Lecture 2: Variability

Variability is often a challenging concept for students, but it doesn't have to be ☺. While mathematical formulas and calculations are included and necessary in this lecture, formulas and calculations are not as important as ***understanding*** (1) what a measure of variability is, and (2) what it tells you about a set of data. You will need to be able to compute the measures of central tendency (mean, median, and mode) as well as measures of variability (variance and standard deviation), both by hand and on SPSS, in order to successfully complete this section of the course.

You may recall from Lecture 1, Central Tendency and Shape, that distributions (sets of data) are fully described by 3 things: a measure of central tendency, shape, and a measure of variability. I hope you recall a fair amount of that lecture. If you do not, then it might be a good idea to review it before focusing on this lecture.

Let's try to get a general sense of what **variability** is. Imagine you had a distribution of scores (or a data set) that you wanted to fully describe. First, you would calculate the appropriate measure of central tendency. We will assume that we calculated a mean for our imagined data set. Once you calculated the mean, you would have an idea about the "middle" of the distribution; specifically, you would have a score, the mean, which best represents the entire data set. Let us also assume that the imagined data set has a symmetrical shape. At this point, you would have 2 out of the 3 necessary things to fully describe a distribution: measure of central tendency and a shape. The only thing left to do is to **calculate the appropriate measure of variability.**

What is Variability?

Variability is a measure of how far or how close all the scores are in relation to the mean, and how "spread out" the scores are from each other. You should read that first sentence again. And maybe even one more time. A measure of variability tells you if the scores in a distribution are all pretty much located very near each other [and the mean (as we will see later)] or if they are pretty much spread out from each other [and the mean (as we will see later)]. Let's look at the following 2 data sets (A and B) to see if we can get a better idea of this thing called variability.

Consider Data Set A----------→2, 4, 6, 5, 3, 7, 8

Consider Data Set B----------→2, 4, 5, 25, 46, 57, 80

Which data set has **HIGHER** variability? Answer = Data Set B.

Why? Because the scores in Data Set A are simply closer together (to each other) than the scores in data set B. Data Set **B** has a **HIGHER** variability because the scores are further apart from each other. There is a bigger **SPREAD** of scores in Data Set B than there is in Data Set A. The basic idea is that simple! ☺ Notice, however, that both Data Set A and Data Set B have the same number of scores. It is not about **how many** scores there are in a data set (a distribution), but about how **spread out or how close together** the scores are from each other in a particular data set. In

data set A, the greatest distance is between 2 and 8; while in data set B, the greatest distance is between 2 and 80. **Get it?** If not, maybe the next example will help.

The Example of Fruit in the Supermarket

Imagine that you are at the supermarket buying fruit. Would you find a higher variability of fruit when the supermarket has apples, bananas, pears, oranges, pineapples, tangerines, and watermelon or when the supermarket has pears and bananas? Clearly, there is more variability when there are 7 types of fruit when compared to only two types of fruit. I hope this example has provided some insight into the fact that variability is much like "variety." The higher the variety of fruit is, the higher the variability of fruit also is.

The same idea applies to variability in numerical data sets. The more the numbers in a data set are far away from each other; that is, the more different they are in value from each other, the more spread out they are on the number line –the bigger the **range** between the lowest number and the highest number, the higher the variability in a set of numerical data.

Imagine these are your exam scores in two courses: Biology and Statistics

Biology course: 90, 92, 94, 87, 85

Statistics course: 40, 60, 50 96, 25

Which data set has higher variability? The range of your scores in biology is from 85 to 94 (10 points), but in your statistics course, the range of your exam scores goes from 25 to 96 (72 points); therefore, your statistics scores have higher variability. **You are a more consistent performer in Biology than in Statistics and a more variable performer in Statistics than in Biology (think about this statement).**

One Last Example (skip it if you don't need it)

Imagine that you wear the **exact same clothing every day** of the week. If you did, then your **weekly outfits have ZERO variability**. If, however, you wear the same clothing all week, but wear something different on just Sundays, then your weekly outfits have, at least, a little variability. **The same applies to numbers:** If your data set is 2,2,2,2,2,2,2 then your data set has ZERO variability. However, if you were to change one number from a 2 to a 3, then your data set would have, at least, a little variability.

