

1.5.2 Measures of Spread

We've now learned about measures of central tendency and found the mean and median of our dataset. Now let's take our analysis a step further and help Louise measure the spread of the dataset. We'll be looking at the range of the dataset as well as adding more statistics to our analysis: standard deviation and variance.

Measures of central tendency distill a lot of information about the distribution of a dataset down to one number. However, two datasets can have the same mean or median but still look very different—that is, the spread of data between the two datasets can vary quite a bit. When considering the distribution of a dataset, we also want to have measures of its spread. **Measures of spread** include range, variance, standard deviation, and quartiles.

Range

The simplest measure of spread is the range of a dataset. The **range** is the difference between the maximum value of the dataset and the minimum value of the dataset. For our purposes, the range does not capture as

much information as we'd like. What we would really like to know is roughly how far each data point is from the center, or mean, or how much of the data is near the center.

Variance

Variance is a measure of how far data points are from the center, or mean. To calculate the variance, do the following:

1. Subtract the mean from each data point.
2. Square the difference so that it's positive.
3. Take the average of those squared differences.

Because we've taken the average of the *squared* differences, the unit of variance doesn't quite match our dataset. To get the unit to match, we take the square root of the variance to standardize it, or get the **standard deviation**. Standard deviation is often represented with a lowercase sigma (σ). You'll also see variance represented as the standard deviation squared (σ^2). Let's look at an example.

Imagine we have a dataset of five backers. We'll signify that this is a dataset by placing the data within brackets: [1, 3, 6, 7, 8], which then makes it into a "set" of numbers. How do we find the standard deviation? Let's begin working through the standard deviation equation:

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

- σ , lowercase sigma, is the symbol for standard deviation.
- \sum , uppercase sigma, is the symbol for summation or the sum.
- X , represents each point of data in the dataset.
- \bar{X} , represents the mean of the dataset and is pronounced "x-bar."

- n , represents the total number of points in the dataset.

This equation looks fairly complicated, so let's talk through what's happening.

1. Find the mean.
2. For each number in the dataset, subtract the mean and square the result.
3. Find the mean of these new numbers.
4. Take the square root of the mean.

Let's apply the equation to our set:

1. Find the mean: $(1 + 3 + 6 + 7 + 8) / 5$ (the sum the values divided by the number of values), or $25/5 = 5$.
2. Next, we find the deviations. The deviations from the *sample* mean are
 - $(1 - 5) = -4$
 - $(3 - 5) = -2$
 - $(6 - 5) = 1$
 - $(7 - 5) = 2$
 - $(8 - 5) = 3$

Ideally, we'd like to know the deviations from the actual population mean, but because we don't know the actual population mean, these deviations have a subtle and slight bias to them. We'll correct that bias in the next step.

3. Find the variance. First, square all of those deviations; this way, we will always be working with positive numbers. This results in 16, 4, 1, 4, and 9. If we take the average of these values, we'll get a slightly smaller variance than the actual population. To correct for this bias, we instead divide by the number of samples minus 1 (if you're

curious, this is known as Bessel's correction). Thus, the unbiased average of these values is $(16 + 4 + 1 + 4 + 9) / (5-1) = 34 / 4 = 8.5$.

4. The square root of 8.5 is about 2.92, which is the standard deviation.

Another method for measuring the spread is to calculate the interquartile range. After organizing the dataset from lowest value to highest value, we can break it into four separate parts known as quartiles.

Quartiles

Like medians, **quartiles** are percentiles. The lower quartile is the 25th percentile, that is, 25% of the data is less than the lower quartile. Similarly, the upper quartile is the 75th percentile, so 75% of the data is less than the upper quartile. You may also see these referred to as the 1st and 3rd quartiles. (The 2nd quartile is the median, so that one already has its own fancy name).

The difference between the upper and lower quartiles is known as the **interquartile range** (IQR). The IQR gives us a sense of how far out you can go from the mean to get 50% of the data.

IMPORTANT

Determining the interquartile range is done with the following formula:

$$\text{IQR} = Q3 - Q1$$

That is, the IQR is equal to the third quartile minus the first quartile. When used to describe a list of data, the IQR tells us how the data is spread around the median.

To learn more about IQR, read [this article that provides additional examples and explanations for the uses of IQR](https://www.thoughtco.com/what-is-the-interquartile-range-rule-3126244) (<https://www.thoughtco.com/what-is-the-interquartile-range-rule-3126244>).

To see these concepts in action, let's add measures of spread to our "Descriptive Statistics" worksheet. Add new rows for the standard deviation, upper and lower quartiles, and IQR.

Here's how your updated table should look:

	Successful	Failed
Mean Goal	\$5,049	\$10,554
Median Goal	\$3,000	\$5,000
Standard Deviation of Goal		
Upper Quartile of Goal		
Lower Quartile of Goal		
IQR of Goal		
Mean Pledged	\$5,602	\$559
Median Pledged	\$3,168	\$103
Standard Deviation of Pledged		
Upper Quartile of Pledged		
Lower Quartile of Pledged		
IQR of Pledged		

The function to calculate the standard deviation of a population in Excel is **STDEV.P**. (The other option is **STDEV.S**, which calculates the standard deviation based on a sample of the whole population. There's a subtle difference between these formulas (one is for the entire population of a dataset while the other is for a sample of the whole) that statisticians care about, but we're going to ignore it. Don't tell any of your statistician friends.

We'll be using the same range and worksheet data as we did with the AVERAGE formula, so the STDEV.P formula we enter into B4 is

```
=STDEV.P('Successful US Kickstarters'!D:D).
```

To calculate the upper and lower quartiles, use the **QUARTILE.EXC** function. QUARTILE.EXC takes two arguments: the first argument is the data array, and the second is the quartile to be calculated. For the upper quartile, put 3:

`=QUARTILE.EXC('Successful US Kickstarters'!D:D, 3)`

For the lower quartile, put 1:

`=QUARTILE.EXC('Successful US Kickstarters'!D:D, 1)`

The IQR cell will be the difference between the upper and lower quartiles, so the two cells would be subtracted. In B7, type `=B5-B6`.

SKILL DRILL

Using the same formulas, update them to complete the standard deviation and quartile calculations in column C.

Your table should now look like this:

1		Successful	Failed
2	Mean Goal	\$5,049	\$10,554
3	Median Goal	\$3,000	\$5,000
4	Standard Deviation of Goal	\$7,749	\$21,968
5	Upper Quartile of Goal	\$5,000	\$10,000
6	Lower Quartile of Goal	\$1,500	\$2,000
7	IQR of Goal	\$3,500	\$8,000
8			
9	Mean Pledged	\$5,602	\$559
10	Median Pledged	\$3,168	\$103
11	Standard Deviation of Pledged	\$8,335	\$1,331
12	Upper Quartile of Pledged	\$5,699	\$501
13	Lower Quartile of Pledged	\$1,717	\$9
14	IQR of Pledged	\$3,982	\$492

FINDING

Based on these statistics, we can determine the following:

- The mean of each distribution is around the 3rd quartile, so the data follows similar distributions in each subset.
- The standard deviations are larger than the mean, which means everything below the mean is considered "close" to the center.
- Some large values are driving all of these distributions. The standard deviations are all roughly twice the IQR in each distribution, except in the failed Kickstarters, where the standard deviation is closer to three times the IQR. There must be some failed Kickstarters with really high goals!

© 2020 - 2022 Trilogy Education Services, a 2U, Inc. brand. All Rights Reserved.