Standard Deviation
A Closer Look at the Video

After watching the *Standard Deviation* video, make sense of the mathematics by reading through the problem situation and solution. Use the comments and questions in bold to help you understand standard deviation.

Problem: Accurate predictions of the weather are important in all sorts of professions and in our lives. In the table below are the high temperatures last week from Elmendorf Air Force Base in Alaska, Thule Air Force Base in Greenland, and Osan Air Force Base in South Korea. We can see from the data that the temperatures were different at each base. Determine the standard deviation of the temperatures at each Air Force base for the week to compare how much the data in each set varies.

	S	M	T	W	T	F	S
Elmendorf AFB:	0	1	3	5	8	8	10
Thule AFB:	4	5	5	5	5	5	6
Osan AFB:	-18	-17	-1	5	10	25	31

What are the mean and median temperatures for each of the three Air Force bases?

The mean temperature for Elmendorf is $\frac{(0 + 1 + 3 + 5 + 8 + 8 + 10)}{7} = 5$ degrees. A similar calculation shows that last week Thule and Osan had the same mean temperature as Elmendorf.

The median is the middle number when the data is arranged in order from smallest to largest, so for Elmendorf, the median is 5. The median is also 5 for both Thule and Osan.

Even though the three Air Force bases have the same mean and median, notice that they are experiencing very different temperatures. The mean and median alone do not tell the whole story. We need to have a measure of how much the data varies.

What is one measure of variation?

One common measure of variation is the standard deviation.

What is standard deviation?

We can think of standard deviation as the spread of the data points; that is, the standard deviation is a number that tells us whether the data clusters closely about the mean or if there are many data points far away from the mean.

In our example, the mean is five degrees, and we need to express how far, on average, the individual temperatures are from five. Find the average of these distances from the mean.

Starting with the Elmendorf data, subtract the mean from each temperature. Then, find the average of those differences. Notice that the sum of the differences below the mean is negative eleven, and the sum of the differences above the mean is positive eleven. If we compute the average, we get zero. This isn't a coincidence. Look at the average of differences for Thule and Osan shown below. They are also zero.

Elmendorf

Temperature	Temp – Mean	Diff. from Mean	
0	0–5	–5.00	–11
1	1–5	–4.00	
3	3–5	–2.00	
5	5–5	0.00	
8	8–5	3.00	11
8	8–5	3.00	
10	10–5	5.00	
AVERAGE =		0	

Thule

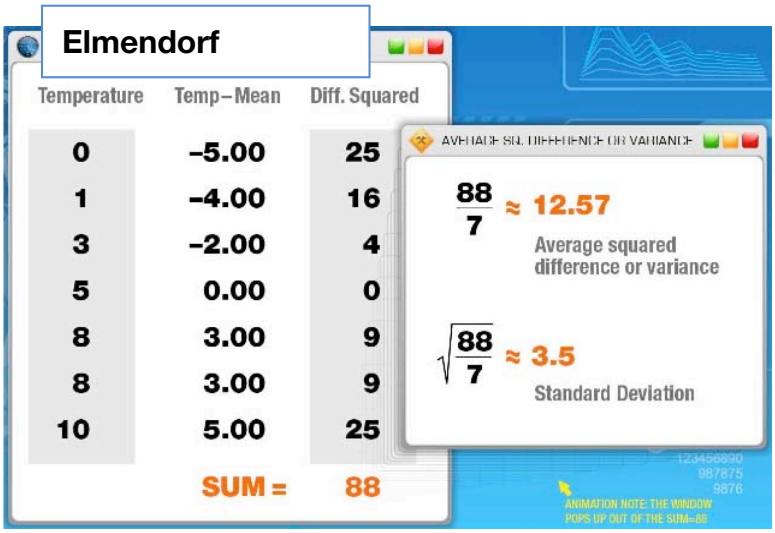
Temperature	Temp – Mean	Diff. from Mean	
4	4–5	-1	–1
5	5–5	0	
5	5–5	0	
5	5–5	0	
5	5–5	0	1
5	5–5	0	
6	6–5	1	
AVERAGE =		0	

Osan

Temperature	Temp – Mean	Diff. from Mean	
-18	-18–5	-23	–51
-17	-17–5	-22	
-1	-1–5	-6	
5	5–5	0	
10	10–5	5	51
25	25–5	20	
31	31–5	26	
AVERAGE =		0	

Finding the average of the distances between the data points and the mean in any set of data results in zero. We need a different approach to determine the spread of our data sets. If we square each difference, then we'll have a collection of positive values to average.

