2.7 Measures of the Spread of Data

Example 1:

Consider the following test score data from 3 small classes of 5 students each. What issues might each set mean for the instructor? Assume each score is out of 100 points possible.

Class 1	Class 2	Class 3
67, 68, 70, 72, 73	40, 60, 70, 80, 100	40, 42, 70, 98, 100

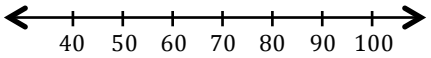
Find the mean of each set:

Class 1	Class 2	Class 3
$\mu =$	$\mu =$	$\mu =$

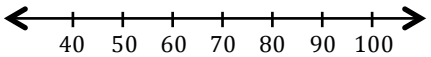
Notice that their means are equal, but the scores tell a different story.

Plot the scores on the number lines at right to see visually how spread out the values are for each set:

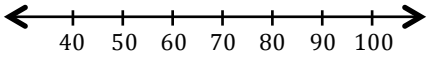
Class 1



Class 2



Class 3



a) Which of the three sets of scores is the least spread out?

b) Between Class 2 and Class 3, which is the least spread out?

The *range* of a data set is a measure of variation determined by the difference between the highest and lowest values of the set.

Find the range of each set:

Class 1	Class 2	Class 3
Range:	Range:	Range:

Finding the range of the classes has answered question (*a*), but not question (*b*).

The range doesn't tell us anything about the data values in between the extremes. We would like to know the "average" spread, or distance, from each data value to the mean.

Variance and Standard Deviation

- The *variance* is the average of the squares of the distances between each data value and the mean.
- The *standard deviation* is the square root of the variance.

	Formulas for Variance and Standard Deviation	
	Variance	**Standard Deviation**
Population	$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$	$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$
Sample	$s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$	$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$

Use the following tables to organize the work steps used to find the variance for each of the three classes above.

X	$X-\mu$	$(X-\mu)^2$
67		
68		
70		
72		
73		
Total:		
$\sigma^2 =$		

X	$X-\mu$	$(X-\mu)^2$
40		
60		
70		
80		
100		
Total:		
$\sigma^2 =$		

X	$X-\mu$	$(X-\mu)^2$
40		
42		
70		
98		
100		
Total:		
$\sigma^2 =$		

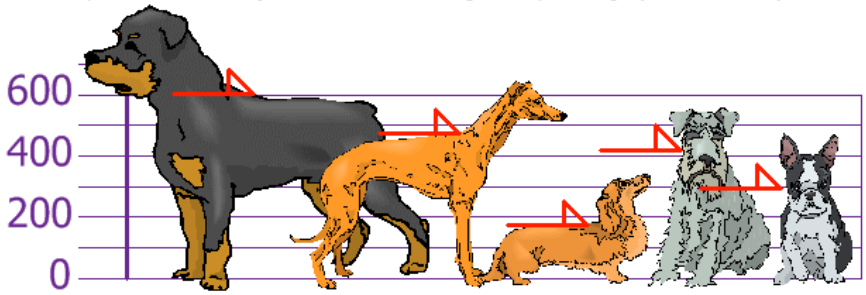
Notice that the ***variance*** of the data of Class 1 is very small compared to the others. Also the variances of Class 2 and Class 3 indicate that the data values of Class 3 are spread farther from the mean than those of Class 2.

The ***standard deviation*** gives the clearest picture of the distances of the data values from the mean.

Class 1	**Class 2**	**Class 3**
Variance: $\sigma^2 = \frac{(67-70)^2+(68-70)^2+(70-70)^2+(72-70)^2+(73-70)^2}{5} = 5.2$	Variance: $\sigma^2 = \frac{(40-70)^2+(60-70)^2+(70-70)^2+(80-70)^2+(100-70)^2}{5} = 400$	Variance: $\sigma^2 = \frac{(40-70)^2+(42-70)^2+(70-70)^2+(98-70)^2+(100-70)^2}{5} = 673.6$
Standard deviation: $\sigma = \sqrt{5.2} = 2.28$	Standard deviation: $\sigma = \sqrt{400} = 20$	Standard deviation: $\sigma = \sqrt{673.6} = 25.95$

Example 3: Let's look at a charming example from an online source, www.mathisfun.com.

