Measures of variability: the range, inter-quartile range and standard deviationar charts

This guide outlines three methods used to summarise the variability in a dataset. It will help you identify which measure is most appropriate to use for a particular set of data. Examples are also given of the use of these measures and how the standard deviation can be calculated using Excel.

Other useful guides: ***Using averages, Working with percentages***

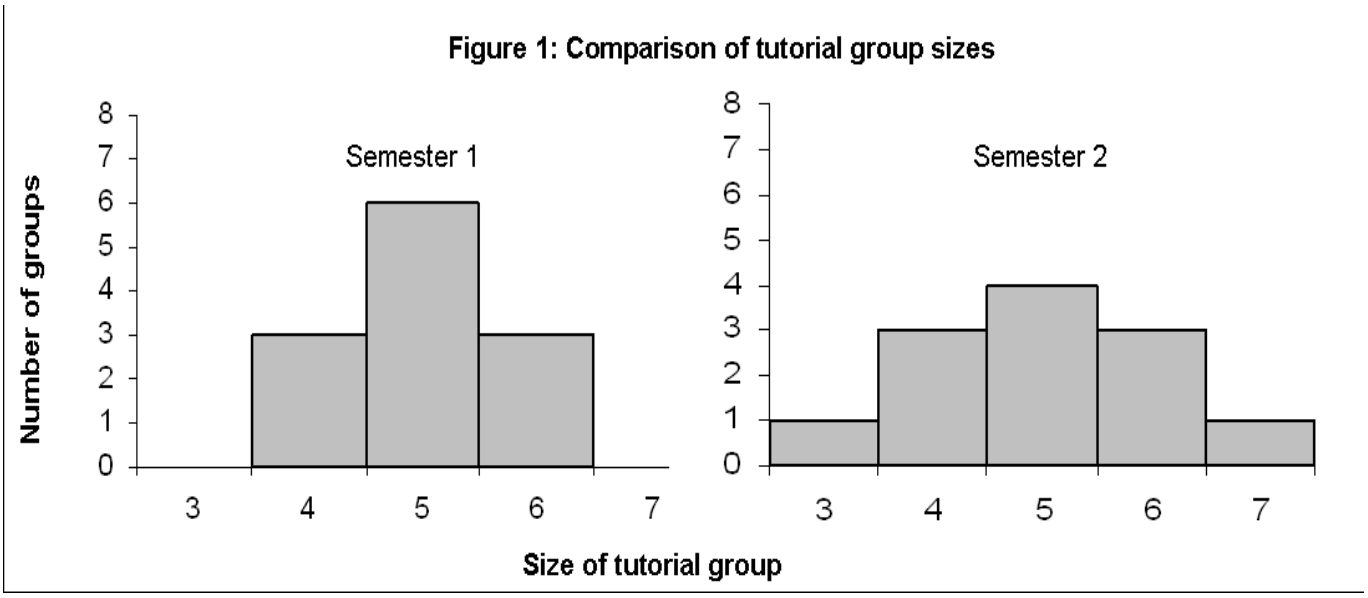
Introduction

Measures of average such as the median and mean represent the typical value for a dataset. Within the dataset the actual values usually differ from one another and from the average value itself. The extent to which the median and mean are good representatives of the values in the original dataset depends upon the variability or dispersion in the original data. Datasets are said to have high dispersion when they contain values considerably higher and lower than the mean value.

In figure 1 the number of different sized tutorial groups in semester 1 and semester 2 are presented. In both semesters the mean and median tutorial group size is 5 students, however the groups in semester 2 show more dispersion (or variability in size) than those in semester 1.

Dispersion within a dataset can be measured or described in several ways including the range, inter-quartile range and standard deviation.

Figure 1: Comparison of tutorial group sizes



The Range

The range is the most obvious measure of dispersion and is the difference between the lowest and highest values in a dataset. In figure 1, the size of the largest semester 1 tutorial group is 6 students and the size of the smallest group is 4 students, resulting in a range of 2 (6-4). In semester 2, the largest tutorial group size is 7 students and the smallest tutorial group contains 3 students, therefore the range is 4 (7-3).

