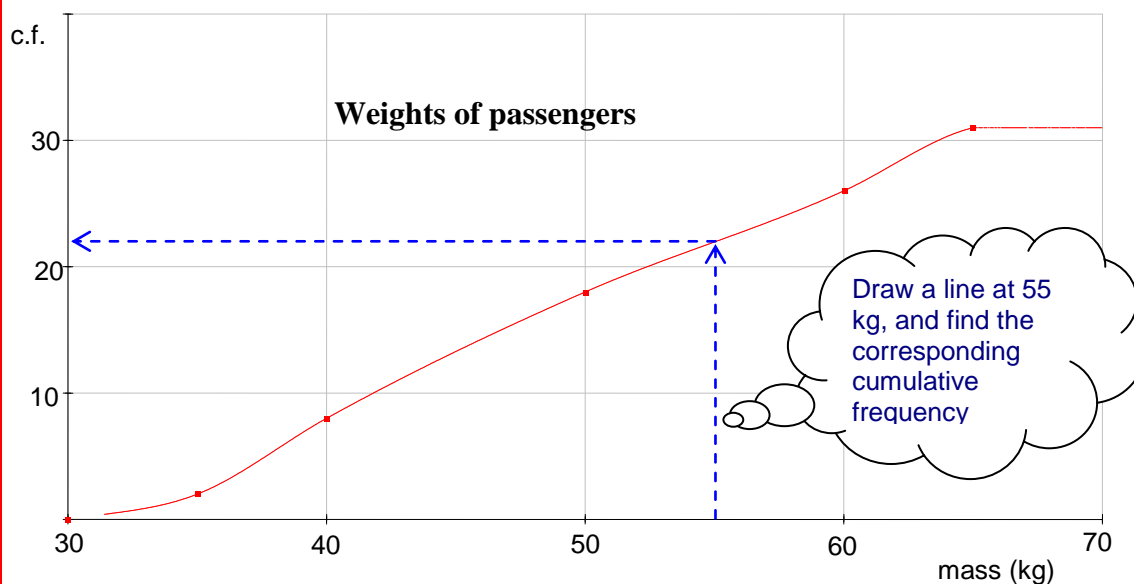
Weight w , (kg)	Frequency
$30 \leq w < 35$	2
$35 \leq w < 40$	6
$40 \leq w < 50$	10
$50 \leq w < 60$	8
$60 \leq w < 65$	5
over 65	0

Solution

Weight (kg)	Frequency	Weight	Cumulative frequency
		$w < 30$	0
$30 \leq w < 35$	2	$w < 35$	2
$35 \leq w < 40$	6	$w < 40$	8
$40 \leq w < 50$	10	$w < 50$	18
$50 \leq w < 60$	8	$w < 60$	26



Edexcel AS Maths Data 2 Notes & Examples



Approximately 22 passengers weigh under 55 kg.
There are 31 passengers altogether, so 9 weigh over 55 kg.

Cumulative frequency curves are useful for estimating the quartiles and the interquartile range of a large data set. The next example shows the eggs data again.



Example 4

Estimate the median and interquartile range of the following dataset, which gives the mass of 100 eggs:

Mass, m (g)	Frequency
$40 \leq m < 45$	4
$45 \leq m < 50$	15
$50 \leq m < 55$	15
$55 \leq m < 60$	22
$60 \leq m < 65$	17
$65 \leq m < 70$	16
$70 \leq m < 75$	11
$75 \leq m < 80$	0

