LESSON 4: MEASURES OF DISPERSION

A measure of central tendency provides information about the “middleness” of a distribution of scores, but not about the width or spread of the distribution. To assess the width of a distribution, we need a measure of variability or dispersion. A measure of variation indicates how scores are dispersed around the mean of the distribution.

Range

The simplest measure of variation is the range—the difference between the lowest and the highest score in a distribution. To find the range, simply subtract the lowest score from the highest score.

Table 4.1 Two – Class Score comparisons

<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>100</td>
<td>55</td>
</tr>
<tr>
<td>Σ = 150</td>
<td>Σ = 150</td>
</tr>
<tr>
<td>μ = 50</td>
<td>μ = 50</td>
</tr>
</tbody>
</table>

In the example above the range for Class 1 is 100 points, whereas the range for Class 2 is 10 points. Thus, the range provides some information concerning the difference in the spread of the distributions. In this simple measure of variation, however, only the highest and lowest scores enter the calculation, and all other scores are ignored.

Percentile

The \( n \)th percentile of an observation variable is the value that cuts off the first \( n \) percent of the data values when it is sorted in ascending order.

> quantile(INCOME, c(.32, .57, .98)) #finds the 32, 57 and 98th percentiles

Quartile

There are several quartiles of an observation variable. The first quartile, or lower quartile, is the value that cuts off the first 25% of the data when it is sorted in ascending order. The second quartile, or median, is the value that cuts off the first 50%. The third quartile, or upper quartile, is the value that cuts off the first 75%.

quantile(INCOME) #gives the first second and third quartiles

The inter-quartile range is the difference between the third quartile and the first quartile.

Average Deviation and Standard Deviation

More sophisticated measures of variation use all of the scores in the distribution in their calculation. The most commonly used measure of variation is the standard deviation. Most people have heard this term before and may even have calculated a standard deviation if they have taken a statistics class. However,
many people who know how to calculate a standard deviation do not really appreciate the information it provides.

To begin, let’s think about what the phrase standard deviation means. Other words that might be substituted for the word standard include average, normal, or usual. The word deviation means to diverge, move away from, or digress. Putting these terms together, we see that the standard deviation means the average movement away from something. But what? It is the average movement away from the center of the distribution—the mean.

The standard deviation, then, is the average distance of all of the scores in the distribution from the mean or central point of the distribution—or, as we shall see shortly, the square root of the average squared deviation from the mean. Think about how you would calculate the average distance of all of the scores from the mean of the distribution. First, you would have to determine how far each score is from the mean; this is the deviation, or difference, score. Then, you would have to average these scores. This is the basic idea behind calculating the standard deviation.

Let’s use these data to calculate the average distance from the mean. We will begin with a calculation that is slightly simpler than the standard deviation, known as the average deviation. The average deviation is essentially what the name implies—the average distance of all of the scores from the mean of the distribution.

\[ X - \mu \]

Then we need to sum the deviation scores. Notice, however, that if we were to sum these scores, they would add to zero. Therefore, we first take the absolute value of the deviation scores (the distance from the mean, irrespective of direction). To calculate the average deviation, we sum the absolute value of each deviation score:

\[ \sum |X - \mu| \]

Then we divide by the total number of scores (N) to find the average deviation.

Although the average deviation is fairly easy to compute, it is not as useful as the standard deviation because, as we will see in later modules, the standard deviation is used in many other statistical procedures.

The standard deviation is very similar to the average deviation. The only difference is that rather than taking the absolute value of the deviation scores, we use another method to “get rid of” the negative deviation scores—we square the deviation scores.

The formula for the standard deviation is:

\[ \sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}} \]

Notice that the formula is similar to that for the average deviation. We determine the deviation scores, square the deviation scores, sum the squared deviation scores, and divide by the number of scores in the distribution. Lastly, we take the square root of that number. Why? Squaring the deviation scores has
inflated them. We now need to bring the squared deviation scores back to the same level of measurement as the mean so that the standard deviation is measured on the same scale as the mean.

If, however, you are using sample data to estimate the population standard deviation, then the standard deviation formula must be slightly modified. The modification provides what is called an “unbiased estimator” of the population standard deviation based on sample data. The modified formula is:

\[ s = \sqrt{\frac{\sum (X - \mu)^2}{N - 1}} \]

s = unbiased estimator of population standard deviation
X = each individual score
\( \bar{X} \) = sample mean
N = number of scores in the sample

The main difference is in the denominator—dividing by \( N - 1 \) versus \( N \). The reason is that the standard deviation within a small sample may not be representative of the population; that is, there may not be as much variability in the sample as there actually is in the population. We, therefore, divide by \( N - 1 \), because dividing by a smaller number increases the standard deviation and thus provides a better estimate of the population standard deviation.

\( \text{sd(INCOME)} \) #standard deviation of INCOME

\( \text{var(INCOME)} \) #variance of INCOME

The Variance is the square of the standard deviation