How spread out is the data? Are all the numbers fairly close to each other or not? So what?

General Education Statistics
Class Notes
Measures of Dispersion: Range, Standard Deviation, and Variance (Section 3.2)

Once we know the mean of a set of data, we might be interested in knowing how close the actual values are to that mean. Are the values spread out or close together and all gathered around the mean? We have a few ways to describe what we will call **dispersion**.

All of these measures require that the data be quantitative.

The simplest (and quickest) is the range.

Definition: The **range, *R*,** of a variable is the *difference* between the largest data value and the smallest data value. That is,

Range = R = Largest Data Value − Smallest Data Value

expl 1: Let's try this out. The following data represent the travel times to work (in minutes) for all seven employees of a start-up web development company. Find the range of these numbers.

23, 36, 23, 18, 5, 26, 43

Another very useful measure is the standard deviation. Its definition below is a little daunting but the standard deviation can be thought of as the average distance each value is from the mean.

Definition: The **population standard deviation** of a variable is the square root of the sum of squared deviations about the population mean divided by the number of observations in the population, N. That is, it is the square root of the mean of the squared deviations about the population mean.

The population standard deviation is symbolically represented by σ (lowercase Greek sigma).

Wow, that's a mouthful. Here is the formula.

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N}}$$

$$= \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}$$

We take each value and find its distance to the mean. We then square those distances, add them up, and divide by N.

That gives us **variance** which we will see later. Square root it and we get the **standard deviation**.

where $x_1, x_2, \ldots, x_N$ are the N observations in the populati
and μ is the population mean.

That is well and good, but you may also see an equivalent (computational) formula for the **population standard deviation** that is sometimes used. It follows.

$$\sigma = \sqrt{\frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{N}}{N}}$$

We square each value and add those up. We separately add all of the values and square that, and then divide by N. Subtract those two results and divide by N again. Finally, square root. Wow!

Now, the above standard deviation concerned data gotten from an entire population. However, often we have sample data. Here, we see *similar but slightly different formulas* for the sample standard deviation.

Definition: The **sample standard deviation**, **s**, of a variable is the square root of the sum of squared deviations about the sample mean divided by $n - 1$, where n is the sample size.

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

$$= \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$$

where $x_1, x_2, \ldots, x_n$ are the n observations in the sample and $\bar{x}$ is the sample mean.

We do almost the same as if it was population data, *but we divide by one less than the sample size.*

Notice how we use s to denote the standard deviation for a sample and σ for that of a population.

The computational formula follows.

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

Is it resistant?:

Since the range is found by subtracting the minimum value from the maximum value, it is affected by extreme values. So, we say the range is *not* resistant.

The standard deviation uses all of the values in its calculation. Therefore, it is also affected by extreme values, so it is *not* considered resistant either.

expl 2: Complete the table to find the standard deviation of this sample data set.

Data Set	$(x_i - \bar{x})^2$
12	
16	
18	
20	
24	
26	
	total = $\sum(x_i - \bar{x})^2 \approx$
	$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \approx$
	$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \approx$

The mean is 19.3333.

Notice how this squared difference gets larger as the value gets farther from the mean.

What is $n - 1$?

The formula is $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$

Doing this on the calculator is the same as the process we used for finding the mean and median.

Enter the data values in column **L1** in the **STAT** editor. We do this by pressing the **STAT** button and then selecting **EDIT > 1: Edit…** from the menu. If necessary, clear out any data in **L1** by arrowing up to the column heading and pressing **CLEAR**. When you arrow back down, any data should be gone. Enter the values of the salaries in **L1,** pressing **ENTER** after each one.

Then press the **STAT** button again. But this time, arrow over to select **CALC > 1: 1-Var Stats**. That will put this expression on the home screen. Press **ENTER** and the calculator will fill with many statistics. (Some newer calculators will have an intermediate screen, where you need to select **L1** for **List** and clear out any entry in the **FreqList:** row. Arrow down and select **Calculate**.)

Look for *Sx* and record it here. (You will also see *σx* which is the population standard deviation. Again, the calculator does *not* know if the data is from a population or sample. You must decide which standard deviation to record.)

Instructions for STATCRUNCH:
Within MSL problems, you will see a little icon that looks like overlapping rectangles next to the data. Click on it and select "Open in StatCrunch". This will open StatCrunch and import the data. Select **Stats** > **Summary Stats** > **Columns**. You will need to tell it where the data is ("Select column(s)" at top). By default, it will calculate lots of stuff including *sample* standard deviation and variance, mean and median, and range. You can select more to display under "Statistics". If you know the data is from a *population*, "unadjusted" (abbreviated "unadj.") variance and standard deviation (abbreviated "std. dev.") is what you need.

Definition: The **variance** of a variable is the square of the standard deviation. The **population variance** is σ^2 and the **sample variance** is s^2.

Notice the table in the last example has us calculate the variance on the way to the standard deviation. The units of the variance are squared units (for instance, if the variable is in feet, the variance would be in square feet). That makes interpreting the variance a little more challenging. We will see the variance later in inferential statistics.

Worksheet: Comparison of two data sets with the same mean:
We will investigate two data sets with the same mean. One data set is more spread out than the other. How do you think their standard deviations will be related?

