Standard Deviation

Introduction

This page looks at the standard deviation: a lot of people, students and practitioners alike, are scared of them; but by the end of this page we are confident that ***you*** will be a master of them!

The standard deviation is a scary name for what is really only the average of the average … honestly, it's really not much more than that. More than that, if we take the calculation of the standard deviation step by step, we'll see that *anyone* can do it. No, really.

The normal distribution on the other hand is just a fact of life: our world is filled with examples that fit into the normal distribution: the heights of men and women, blood pressures in healthy people, the lengths of string on reels of string that comes out of a string factory.

The standard deviation is not necessarily of any use if all we do is calculate it: however, we need to know what it is and how to calculate it when we look at the normal distribution. This page, therefore, takes us through the standard deviation first: what it is, how to calculate it and what it means. Once we've understood the standard deviation, we will look at the normal distribution of data. Finally, we'll put both the standard deviation together as we work on a variety of examples and situations.

A subsequent page deals with the means, standard deviations and coefficients of variation of grouped data: see the link in the menu on the left of this page.

Averages

Perhaps the most common thing we do to data once we have collected it and put it into a table or spreadsheet is to work out the average value of it. We know that we can calculate the arithmetic mean, the median and the mode as different versions of what we can call an average. Once we've calculated an average figure, we can talk about average earnings, average marks in an exam and average height of sixteen year olds. Nothing wrong with that: we all do that and it's fine.

The average does have a problem, though, which is that if there are some exceptionally small or large values in our data set, they can have an undue influence on the values in the middle of our data set. For example, look at the averages of the data in this table:

Value	**Set 1**	**Set 2**
1	92	56
2	50	42
3	49	51
4	44	50
5	48	54
Average	**56.6**	**50.6**

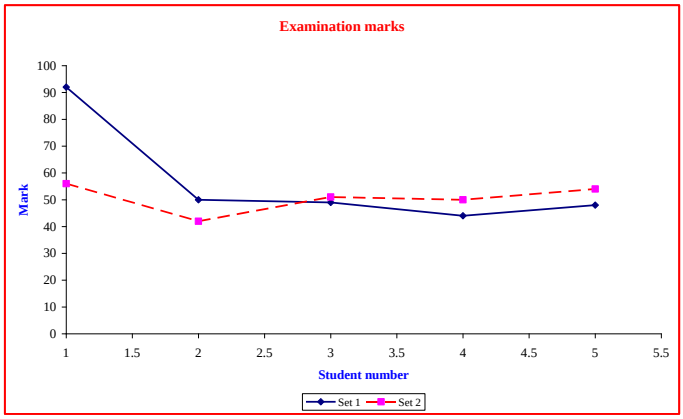
See, value 1 in set 1 is exceptionally large and it has dragged the average (arithmetic

mean in this case) up to 56.6 for set 1 whereas set 2 has no exceptional value (we call these exceptional values outliers, by the way) and so the average is much smaller and much more representative of all of the values in the data set. However, apart from the outlier, set 1 has poorer results than set 2

> *A major disadvantage of the mean is that it is sensitive to outlying points*
> http://bmj.com/collections/statsbk/2.shtml

Imagine that the table above shows the exam results for two groups of students: on the basis of the averages, set 1 would be classed as the better of the two groups, right? Still, it's only because of the performance of one very talented student that their average is so much better than the other set.

Graphically, we can see the influence of this outlier on set 1 and we can also see how stable set 2's data are:



The formula for calculation the average, in this case the arithmetic mean was, is and always will be

$$\text{arithmetic mean} = \bar{x} = \frac{\sum x}{n}$$

where

$\bar{x}$ is the standard way of representing the arithmetic mean: pronounced x bar
∑ is the Greek capital letter Sigma that means total of or sum of
x is the values of the variable 'x'
n is the number of values or data points

For set 1, we use the arithmetic mean formula as follows:

$$\bar{x}_1 = \frac{\sum 92 + 50 + 49 + 44 + 48}{5} = \frac{283}{5} = 56.6$$

The formula works with set 2 works in the same way … the average for set 2, remember, is 50.6

There is a way to sort out the problem where we know the average of a data set but we don't yet know some of the other characteristics of that set: whether all of the values are near to the average, some are near to the average, most are near but one or two are far away or outliers … it's the standard deviation.

The standard deviation

> *The standard deviation is a statistic that tells you how tightly all the various examples are clustered around the mean in a set of data. When the examples are pretty tightly bunched together… the standard deviation is small. When the examples are spread apart … that tells you have a relatively large standard deviation.*
> http://www.robertniles.com/stats/stdev.shtml

The standard deviation is really the average of the average: more than that, though, it tells us instantly whether we are dealing with data like set 1 or set 2: to prove that before we do any calculations, the standard deviation for set 1 is 19.92, for set 2 it is 5.37; because the standard deviation for set 1 is so large, it tells us that set 1's data are more spread out than set 2's data. And that's true!

