Questions To Engage during the discussion:

What does it mean to be at the center of something?

What is centerness?

Is there a clear and definitive definition of centrality?

Why does it help to define a middle?

How does spread play into this?

Why is spread an important thing to keep track of?

Explain what Pandas allows us to do, and explain the high level nature of it

Explain that the graphing library behind Pandas is MatPlotLib and that it’s a separate package from Pandas, and should be used as such. Personally I prefer to use it separately from Pandas so that I can keep track of code more clearly.

Let’s try a few ways to define the middle.

The “average” value.

This is known as the arithmetic mean. Let’s shy away from average. All of these values can be seen as the average value.

The middle value in terms of location. This is known as the median.

The most commonly appearing value. This is known as the mode. This is the most rarely used value, but it is helpful when we are using qualitative data, data that talks about the quality of something, rather than the numerical value of something.

Let’s start with how to calculate the mode. We simply calculate the value that appears the most. This one isn’t too complicated.

{1,2,3,3,4,5} The mode is 3. That is because that is the value that appears the most.

A data set could also be said to have more than one mode or no modes. A no mode dataset would be a dataset that has every object appear at the same frequency.

{1,2,3,4} This would be an example of a no mode data set.

{1,2,2,3,3,4} This would be an example of a 2 mode data set. The modes are 2 and 3, because they each appear twice.

Let’s now work to calculate the median. The median is said to be the value that appears at the “middle” of a data set. To calculate the median we have to find the midpoint.

{1,2,3,4} This set seems to have no median, but the median is special. The median can appear outside of the data set. If the set is even, then the media is calculated by adding the two middle most values, and then dividing them by two. In this set, the median is 2.5, because 2+3 = 5 and 5/2 = 2.5.

{1,2,3,4,5} The median of this data set is 3. There are two values to the left, and two values to the right of it. It’s the “middle”.

The median gives us a robust number. What does robustness mean in terms of statistics? It means that the value isn’t susceptible to outliers. We will more clearly define outliers later, but let’s see what happens when we add a value that is seemingly much different from the other values.

{1,2,3,4} = We said that the median of this data set was 2.5. Let’s see what happens when we add 100 to the data set.

{1,2,3,4,100} median = 3

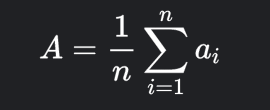
Yes, we can see that our median did indeed change, but did it change that much? No, it changed by a measly 0.5

That is what “robustness” means.

Lastly, we have the arithmetic mean. I know that we have become comfortable with saying the average, but in reality, any of these 3 measures of central tendency can be said to be the “average” of a data set, so saying the average isn’t helpful. The arithmetic mean serves us better. We say arithmetic mean, because there are a few different means.

Before we get into the next bit, let’s take some time to talk about Math notation. I don’t expect everyone to become masters at it, and I myself can’t read it particularly well, but a base level knowledge of what the symbols mean and what they represent is important.

Let’s look at the formula for the arithmetic mean in math notation.



Now this looks like nothing, but let’s define each term.

A stands for the arithmetic mean.

1\n is 1 over n, n represents the number of values in our dataset.

It’s MULTIPLYING by Sigma. What does that mean.

Sigma is a special Symbol, that means the sum of. So it’s the sum of ai.

Well what’s Ai? Well that represents every term in our data set. How does it do thgat?

Sigma has two things that guide us. I = 1 under sigma, and n over sigma.

That means that ai is going to take every value between a1, an. It’s going to be the sum of all the data points in our data set.

So now let’s combine that all together.

The arithmetic mean is equal to one over the number of points in our dataset, times the sum of all the datapoints in our data set.

That checks out with our “more normal” understanding of what the arithmetic mean is.

Add up all the values, and divide it by the number of values in our data set.

I know it might seem complicated to have written it out in math notation and to have explained it, but it only serve you. In the future, now when you see those terms, they’ll make more sense to you.

Alright, so let’s work through calculating the arithmetic mean in a few data sets.

{1,2,3,4} mean = 2.5

{1,2,3,4,5} mean = 3

Median is said to be robust, but the arithmetic mean is said to NOT be robust. The mean is pretty sensitive to outliers. Let’s try it out. Let’s add 100 to our first data set.

{1,2,3,4,100} mean = 22. This is a drastically different arithmetic mean. The median didn’t shift as much as the mean did when we did the same operation. This helps us evaluate different types of data sets.

There isn’t really a right or wrong answer to which measure of central tendency we should use. It’s up to the discretion of the person implementing the formula. It should be noted though that robustness helps us know whether or not a median or a mean is a best fit. Oftentimes it helps to calculate both, and then to measure them against each other.

~~Temperature Check ~~

Is everyone clear as to what these key values are and how to calculate them?

What needs clarity?

Now that we’ve defined middles, we can define spread. A measure of how far things tend to be from a middle value. We have multiple definitions of middle though, and so similarly we have multiple definitions of spread.

Spread, Variability, or Dispersion, tries to answer the question of where things more or less tend to be relative to the center.

Let’s start with using Median as our Center.