- The range is simple to compute and is useful when you wish to evaluate the whole of a dataset.
- The range is useful for showing the spread within a dataset and for comparing the spread between similar datasets.

An example of the use of the range to compare spread within datasets is provided in table 1. The scores of individual students in the examination and coursework component of a module are shown.

Table 1: Comparison of coursework and examination marks for 14 students

Student	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Coursework mark	27	44	39	23	41	48	37	34	40	43	30	43	29	27
Examination mark	12	47	26	25	38	45	35	35	41	39	32	25	18	30

To find the range in marks the highest and lowest values need to be found from the table. The highest coursework mark was 48 and the lowest was 27 giving a range of 21. In the examination, the highest mark was 45 and the lowest 12 producing a range of 33. This indicates that there was wider variation in the students' performance in the examination than in the coursework for this module.

Since the range is based solely on the two most extreme values within the dataset, if one of these is either exceptionally high or low (sometimes referred to as outlier) it will result in a range that is not typical of the variability within the dataset. For example, imagine in the above example that one student failed to hand in any coursework and was awarded a mark of zero, however they sat the exam and scored 40. The range for the coursework marks would now become 48 (48-0), rather than 21, however the new range is not typical of the dataset as a whole and is distorted by the outlier in the coursework marks. In order to reduce the problems caused by outliers in a dataset, the inter-quartile range is often calculated instead of the range.

The Inter-quartile Range

The inter-quartile range is a measure that indicates the extent to which the central 50% of values within the dataset are dispersed. It is based upon, and related to, the median.

In the same way that the median divides a dataset into two halves, it can be further divided into quarters by identifying the upper and lower quartiles. The lower quartile is found one quarter of the way along a dataset when the values have been arranged in order of magnitude; the upper quartile is found three quarters along the dataset.

Therefore, the upper quartile lies half way between the median and the highest value in the dataset whilst the lower quartile lies halfway between the median and the lowest value in the dataset. The inter-quartile range is found by subtracting the lower quartile from the upper quartile.

For example, the examination marks for 20 students following a particular module are arranged in order of magnitude.

					Lower quartile						Median				Upper quartile					
Student	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Mark	43	48	50	50	52	53	56	58	59	60	62	65	66	68	70	71	74	76	78	80

The median lies at the mid-point between the two central values (10th and 11th)

= half-way between 60 and 62 = 61

The lower quartile lies at the mid-point between the 5th and 6th values

= half-way between 52 and 53 = 52.5

The upper quartile lies at the mid-point between the 15th and 16th values

= half-way between 70 and 71 = 70.5

The inter-quartile range for this dataset is therefore **70.5 - 52.5 = 18** whereas the range is: **80 - 43 = 37**.

The inter-quartile range provides a clearer picture of the overall dataset by removing/ignoring the outlying values.

Like the range however, the inter-quartile range is a measure of dispersion that is based upon only two values from the dataset. Statistically, the standard deviation is a more powerful measure of dispersion because it takes into account every value in the dataset. The standard deviation is explored in the next section of this guide.

Calculating the Inter-quartile range using Excel

The method Excel uses to calculate quartiles is not commonly used and tends to produce unusual results particularly when the dataset contains only a few values. For this reason you may be best to calculate the inter-quartile range by hand.

