Slide 1 – Title Slide

Hello and welcome to Week 5, Part 5 of EGM101: Descriptive Statistics. In this lesson, we'll continue the theme of summarizing our data by talking about the different ways we can measure the "middle" and "spread" of the data.

Slide 2 – Summarizing Data

So far this week, we've seen a number of different ways that we can summarize or view our data, such as stem-and-leaf plots, tally charts, frequency tables, and histograms or frequency polygons.

Despite the advantages that these offer us, it can still be difficult to make effective comparisons between datasets.

For example, if we wanted to answer a question like is there a difference between these two samples? Or is the population changing over time?, it would be kind of cumbersome to try to make these comparisons using tally charts or stem-and-leaf plots.

What we want, instead, is a way to quickly compare two sets of data – and that's where descriptive statistics really proves to be useful.

Slide 3 – Central Tendency

So, if our goal is to describe our dataset using a single representative value, what can we do?

Well, one way that we can do this is by looking at the "central tendency" of the data, which tells us where the middle of the data are, or where the data are clustered. Another term you might run into is "measures of location", or more likely, an average.

One thing to keep in mind here, though, is that all of the measures of central tendency that we're going to consider are "averages" – it's important to specify what you mean when you say "average", because as we will see, these different measures can end up telling us very different things.

Slide 4 – The Arithmetic Mean

We'll start off by considering the "arithmetic mean", which is what most people think of when they say "average."

Let's say we have some data of n values, which we'll denote "x", and these values are organized into a list of values. We can think of individual value of as having its own identifying value – sort of like the row number in a table. So, the first value of is x_1, the second is x_2, and so on, all the way up to the last value, x_n.

The way that we calculate the arithmetic mean of x, denoted as "bar", or with a line over it, is by taking the sum of each of the individual values of x, and dividing by the total number of values. The way that you read this formula is: "bar is the sum of the values of x, from inde1 to n, divided by n."

Don’t feel intimidated by the notation here – this symbol, which is the uppercase greek letter sigma, is just how we write “sum” in mathematical notation, as a way of saving space. It means that we don’t have to write out all of the individual values of in the formula. I’m not going to expect you to memorize this formula, but you should have some idea of what it’s telling you, because this sort of notation does tend to come up a lot.

The way that we calculate some descriptive statistics differs slightly depending on whether we’re working with a sample or a population. The reason for this is something that we’ll talk a bit more about as we make our way through the module. For a sample, we only have small n values of to use in calculating the mean, but for the population, we use the entire population. Here, we use a capital N to indicate that we’re using the entire population, and we denote the population mean with the Greek letter μ.

If we stick with our fish lengths data that we’ve used so far, we can calculate the sample mean using the first formula here. If we plug in the individual values for here in the numerator, and the sample size of 25 in the denominator, the calculation would look like this. The numerator, which is just the sum of all of the values of x, ends up being 15,024, and when we divide this by 25, we end up with 600.96, which we might round to 601 mm.

One of the main benefits of the arithmetic mean is that it is fully representative of the data, because we use all of the values of the data in calculating the mean. It’s also very easy to calculate – just add all the values together and divide by the total number of values – and it’s fairly easy to use when we’re doing algebra.

On the other hand, the arithmetic mean is very sensitive to large outliers in the data – that is to say, values that are very unlike the other values in the data; and, the value of the arithmetic mean might not be a possible value of our data. For example, if we have discrete data such as “number of children per household”, the mean number of children might not be a whole number.

Slide 5 – Mode

The next measure of central tendency that we’ll discuss is the mode, or “most frequent value.” If we look at an example dataset here, we can calculate the arithmetic mean like we saw on the previous slide, and we get a value of 5.9.

Looking at the values of the data, we see that the value 8 appears 3 times, while 7 and 4 appear 2 times, and the other values only appear once. So, for this dataset, 8 is the modal value.

The mode is straightforward to find for discrete observations – all we have to do is count how many times each value appears in the data. It’s less straightforward to calculate for grouped or continuous data, and we won’t get into the specifics of how to calculate it here. When we look at a histogram or frequency plot, though, we can see it easily – the mode is the peak values of the histogram.

The main advantages of the mode are that because we’re counting values that are present in the dataset, the mode is always a possible value – we won’t end up with any partial numbers of people with the

mode. The mode is also unaffected by extreme values – we could replace the first value here with minus 10000, and it wouldn't change the way that we calculate the mode.

Unfortunately, as this graph shows, the mode is not necessarily unique – we could have two modes, in which case we could say that the dataset is "bimodal", and we can even have more than two modes, in which case the dataset is "multi-modal."

The mode doesn't necessarily exist, either – if we have a dataset where all of the values are unique, for example, we wouldn't be able to calculate the modal value.

Slide 6 – The Median

The last measure of central tendency that we'll consider is the median, or the exact middle, of the distribution. Exactly 50% of our data has a value less than the median, and exactly 50% of the data has a value greater than the median.