You and your friends have just measured the heights of your dogs (in millimeters):



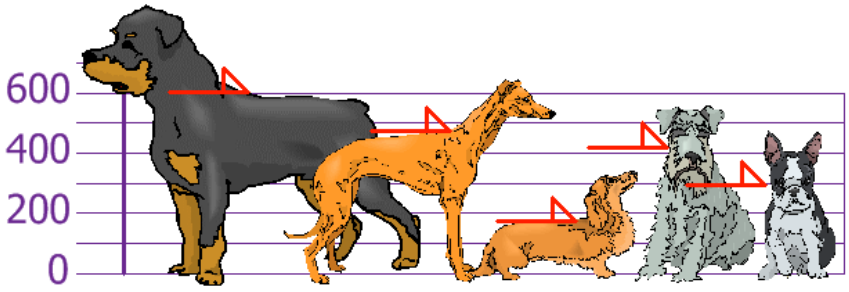
The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

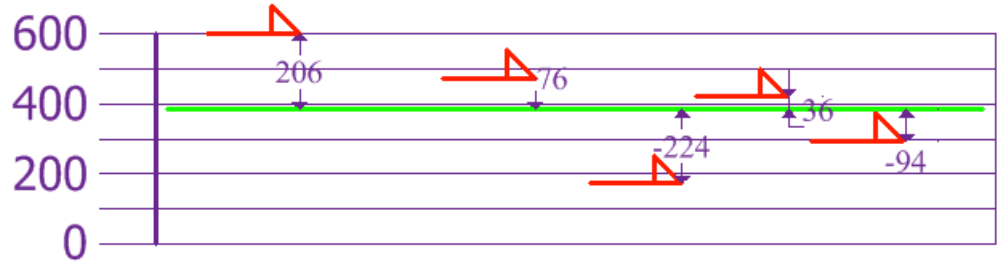
Note that 600 mm = 24 inches, 400 mm = 16 inches, and 200 mm = 8

	X	$(X-\mu)^2$
Rottweiler	600	
Whippet	470	
Dachsund	170	
Schnauser	430	
Bulldog	300	
	$\Sigma X =$	$\Sigma(X-\mu)^2 =$
	$\mu = \frac{\Sigma X}{N}$	$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} =$
		$\sigma =$

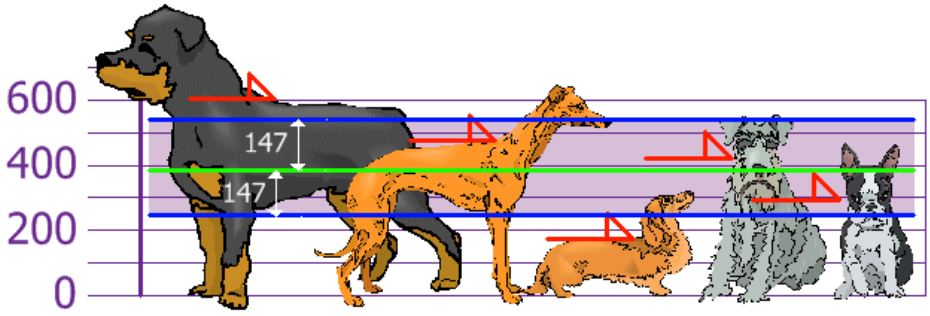
Draw a line through the mean value of the heights.



Notice the differences between the each height and the mean:



And the good thing about the Standard Deviation is that it is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:



So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.

Rottweilers **are** tall dogs. And Dachshunds **are** a bit short ... but don't tell them!

Example 4: John F. Kennedy's IQ is reported to have been **158**. The mean of all human IQ scores is **100** with a standard deviation is **15**. How significant is JFK's IQ of 158? Was he a little above average, way above average, or extremely gifted?

We can answer this question by considering how many standard deviations his IQ score falls above the mean.

A *standard score* or *z score* is obtained by dividing the difference between the data value and the mean by the standard deviation.

We use the formula

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

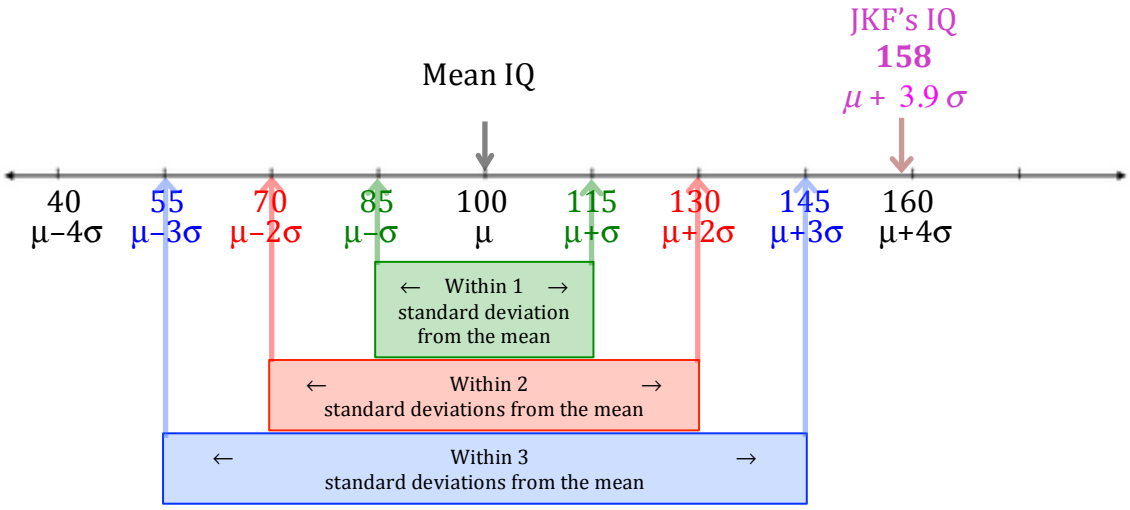
z-score	
Population	*Sample*
$z = \dfrac{X - \mu}{\sigma}$	$z = \dfrac{X - \bar{X}}{s}$