We can use the median to calculate the IQR, the Interquartile Range.

The Inner quartile range is built from the median as follows.

First we get the median of a data set.

{1,2,3,4,5,6,7,8,9,10} median = 5.5

Then we remove the median from the data set, and we are left with two halves. 50% of the data on the left of the median, and 50% of the data to the right.

Then we get the medians of those 2 data sets.

{1,2,3,4,5} median = 3

{6,7,8,9,10} median = 8

Q1 = 3

Q2 = 5.5

Q3 = 8

There are our quartiles.

To get the IQR, we just do Q3-Q1. In this dataset it is equal to 5. This IQR shows us where 50% of our data lives within our dataset. Now that we’ve started to define spread, we can come back to those pesky outliers. Now we can concretely define outliers. When using IQR, we can say that an outlier is a value that is outside of the maximum and minimum bounds of our data set.

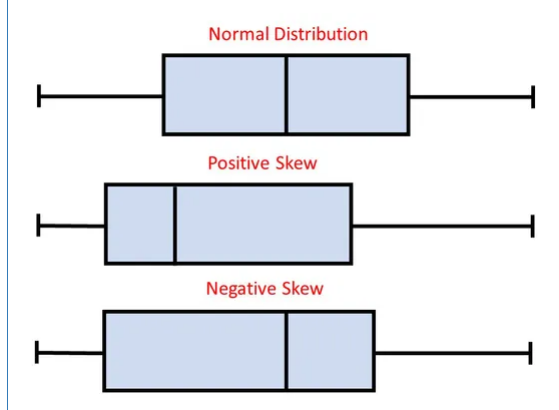
The minimum bound is Q1 -1.5\*IQR

The maximum bound is Q3 + 1.5\*IQR

Any values that are found outside of these bounds are called outliers. Are these somewhat arbitrary? Yes, but we at least have some mathematical way of defining a value as “too far”.

What does the IQR let us do? It let’s us make some box and whiskers plots.

Box and whisker plots can be useful because they help us visualize some aspects about our data. They can tell us whether or not a data set is skewed right, or skewed left.



Let’s try to think about what these three box plots are saying. If the median is to the left of the “middle” of the boxplot, that means that of the middle 50%, values are more dispersed to the right of the data set. This means that our data set is RIGHT skewed, or positively skewed.

Now, if the opposite is true, that means that data tends to be more dispersed to the LEFT of the dataset, this would mean that our data is LEFT skewed.

Skew let’s us get a visual as to where the data points tend to be clustered. It can be helpful to us to know this, because when we’re doing analysis, and training models, we know what our model will have more access to. Higher quality results will be against the skew.

Q3 - Q2 vs Q2 vs Q1, length of the whiskers.

Minimum lowest score excluding outliers, maximum score excluding outliers.

