Course Section 2/Unit 2, Lecture 1: Central Tendency and Shape

The main goal in Course Section 2/ Unit 2 is to understand that a distribution of scores (a set of data) can be fully described by the following three "things:" (1) an appropriate measures of central tendency, (2) an appropriate measure of variability, and (3) the shape of the distribution. This lecture, Lecture 1, focuses on shape and central tendency. Lecture 2 focuses on variability. Again, your overall goal in this Course Section is to learn how to fully describe a set of data.

Central Tendency

A measure of central tendency aims to find the "middle" of a distribution. A measure of central tendency identifies one single measure that best describes a set of data. There are 3 measures of central tendency: mean, median, and mode. Each of these measures is considered an average, but they are *very different averages*. Each measure of central tendency has a specific definition. No one measure of central tendency is the "best" for a given set of data. It is important to choose the appropriate measure of central tendency for a given set of data. Let us use some data to further discuss that which has been written to this point.

Imagine that you had the following set of exam scores in one of your courses:

70, 60, 40, 90, 85

A measure of central tendency would identify **ONE** score that would best represent your data set (all of your exam scores). There are three possible measures of central tendency that you can choose from to come up with the one score that would best represent your exam performance. Therefore, we will turn our attention to the **3 measures of central tendency:** mean, median, and mode. We will discuss how to go about choosing the most appropriate measure of central tendency, a mean, a median, or a mode for your exam scores…pay attention because you may be able to argue for a better grade in some future course ☺

Most of you are probably very familiar with the three measures of central tendency: the **mean, median, and mode**. But some of you may not be very familiar with the fact that they are **very different types of averages**.

The Mean

The "**mean**" is defined as the **arithmetic average** because in order to calculate a mean you must perform mathematical operations. If you wanted to calculate a mean for your exam scores, 70, 60, 40, 90, and 85, you would need to add up all of your exam scores (70+60+40+90+85) = 345 (which is EX, sum the scores) and then divide by the number of scores in your data set of exam scores, which is 5 (N). Therefore, the mean of your exam scores is EX/N or 345/5, which = 69.

The formulas for calculating a mean are as follows:

If the data set were from a **population** of scores, then $\mu = \sum X / N$

If the data set were from a **sample** of scores, then $M = \sum X / n$

Notice that the mean is calculated exactly the same way whether the data set is from a population or from a sample, but the statistical notation is different in order to differentiate whether you have calculated a mean from a population or from a sample. While most students are familiar with mathematically calculating a mean, they are often less familiar with the *balancing* aspect of the mean.

Mean as the "Balance Point"

Let me try to illustrate the *balancing characteristic* of the mean by returning to your exam data:

70-60-40-90-85

As we have already calculated, your exam mean is = 69. The balancing characteristic of the mean is that the mean (69) will always be **precisely located** at the exact point in a distribution **where the distance between the mean and the scores below it will equal the distance between the mean and the scores above it**. In other words, the mean of 69 would be located exactly at the point where the scores below the mean (the 60, and 40) would have the same total **distance** from the mean as the scores above the mean (70, 90, 85). Here is what it would look like:

Mean=69

Distance Below Distance Above

The score of 60 is 9 units below the mean……9

The score of 40 is 29 units below the mean……..29

The score of 70 is above the mean by 1 unit……. 1

The score of 90 is above the mean by 21 units……..21

The score of 85 is above the mean by 16 units……..16

If you added the distances for each score below the mean (9 +29) and added the distances for each score above the mean (1+21+16), you would wind up with equal distances =38. The mean MUST be located at the point in the distribution where the total distance of all the scores below it equals the total distance of all the scores above it. This is why we refer to the mean as the "balance point" of a distribution: It balances (makes equal) the **DISTANCE** between the scores below it and the scores above it.

Notice that the mean is NOT located at the point where exactly half the scores are below it and exactly half the scores are above it, although it could be; but, it doesn't have to be. In our example, notice that there are two scores below the mean and three scores above the mean. But it **IS** located, again, at the precise location where the total distance of scores below it equals the

total distance of scores above it. And, once again, this is why we call the mean, the average that balances the distance between the scores.