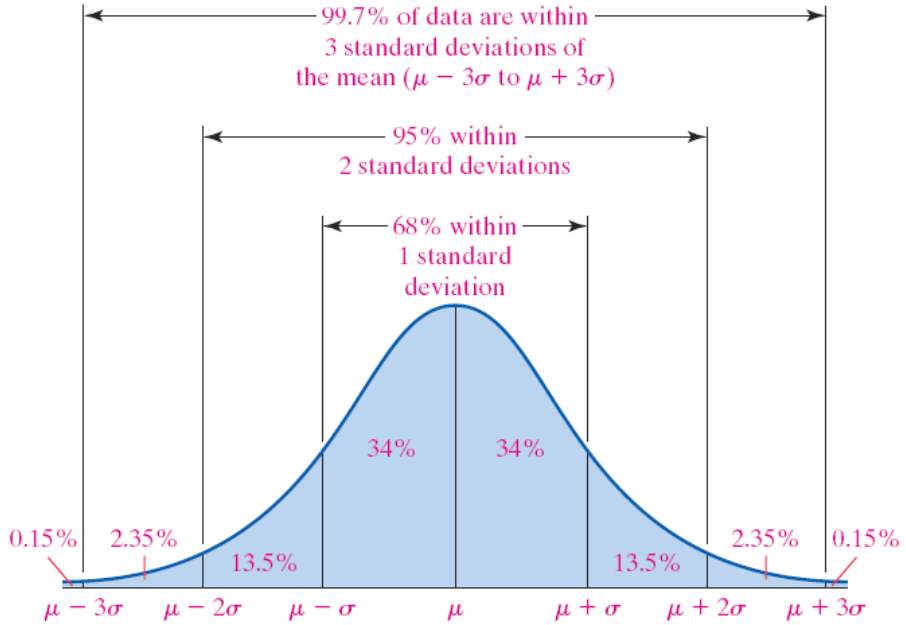
The Empirical Rule:

If a distribution is roughly bell shaped, then

- Approximately 68% of the data will lie within 1 standard deviation of the mean. That is, approximately 68% of the data lie between $\mu - 1\sigma$ and $\mu + 1\sigma$.
- Approximately 95% of the data will lie within 2 standard deviations of the mean. That is, approximately 95% of the data lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean. That is, approximately 99.7% of the data lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Note: We also use the Empirical Rule based on sample data with $\bar{x}$ and s used in place of μ and σ.

Here is a picture that illustrates the Empirical Rule.



Notice how μ is positioned in the exact middle, showing the center of the data. Let's see this in action to get a better understanding.

expl 3: The following data represent the serum HDL cholesterol of the 54 female patients of a family doctor.

41	48	43	38	35	37	44	44	44
62	75	77	58	82	39	85	55	54
67	69	69	70	65	72	74	74	74
60	60	60	61	62	63	64	64	64
54	54	55	56	56	56	57	58	59
45	47	47	48	48	50	52	52	53

a) Compute the *population* mean and standard deviation. Use a calculator.

b) Draw a histogram to verify the data is bell-shaped.

c.) Draw a quick sketch of a bell-shaped curve, labeling the mean in the middle. Then mark the various standard deviations (plus or minus 1, 2, and 3) from the mean using a reasonable scale. You must calculate these values and label them on the horizontal axis.

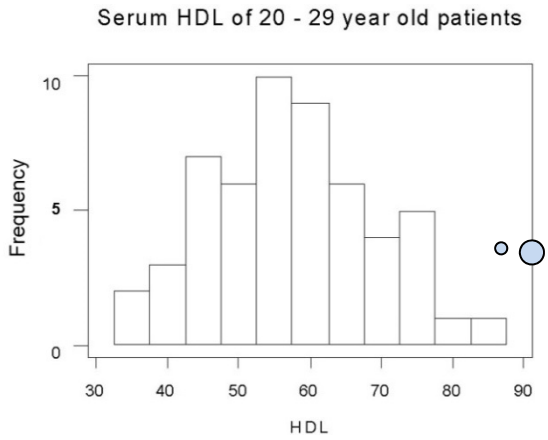
Go on to the next page.

We complete example 3 here.

a) Compute the *population* mean and standard deviation. Use a calculator.

If you want, you can verify this information. The calculator tells us that $\mu = 57.4$ and $\sigma = 11.7$.

b) Draw a histogram to verify the data is bell-shaped.



Parts *a* and *b* are done for us.

Does the histogram look roughly bell-shaped?

c.) Draw a quick sketch of a bell-shaped curve for this data, labeling the mean ($\mu = 57.4$) in the middle. Then mark the various standard deviations (plus or minus 1, 2, and 3) from the mean using a reasonable scale. Calculate these values and label them on the horizontal axis.

We will use this graph to answer several questions.

expl 3 additional questions:

d.) What is the percentage of all patients that have serum HDL within 1, 2, and 3 standard deviations of the mean according to the Empirical Rule?

within 1 standard deviation from mean: _____

within 2 standard deviations from mean: _____

within 3 standard deviations from mean: _____

e.) Determine the percentage of all patients that have serum HDL between 22.3 and 92.5 according to the Empirical Rule. Refer to the gaph you produced on the previous page and your answer to part *d*.

f.) Determine the percentage of all patients that have serum HDL between 45.7 and 69.1 according to the Empirical Rule. Refer to the gaph you produced on the previous page and your answer to part *d*.

g.) Determine the actual percentage of patients that have serum HDL between 45.7 and 69.1. Do this by looking at the *actual data*. Compare this to the answer in part *f.*

h.) Determine the percentage of all patients that have serum HDL between 34 and 69.1 according to the Empirical Rule. Refer to the gaph you produced on the previous page and your answer to part *d*. You will also use the fact that this bell-shaped graph is symmetrical.

Definition: We call $n - 1$ the **degrees of freedom** because the first $n - 1$ observations have freedom to be whatever value they wish, but the n^{th} value has no freedom. It must be whatever value forces the sum of the deviations about the mean to equal zero.

This does *not* play a large part now but we may see degrees of freedom later in inferential statistics.