The Standard Deviation

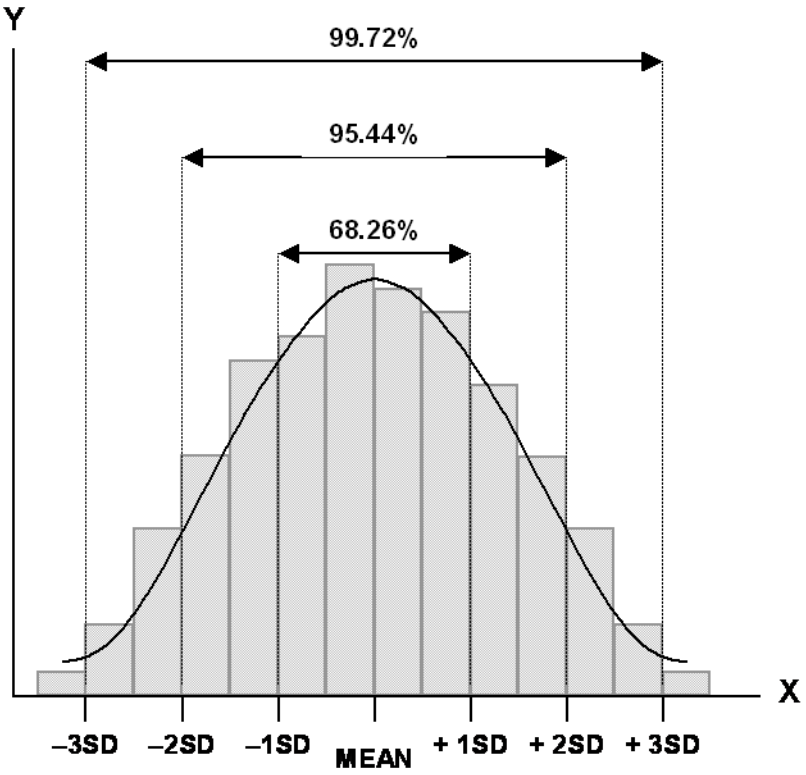
The standard deviation is a measure that summarises the amount by which every value within a dataset varies from the mean. Effectively it indicates how tightly the values in the dataset are bunched around the mean value. It is the most robust and widely used measure of dispersion since, unlike the range and inter-quartile range, it takes into account every variable in the dataset. When the values in a dataset are pretty tightly bunched together the standard deviation is small. When the values are spread apart the standard deviation will be relatively large. The standard deviation is usually presented in conjunction with the mean and is measured in the same units.

In many datasets the values deviate from the mean value due to chance and such datasets are said to display a normal distribution. In a dataset with a normal distribution most of the values are clustered around the mean while relatively few values tend to be extremely high or extremely low. Many natural phenomena display a normal distribution.

For datasets that have a normal distribution the standard deviation can be used to determine the proportion of values that lie within a particular range of the mean value. For such distributions it is always the case that 68% of values are less than one standard deviation (1SD) away from the mean value, that 95% of values are less than two standard deviations (2SD) away from the mean and that 99% of values are less than three standard deviations (3SD) away from the mean. Figure 3 shows this concept in diagrammatical form.

Figure 3: A frequency distribution with a normal distribution: the relative location of the standard deviation and the mean is indicated

The X axis (horizontal) shows the value of something – such as height, calories consumed, number of books read per year; and the Y axis (vertical) indicates the number of times that value was observed (or its frequency).