The Median

The median is also a measure of central tendency. It is also an average, but it is a very different average than the mean. The definition of the median is that it **is** the single score in a distribution that marks the exact location where the **number of scores that are below it must equal the number of scores that are above it**. It is the only measure of central tendency that MUST always divide the number of scores in a distribution exactly in half.

Let us look at an example by returning to your exam data:

70-60-40-90-85

Unlike the mean, in order to calculate a median, you **MUST** first **order** the data:

40, 60, 70, 85, 90

While there are some unnecessarily complicated formulas to calculate the median, it is most important to "see" what a median is: It is the score that divides the distribution exactly in half so that there are an equal number of scores below it and an equal number of scores above it. The median here is 70. Notice that while the mean and median for your exam scores are very close, M=69 and median =70, they are, nonetheless different. Remember that while the mean balances the **distance** between itself and the scores above and below it, the median balances the NUMBER of scores above it and below it.

Building on the above examples, let us assume you took another exam and scored a 95. Your exam data set would look like this:

40, 60, 70, 85, 90, 95

Now there is no score in the data set that is at the midpoint to divide the distribution exactly in half. We can see that there are 6 scores and so three must be below the median and three must be above the median. In this case, we can see that the midpoint is between the 70 and 85 and we can simply take the mean of the two (70+85/2) to get the median: 77.5

The Mode

The mode is the score with the highest frequency. In other words, the mode is the score that is recorded most often. Let us return to your exam data:

40, 60, 70, 85, 90

Your set of exam scores, 40, 60, 70, 85, 90, has **NO mode** because there is no one score that occurs with a higher frequency than any other score. But, if your exam scores were:

40, 40, 40, 60, 70, 85, 90

then, the mode would be 40 because 40 is the score with the highest frequency.

It is possible for a set of data to have more than one mode; for example:

40, 40, 60, 60, 70, 85, 90

the mode would be both 40 and 60. It is not, however, possible to have more than one mean or more than one median in a distribution.

Calculating a Mean, Median, and Mode from a Frequency Table

You should be able to calculate a mean, a median, and a mode from a frequency distribution. Let us look at the following frequency table (you should be familiar with frequency tables from Course Section 1/Unit 1):

X	f
5	2
4	1
3	6

Calculating the Mean

Remember that in order to calculate a **mean**, you need to add the scores and divide by the number of scores. Therefore, you need EX (add the scores) and N (the number of scores).

When you are working with a frequency distribution, you cannot just add the 5+4+3 and say that EX=12 because you DO NOT ONLY have ONE 5, ONE 4, and ONE 3, but rather you have TWO 5's, ONE 4, and SIX 3's.

Your data set is actually 5+5+4+3+3+3+3+3+3; therefore, EX= 32. Similarly you cannot say that N=3 for the same reason (you do not only have one score of 5, one of 4, and one of 3), but

rather two 5's, one 4, and six 3's. Therefore, the number of scores is 9, N=9. N can always be calculated by adding up the frequencies (2+1+6) = 9. To calculate the mean you divide EX by N, 32/9, and the mean =3.55.

Calculating the Median

Calculating the median can also be done in a number of ways as described in the required reading (assuming you are reading a textbook), but perhaps the simplest way is to order the data represented in the table:

X	f
5	2
4	1
3	6

Begin by ordering the data: 3, 3, 3, 3, **3,** 3, 4, 5, 5,

After ordering the data, you must locate the midpoint at which the number of scores below the median = the number of scores above the median. In this case, the median =3 because four scores are below the median (3, 3, 3, 3,) and four scores are above the median (3, 4, 5, 5).

Calculating the Mode

X	f
5	2
4	1
3	6

Calculating the mode from the same frequency table only requires you to recognize the score with the highest frequency. Hopefully, you can see that the highest frequency is 6. Therefore, the score that occurs the most, the mode = 3.

Why 3 Measures of Central Tendency?

