Video: “Measures of Spread” (8:50 min)

Measures of Spread (0:04):

Whenever we meet a new data set, we calculate a mean and median to get a sense of where the data is located. We also need measures of spread. In this video we will explain how to calculate some common measure of spread including average absolute deviation, median absolute deviation, median, and interquartile range. We will give some geometric interpretations of these measures and talk about what they tell us about the data. Let’s talk about how we can calculate spread. One way to think about it is that the data points are clustered around their mean. We want to know how far apart the data points are on average from the mean. In other words we use the mean as a prediction and we calculate the average error, and use the mean to predict the data points. Let’s calculate average error for are example. The table shows the data points in the left column and the difference between the data and the average in the right column. For example, the error in using the mean of 1 to predict 4 is 4 - 1 which is 3. When we compute the average of these errors we find the result is 0. That is not a good estimate of spread, clearly the mean does not exactly predict these points. The problem here is that some of the errors here are positive and some are negative, and the effects cancel. We’ll have to try another approach.

Average Absolute Deviation (1:28):

One way is to take the absolute values before averaging, this is called the average absolute deviation or AAD. Let’s look at the example. We add a third column to the table where we take the absolute values of the errors. Now when we compute the average we find the deviation is 2, which is a much better estimate. Here is the math formula for AAD, it simply says we average the absolute errors. In matlab, the average absolute deviation is computed with the mad function. MAD you say? Why not AAD? Well, some people refer to average absolute deviation as mean absolute deviation. In a way this is unfortunate because MAD also refers to median absolute deviation which we will discuss next. Notice in this case that the mad function has a second argument of 0. If the second argument is 0, it’s the average absolute deviation. If it’s 1 then its the median absolute deviation.

Median Absolute Deviation (2:36):

So let’s calculate the median absolute deviation which we will call MAD. The strategy is to use the median as the measure of central tendency and to summarize the errors by taking the median of the absolute errors from the median. Let’s calculate the mad for the example. The median of these data points are 0. When we produce are table, we calculate errors by subtracting the data points from the median rather than from the mean. In order to take the median of the last column, we are going to have to sort first. The median is just the middle value, so the median absolute deviation is just 1. In matlab, median absolute deviation is computed using the mad function with the second argument being 1.

SD (3:20):

The average and median absolute deviations are calculated using the absolute values or the errors. Another common measure of spread, the standard deviation, uses the squares of the

errors instead. Let's calculate the squared errors for are example. Notice the points that are farther from the mean are amplified when they are squared. The average of the squared errors or the mean squared error can be referred to as the variance. To calculate it we simply average the squared errors and it's 22/5 or 4.4. The variance however is not in the right units because we squared the error, so you must take the square root of the variance to get it in the proper values. This is called the standard deviation or root mean squared error. That is the standard deviation is the square root of the variance. We calculate the standard deviation in matlab by using the std function.The standard deviation of the sample has a second argument of 1. 1 you say? What does matlab do when the second argument is 0?

Standard Deviation (SD pop est) (4:23):

This brings us to some fundamental ideas from statistics, populations and samples. Often when we make a group of measurements we are not interested in just those values. We want to predict what happens in general. In other words we make a small number of measurements of the sample and try to make predictions about a population in general. The ideas of population and sampling are the foundation of statistical analysis, it's a complicated topic. We'll explore the ideas of population and sampling in more detail later in the course for now its fine to say that the sample standard deviation is not a good estimate of what the population standard deviation is. Luckily, a small correction factor can improve this estimate. Let's review are original calculation of variance and standard deviation first. They are based on the squared error. Variance is just the average of the squared error or mean squared error. Unfortunately the variance of the sample doesn't generalize. It consistently underestimates the population variance, that is it is biased towards the lower side. We can get a better estimate of the population variance by dividing by n -1 after summing the squares. This is called the unbiased estimator of population variance. It's easy to calculate, we simply add up the squared errors and divide by n-1. In this case that is 4 and we get 5.5. The unbiased estimator of the population standard deviation is just the square root of this value. We compute the estimate of the population standard deviation in matlab by calling the std function with the second argument being 0.

Standard Deviation Summary (6:10):

There are a lot of variations here, so we better summarize. The sample variance is just the mean squared error of estimating individual data points in the sample by the mean of the sample. You can compute the sample variance in matlab by calling the var function with a second argument of 1. The sample standard deviation is just the square root of the sample variance. You can compute it in matlab by calling the std function with a second argument of 1. However, when you want to use your data, the sample to make more general predictions you should probably use the population estimates. These estimates use a factor of n-1 rather then n in the formulas. You can calculate these estimates in matlab by calling var or std with a second argument of 0.

Interquartile Range (7:03):

The last measure of spread we are going to be looking at is the interquartile range or the IQR. The idea is we begin by ordering the data values from smallest to largest. The 25th

percentile is the value such that 1/4th of the values are smaller. The 75th percentile is the value such that 3/4ths of the values are smaller. The IQR is simply the 75th percentile - the 25th percentile. Sometimes the 25th percentile is called Q1 and the 75th percentile is called Q3. In matlab, we compute IQR by calling the IQR function. IQRs are often reported in standardized testing.

Final Word (7:47):

Let's summarize what we have done. We have introduced 4 different ways of calculating spread: average absolute deviation which is AAD or MAD for mean absolute deviation, median absolute deviation, standard deviation and its population estimate, and interquartile range. The unbiased estimator of standard deviation is the most common, but both SD measures are sensitive to outliers because we are squaring the errors. Squaring amplifies larger values, so they make a larger contribution to the overall average. MAD and IQR are the least sensitive to outliers. If you have data with outliers, you should consider using them in addition to standard deviation. Finally, measures of spread are often displayed as the wings of error bars with the measure of central tendency being the central point of the error bars. Calculating measures of central tendency such as mean and median as well as measures of spread should be your standard operating procedures when encountering new data sets.