Slide 1 – Title Slide

Hello and welcome to Week 7, Part 6 of EGM101: The Central Limit Theorem. In this lesson, we'll build on what know about the law of large numbers, and learn about another of the most important foundation stones of probability theory – the Central Limit Theorem.

Slide 2 – The Central Limit Theorem: Setup

Let's assume that we have a population with a known population mean, mu, and a finite variance, denoted sigma squared. Next, let's say that we take a random sample, Si, of size n from the population.

For the ease of fitting it on a slide, let's say that n is equal to 5, so our sample looks like this. Next, we can take a different sample of the same size, and it might look something like this. And we can keep going like this, taking all the way to big N number of random samples like this.

Next, we can calculate the sample mean of each of our individual samples, denoted X-bar n, to indicate that it's a mean of small n values.

Question: what do we know about the sampling error of each of these samples? That is, do we know anything about how far away from the population mean each of these sample means is?

It turns out that we do know something – at the beginning of the week, when we learned about the law of large numbers, we learned that as we increase the size of the sample, the sample mean gets closer and closer to the expected value – that is, the sample mean is going to get closer and closer to the population mean.

What about if we take lots of different random samples of a given size, and calculate their sample means? What does the distribution of the sample means look like?

Slide 3 – The Central Limit Theorem (CLT)

This figure shows how the distribution of sample means changes, based on the size of each individual sample. The gray bars in this figure show the probability distribution of sample means of a given size, and the red line shows the normal distribution based on the mean and standard deviation of the sample means.

So, when we take samples of size 1, we see that the spread of the sample means looks almost uniform – the bars are almost all the same height, and they don't really match the normal distribution at all. As we increase the sample size, though, we notice two things – first, the distribution looks more like the normal distribution; and second, the peak of the distribution gets higher, and the spread of values gets narrower and narrower – for this example, the standard deviation of the sample means drops from 23.18 to 5.19 as we increase the sample size from 1 to 20.

It turns out, this is exactly what the Central Limit Theorem tells us. As we increase the sample size, the distribution of the sample means approximates a normal distribution, with a mean equal to the population mean, and a variance equal to the population variance.

In other words, as we take larger and larger samples, the distribution of the sample means starts to look more like a normal distribution, as we see here, but also, it has a mean and variance that looks like the population.

Most importantly, though – this does not depend on the distribution of the population (mostly). Even if the population is not normally distributed, even if it's skewed (to a point), this fact about the distribution of the sample means is true. By "up to a point", I mean that for skewed data, we might have to take larger samples to get this to work.

Slide 4 – How Can We Spot Normal?

Before we look further at what this means and how important it really is, let's stop to talk about normal. How can we determine whether or not a distribution is actually normal or not? We've already looked at one example: plotting a normal curve on top of a histogram. We can look at the histogram, and look at the normal curve, and see how different they are. As you might guess, though, this is neither the only way, nor the best way, to determine whether a distribution is normal.

Instead, we can use something called a Q-Q plot, or a quantile-quantile plot. In a Q-Q plot, we plot the quantiles of two datasets against each other. On the plot here, the actual quantile value is plotted on the y axis, and the theoretical value for a normal distribution is plotted on the x axis. If the two distributions are similar, then we should see that all of the points lie on the red line shown here.

And, just like we saw by plotting the normal distribution on top of the histogram on the previous slide, when the sample size is equal to 1, the Q-Q plot tells us that the distribution is not normal. Instead of a straight line, we have a nice s-shaped curve, which indicates that values far away from the mean are more frequent, or more probable, than they should be in a normal distribution. As we increase the sample size, we see that the lines get straighter and straighter, confirming what we saw on the previous slide.

Slide 5 – Skew and Kurtosis

Another way that we can judge whether a dataset is "normal" is using statistics such as skew and kurtosis. We've seen skew before, in week 5 – this tells us how much the data "leans" compared to a symmetrical distribution – if we have positive, or right, skew, it means that the distribution looks like it's leaning to the left; if we have negative, or left, skew, it means that the distribution looks like it's leaning to the right.

Kurtosis, on the other hand, is a new one. This is a measure of how much of the distribution is located in the "tails" off to the sides vs the peak in the middle of the distribution, relative to a normal distribution.

