4. Variance

In addition to measuring the mean of a collection of numbers, it is usually also necessary to measure how much *variability* there is in the numbers. Since the mean is the "average" of the numbers, we might first calculate how much each observation differs from the mean:

x_i	$x_i - \mu$	
2	2-18	-16
2	2-18	-16
2	2-18	-16
10	10-18	-8
17	17-18	-1
24	21-18	6
26	26-18	8
34	34-18	16
45	45-18	27

If you add up how much each observation varies from the mean, *you will get zero.* In fact, you will get zero *every time no matter what the original numbers*. Since

$$\text{mean} = \frac{\text{sum of observations}}{n}$$

it follows that

$$n \times \text{mean} = \text{sum of observations}.$$

Adding the positive terms and the negative terms together in the table then exactly cancel out!

x_i	$x_i - \mu$	
2	2-18	-16
2	2-18	-16
2	2-18	-16
10	10-18	-8
17	17-18	-1
24	21-18	6
26	26-18	8
34	34-18	16
45	45-18	27

Because this naive approach fails, it is standard to instead *square* the differences between the observations and the mean.

x_i	$x_i - \mu$		$(x_i - \mu)^2$
2	2-18	-16	256
2	2-18	-16	256
2	2-18	-16	256
10	10-18	-8	64
17	17-18	-1	1
24	21-18	6	36
26	26-18	8	64
34	34-18	16	256
45	45-18	27	729

This approach gives the following average:

$$\text{sum of the differences squared} = 1918$$
$$\text{average of the differences squared} = 213.1$$

This average is called the *mean square error* or the *variance*.

Remember that these data are on savings of retirees. Thus we have computed a "variability" of

$$213.1(\text{thousands of dollars})^2$$

When we squared the differences to do our computations, we also squared the labels (dollars). Since "dollars^2" has no intuitive meaning, we should probably now take the square root (so that the units return to dollars, the same as those used for the mean). The result is a number called the *standard deviation*:

$$\begin{aligned}\text{standard deviation} &= \sqrt{\text{variance}} \\ &= \sqrt{213.1} \\ &= 14.6\end{aligned}$$

The *standard deviation* for the July retirement data is thus \$14,600.

The steps we just went through found first the *average of the squared differences* or the **variance**. Then we took the square root of the variance to find the **standard deviation**.
For the mean, the calculations for the *sample mean* and the *population mean* are identical. *This is not true for the variance and the standard deviation: the calculation is different depending on whether we started with census data or sample data.*
The calculations we just did were for *census data.* To summarize these:

$$\text{population variance} = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

and

$$\text{population standard deviation} = \sigma = \sqrt{\sigma^2}$$

For sample populations, the calculation is almost the same. First, for sample data the symbol for the variance is

$$\text{variance} = s^2$$

and the symbol for the standard deviation is

$$\text{standard deviation} = s.$$

The formulae are:

$$\text{sample variance} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

and

$$\text{sample standard deviation} = s = \sqrt{s^2}$$

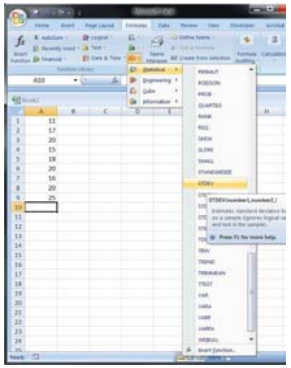
In practice, calculators and spreadsheets have the formulae for the mean, the population standard deviation and the sample standard deviation built in. Thus it is usually not necessary to use these formulae directly.

4.1. Example.

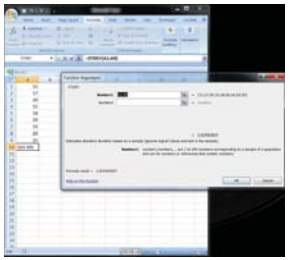
Using the built-in functions of Excel, find the mean, standard deviation and variance of the following sample.

11	15	16
17	18	20
20	20	25

Solution. **Step 1.** Enter the nine numbers into nine different cells on an Excel spreadsheet.
Step 2. Now position the cursor in a cell that does not include data–for example, the cell just below where you've entered your numbers.
Step 3. Select the *Formulas* tab at the top of the window. Then, click on the lower right icon in the function library. Select *Statistical Functions*, and a drop-down menu will appear. Select *StDev*.



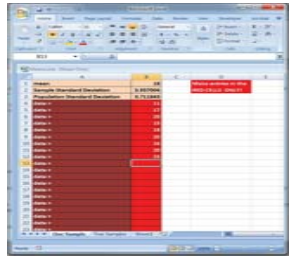
Step 4. When you select *StdDev*, a pop-up window appears asking you for the location of the cells that contain your data. Excel will pre-fill this with the non-zero cells adjacent to the location of the cursor. You can this by over-typing, or by selecting the appropriate cells with your mouse within the spreadsheet.



Step 5. Once you have the proper cells selected in the pop-up window, press enter or click on OK. The spreadsheet will then display the sample standard deviation. **Note:***If you wanted the population standard deviation, you would select the function StDevP.*

Excel reports that the sample standard deviation is 3.937.

Alternatively, you could use the spreadsheet MEANS.XLSX found in the online course resources. This spreadsheet is pre-configured to calculate the mean, sample standard deviation and population standard deviation of up to 60 data points.



4.2. Example.

Use Excel to find the mean and standard deviation of the following census data.

8	6	10	12
11	12	15	3

Interpretation of Standard Deviation.

The standard deviation is just a measure of how much the data deviate from the mean. In general the standard deviation has no intrinsic meaning beyond the concept of "mean square error." However, under "normal" conditions – we will define normal conditions shortly – there are some numerical inferences possible from the standard deviation.

For example

Approximately 68% of all observations will "normally" fall between
mean - one standard deviation
and
mean + one standard deviation

Similarly,

About 95% of all observations will "normally" fall between
mean - $2\times$standard deviation
and
mean + $2\times$ standard deviation

4.3. Example.

GRE scores are "normally distributed" with a mean of 500 and a standard deviation of 100. Thus approximately 68% of all GRE scores fall between

$$500 - 100 \quad \textit{and} \quad 500 + 100$$

or approximately 68% of all GRE scores fall between

$$400 \quad \textit{and} \quad 600$$

Similarly, approximately 95% of all GRE scores fall between

$$500 - 200 \quad \textit{and} \quad 500 + 200$$

or approximately 95% of all GRE scores fall between

$$300 \quad \textit{and} \quad 700$$

4.4. Example.

The scores from the Stanford-Benet IQ test are "normally distributed" with a mean of 100 and a standard deviation of 15. Approximately 68% of all IQ scores fall between

$$100 - 15 \quad \textit{and} \quad 100 + 15$$

or approximately 68% of all IQ scores fall between

$$85 \quad \textit{and} \quad 115$$

Similarly, approximately 95% of all IQ scores fall between

$$100 - 30 \quad \textit{and} \quad 100 + 30$$

or approximately 95% of all IQ scores fall between

$$70 \quad \textit{and} \quad 130$$