Take a close look at the following data sets. Can you ***"see"*** how variability is changing from one data set to another (going from data set 1 to data set 5)?

Data set 1……..2, 2, 2, 2, 2, 2, 2, 2, 2, 2

Data set 2……..2, 2, 2, 2, 2, 2, 2, 2, 2, 3

Data set 3……..2, 2, 2, 2, 2, 2, 2, 3, 4, 5

Data set 4……..2, 2, 3, 5, 6, 8, 19, 29,45

Data set 5……. 2, 75, 101

Each data set INCREASES in variability. Again, take note that the number of scores does not matter. Notice that Data Set 5 has the least number of scores, but the highest variability. It is all about how close together or how far apart the numbers are from each other, not about how many numbers there are. I hope this was helpful because that is the easier part ☺

Re-read the above part of the lecture as many times as you need in order to clearly understand the basic idea of variability before you go on to the next section.

Measures of Variability

There are three measures of variability: Range, Standard Deviation, and Variance.

Range

The range is a **descriptive measure** of variability. It is simply the distance between the highest score and the lowest score (plus 1). Its formula is not important at this point, and we will keep to a "gross" measure of range—just a general understanding. For the following set of data:

2, 5, 9

The range would be calculated as H-L+1. H=highest score, L=lowest score. Therefore, the range would be 9-2 +1 = 8. We add the 1 because of "real limits" which you can read about in the course text, but we are not focusing on that ☺ -at least not right now.

The Standard Deviation

The statistical notation for the standard deviation of a population is **σ (sigma)**, and the statistical notation for the standard deviation of a sample is (**s**). To calculate either one requires that you use the appropriate formulas because there are slight differences in the formulas to make up for the fact that populations are more variable (have more variability) than do samples—something that we will discuss later in the course.

Our primary focus is the standard deviation. The standard deviation is also a **descriptive measure of variability.** The standard deviation is a measure of "the average distance of the scores from the mean." Therefore, like the variability examples we already discussed (remember the fruit? The clothes?), the more spread out the scores are in a distribution, the higher the standard deviation will be (remember that the standard deviation is a measure of variability so that all we said about variability is also true of the standard deviation). But, for now, let us focus on how to calculate the standard deviation.

Calculating the Standard Deviation

Calculating the standard deviation, whether for a data set from a sample **(s)** or a data set for a population **(σ),** requires us to first calculate "the sum of the squared deviations" known as **SS**. SS is the "sum of the squared deviations," which directs us to carry out three steps: (1) calculate deviation scores, (2) square those deviation scores, and (3) add up the squared deviation scores. If we do steps 1, 2, and 3, then we will have the "sum of the squared deviations" or **SS**. Once we have calculated SS, then we can insert it in a simple formula to calculate the standard deviation for a sample **(s)** or for a population **(σ).** Read this paragraph again ☺

Let us use the following set of exam data to complete the 3 steps of calculating SS

60, 70, 80, 90

Step 1: Calculate the deviation scores

Each individual **distance** from a single score to the mean is a **deviation score**. A deviation score is the distance between any **one** score in the data set (60, 70, 80, or 90) and the mean. It should make sense, then, that the statistical notation for a deviation score is X-M: the difference, the distance, between a single exam score (X) and the mean (M).

I hope that it also makes sense that the first thing we need to do is to calculate the mean so that we can calculate a deviation score for each of the exam scores. The mean for our data set (the red data set above) is 75.

The **deviation score** for the exam score of 60 is calculated as X-M (60-75) = -15.

