Slide 1 – Title Slide

Hello and welcome to Week 5, Part 6 of EGM101: Data Distributions. In this lesson, we'll discuss what the shape of our data can tell us.

Slide 2 – Frequency Distributions

As we saw earlier this week when looking at histograms, the shape of the histogram can tell us something about the data – by looking at the histogram, we see where the approximate middle of the dataset is, we can see how spread out the values are, and so on.

Looking at the frequency distribution of different samples or datasets can also help us compare them, and eventually, we will see how we can use some of this information to help us draw conclusions about the data that we have: about whether a particular medical treatment has a noticeable effect on outcomes, or what the outcomes of different lab experiements are, or even if there's been some kind of fraud or tampering.

Slide 3 – Symmetry

For now, though, we'll start by introducing a few more terms to keep in mind, starting with symmetry. If we have a symmetrical distribution, then we can draw a line through the middle of the distribution, and we should see that the left side of the distribution "mirrors" the right side – like in this example here.

If a distribution is perfectly symmetrical, the mean and median of the dataset are the same. In a unimodal distribution, like this one here, then the mode is equal to the mean and the median – in the plot here, you can see that this is the case, with the mode, median, and mean all equal to 7.

In a multimodal distribution, the mode(s) won't be equal to the median or mean – the example shown here has a mean and median of 11.5, but modes equal to 7 and 16. The distribution is still symmetrical, because we can draw a line through the middle and see how the left side mirrors the right side, but the symmetry is not perfect.

If we don't have a symmetrical distribution, then we say that the distribution is skewed – and this is something that we'll come back to in a few slides.

Slide 4 – The Normal Distribution

There are a number of named distributions out there, and we'll look at a few more in Week 7 when we discuss some topics in probability. The normal distribution, though, is important enough that we're going to introduce it earlier on. The normal distribution is probably the most important distribution in statistics/probability, and it's often something that we use to approximate unknown distributions.

You may also have heard of this as the “bell curve” or “bell-shaped curve”, but it’s important to note that there are actually a number of bell-shaped distributions – just because a distribution has a “bell shape” does not mean that it’s a normal distribution.

The normal distribution is perfectly symmetrical – the mean, median, and mode coincide at the peak of the distribution, plotted here as a dashed line.

We’ll have a lot more to say about this distribution as we go through the module, but just know that any time you see the word “normal” in a statistics or probability context, this is the picture that you should have in your head.

Slide 5 – Dispersion

As we’ve touched on briefly, we can infer the dispersion, or spread, of a dataset based on the shape of the distribution.

In this example, we can see how changing the value of the standard deviation changes the shape of these normal distributions – the red line, with sigma equal to 0.5, has a much taller, narrower peak. As we increase the value of sigma, we see how the peak of the distribution gets lower and lower, and how it spreads out as the curve gets wider.

In general, if a distribution is very narrow with a tall peak, it means it has low dispersion, and as the peak gets lower and more spread out, it means that we have more dispersion in the data.

Slide 6 – Positive Skew

When I introduced the term “symmetrical”, I also said that non-symmetric distributions are skewed.

If a distribution has what is known as “positive skew”, it means that the so-called “tail” of excess values is on the right side of the peak (making the distribution more positive). In the example shown here, the shaded area represents the “extra” values on top of the symmetric distribution, or the “tail”. “Positive skew” is also known as “right-skew”, because the tail is to the right, or “left-leaning”, because it looks like the whole distribution is leaning toward the left.

In a positively-skewed dataset, the mean is typically greater than the median, which is greater than the mode. In this example, we can see that the line representing the mean is farthest to the right, the mode is farthest to the left, and the median is somewhere in between them.

Slide 7 – Negative Skew

The opposite of positive skew is helpfully known as “negative skew.”

If a distribution has what is known as “negative skew”, it means that the so-called “tail” of excess values is on the left side of the peak (making the distribution more negative). “Negative skew” is also known as “left-skew”, because the tail is to the left, or “right-leaning”, because it looks like the whole distribution is leaning toward the right.

In a negatively-skewed dataset, the mean is typically less than the median, which is less than the mode. In this example, we can see that the line representing the mean is farthest to the left, the mode is farthest to the right, and the median is somewhere in between them.

Slide 8 – Measures of Skewness

Sometimes, we might want to be able to compare the skewness of distributions. In the example here, we can see that the distributions in the top row are less skewed than the distributions in the bottom row – but how can we quantify this?

There are two common ways you'll run into for estimating how skewed a dataset is. The first is Pearson's first coefficient, also known as the mode skewness. This is calculated as the difference between the mean and the mode, divided by the standard deviation. As you can probably guess, the sign depends on whether we have positive or negative skew – positive skew means a positive coefficient, and negative skew means a negative coefficient.

The second method is known as Pearson's second coefficient, or the median skewness. This is calculated as three times the difference between the mean and median, divided by the standard deviation. Just like with mode skewness, positive skew means a positive value here, and negative skew means a negative value.

The values of these coefficients range from -1 to 1 for mode skewness, or -3 to 3 for the median skewness. If our data are perfectly symmetrical, you can see how both the first and second coefficient will be equal to zero – but not necessarily if we have a multi-modal distribution.

In general, these values will be similar, but as you can see from the figure here, they aren't completely interchangeable. For the datasets considered here, the mode skewness gives a much higher skew value for the bottom row compared to the median skewness, even though the values are basically the same in the top row here.

Slide 9 – Summary

In this lesson, we've seen how the shape of the data or frequency distribution is important, and tells us something about the data.

If our data our symmetrical, it means that we can draw a line down the middle, and the two halves will look the same. If the data are perfectly symmetrical, it also means that the mean and median values are the same.

If the data are not symmetrical, though, it means that they are skewed. In this case, we can use the relationship between the mean, median, and mode to learn more about how far from symmetrical the distribution is.

Slide 10 – Additional resources

You can read more about the topics we've discussed here in the textbooks – Illowsky and Dean, Chapters 2.2 and 2.6; Caswell, Chapters 7 and 8; and Weiss, Chapter 2.4.

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!