There are three measures of central tendency because no one measure is always the best representative measure for a given set of data. At times, the case becomes that one or another measure of central tendency **cannot** be used for a particular data set. Let us discuss this further.

Nominal Data Set

When possible, the mean is the most often preferred measure of central tendency. However, there are particular data sets, for which the mean is neither possible nor the "best" choice. Recall that the mean is an arithmetic average; that is, it requires addition and division. Therefore, if the

data set is a nominal one, then the mean simply cannot be computed. If the data set is nominal, then you would have "word" categories; and, you can't add and divide words because what exactly is purple + blue / 2? In the case of a nominal data set, the best measure of central tendency would be the mode. In the case of color categories, the name of the color most often recorded would be the mode.

Case of the Outlier

Another data set for which the mean is not appropriate is a data set that has an **outlier**. An outlier is a score that is very different (located far away) from most of the other scores in the data set:

Data set 1…… 2, 3, 3, 4, 5, 10 ----------→mean = 4.5

Data set 2…….. 2, 3, 3, 4, 5, 150 ----------→mean = 27.83

Notice that in Data Set 1, the mean (4.5) is a representative measure of central tendency; that is, the single score of 4.5 is in the "middle" of the data and represents the data set quite well. But, in Data Set 2, the mean (27.83) is not a good representative measure. The mean is PULLED toward the outlier, the extreme score of 150; and, thus, the mean becomes a misleading representation of the "average." There is a very short video on this that you can access in this Unit's resources.

A More Personal Example of an Outlier's Effect on the Mean

Imagine that the following scores represent your exam scores for a given course:

95, 95, 92, 90, 23

The **mean** of the exam scores = 79, but is 79 a good representative of your "average" exam performance? No, it isn't because your "tendency" is to score in the 90's and not the upper 70's. The outlier (23) distorts your "average" exam performance.

In a case like this, where there is clearly an outlier, the **MEDIAN** would be the **preferred** measure of central tendency. The median is 92, which, by far, is a better representative of your exam performance. Would it be fair for a professor to give you a final course grade of C+ (based on the mean, 79) rather than an A- (based on the median, 92)?

Shape of Distributions

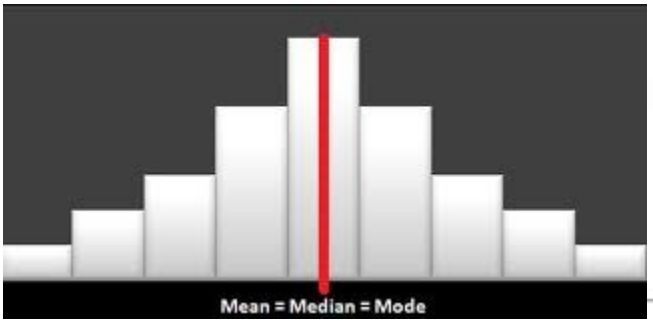
It is possible to spend all semester just discussing the shapes of distributions because there is a lot to know. However, in this introductory course, we will focus on two main shapes of distributions: symmetrical and skewed.

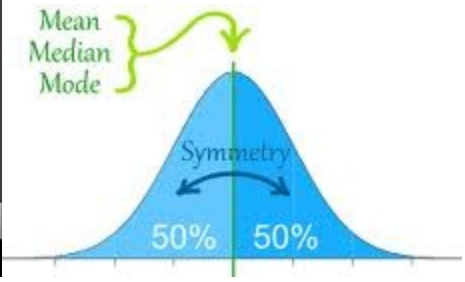
Symmetrical Distributions

A symmetrical distribution is one where drawing a line directly down its center, leaves two equal, or mirror imaged, sides. If, after drawing a line down the middle, one side would be the mirror image of the other side, then you would have a symmetrical distribution.

If you know the shape of the distribution, you will have an idea of where the measures of central tendency will be located.