To be clear, the deviation score for X=60 is -15. That is, the deviation score, the distance between 60 and the mean, is 15 units. Notice, however, that we have a NEGATIVE 15 (-15). Importantly, the negative sign indicates that the LOCATION of the exam score (60) is located BELOW the mean—not that the exam score is negative. Clearly, the exam score (60) is a positive number.

The **deviation score** for the exam score of 80 is calculated as X-M (80-75) = 15.

The deviation score of 15, in this case, is POSITIVE. A positive deviation score informs you that the exam score is ABOVE the mean.

Important to Take a Moment and Note the Following:

The sign (+) or (-) of any deviation score tells you whether the exam score (the X-score) is located above or below the mean. If the deviation score is negative, then the X-score is located below the mean. If the deviation score is positive, then the X-score is located above the mean.

The numerical value of the deviation score tells you how far (in single units) the X-score (exam score) is from the mean. Therefore, a deviation score of -15 informs you that the exam score is 15 points ***below*** the mean. A deviation score of 15 informs you that the exam score is 15 points ***above*** the mean.

If you have a handle on what a deviation score is, how to calculate it, and how to read it, then you are ready to move forward. If you are confused, then do not move forward. Rather, review this lecture to this point, and when you believe you have a solid understanding, then move forward to the next part of this lecture.

Are you ready to move forward? If so….focus! Don't get discouraged. You CAN do this!

Let us continue with what we were doing. In a more simple format, can you see that we can calculate all the deviation scores for the data set of exam scores (**60, 70, 80, 90)** as follows:

X	X-M (Deviation Score)	
60	60-75 =	-15
70	70-75 =	-5
80	80-75 =	5
90	90-75 =	15
		0 = problem

At this point, we have all the deviation scores for our exam scores. Remember that we are working our way to calculating SS, the sum of the squared deviations. At this point, all we have are deviations. The deviations are not yet squared. The reason we have to square them is because, if we added up the non-squared deviations (the ones we have above: -15, -5, 5, 15), we would wind up with a sum of zero (0), which is a problem when calculating the standard deviation (our ultimate goal). It is a problem because the standard deviation is the "average" of all the deviation scores of a data set; And, if the total of the deviation scores equals zero, then diving anything by zero would always result in an answer of zero. As I already noted, if we added up all the deviations above (-15, -5, 5, 15), we would get an answer = 0. Again, this is a problem because 0 divided by anything would always give us an answer of 0, making it impossible to calculate the average deviation (the standard deviation) that we ultimately want. To rid ourselves of this zero problem, we will move on to step 2 of calculating SS on our way to, ultimately, calculating the standard deviation.

Step 2: Square all the Deviation Scores

X	X-M	$(X-M)^2$
60	60-75 = -15	$(-15)^2$ = 225
70	70-75 = -5	$(-5)^2$ = 25
80	80-75 = 5	$(5)^2$ = 25
90	90-75 = 15	$(15)^2$ = 225
	0 = problem	500----------→ SS

In the diagram above, I hope you can see that we took each deviation (-15, -5, 5, 15, highlighted in yellow), squared it (highlighted in green), and then added up the squared deviations (all the highlighted green) to get SS=500 (highlighted in pink). Remember that SS is the sum of the squared deviations, which in our case is 500.

Step 3 of Calculating SS

One formula for calculating SS, the conceptual/definitional formula, which we already have calculated in Step 2 above is as follows:

$$SS= E(X-M)^2$$

Notice that this formula reads as: SS is equal to the sum of the squared deviations; that is, SS is the sum of the squared deviations. In step 2, you can see that have already calculated this formula when we summed up the squared deviations and got an SS=500.

Another formula you can use to calculate SS is the computational formula provided in the text. The conceptual formula is used here because it is a better way to understand what SS actually is, but you can calculate SS any way you like. The computational formula is easier to use when the mean is not a whole number. When the mean is a whole number, the conceptual/definitional formula (the one that I have used here) may be easier for some students.