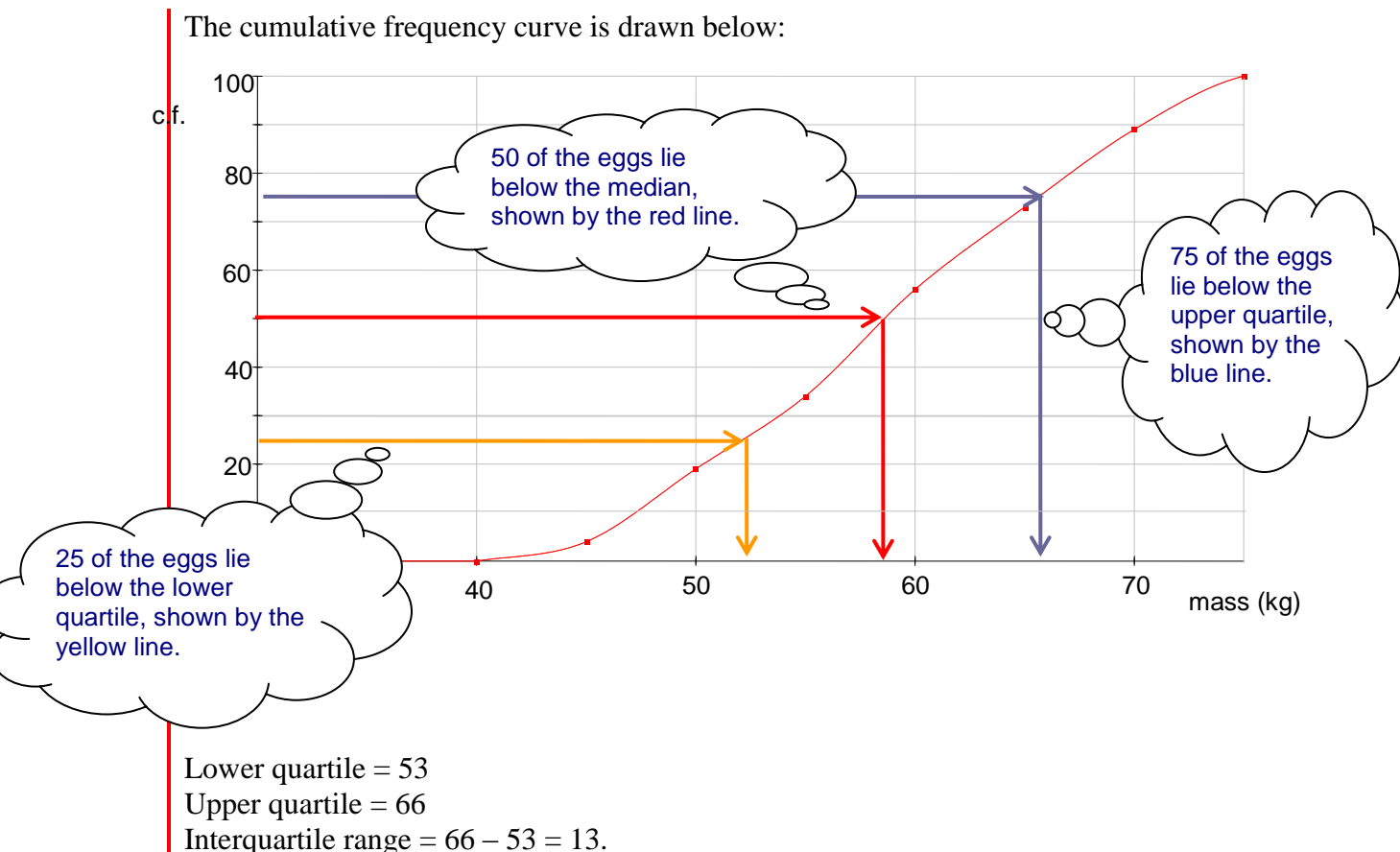
Solution

Mass, m (g)	Frequency	Mass	Cumulative frequency
		$m < 40$	0
$40 \leq m < 45$	4	$m < 45$	4
$45 \leq m < 50$	15	$m < 50$	19
$50 \leq m < 55$	15	$m < 55$	34
$55 \leq m < 60$	22	$m < 60$	56
$60 \leq m < 65$	17	$m < 65$	73
$65 \leq m < 70$	16	$m < 70$	89
$70 \leq m < 75$	11	$m < 75$	100



Edexcel AS Maths Data 2 Notes & Examples

The cumulative frequency curve is drawn below:



Percentiles

75% percent of the data lies below the upper quartile. 25% of the data lies below the lower quartile. This concept can be generalised to give the value below which any percentage of the data lies. These are called percentiles.

