Learning Enhancement Team

University of East Anglia

STUDENT SUPPORT SERVICE

Steps into Statistics

Measurements of Spread II: Variance and Standard Deviation

This guide explains the rationale behind two of the most important measurements of the spread of a data set: the variance and the standard deviation. It also gives an example of their calculation.

Introduction

Measurements of spread prove to be crucial when comparing data sets. They are part of the set of **descriptive statistics** and are often referred to as **measures of variation** or **dispersion**. The study guide: *Measurements of Spread I: Range and Interquartile Range* introduces two common measurements of spread and explains that the **range** and **inter-quartile range** (IQR) have their own disadvantages: the range is sensitive to outliers and the IQR is limited by not taking into account every data point.

To get a clearer representation of the spread of a data set, a measure is needed which takes into account each data point. Also, as you are looking to quantify the spread of the data set, it would be useful to ask "What value are the data points spread around?". It makes sense that data points are spread around the middle of a data set and for these purposes the **mean** is a useful measure of the middle (see study guide: *Measurements of Central Tendency*). This guide introduces two common and very important measures of spread, called the **variance** and the **standard deviation**. These descriptive statistics quantify the variation around the mean of a data set which is continuous, interval or ratio, but not nominal or ordinal, level data (see study guide: *Levels of Data*).

Differences from the mean: Using deviations

In statistics, the measure of difference between the mean and a data point is called the **deviation**. In order to calculate this deviation of a data point from the mean, you subtract the mean of the data set from that particular point. A particular data point is given the symbol x_i (where the subscript *i* helps you to keep track of the different points) and the mean is given the symbol $\bar{x}$ (pronounced "*x*-bar"). So:

deviation of a data point $i = x_i - \bar{x}$

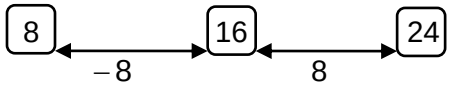
If you calculate the deviation of every data point and add them together you get the

statistical measure the **sum of deviations**. Statistics uses the upper case Greek letter sigma Σ to represent "add them up" and so:

$$\textbf{sum of deviations} = \sum_i (x_i - \bar{x})$$

where the deviation is bracketed because these are what are being added together. You might think that if you divided the total deviations by the number of data points you have, then you would have an average deviation. This is a reasonable idea but, as the mean is in the middle of the data set, **the sum of deviations is always zero for any data set**. This is an important result in statistics and it occurs because the deviations due to points greater than the mean (which are positive) perfectly cancel out the deviations due to points less than the mean (which are negative). This does not help you to measure the spread of your data set, because the sum of deviations is zero you would get a result of zero which is not helpful. Let's use a simple case to see this more clearly. Assume that your data set is the two data points 8 and 24, as shown below:

The mean of these two points is $(8+24)\div 2 = 16$ and you can use this to calculate the deviations for both points which are $8-16=-8$ and $24-16=8$ respectively.



If you add the deviations from both points the sum is zero. You should notice that the size of the deviation is the same for each point, but not the sign. In calculating a measurement of spread, it is the signs of deviations that cause problems.

In statistics the problem of the sum of the deviations being equal to zero is overcome by **squaring the deviations**. When you square a number (i.e. multiply it by itself), you always get a positive number. This solves the problem of the negative sign. For example, for the deviations above, both $(-8)^2$ and 8^2 equal 64. Adding all of the **squared deviations** together gives the **sum of the squared deviations** which is an important measure in statistics as it forms the basis of calculating both the **variance** and the **standard deviation**. You write it mathematically as:

$$\textbf{sum of the squared deviations} = \sum_i (x_i - \bar{x})^2$$

Take notice of the superscript 2 which denotes squaring. So, for the data points above:

$$\text{Sum of the squared deviations} = \sum_i (x_i - \bar{x})^2 = 64 + 64 = 128$$

The variance

The variance is one of the most important descriptive statistics of a data set and one of the measurements of spread most often quoted. **The larger the spread of a data set relative to the mean, the larger the variance is**. It plays a crucial role in inferential statistics, the statistics concerned with comparing data sets. The variance is fundamental to the *F*-test and therefore ANOVA, which stands for **AN**alysis **O**f **VA**riance.

The variance is the 'average' of the sum of the squared deviations. The way you determine this average slightly differs if you have a population as your data set or if you have a sample from a population. Remember that a population is every example of something that has ever existed. It may be difficult to obtain data for whole populations and most often, statistical measures are based on samples taken from populations.

If your data set is a **sample from a population**, the variance of that sample is calculated by dividing the sum of the squared deviations by the number of data points minus one ($n-1$). Written as an equation:

$$\textbf{sample variance} = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

If you have the **whole population** as your data set, the variance of the population is calculated by dividing the sum of squared deviations by the number of data points (n). Written as an equation:

$$\textbf{population variance} = \frac{\sum_i (x_i - \mu)^2}{n}$$

In statistics it is customary to use the lower case Greek letter μ (mu) to represent the population mean. The sample variance of a data set is a good indicator of the population variance from which the sample is taken. Furthermore, the larger your sample is, the closer the sample variance is to the population variance. In statistical language the sample variance is said to be an **unbiased estimator** of the population variance.