For example, if a distribution is perfectly symmetrical, then all three measures of central tendency will be located exactly dead center and all three measures will have the exact same value. Here are two examples of a perfectly symmetrical distribution:





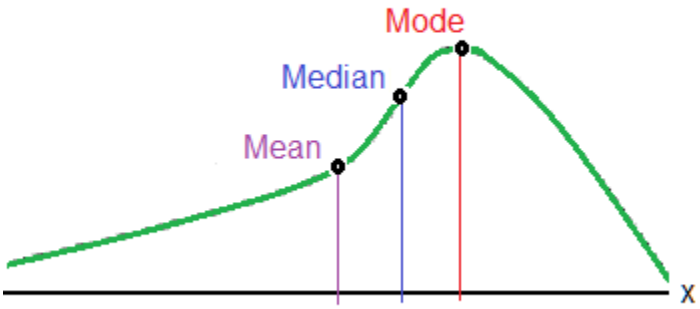
Skewed Distributions

Distributions can also be skewed, which simply means not symmetrical. Distributions can be **negatively** skewed or **positively** skewed. Skewed distributions are not shaped in a way that you can draw a line down the middle and see "equal" sides like you would see when the distribution is symmetrical.

See next page…..

Negatively Skewed Distributions

Here is an example of what a negatively distribution looks like:



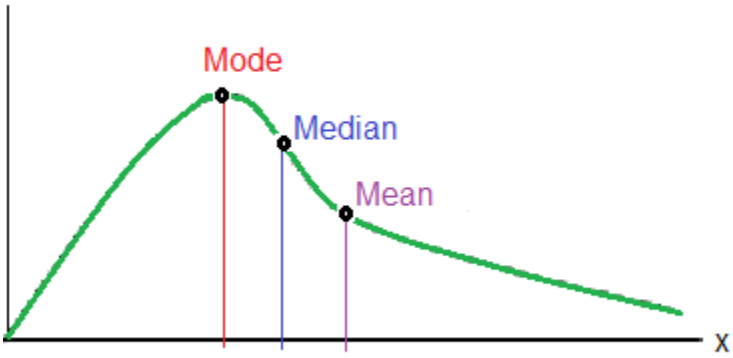
Negatively skewed distributions have **a lot of HIGH** scores and a **few LOW** scores (outliers). This is counter-intuitive to some students; therefore, read it again: Negatively skewed distributions have a lot of HIGH scores and few LOW scores. "Negative" means that the "tail" of the distribution is to the left, and that the outliers are low scores.

If a professor gave a very, very, very, **EASY** exam, the expected shape of the exam distribution would be negatively skewed. This is because while most students would likely get high grades, a few students would likely get low grades. "Negative" doesn't mean "bad," but rather it simply means, again, that the outliers are low scores.

As you can hopefully "see" in the negatively skewed distribution above, it has more high scores than low scores. Because most scores are high, the mode will tend to be to right, the median in the middle, and the **mean will be PULLED toward the low scores**, toward the tail end of the distribution.

Go to next page….

Positively Skewed Distribution



You should notice that, in a **positively skewed** distribution, there are **a lot of LOW** scores and **FEW high** scores. Therefore, the mode will be located to the left (where most scores are piled up), the median will be located in the middle, and the mean will be PULLED toward the FEW high scores (outliers).

Importantly, if you know the SHAPE of a distribution, you WILL also know where to most often expect the relative locations of the measures of central tendency. Imagine that I asked you the following questions:

1. In a perfectly symmetrical distribution with a mean of 80, what is the mode?
2. If a professor gives a very, very difficult exam, who would have the highest score, the person who scores at the mode or the mean?
3. If a professor gives a very, very easy exam, who would have the highest score, the person who scores at the mode or at the mean?

Think about the answers to the above questions before reading my answers below.

The answer to (#1) is 80.
It is 80 because if the distribution is perfectly symmetrical, then all three measures of central tendency would be at the same location and would have the same value. Therefore, the mean, median, and mode would all be 80.

The answer to (#2) is the person whose score is located near the mean because the mean would be pulled to the extreme and rare high scores. Most people would score low on a very, very, difficult exam.

The answer to (#3) is the person whose score is located near the mode because on a very, very, easy exam, most scores would be HIGH. You should expect very few low scores on a difficult exam. The mean, remember, would follow the outliers (the low scores, in this case).