For example, the 10th percentile is the value below which 10% of the data lie.

Example 5

For the 'eggs' data from Example 4, estimate the 20th percentile and the 70th percentile.

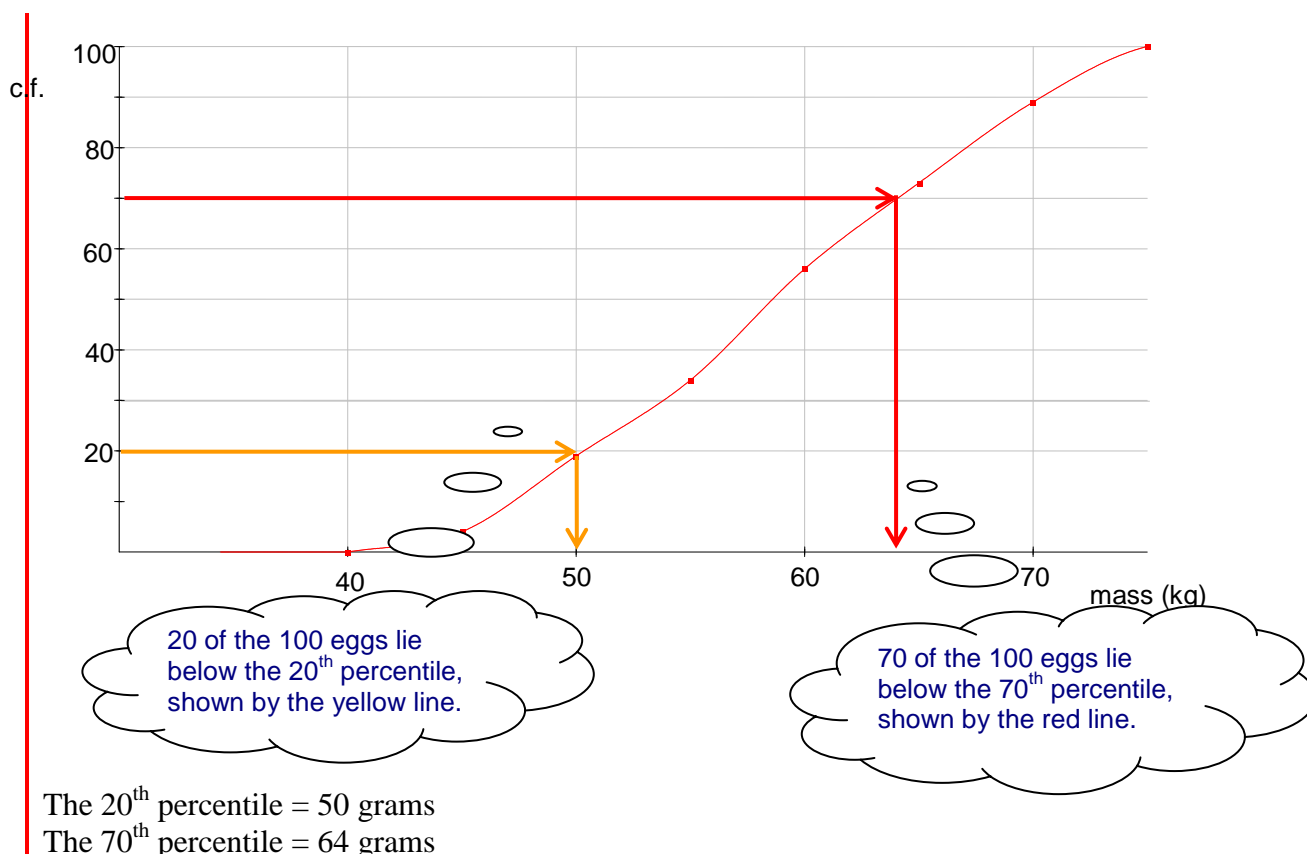
Solution

(See Example 4 for the cumulative frequency tables)

The cumulative frequency curve is drawn below:



Edexcel AS Maths Data 2 Notes & Examples



Notice that the median is the 50th percentile, the lower quartile is the 25th percentile and the upper quartile the 75th percentile.

