We went over measures of central tendency and we can start to see how this type of data is useful. We should understand now that from a data set we can gather even more information. Measures of central tendency help us understand “middleness”. We have the median, which is a robust measure of “middleness.” This means that the median isn’t as affected by outliers in comparison to our other measures of central tendency.

We get the median by ordering our dataset, finding the value at the midpoint of our dataset. If there are two values there, we then get the arithmetic mean of the two values.

We have the mode. This is the least used way of measuring central tendency, but it can be important. We calculate the mode by enumerating the amount of times a value appears within a dataset. The data value with the highest appearance rate becomes the mode. If there are multiple values tied for highest frequency, then it can be said that there are more than 1 modes. If a dataset has equal frequencies for all values, then the data set is said to have no mode.

Lastly, we have the arithmetic mean. This is generated by adding the total value of all data points, and then dividing by the total number of data points. This measure of central tendency is not considered robust because it can be heavily influenced by outliers, as we’ve seen.

It’s important to ask ourselves now, What else can we pull from our data? What type of important numbers can we take from our data sets?

Dispersion is the next step in data analysis. We want to know how “far” from the “middle” values tend to lie. Why is this important? Well it helps us understand our dataset’s “growth” or “decay” so to speak. We can start to put into our head what values “should” look like so to speak. We can also start to see “clusters” and “groups” in our data set. Values might be relatively “close” or “far” from the mean, and then we can start to ask ourselves, well why is that? When we talk about dispersion, we are talking about distance relative to the center. That means that when we are calculating dispersion, we are using our measurements of central tendency.

We’re going to start with the variance. Variance is a measure of spread that is built on top of the mean.

Before we dive in, let us touch once again on that idea of sample vs population.

The measures of a data set have different names when we reference the population or the sample. The reason being is that one is representative of the other. The sample mean and the same variance are in reference to the population mean and sample. One is a parameter and one is a statistic. The sample mean and variance are supposed to be representative of population mean and variance, and so we can say it’s an approximation. In that approximation, we lose some value. Let’s plot this. And then maybe we can see something interesting.