The standard deviation actually comes in two parts:

- The variance
- The standard deviation

The variance is a measure of dispersion of values based on their deviation from the mean. In general, the formula for the variance is

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

where

$\bar{x}$ is the standard way of representing the arithmetic mean: pronounced x bar
∑ is the Greek capital letter Sigma that mean total of or sum of
x is the values of the variable 'x'
n is the number of values or data points

This formula calculates the variance for sample data, or where the number of values is less than 30. If we are working with population data, or more than 30 values, the formula would change and become

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

In fact, if n is bigger than 30, the two formulae tend to give the same, or almost the same, values.

The Variance for Sets 1 and 2

We'll work out the variance for set 1 and you can work out the variance for set 2:

Value	Set 1	$(x - \bar{x})$	$(x-\bar{x})^2$	Set 2	$(x - \bar{x})$	$(x-\bar{x})^2$
1	92	35.40	1253.16	56		
2	50	-6.60	43.56	42		
3	49	-7.60	57.76	51		
4	44	-12.60	158.76	50		
5	48	-8.60	73.96	54		
Total	**283.00**	**0.00**	**1587.20**	**253.00**		
Average	**56.60**	**0.00**				
Variance			**396.80**			

Notice we have to calculate $(x-\bar{x})^2$ in order to work out the variance because if we used $(x-\bar{x})$ the answer would automatically be zero because it will always be the case that the difference will add up to 0 … you can see this in the table above where the total of the $(x-\bar{x})$ column is 0 and therefore the average of that column must also be 0: try this and prove it with all other data sets you come across: the differences will always add up to 0. When we square them, the minus signs disappear and we have some values that we can work with.

Confirmation: $\text{Variance}_{\text{set1}} = \frac{1587.2}{5-1} = 396.80$

The value of the variance for set 2 is 28.80, did you get that? If you disagree, check your workings carefully and correct your mistakes.

What does the Variance Mean?

We now know that the variances for the two data sets are 396.80 and 28.80 respectively: set 1's variance is almost 14 times bigger than set 2's variance. That tells us that set 1's data is wider spread, more dispersed, than set 2's data. That's all it means: nothing complicated!

However, the variance is really just a number that isn't really real! Look at the calculation and see that the variance is the average of the differences between each value and the average SQUARED: for set 1, the variance of 396.80 is really the average of the differences SQUARED. This is where the standard deviation comes in … yep, the standard deviation is actually helpful!

Definition of the standard deviation

The simple definition of the **standard deviation is that it is the square root of the variance**: that's true but since we might want to use the standard deviation alone, without the variance, it's best to think of it as the average of the average.

The formula for the standard deviation looks as dreadful as the formula for the variance but by working through some examples we'll turn it into mincemeat!!

$$\text{Standard Deviation}_{\text{set1}} = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\text{variance}} = \sqrt{396.80} = 19.92$$

That's it! The value of 19.92 is much more meaningful than 396.80 because it is now equivalent to the raw data … it is 19.92 marks and not 396.80 square marks.

Your Turn: calculate the standard deviation for set 2 now

Did you get 5.37? You should have: check your workings if you didn't.

Comparing the standard deviations now, we can see that set 1's standard deviation is 19.92 and set 2's is 5.37: 19.92 is a bit less than four times bigger than set 2's standard deviation. So, we know that the set 1 data set is more variable or spread or disperse than set 2.

Please note: there is no single, absolute, value of the SD that tells us whether data is dispersed: all we can do is compare SDs from different data sets, combine an SD with the normal distribution and so on and then we can begin to determine dispersion.

Your Turn

Calculate

- the variance
- the standard deviation

for each of set 3 and set 4 from the table below; and say which set of data is the more dispersed.

Value	**Set 3**	**Set 4**
1	6	3
2	8	1
3	3	12
4	5	2
5	4	8
Total	**26**	**26**
Average	**5.2**	**5.2**

Did you get these? The answers you are looking for are

	Set 3	**Set 4**
Total	26.00	26.00
Average	5.20	5.20
Variance	3.70	21.70
Standard deviation	1.92	4.66

Set 4 is the more disperse of the two sets of data because both its variance and standard deviations are higher than those for set 3.

Basic Rule of the Variance and the standard deviation

If the Variance of data set A is greater than the Variance of data set B, then the standard deviation of data set A will be larger than the standard deviation of data set B

Want to make this look fancy to impress your friends?

if $Var_A > Var_B$ then $SD_A > SD_B$; and the converse is true

or even

if $Var_A > Var_B$ then $\sigma_A > \sigma_B$; and the converse is true

Variance and standard deviation Template

Here is a template that you can use whenever you want to calculate some variances and/or standard deviations. Just fill in your raw data and then do the arithmetic … a spreadsheet is brilliant at doing this kind of work for you!

			Difference		**Difference Squared**
	Data		$(x-\bar{x})$		$(x-\bar{x})^2$
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
Total		$\sum(x-\bar{x})$		$\sum(x-\bar{x})^2$	
Number of values		n		n	
Average		$\bar{x}$			
Variance				$\frac{\sum(x-\bar{x})^2}{n-1}$	
Standard deviation		σ		$\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$	