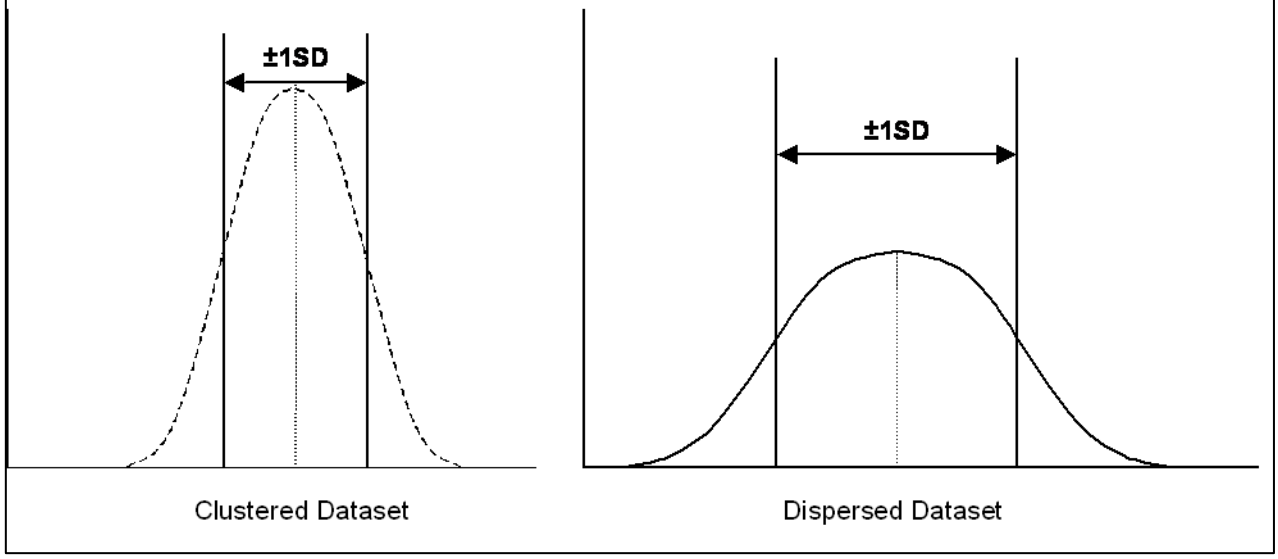


If the mean of a dataset is 25 and its standard deviation is 1.6, then

- 68% of the values in the dataset will lie between **MEAN-1SD** (25-1.6=23.4) and **MEAN+1SD** (25+1.6=26.6)
- 99% of the values will lie between **MEAN-3SD** (25-4.8=20.2) and **MEAN+3SD** (25+4.8=29.8).

If the dataset had the same mean of 25 but a larger standard deviation (for example, 2.3) it would indicate that the values were more dispersed. The frequency distribution for a dispersed dataset would still show a normal distribution but when plotted on a graph the shape of the curve will be flatter as in figure 4.

Example 4: Typical frequency distributions for clustered and dispersed datasets



Population and sample standard deviations

There are two different calculations for the Standard Deviation. Which formula you use depends upon whether the values in your dataset represent an *entire population* or whether they form a *sample* of a larger population. For example, if *all* student users of the library were asked how many books they had borrowed in the past month then the entire population has been studied since *all* the students have been asked. In such cases the population standard deviation should be used. Sometimes it is not possible to find information about an entire population and it might be more realistic to ask a sample of 150 students about their library borrowing and use these results to estimate library borrowing habits for the entire population of students. In such cases the sample standard deviation should be used.

Formulae for the standard deviation

Whilst it is not necessary to learn the formula for calculating the standard deviation, there may be times when you wish to include it in a report or dissertation.

The standard deviation of an *entire population* is known as **σ** (sigma) and is calculated using:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Where **x** represents each value in the population, **μ** is the mean value of the population, **Σ** is the summation (or total), and **N** is the number of values in the population.

The standard deviation of a *sample* is known as **S** and is calculated using:

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

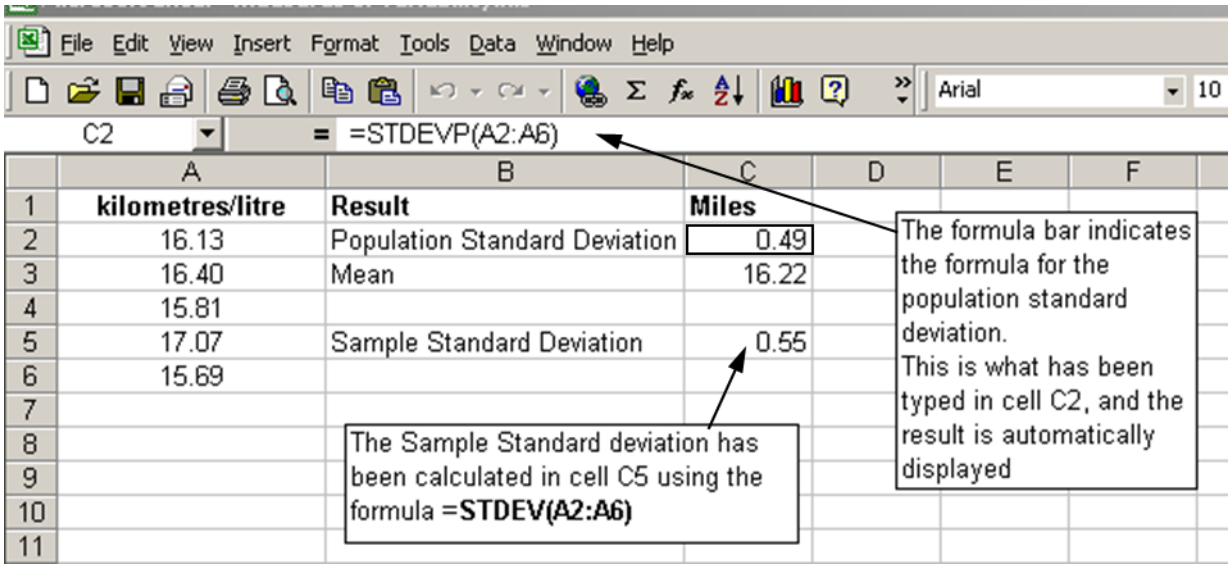
Where **x** represents each value in the population, **x** is the mean value of the sample, **Σ** is the summation (or total), and **n-1** is the number of values in the sample minus 1.