Sometimes you need to think carefully about which percentile you need. In the example below, because 70% of the students passed the test, it is tempting to think that you need the 70th percentile. In fact, because cumulative frequency tells you how many are below a certain point, you need to look at the 30th percentile since 30% scored below the pass mark.

Example 6

The marks scored by 200 students in a test were as follows:

Mark (%)	Frequency
1 – 10	1
11 – 20	5
21 – 30	12
31 – 40	23
41 – 50	45
51 – 60	64
61 – 70	25
71 – 80	13
81 – 90	8
91 – 100	4

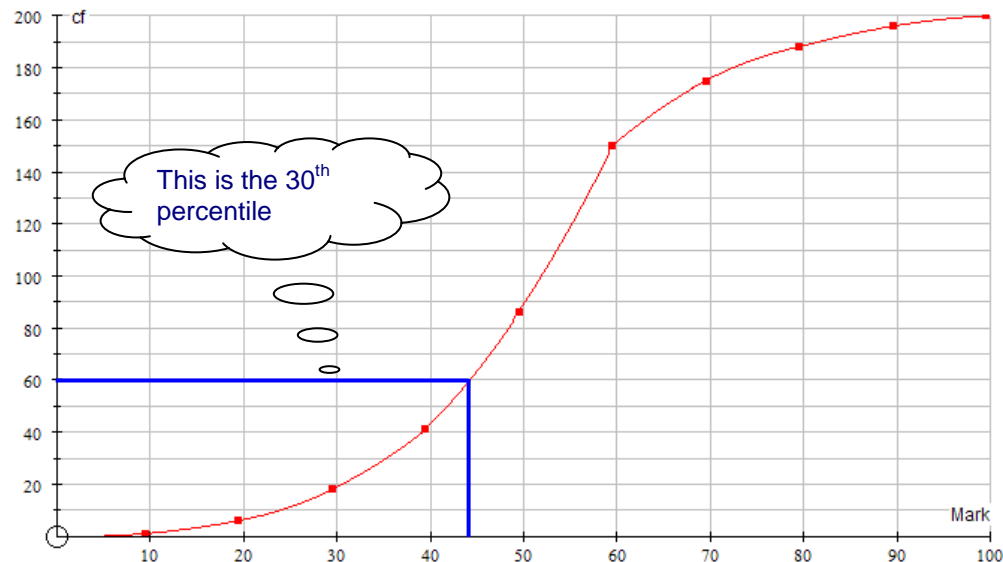
Edexcel AS Maths Data 2 Notes & Examples

70% of the students passed the test. What was the pass mark?

Solution

Mark (%)	Frequency	Mark m	Cumulative frequency
1 – 10	1	$m \leq 10$	1
11 – 20	5	$m \leq 20$	6
21 – 30	12	$m \leq 30$	18
31 – 40	23	$m \leq 40$	41
41 – 50	45	$m \leq 50$	86
51 – 60	64	$m \leq 60$	150
61 – 70	25	$m \leq 70$	175
71 – 80	13	$m \leq 80$	188
81 – 90	8	$m \leq 90$	196
91 – 100	4	$m \leq 100$	200

The cumulative frequency diagram is shown below:



70% of the students passed, so 30% scored less than the pass mark.

30% of 200 is 60.

From the graph, the 30th percentile is 44.

The pass mark is 44%.

Notice that in Example 6 above, you are dealing with discrete data, so that the cumulative frequencies relate to frequencies **less than or equal to** a particular mark.

Variance and standard deviation

Another way of measuring the spread of data is using the standard deviation. This is calculated from the mean of the data, so it is usually used alongside the mean, just as the interquartile range is usually used alongside the median.