If the kurtosis value of a dataset is positive, it means that more of the values are in the peak relative to the normal distribution – the peak is "heavier"; it's negative, it means that more of the values are located in the tail – in other words, it means that the tails are heavier.

For our distribution of sample means, we can see that the skew is small and fairly similar for each value of n – the distribution for each sample size is fairly symmetrical. But, we can see that the value of kurtosis goes down by quite a lot as we increase the sample size – it approaches zero, which again confirms what we saw previously: as we increase the sample size, the distribution of the sample means becomes more normal.

Slide 6 – Standard Error of the Mean

As we have seen, as we increase the sample size, we still have some dispersion of the sample means around the population mean. If we think of each sample mean as an estimate of the population mean, we still have some uncertainty in our estimate.

But, as n increases, the mean of the sample means gets closer to the population mean, and the dispersion decreases. In other words, as we increase our sample size, the uncertainty in our estimate of the population means decreases. We can quantify this using something called the standard error of the sample mean, denoted as “sigma x bar”, equal to the population standard deviation divided by the square root of the sample size. As you can see on this graph, as we increase the sample size, the standard error drops quickly at first, but as the sample size gets larger, the drop in the standard error is smaller and smaller.

Slide 7 – Confidence Intervals

Most of the time, we don’t actually know the population mean – as we have discussed, this is very often something that we’re trying to estimate by sampling. So, how can we use what we’ve learned in this lesson to help estimate the unknown population mean, and what can we say about the uncertainty in that estimate?

By the empirical rule, we know that for 95.45% of the random samples of a given size that we take, the sample mean will be less than two times the standard error away from the population mean. Along the same line, 95% of the random samples that we take will have a sample mean within 1.96 standard errors of the population mean.

In other words, if we take a random sample, and calculate its mean, x bar, there is a 95% probability that the population mean is within the interval from the sample mean minus 1.96 times the standard error to the sample mean plus 1.96 times the sample error. This also means that there is only a 5% chance that the population mean is outside of this interval.

This is an example of a confidence interval – a way of expressing the uncertainty of our estimate of the population mean. By calculating the sample mean and the standard error of the mean, we can estimate, with 95% confidence, that the population mean is within the range x bar plus or minus 1.96 times the standard error. And, using the empirical rule, we can construct similar intervals for 90% confidence, 99% confidence, 99.5% confidence, 99.9% confidence, and so on.

Slide 8 – Consequences of the CLT

Next week, we will see how we can use hypothesis tests with the distribution of sample means of non-normally distributed data, just as long as the sample size that we use is “large enough”.

By “large enough”, we typically mean a sample size greater than about 30 – if we look at this plot of kurtosis vs sample size, we can see that after about 30 or so, the kurtosis doesn’t vary too far away from zero, meaning that the distribution is a good approximation of normal.

We have also seen that as the sample size increases, the standard error of the mean – one way that we can measure the sampling error – decreases. Another way of saying this is that the average of many measurements is more accurate than a single measurement – if we take the average of a large sample, it gives us a better representation of the “true” value than a single sample does.

Slide 9 – Summary

In this lesson, we’ve seen how the Central Limit Theorem tells us that the distribution of sample means from a population starts to approximate a normal distribution as we increase the sample size.

Most importantly, this means that with a “large enough” sample size, typically larger than about 30 or so, we can even treat non-normal data as if it were normally distributed.

We have also seen that we can use this fact to estimate sampling error, using the standard error of the mean, and that the average of many individual estimates is more accurate than a single estimate.

Slide 10 – Additional resources

You can read more about the topics we’ve discussed here in the textbooks – Illowsky and Dean, Chapter 7; Caswell, Chapters 14.1 and 14.2; and Weiss, Chapters 7.2 and 7.3.

I’ve also included links to a pair of articles about the Central Limit Theorem and ways that we can assess whether a dataset is normal, and links to videos from Khan Academy that cover both the Central Limit Theorem, and the sampling distribution of the sample means.

Finally, I’ve included a link to an online simulator that you can use to simulate the distribution of sample means by changing the sample size, similar to the figures I showed earlier in the lesson.

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