Calculating the standard deviation using Excel

Excel has functions to calculate the population and sample standard deviations. The appropriate commands are entered into the formula bar towards the top of the spreadsheet and the corresponding cells in the spreadsheet are updated to show the result.

For an example of calculating the population standard deviation, imagine you wish to know how fuel-efficient a new car that you have just purchased is. You calculate how many kilometres you have done per litre on your first five trips. This information is presented as column A of the spreadsheet (figure 5). As you have only made 5 trips you do not have any further information and you are therefore measuring the whole population at this point in time. The command to find the population standard deviation in Excel is **=STDEVP(VALUES)** and in this case the command is **=STDEVP(A2:A6)** which gives an answer of 0.49.

Basing your results on the population standard deviation and assuming that your first 5 trips in your new car have been typical of your usual journeys, you can be 99% confident that your new car will do between 14.75 (**MEAN-3SD**) and 17.69 (**MEAN+3SD**) kilometres per litre .



The same data can be used to demonstrate how to calculate the sample standard deviation in Excel. In this case, imagine that the data in column A represent the kilometres per litre found for a sample of 5 new cars tested by the manufacturer. The population standard deviation is calculated using **=STDEV(VALUES)** and in this case the command is **=STDEV(A2:A6)** which produces an answer of 0.55.

The sample standard deviation will always be greater than the population standard deviation when they are calculated for the same dataset. This is because the formula for the sample standard deviation has to take into account the possibility of there being more variation in the true population than has been measured in the sample.

Based on their sample of 5 cars, and therefore using the sample standard deviation, the manufacturers could state with 99% confidence that similar cars will do between 14.57 (**MEAN-3SD**) and 17.87 (**MEAN+3SD**) kilometres per litre .

These examples show the quick method of calculating standard deviations using a cell range. Each of the commands can also be written out in a longer format with the individual kilometres/litre entered.

For example entering: **=STDEV(16.13,16.40,15.81,17.07,15.69)** produces an identical result to **=STDEV(A2:A6).** However, if one of the values in column A was found to be incorrect and adjusted, the cell range method would automatically update the calculation of the standard deviation whereas the longer format will require manual adjustment of the command.

Further information about using Excel to perform calculations can be found in the online tutorials available at http://www.le.ac.uk/slc/resources/it/excel.

Summary

The range, inter-quartile range and standard deviation are all measures that indicate the amount of variability within a dataset. The range is the simplest measure of variability to calculate but can be misleading if the dataset contains extreme values. The inter-quartile range reduces this problem by considering the variability within the middle 50% of the dataset. The standard deviation is the most robust measure of variability since it takes into account a measure of how every value in the dataset varies from the mean. However, care must be taken when calculating the standard deviation to consider whether the entire population or a sample is being examined and to use the appropriate formula.

This study guide is one of a series produced by Student Learning Development at the University of Leicester. As part of our services we provide a range of resources for students wishing to develop their academic and transferable skills.

studyhelp@le.ac.uk | www.le.ac.uk/succeedinyourstudies