Once you Complete SS, you are Ready to Calculate the Standard Deviation

At this point, many students have great difficulty; therefore, read carefully. Thus far, all that we have calculated is SS. We have calculated the deviation scores, squared the deviations, and summed the squared deviations, all of which was necessary to calculate SS, which is the first step in calculating the standard deviation, to which we turn next.

Calculating the Standard Deviation for Sample or Population Data

Once you calculate SS, you need to be clear about whether your data is from a population or a sample because you must use the appropriate formula to calculate either a standard deviation for a sample (s) or a standard deviation for a population (σ). Take a minute to note the **statistical formulas for each:**

Standard Deviation for a Population: $\sigma = \sqrt{SS/N}$

Standard Deviation for a Sample: $s = \sqrt{SS/n-1}$

Remember our original data of exam scores (60, 70, 80, 90)? Let us assume this is a SAMPLE data set. To calculate the standard deviation, we can insert SS into the appropriate formula and calculate the sample standard deviation (s) as follows:

$s = \sqrt{SS / n-1}$

$s = \sqrt{500 / 4-1}$

$s = \sqrt{166.66}$

$s = 12.90$

If we assume a population, then

$\sigma = \sqrt{SS/N}$

$\sigma = \sqrt{500/4}$

$\sigma = \sqrt{125}$

$\sigma = 11.18$

Variance

Let us begin by getting clear on our statistical notation:

Sample standard deviation: s (calculated above)

Sample variance: s^2 (to get the variance, you have to square the standard deviation)

Population standard deviation: σ (calculated above)

Population variance: σ^2 (to get the variance, you have to square the standard deviation)

The math we calculated in the previous section of this lecture was to calculate SAMPLE standard deviation (s), and our answer was 12.90. We also calculated the POPULATION standard deviation (σ) and our answer was 11.18. We did this by first assuming that our data (60 70 80 90) was from a sample, and then we assumed the same data came from a population.

In the previous section of the lecture, we reviewed and calculated the ***standard deviation***, defined as the **average distance of all the scores from the mean** (or, the average deviation).

In this section, the focus is on **variance. Variance** is another measure of variability that is **defined as the mean of the squared deviations** (or the average distance of all the scores from the mean, but in squared units). It is the average of squared distance while the standard deviation is the average of the "unsquared" or "standard distance" between the scores in a distribution and the mean.

Notice the statistical notation: Sample standard deviation is (s) and that variance is ($s^{2)}$

Population standard deviation is (σ) and variance is (σ^2)

We will cover **variance** later in the course when we get to inferential statistics. I am sure you are relieved ☺ But you need to know the following about the relationship between variance and standard deviation. Understanding the difference in the statistical notation above will help you understand the last part of this lecture.

Relationship Between Variance and Standard Deviation

There is an important relationship between variance and standard deviation. If you have calculated variance, you can take its square root to get the standard deviation.

So, for example, if you know the variance for a population (σ^2) is 9, then you can take the square root of that variance ($\sqrt{9}$) and also know that the population standard deviation is 3 (σ = 3). Conversely, if you knew that the standard deviation for a population (σ) was 10, then you can square that standard deviation (σ) $(10)^2$ and know that the population variance (σ^2) = 100.

The same goes for samples. If you knew the sample variance (s^2) was 25, then you can take the square root of the sample variance (s^2) and know that the sample standard deviation (s) = 5. If you knew that the sample standard deviation (s) was 6, then you can square the sample standard deviation $(6)^2$ and know that the sample variance (s^2) = 36. This is a very difficult point for many students.

In numbers to reflect what was said above:

If σ^2 =9, then σ =3

If σ = 10, then σ^2 100

If s^2 =25, then s = 5

If s = 6, then s^2 =36

Notice that the standard deviation is either (s) for a sample or (σ) for a population and that the **VARIANCE** is the standard deviations **squared:** The variance for a sample is (s^2) and the variance for a population is (σ^2).