# M1.10 – Understanding measures of dispersion including standard deviation and range

### Tutorials

Learners may be tested on their ability to:

* calculate the standard deviation
* understand why standard deviation might be a more useful measure of dispersion for a given set of data e.g. where there is an outlying result

### Standard Deviation

You will already be familiar with calculating the “mean” of a dataset. This is a very useful summary value, but it can also be interesting to look at the distribution of scores in the data. One way to do this is to calculate the range of scores in the data. Calculating the range of scores is easy – simply subtract the smallest score from the largest score.

Here is an example - There are 11 people on a bus and their ages are as follows: 15,17, 19, 22, 36, 39, 40, 44, 53, 54 & 90

The oldest person on the bus is 90, and the youngest person is 15

The **interval** within which all the data in the set falls is 15 years to 90 years. This sometimes gets confused with the range, as it is natural in everyday speech to say ‘the ages of people on the bus ranged from 15 to 90 years’. However, to find the range, by its strict definition, we need to do that very simple calculation:

Subtract smallest score from largest score

the **range** is 90 - 15 = 75 years.

One problem with using the range is that because it only uses the highest score and lowest score it is dramatically affected by extreme scores.

For example, the age range on the bus was 75 years, but if we compute the range while excluding the score of 90 and use the next highest age score, which was 54, we see that the range drops dramatically from 75 years to 39 years. Almost half the size!

Another problem with the range is that it is little help in making inferences about a whole population when the data we have is just a sample.

In this case if we were using the people on the bus as a sample of the whole population of a town, knowing the range in our sample is 75 only really tells us that the range in the whole population must be at least 75.

A different, and much more useful, way of measuring dispersion in sample data is to calculate the sample standard deviation. This is less influenced by outlying results and it allows us to make inferences about the whole population from which our sample is drawn.

The greater the standard deviation, the greater the spread of the data.

The standard deviation is calculated in a few steps.

One simple way to describe what we are aiming to get out of these steps is that we are going to work out how different from the mean each of our data points is. If the data are widely spread we’ll get a big number, whereas if they are generally close to the mean we’ll get a small number. But we’re going to do this in two slightly strange ways, which give us a far more useful statistic than a ‘straightforward’ mean of the differences:

* We will square the differences, to deliberately accentuate the effect of the larger ones
* We will divide by one less than the number of data points (i.e. n-1 instead of n)

So first we need to know the mean of our data. Let’s use the ages of the people on the bus again.

$\overbar{x}$ = $\frac{15+17+ 19+ 22+ 36+ 39+ 40+ 44+ 53+ 54+ 90 }{11}$= 39

Now we subtract the mean value from each of our observed values (the individual ages of people on the bus)

$(x\_{i}-\overbar{x})$ = (15-39) (17-39) (19-39) (22-39) (36-39) (39-39) (40-39) (44-39) (53-39) (54-39) (90-39)

$(x\_{i}-\overbar{x})$ = -24, -22, -20, -17, -3, 0, 1, 5, 14, 15, 51

Next we square all these values. This increases the effect on the final result of the numbers further away from the mean (which we want to do to give us the most useful statistic for drawing inferences about the population – it is **not** done simply to make all the deviations positive!). We sum the values next and you can see that the ‘51’ makes a big contribution (2601) to the resulting sum now that it is squared.

∑$(x\_{i}-\overbar{x})$2 = (-24)2+(-22)2+(-20)2+(-17)2+(-3)2+(0)2+(1)2+(5)2+(14)2+(15)2+(51)2 = 4806

This sum of the squares is a good measure of the precision of our model, but it needs to reflect how much data we collected (i.e. we are getting a **mean** squared difference) and so we divide the sum of squares by the number of observations minus 1 (n-1), (More on this in a moment)

$\frac{∑(x\_{i}-\overbar{x})2 }{n-1}= \frac{4806}{10}$ = 480.6

There is only one problem with this number – it gives us the mean difference between the mean and the observations made, but it gives us this in units squared. So if we square root this value it will turn back into the original unit of measurement (age in years).

$$\sqrt{\frac{∑(x\_{i}-\overbar{x})2 }{n-1}}= \sqrt{\frac{4806}{10}}=\sqrt{480.6}=21.9$$

The standard deviation (s) is a very useful tool; the smaller this value is, on average the closer our individual data points sit towards the mean.

With a normal distribution of data we would expect the following to hold true:

* 68% of the data lie within one standard deviation either side of the mean
* 95% of the data lie within 2 standard deviations either side of the mean.

**n-1**

Why do we divide by n-1 rather than n?

If we really want to know the ‘root mean square difference from the mean’ (which is a pretty good description of what the standard deviation is) why don’t we divide by n just like calculating other means?

In fact if we were calculating the standard deviation of a whole population (e.g. if we were only interested in the people on the bus as a (tiny) whole population and did not want to use them as a sample to make estimates about a larger population) we **would** divide by n.

But in most biological data handling situations we actually have a sample (because we don’t have time to measure every single individual in the whole population) and we use our calculation of the sample standard deviation to make inferences about the population. Using n-1 in our calculation is essentially a correction factor that makes the number we come up with from our sample a better estimate of the standard deviation of the population.

So for AS and A Level Biology we stick to using the n-1 formula.

If you are using a calculator or a spreadsheet to calculate standard deviation make sure you press the right button, or use the right embedded formula.

**Document updates**

 v1.0 April 2017 Original version.

 v1.1 June 2019 Changed how the word accuracy and mean were used in order to be in line with the ‘Language of measurement’

**OCR Resources**: *the small print*OCR’s resources are provided to support the delivery of OCR qualifications, but in no way constitute an endorsed teaching method that is required by the Board, and the decision to use them lies with the individual teacher. Whilst every effort is made to ensure the accuracy of the content, OCR cannot be held responsible for any errors or omissions within these resources.
© OCR 2019 - This resource may be freely copied and distributed, as long as the OCR logo and this message remain intact and OCR is acknowledged as the originator of this work.

OCR acknowledges the use of the following content: n/a

Please get in touch if you want to discuss the accessibility of resources we offer to support delivery of our qualifications: resources.feedback@ocr.org.uk

### Produced in collaboration with the University of East Anglia

We’d like to know your view on the resources we produce. By clicking on ‘Like’ or ‘Dislike’ you can help us to ensure that our resources work for you. When the email template pops up please add additional comments if you wish and then just click ‘Send’. Thank you.

If you do not currently offer this OCR qualification but would like to do so, please complete the Expression of Interest Form which can be found here: [www.ocr.org.uk/expression-of-interest](http://www.ocr.org.uk/expression-of-interest)

Looking for a resource? There is now a quick and easy search tool to help find free resources for your qualification:
[www.ocr.org.uk/i-want-to/find-resources/](http://www.ocr.org.uk/i-want-to/find-resources/)

This resource has been produced as part of our free A Level teaching and learning support package. All the A Level teaching and learning resources, including delivery guides, topic exploration packs, lesson elements and more are available on the qualification webpages.

 If you are looking for examination practice materials, you can find the Sample Assessment Materials (SAMs) on the qualification webpages: [Biology A](http://www.ocr.org.uk/qualifications/as-a-level-gce-biology-a-h020-h420-from-2015/) / [Biology B](http://www.ocr.org.uk/qualifications/as-a-level-gce-biology-b-advancing-biology-h022-h422-from-2015/)