Lesson 2 Additional Information.

Why do we care so much about math in Statistics?

We can infer and pontificate as much as we want over data sets, and claim that certain trends and patterns appear in our data, but all of that is bunk without any “proof”. We need real world, tangible ways of explaining why our analysis of the data is accurate. Never believe anything that is told to you without a bevy of additional information. We need to build a “fact” base we can work through that helps to concretely show that what we are trying to say is true. The best “facts” to pull out of data sets are statistics or parameters.

Let’s take a step back to explain a core concept in statistics.

The idea that a population that is studied is able to produce more accurate, higher quality data; that is irrefutable. Data taken from a population data set are called parameters (mean, mode, median). We have all the individuals in our population, we don’t need to infer *as much(human error and other inconsistencies make even this hard)* because all the data we need is right in front of us. A set that is a population could be “all the students in our class”. We could talk about stats of our class, and our class only, with no intention to use it to represent all classes in every class in the world. We could then interpret data from that. (hint: TKH does)

Sometimes, we aren’t able to get all the individuals we need to form population data. If for example, we were doing a study of all men in the Tri-State region, it would be impossible to get data from every man in the region. Instead, we can SAMPLE, the population. Take CERTAIN individuals from that population, and use THEM to represent the population as a whole. Let’s say we gather data from 100 random individuals from NJ, 100 from NY, and 100 from CN. From that data set of 300, now we take 33 random individuals from each of those groups of 100. Now we have a set of 99. This set of 99, could be used to represent the population of men in the tristate area as a whole(a sample of 100, or even 300 is a bit too small to represent the population of the men in the area as a whole. Rule of thumb says 10% is best if possible). From our SAMPLE of 100, we can get STATISTICS. Our mean, mode, and median are STATISTICS, they are supposed to represent the population as a whole.

Knowing the difference between them and understanding it will be more important later, but for now, know this key difference. What we need to be able to do, is to be able to pull “important” data points from them. Points of reference that we can then build on top of to further help our understanding of the data. Some of the most important data points are the mean, the median, and the mode.

Let’s start with the Median. The median is the “middle” number of the data set, if the data set was ordered.

For example a Data Set of

1,2,3

Would have a median of 2.

A Data Set of

1,2,3,4,

Would have a median of 2.5 (The median does not have to exist in the data set)

It’s the middlemost value, so if the set has “two middles”, you add them up, and then divide them by 2.

The median is a good measure of central tendency. It’s ROBUST as well, meaning that is not as susceptible to outliers. Let us do a proof of concept.

1,2,2,2,4,10

What is our Median? The median is 2. Even though 10 is very far away from the other values, the median is still relatively close to the other values.

Let’s add 1000 to the data set.

1,2,2,2,4,10,1000

Our median is still 2.

What if we add 500?

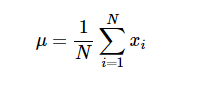
1,2,2,2,4,10,500,1000

Now our median is 3. It has increased, but not by much. We can argue that 3 is still closer to the “middle” of our values, since 3 is RELATIVELY close to 1,2,2,2,4 and 10, which make up the majority of our data set.

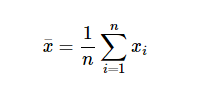
The median is a core concept, and one that will appear frequently in statistics, so let us make sure to memorize this. It’s usually represented as M or m or x-tilde.

The next measure of central tendency is the Mean. We want to stay away from using the word “average.” The median can be considered an “average” as can the mode, so when we talk about the “average” we aren’t really saying much. The mean is the “middle-most” value again of our data set. If we could choose one singular value to represent the data-set, the mean would be it. The mean of a population is represented by mu(pronounced myu), while the sample mean is represented by xbar.

Here are the formula. For Population mean



For Sample Mean



Notice the difference. We use N(our population) in population mean, and n in our sample mean.

This fancy notation might seem daunting, but it’s quite simple. The Large E looking thing is sigma. It means sum up. Sum up xi. Sum up xi, between xsub1 and x sub n. This is a fancy way of saying sum up all the values in our data set. Once we have that sum, divide it by the amount of data points in our data set.

Let’s get some rudimentary practice.

1,2,2,2,4,10

What is the mean? 3.5

What happens if we add another large number?

1,2,2,2,4,10,100

Mean goes to 17.2

And another?

1,2,2,2,4,10,100,500

Mean goes to 77.

We can see that the mean can be heavily influenced by values that are far away from(relatively) the rest of the values in the data set. The mean is NOT robust. It IS influenced by outliers.

Lastly, we have the mode. A rarely used, but sometimes relevant measure of central tendency. This is the value that appears most in a given data set. This one has no real fancy calculation.

In the same dataset of

1,2,2,2,4,10

The mode is 2. If a dataset has every value occurring at the same frequency, then the set is said to have no mode. If a dataset contains more than one value at the same frequency, the set is said to have multiple modes.

Now that we have 3 measures of central tendency. We might ask ourselves, Why? Why do we need all these measures and what can we do with them? The answer is that we can use them for some very precursory analysis, and then, more importantly, we can use them to derive even more data about our dataset.

The truth is that we will often use the Median AND/OR the Mean. The mode is used as often. The Median and Mean can be just either in conjunction, or on their own. When and where we use them, actually comes down to knowing and understanding your data set. Sometimes we have data sets that are going to be heavily weighted by outliers. That would be a time to use the median. Sometimes we have a dataset where outliers are the things we are looking for and the values we care about. Then there are times where we have a super tight data set. Our median might not be best then for accurately describing the “middle” of the data set, in those cases we use the mean. Then, there are times where we can and want to use both of them. It comes with practice knowing which and when to use, and which and how to use them to describe the data you’re working with, and how to extrapolate further data from your data sets.