Larger ranges show larger spread of data.

~~~~~

The IQR is built out of the median, so it follows that the arithmetic mean will also build into something. That something is the variance(standard deviation).

Before we dive deep into it, we have to cover something called the sample, and the population.

The population is a data set that represents the entirety of something. Be it all people in the world, all cats in a zoo(in a particular zoo, not all zoos around the world), or all students in the Data Science cohort. They are not representative of anything, they simply are, and encapsulate some data about that population.

A sample however, IS representative. It’s representative of the population. It serves not as a data set meant to cover the entirety of something, but as a working example about a slice of a population. Things can be samples or populations in different contexts. Using that all cats in a zoo example, we can say that it is a population when we are referring to all cats in a specific zoo, and we have all data on them, or we can say that it is a sample, and it is a sample of the population of all cats in all zoos around the world.

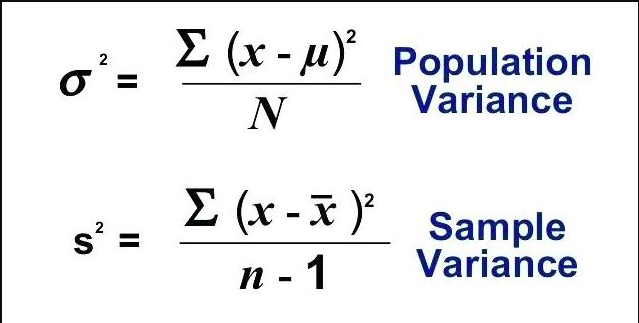
This may seem like an arbitrary distinction, but it is an important one. The idea of “representing” the population has a few implications. It means that the sample itself is only a segment of the entirety of some data set, and that for one reason or another, we cannot access all individuals in a data set.

This means that when we talk about values derived from a population, we refer to them as parameters. The Mean, Median, etc of a Population data set are parameters.

When we get these values from a sample, then we refer to these values as a statistic. They are meant to be representative of the population that we don’t have access to.

The implications are major.

It involves how we calculate the variance.



Let’s break things down and then talk about the reasoning.

Population Variance is

sigma^s = Sigma(x - mu(pronounced myu)) squared divided by N.

Sample Variance is

S squared = Sigma( x - xbar) squared over n - 1

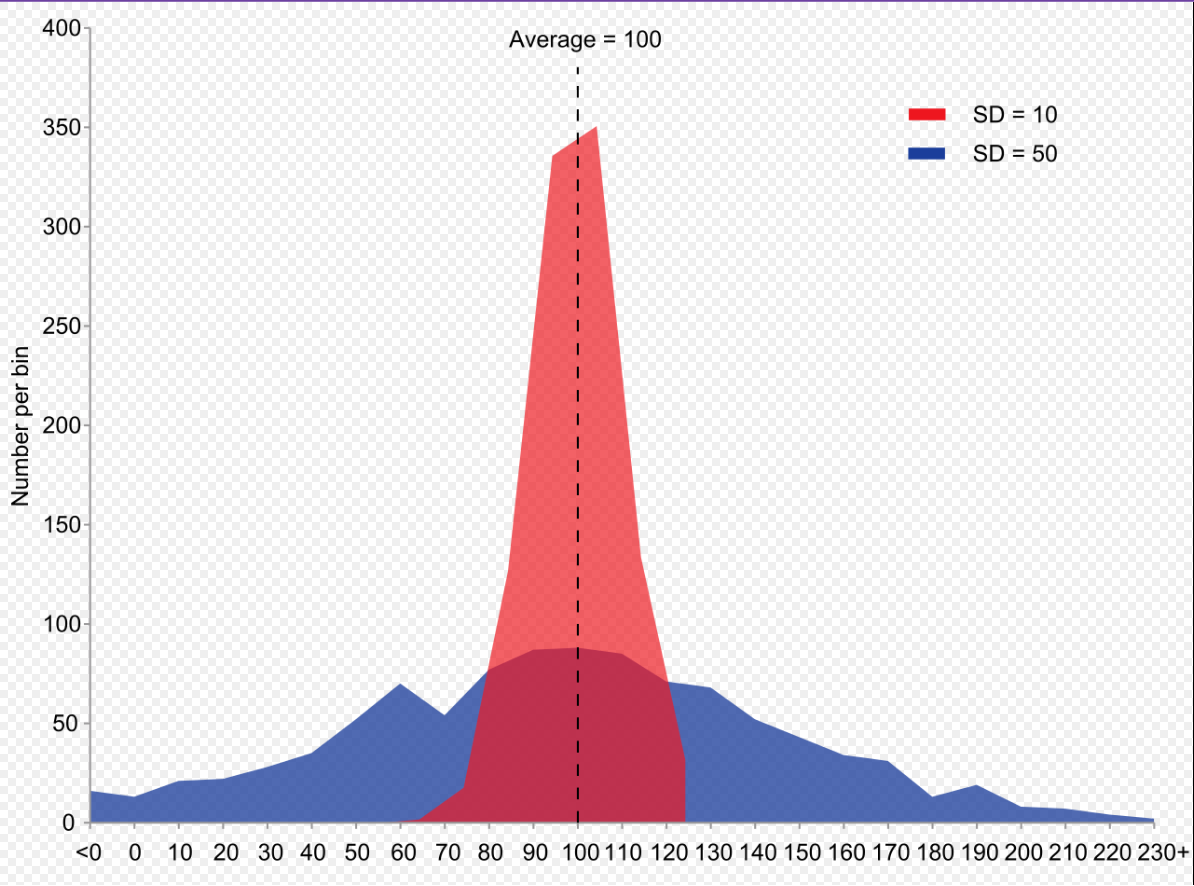
What is that pivotal difference? It is known as Bessel’s correction.

N-1. Why are we doing n-1? It is because sample variance tends to UNDERESTIMATE the true variance. This is not something we can go deep into, but let’s think of it intuitively. The mean of a data set is the value that produces the least squared distance between the points and singular value. If you move from the mean, the resultant total squared distance is larger. A sample’s mean works the same. It’s works to produce the smallest possible squared distance. That means that if the sample mean and the population mean are the same, that they will produce the exact same squared distance, if however, the population mean is any other value, that means that it will be larger than the sample mean. That makes this equality true. Pop mean >= sample. Think of all the possible values that the population mean can be, it’s infinite. The sample mean however, because it’s generated only from our sample values, is not infinite, it is a singular finite value. This means that the chance of the sample mean being the population mean is significantly smaller than the sample mean being smaller than the population mean.

The true math proof is a bit more complicated, but I will link it for those who want to try to understand it at a deeper level.

<https://en.wikipedia.org/wiki/Bessel%27s_correction>

The variance is usually a number that seems somewhat arbitrary. We squared values in it to get rid of the negatives, and so the value is in a form we can’t use. This means that we should get the root of it to get a “usable” value. This is what is known as the standard deviation. The standard deviation gives us a good look into the dispersion of the data. A low standard deviation means that values tend to lie fairly closely to the mean. A high standard deviation means that values tend to be far away from the arithmetic mean. It gives you a way to think about your data. Where “values” tend to be.



Temperature Check

~~~~~~~~~~~

Is everyone following along? Is everything clear now?

Take a break.

Move onto the Pandas.

Now we are going to be working on using Pandas to help us get these key values, and then using these key values to paint a picture.