JFK's IQ z-score is $\dfrac{158 - 100}{15} = \dfrac{58}{15} \approx 3.9$.

Thus, his IQ score falls about 3.9 standard deviations above the mean of 100.

Paftnuty Chebychev, a Russian mathematician ascertained that at least 94% of all data lie within 4 standard deviations of the mean. From this we know that JFK's IQ score is higher than at least 97% of all other IQ scores, considerably above average. We will find in our study of normally distributed data in chapter 6 that his score is actually even much higher than the 97th percentile!

We call this the *measure of position* of JKF's IQ score relative to the world population.



Footnote:

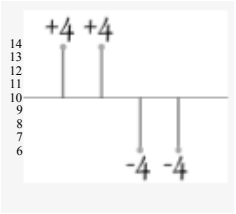
Question #1: When calculating the variance, $\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$,why do we square the difference between each data value and the mean?

Since the formula for the mean of a data set is $\mu = \frac{\sum x}{N}$,

it would seem reasonable that the formula for the spread would be $\frac{\sum(x-\mu)}{N}$.

However, if we add the differences from the mean … the sum will always be 0, as shown below:

Data set:
{6, 6, 14, 14}

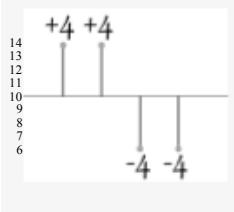


$$\frac{4+4-4-4}{4} = 0$$

So that won't work.

Then how about defining the formula for the spread to be $\frac{\sum|x-\mu|}{N}$?

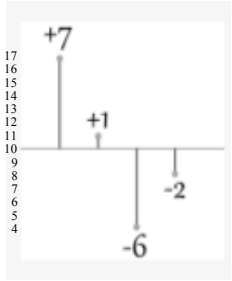
Data set:
{6, 6, 14, 14}



$$\frac{|4|+|4|+|-4|+|-4|}{4} = \frac{4+4+4+4}{4} = 4$$

That looks good, but what about the following case?

Data set:
{4, 8, 11, 17}

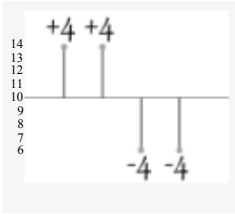


$$\frac{|7|+|1|+|-6|+|-2|}{4} = \frac{7+1+6+2}{4} = 4$$

Hmm… This calculation has a value of 4 even though the values are more spread out than the data set above it.

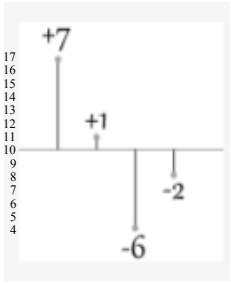
Let's use the actual formula for standard deviation, $\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$.

Data set:
{6, 6, 14, 14}



$$\sqrt{\frac{4^2+4^2+(-4)^2+(-4)^2}{4}} = \sqrt{\frac{64}{4}} = 4$$

Data set:
{4, 8, 11, 17}



$$\sqrt{\frac{7^2+1^2+(-6)^2+(-2)^2}{4}} = \sqrt{\frac{90}{4}} = 4.74\ldots$$

Notice the standard deviation is bigger when the data values are more spread out. Perfect!

Question #2: Why do we divide by $n - 1$ when calculating the standard deviation for the sample instead of dividing by n?

Watch the video at Khan Academy, linked below if you want to know more about why the sample standard deviation is calculated differently than the population standard deviation at

https://www.khanacademy.org/math/ap-statistics/quantitative-data-ap/measuring-spread-quantitative/v/review-and-intuition-why-we-divide-by-n-1-for-the-unbiased-sample-variance