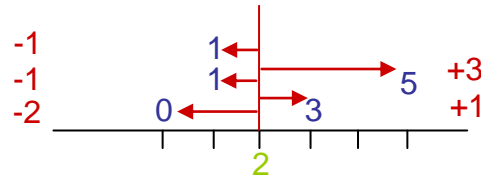
Edexcel AS Maths Data 2 Notes & Examples

Consider a small set of data: $\{0, 1, 1, 3, 5\}$

The mean of this data is given by $\bar{x} = \frac{0+1+1+3+5}{5} = 2$

The **deviation** of an item of data from the mean is the difference between the data item and the mean, i.e. $x - \bar{x}$.

The set of deviations for this set of data is:
 $\{-2, -1, -1, 1, 3\}$



These deviations give a measure of spread. However, there is no point in just adding them up, because their sum is always zero! Instead, square each deviation and add them up. The sum of their squares is denoted S_{xx} :

For the set of data above:

$$S_{xx} = (-2)^2 + (-1)^2 + (-1)^2 + 1^2 + 3^2 = 16$$

In general:

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{or} \quad S_{xx} = \sum (x - \bar{x})^2$$

To use S_{xx} as a measure of spread, it is necessary to take into account the number of data items, so that the spread of two data sets of different sizes can be compared.

The quantity $\frac{S_{xx}}{n}$ for a sample of data is called the *variance*. It is usually denoted by s^2 .

The *standard deviation* is the square root of the sample variance and is given by

$$s = \sqrt{\frac{S_{xx}}{n}}$$

In general:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n}$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Sometimes the divisor $n - 1$ is used rather than n , so that the standard deviation is

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

In this work you are expected to use n . Be careful when using calculator or spreadsheet functions to find standard deviation – there are usually functions for both versions, so make sure you use the one with divisor n .

Edexcel AS Maths Data 2 Notes & Examples

The alternative form of the sum of squares

When the mean does not work out neatly, the deviations will also be difficult to work with. In this case, it is easier to work with an alternative formula for S_{xx} :

$$S_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - n\bar{x}^2$$

For the first dataset {0, 1, 1, 3, 5}:

$$\bar{x} = 2$$

$$\sum x^2 = 0^2 + 1^2 + 1^2 + 3^2 + 5^2 = 0 + 1 + 1 + 9 + 25 = 36$$

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 36 - 5 \times 2^2 = 36 - 20 = 16 \text{ as before.}$$

The measures of spread can now be written in the alternative forms:

$$s^2 = \frac{\sum x^2 - n\bar{x}^2}{n}$$

$$s = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n}}$$

Example 7

A group of children were asked how many pets they own.

The results were {0, 0, 1, 1, 1, 2, 3}.

Calculate the standard deviation of the number of pets owned.

Solution

$$\bar{x} = \frac{0+0+1+1+1+2+3}{7} = \frac{8}{7}$$

$$\sum x^2 = 0^2 + 0^2 + 1^2 + 1^2 + 1^2 + 2^2 + 3^2 = 16$$

$$\text{Standard deviation} = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n}} = \sqrt{\frac{16 - 7 \times \left(\frac{8}{7}\right)^2}{7}} = 0.990 \text{ (3 s.f.)}$$

Since the mean is not a round number, it is easier to use the second form of the formula.

Always do the whole calculation at once. Do not use a rounded version of the mean!

For large sets of data, you are sometimes given a summary of the data: the values of n , $\sum x$ and $\sum x^2$.

Example 8

A set of sample data is summarised as:

$$n = 100 \quad \sum x = 1420 \quad \sum x^2 = 22125.$$

Find

- (i) the mean
- (ii) the standard deviation

Solution

$$(i) \quad \bar{x} = \frac{\sum x}{n} = \frac{1420}{100} = 14.2$$

$$(ii) \quad \text{standard deviation} = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n}} = \sqrt{\frac{22125 - 100 \times 14.2^2}{100}} = 4.43$$

Edexcel AS Maths Data 2 Notes & Examples

Measures of spread using frequency tables

In the previous section, you saw how the formula for the mean: $\bar{x} = \frac{\sum x}{n}$

can be adapted for use with data given in a frequency table: $\bar{x} = \frac{\sum fx}{\sum f}$.