The median is also straightforward to calculate, even for grouped data – though we'll stick with the simple example on the slide for now. We start by sorting the data from smallest to largest. If we have an odd number of data values, we take the value that is exactly halfway through the sorted values. For this example data with 9 values, that's just the 5^{th} value, which corresponds to 6 in this example.

If we have an even number of values, it's slightly more complicated: we take the arithmetic mean of the middle two values. In this second example here, we add 6 and 7 together, divide by 2, and get a value of 6.5.

For our fish lengths, we get a value of 601 mm, almost exactly the same as we got for the arithmetic mean.

Because the median only depends on the middle values of the data, it's less affected by extreme values than the arithmetic mean. For example, we could switch one, or even two of the values of 8 here to be 8 billion, and the median value would stay exactly the same.

We can also calculate the median without measuring all of the values, which can be quite handy.

Unfortunately, like with the mode, the median might not be a possible value of our dataset, illustrated by the even example here. Because it's not based on all the observations, it's not fully representative of the data, either – though, as we will see, it can be a better choice for describing datasets where there are lots of extreme values.

Slide 7 – Dispersion

One problem with only considering central tendency is that it doesn't necessarily say much about the data that we're working with. For example, two datasets can have the same mean, or median, or mode, and be very different.

Look at these two frequency distributions, which both have an arithmetic mean of 0. The top dataset has values that range from -6 to +6, while the dataset on the bottom only has values that range from -2

to +2. You can also see that the “peak” of the dataset on the bottom is much higher than the dataset on the top, while the data on the top look much more spread out.

All of this is to illustrate that the spread of the data, also known as the dispersion, is also important for us to consider. For the rest of this lesson, we’ll look at ways that we can describe the dispersion of our data.

Slide 8 – Range

The first measure of dispersion that we’ll consider is the simplest measure of dispersion that we have: the range, which is just the difference between the largest and smallest value in the dataset.

In this example, the largest value is 8, and the smallest value is 2, so the range is equal to 8 minus 2, which is 6.

The biggest advantage of the range is that it is simple to calculate, and to understand.

Unfortunately, even more so than with the arithmetic mean, extreme values can be very misleading, which we can illustrate with a simple example.

Let’s say that we changed one of these 8s to a 100 – the new range is 100 minus 2, or 98. But, these datasets are almost exactly the same, with only a single value different. The range doesn’t provide us with a very useful way to distinguish between these two datasets.

Slide 9 – Interquartile Range

A better option that we have available is the interquartile range, or IQR. As the name suggests, a “quartile” is just one fourth, or 25%, of the data. Q1 is the value that distinguishes the smallest 25% of values; Q2, also known as the “median”, splits the data into two equal halves; and Q3 is the value that distinguishes the largest 25% of the data.

Looking at our example data again, the median value is equal to 6.5, as we calculated earlier.

We can find Q1 by looking for the middle value of the dataset that’s less than the median – in this example, because we have 5 values less than the median, Q1 is the 3rd value here, equal to 4.

We find Q3 in the same way: it’s the middle value of the dataset that’s larger than the median – here, equal to 8.

The interquartile range is just the difference between Q3 and Q1, or the middle 50% of the data. In this example, the IQR is 8 minus 4, or 4.

You might also see something called the semi-interquartile range, or “quartile deviation” – this is just the interquartile range divided by 2.

The advantages of the IQR are similar to the advantages that we saw for the median: we can use it even if we don’t have exact values, like if we have grouped frequency distributions; it’s also less affected by extreme values, especially compared to the range.

One of the big drawbacks to the IQR is that it's less useful for mathematical manipulation, which means it doesn't show up in a lot of the more advanced topics that we'll be seeing later on. But, as a way to describe the spread of your data, it's a very useful thing to have.

Slide 10 – Mean deviation

The goal with the next three measures that we'll look at is to describe how far away from the "middle" of the dataset most of the values are. So, we start by calculating the arithmetic mean value, and subtracting that from each of our values.

On the plot here, the mean value of 5.9 is plotted as a dashed line, and each of the values is plotted as a dot. If we subtract the mean value from each of these, we end up with the following values. We can see that the difference to the mean ranges from a low value of -3.9, up to a high value of 2.1. One thing that we could then try to do is calculate the sum, or the mean, of these differences.

Now, one problem that we have here is that the total deviation from the mean is always going to be zero. That is, if sum up each of these values, we will always get 0 (and so the mean value will also be 0), because all of the values above the mean are balanced by values below the mean.

So, one thing that we can do is take the absolute value of the differences – we just ignore the minus sign here, and then calculate the mean of the differences. That's the idea behind the mean deviation – we're just calculating the arithmetic mean of the total distance away from the mean, ignoring whether that difference is positive or negative. In this formula, the vertical bars here indicate that we're taking the absolute value of whatever is inside of the bars.

For our example here, if we take the mean of the absolute value of these differences, we end up with a mean deviation of 1.72. This means that on average, each of our values is 1.72 away from the mean value.