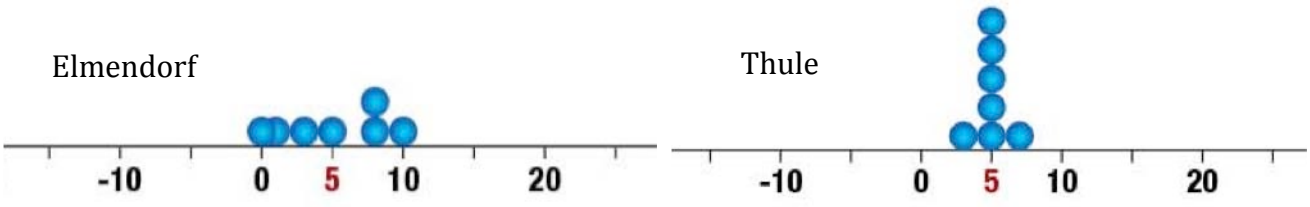
What do we get when we average the squared differences for Elmendorf?
The sum of the squared differences is 88 for Elmendorf. Since there were 7 total values, we divide 88 by 7 to find the average square difference. This rounds to 12.57. Since we squared the values, we take the square root of 12.57, which rounds to 3.5. This value is our standard deviation.



Why do we need to take the square root of the average squared differences to find the standard deviation?
Remember we're working with degrees, but when we square the differences, our units became squared degrees. When we find the average, we still have squared degrees. By taking the square root, our standard deviation goes back to degrees.

The standard deviation provides us a single value that indicates how much the data in the set varies from the mean.

Do the Thule Air Force Base temperatures have a smaller or larger standard deviation than Elmendorf?
Just by looking at the data, we can tell the daily temperatures are closer to the mean. That indicates it would have a smaller deviation; the data in the set varies less.



MakingSense

Let's go through the steps of finding the standard deviation, but this time using mathematical notation. Here is a quick look at the symbols we will be using.

x_i	**is each piece of data (x_1, x_2, …, x_n)**
μ	**is the mean of the data**
n	**is the total number of pieces of data**
$\sum_{i=1}^{n}$	**is the sum of the expression from the first piece of data to the last piece of data**

To compute the standard deviation:

- Find the mean (in our example the mean=5).
- Find each value's difference from the mean.
- Square the differences.
- Find the sum of the squared differences (in our example this sum is 2).
- Divide by the number of pieces of data. This is $\frac{2}{7}$ or approximately 0.286.
- Take the square root of the result. This is approximately 0.535.

Temperature	Temp-mean $x_i-\mu$	Diff. Squared $(x_i-\mu)^2$
4	−1	1
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
6	1	1
	SUM =	2

$\mu = 5$

$x_i - \mu$

$(x_i - \mu)^2$

$\sum_{i=1}^{n}(x_i-\mu)^2 = 2$

$\frac{2}{7} \approx .286$

$\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} \approx .286$

$\sqrt{\frac{2}{7}} \approx .535$

$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}} \approx .535$

Just as we suspected, the standard deviation for Thule is quite a bit smaller than the standard deviation for Elmendorf.

How does the standard deviation of Osan compare to those of Elmendorf and Thule?
Since Osan has the data that varies the most, Osan should have the largest standard deviation. Calculating, as we did for other data sets, we get about 17.5 for the standard deviation, so Osan does have the largest standard deviation.

	S	M	T	W	T	F	S
Elmendorf AFB:	0	1	3	5	8	8	10
Mean: 5	Median: 5				Standard Deviation: 3.5		
Thule AFB:	4	5	5	5	5	5	6
Mean: 5	Median: 5				Standard Deviation: 0.535		
Osan AFB:	-18	-17	-1	5	10	25	31
Mean: 5	Median: 5				Standard Deviation: 17.05		

In our temperature example, we used all of the data. Most of the time, statisticians calculate the standard deviation for a sample of the data instead of all of the data.

What is the entire set of data called?
The entire set of data is called the population.

When a sample is used, dividing by n could under-estimate the true standard deviation. What do statisticians divide by, to compensate for this when using a sample of the population?
Statisticians divide by n - 1 rather than n when calculating the standard deviation for a sample of the population.

Many calculators and computer software programs calculate both the standard deviation for a population and the standard deviation for a sample of the population.

POPULATION

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

SAMPLE

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}}$$