In the same way, the formulae for the measures of spread can be adapted for data given in a frequency table.

$$S_{xx} = \sum fx^2 - n\bar{x}^2$$
$$s^2 = \frac{S_{xx}}{\sum f}$$
$$s = \sqrt{\frac{S_{xx}}{\sum f}}$$

Be careful: fx^2 means square x , then multiply by f .

It is often convenient to set out the calculation in columns, as shown in the following example:



Example 9

The table below shows the number of occupants of each house in a small village.

Number of occupants	Frequency
1	26
2	34
3	19
4	57
5	42
6	12
7	3
8	1
Total	194

Find the mean and standard deviation of the number of occupants.

Solution

x	f	fx	x^2	fx^2
1	26	26	1	26
2	34	68	4	136
3	19	57	9	171
4	57	228	16	912
5	42	210	25	1050
6	12	72	36	432
7	3	21	49	147
8	1	8	64	64
	$\sum f = 194$	$\sum fx = 690$		$\sum fx^2 = 2938$



Edexcel AS Maths Data 2 Notes & Examples

$$\text{Mean} = \frac{\sum fx}{\sum f} = \frac{690}{194} = 3.557$$

$$\sigma = \sqrt{\frac{\sum fx^2 - n\bar{x}^2}{n}} = \sqrt{\frac{2938 - 194 \times \left(\frac{690}{194}\right)^2}{194}} = 1.58$$

In practice, of course, calculations like these can be carried out much more easily by entering the data into a calculator (most calculators allow you to enter either raw data or frequencies, and then will calculate the various statistical measures for you).



Example 10

Estimate the mean and standard deviation of the data with the following frequency distribution:

Weight, w , (grams)	Frequency, f
$0 \leq w < 10$	4
$10 \leq w < 20$	6
$20 \leq w < 30$	9
$30 \leq w < 40$	7
$40 \leq w < 50$	4

Solution

w	Mid-interval value, x	f	fx	x^2	fx^2
$0 \leq w < 10$	5	4	20	25	100
$10 \leq w < 20$	15	6	90	225	1350
$20 \leq w < 30$	25	9	225	625	5625
$30 \leq w < 40$	35	7	245	1225	8575
$40 \leq w < 50$	45	4	180	2025	8100
		$\sum f = 30$	$\sum fx = 760$		$\sum fx^2 = 23750$

$$\text{Mean} = \frac{760}{30} = 25.33$$

$$\text{Standard deviation} = \sqrt{\frac{\sum fx^2 - n\bar{x}^2}{n}} = \sqrt{\frac{23750 - 30 \times \left(\frac{760}{30}\right)^2}{30}} = 12.24$$

Using standard deviation to identify outliers

Standard deviation can be used to identify outliers, using the following rule:

All data which are over 2 standard deviations away from the mean are identified as outliers.

Edexcel AS Maths Data 2 Notes & Examples



Example 11

Use the standard deviation to identify any outliers in the following set of data which gives the ages of the people at a golf club dinner.

45 34 12 56 56 73 99 33 25 45 60 56 30 32 21 35 56 40 30 28

Solution

$$n = 20$$

$$\sum x = 866$$

$$\sum x^2 = 45212$$

$$\bar{x} = \frac{866}{20} = 43.3$$

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 45212 - 20 \times 43.3^2 = 7714.2$$

$$s = \sqrt{\frac{S_{xx}}{n}} = \sqrt{\frac{7714.2}{20}} = 19.64$$

2 standard deviations below the mean is $43.3 - 2 \times 19.64 = 4.02$.

2 standard deviations above the mean is $43.3 + 2 \times 19.64 = 82.58$

So any outliers are below 4.02 or above 82.58.

The only value outside this range is 99; so this is the only outlier.