The mean deviation is useful for comparing the variation of different datasets, because the values tell us about the average distance away from the mean, rather than the actual values of the data.

On the other hand, just like the arithmetic mean, it's very sensitive to extreme values. It can also be more difficult to calculate the mean deviation when we have grouped data, at least by hand.

Slide 11 – Variance

The variance works in much the same way as the mean deviation. Here, though, instead of the absolute value, we instead square the differences (or raise them to the power of 2).

Like the arithmetic mean, we have two different formulas here – for a population, variance is denoted using the lower-case greek letter sigma, raised to the power of 2 (or "sigma squared"). When calculating the population variance, we're comparing the values to the population mean, μ, and we divide the sum of the differences by the population size, capital N.

Squaring the differences means that more "weight" on larger differences, as you can see in the table here. We're not going to go into the details for why, but squaring the differences also makes some kinds

of analysis easier, which is why variance and the next measure of dispersion that we'll look at tend to be used far more frequently than mean deviation.

For a sample, we denote the variance as lower-case s, raised to the power of two (or "s squared"). Most importantly, though, we divide by n minus 1, rather than the sample size. The reason that we do this is that it's an attempt to correct for sampling bias – by making the denominator smaller, the variance is slightly larger for the same values of n.

We can see this by comparing the mean deviation, population variance, and sample variance, all calculated from the same set of values here. The largest of these values is the sample variance at 4.32, followed by the population variance at 3.89.

One of the big advantages of the variance is that, like the mean deviation, the direction (±, or above/below the mean) doesn't matter – what matters more is how far away from the mean each value is. Again, this is outside of the scope of this module, but the variance can also be easily mathematically manipulated, which means that it tends to show up in a lot of places.

Unfortunately, the additional weight given to the larger deviations means that the variance tends to be more sensitive to outliers. It also has different units than the data, which means it's not immediately clear what those values represent in terms of our data.

Slide 12 – Standard Deviation

This brings us to the last measure of dispersion that we'll consider, the standard deviation. The fact that the variance has different units can be a big problem. Fortunately, there's a simple solution: take the square root. This gives us something called the standard deviation. For the population, the standard deviation, denoted using a lower-case sigma, is just the square root of the population variance; we have the same thing for the sample standard deviation, denoted using a lower-case s.

If we then compare the standard deviation to the mean deviation, we see that these measures give us similar values, with the two standard deviation values being slightly higher due to the additional emphasis on the larger differences.

The standard deviation has the same advantages as the variance, with the additonal benefit of being easier to compare to our data, since they now have the same units.

Unfortunately, simple solutions do not always solve all of our problems: we still have the problem of more "weight" being placed on larger deviations, and dividing by n – 1 for the sample isn't as good of a correction for sample bias, since we're now dividing by the square root of n – 1. Despite these issues, the standard deviation is probably the most commonly-used measure of dispersion, and we will be seeing it a lot more as we continue in the module.

Slide 13 – Coefficient of Variation

Before wrapping up, I want to mention something called the coefficient of variation. Sometimes, we might want to compare data that have different units – for example, we might compare economic

outputs from countries that use different currency. But, what's the right way to compare dollars, pounds, and euros?

The coefficient of variation gives us a dimensionless, or unitless, measure: to calculate it, we divide the standard deviation by the arithmetic mean.

This is something that comes up in fields like chemistry or engineering, where it is used to describe the repeatability or precision of laboratory tests, but especially in economics, where it is used to compare economic inequality.

The big advantage of the coefficient of variation is that it is dimensionless, which means we can compare quantities with different units, like currency.

On the other hand, for small means, the coefficient of variation is very large and tends toward infinity as the mean approaches zero; it's also not the best measure of certainty that we have available, and it's sensitive to extreme values.

Slide 14 – Summary

In this lesson, we've discussed how one of our goals is to describe datasets using single values, to help facilitate comparisons.

Measures of central tendency are used to describe "where" the data are, or where the middle of the data is located;

Measures of dispersion help us understand how "spread out" the data are.

As you have now seen, there are many, many ways to achieve our goal of describing our data using a single number, and they all have their own unique advantages and disadvantages – because of this, very often, we use multiple measures instead of a single measure.

Slide 15 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Illowsky and Dean, Chapters 2.3, 2.5, and 2.7; Caswell, Chapters 7 and 8; and Weiss, Chapters 3.1 – 3.2 and 3.5.

I've also included a link to a 2004 paper that discusses the differences between the mean deviation and standard deviation in more detail, and makes an argument that for many applications, the mean deviation is a better statistic to use. It's written at a fairly accessible level, so if this is something that interests you, I recommend having a look.

Finally, I've included links to three youtube videos here – the first, on means and medians, discusses some of the ways that extreme values can have an impact on how we describe our data and draw conclusions from it; the other two, from Khan Academy, provide a bit more reinforcement for how to calculate some of these measures by hand.

That's all for this lesson – I hope you found it interesting, and you have any questions, please don't hesitate to e-mail me or post in the discussion forum on blackboard. Bye!