The variance is measured in squared units due to the squaring of the deviations. So if you are measuring heights in metres for example, the variance would be measured in metres squared which is an area. This makes it difficult to directly relate the variance to the data points themselves. To overcome this problem, you can take the square root of the variance which gives a statistic with the same units as the data do. The statistic which results from taking the square root of the variance is called the **standard**

deviation. So, returning to the example of heights, the variance has units of metres squared but the standard deviation will have the units of metres, the same as the data.

The standard deviation

The standard deviation is probably the most commonly reported and important measurement of spread of a data set. It is closely related to the variance as it is calculated by taking its square root. As variance is calculated differently for population and for sample data, so is the standard deviation. For a **sample from a population**, the standard deviation is usually denoted by the letter *s* (sometimes *sd*) and is **the square root of the sample variance** given above:

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

You can also say that:

$$\textbf{sample variance} = s^2$$

You should be aware that if you are given a standard deviation you can always calculate the associated variance and vice-versa.

The **population standard deviation** which is given by the lower case Greek letter σ (sigma) is calculated by taking **the square root of the population variance**:

$$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{n}}$$

It is common in statistics for sample statistics to be represented by Latin letters and for population statistics by Greek letters. In statistical analysis it is sometimes difficult to measure the standard deviation of the population as it is only possible if you have data on the **whole** population. However, as with the sample variance, a sample standard deviation is an unbiased estimator of the population standard deviation. The reason for the difference between the calculations of the sample and population standard deviations is beyond the scope of this study guide. However, a **Learning Enhancement Tutor** will be happy to discuss it with you.

Calculating a variance and standard deviation

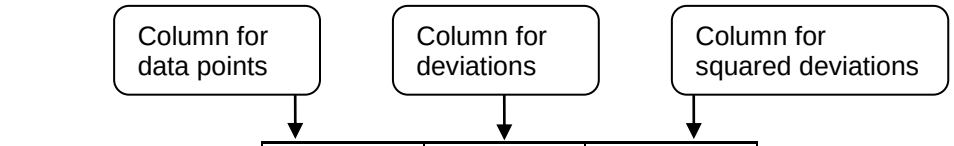
Since the advent of computer programs, it has become rare to have to calculate a

standard deviation or variance by hand. However, calculations by hand help you to understand and engage with statistics at a deeper level as they help you to see how the parts of equations are built. Also, understanding how computer programs obtain their results can help when you are interpreting data.

Example: Calculate the variance and standard deviation of the following sample of the test results of 10 students taken from a population of 100 students.

23 56 45 65 59 55 62 54 85 25

The basis of both these calculations follow a method, first calculate the mean, then the deviations, then the squared deviations. Also you need to sum the data points to find the mean and the square deviations to find the variance. All of these procedures can be facilitated by constructing a well-designed table:



x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
23	-29.9	894.01
56	3.1	9.61
45	-7.9	62.41
65	12.1	146.41
59	6.1	37.21
55	2.1	4.41
62	9.1	82.81
54	1.1	1.21
85	32.1	1030.41
25	-27.9	778.41
Σ	**529**	**0**	**3046.90**

Row for sums of columns

This table is designed to accommodate the calculation of both variance and standard deviation. It has columns which facilitate a stepwise calculation of the squared deviation and a row in which the sums of the columns can be inputted.

Method for calculation of variance and standard deviation:

1. Use the sum of the first column to calculate the mean:

$$\bar{x} = \frac{\sum_i x_i}{n} = \frac{529}{10} = 52.9 \text{ marks}$$

2. Calculate the deviations in the second column of the table. Each entry in the second column is calculated by subtracting the mean 52.9 marks from the relevant entry in column 1. For example the first entry is $23 - 52.9 = -29.9$ marks. Check you have done this properly by summing the deviations which should give a total of 0.

3. Square each of the results in column 2 to give the squared deviations. For example, the first entry is $(-29.9)^2 = 894.01$ squared marks.

4. Add the squared deviations to find their sum, which is part of the variance calculation.

   Sum of the squared deviations $= \sum_i (x_i - \bar{x})^2 = 3046.9$ squared marks

5. You now need to decide if you need to use the formula for sample variance or population variance. As the question states that this is a sample of marks from a population, you use the sample variance equation with $n = 10$:

   $$\text{variance} = \frac{\sum_i (x_i - \bar{x})^2}{n-1} = \frac{3046.9}{9} = 338.45 \text{ squared marks}$$

   This value can be used as a reliable estimate of the population variance.

6. The standard deviation is the square root of the variance and so:

   $s = \sqrt{\text{variance}} = \sqrt{338.54} = 18.40$ marks

   You can see that the units of *s* are "marks" which allows you to directly relate the standard deviation to the mean. So you could say that, on average, marks for this sample are 18.4 marks away from the mean of 52.9.

Want to know more?

If you have any further questions about this topic you can make an appointment to see a **Learning Enhancement Tutor** in the **Student Support Service**, as well as speaking to your lecturer or adviser.

Call: 01603 592761
Ask: ask.let@uea.ac.uk
Click: https://portal.uea.ac.uk/student-support-service/learning-enhancement

There are many other resources to help you with your studies on our website.
For this topic there is a webcast.

Your comments or suggestions about our resources are very welcome.

	Scan the QR-code with a smartphone app for a webcast of this study guide.	