In this section we will introduce a very important concept in hypothesis testing which is the idea of a central limit theorem.

>>>

The central limit theorem has a lot of application in hypothesis testing and in many ways makes hypothesis testing a relatively straight forward exercise. Supposing that we had a population, respondents to a survey and we wish to pick a sample of 500 respondents. Let's say that one of the attributes in the survey response is income.

Supposing we were picking random samples of size 500 from a population of 10,000 respondents and we looked at the average income in each sample. Let's also assume that the average of income in the population of the 10,000 respondents as a whole is $80,000. In every sample would we expect to see a sample income average of $80,000? If we are picking samples completely at random, then ideally most of those samples will be representative especially with the sample size of 500. But even though we are picking completely at random every sample may not be representative. In fact every sample may have a slightly different average.

We will expect a lot of the samples to have averages close to $80,000 in terms of income. But we do understand that some samples may have averages that are very different from 80,000, may be 75,000, and may be 95,000. Why does that happen? Because again random chance variation.

Even though we are picking samples completely at random, sometimes simply because of random variation not because we are picking the sample wrongly you may end up with the sample average that is different from the expected population average.

What happens if we do a frequency distribution of these sample averages? To understand let's simulate an example. In excel we can use a RANDBETWEEN function to generate a random number between 1 and 100. So if I say RANDBETWEEN (1,100) and I say ENTER then excel will generate a random number between 1 and 100. Now I am going to this for multiple columns. So this is the first number. If I drag this then essentially I am creating multiple random numbers and each cell has one random number between 1 and 100.

Imagine that this row is one sample. Imagine that each row is one sample. If you had accessed to a glass bowl with these numbers written on it, mix them up and you were picking numbers completely at random with replacement. If you pick this many numbers, this would be one sample.

Imagine that we had multiple samples like this. I am just going to drag this formula down all the way to many rows. Let's say we have a 100 samples. Remember each row is one sample and we have a 100 such samples. What happens to the sample averages? We can calculate sample averages using the AVERAGE function in excel. I will say AVERAGE (B2:BX2), this will give me the average across all the different numbers in that one sample which is every row.

If we are picking random numbers between 1 and 100, we would expect the sample averages to be around 50. In fact if you look at the sample averages which have now calculated and I am just going to recalculate the average because every time when we hit enter when we use a RANDBETWEEN function the numbers get automatically updated.

You can see that a lot of these sample averages are close to 50. However in between we will also see some sample averages that are not close to 50. For example 45 or 54. You can see that when we look at the sample averages, many of them may be close to 50 but some of them may be quite far away from 50. If I look at the frequency distribution of the sample averages, remember in a frequency distribution we are simply doing count.

Supposing that I am going to put these numbers here and I am going to count how many of my samples have averages of less than 30. How many have averages of between 31 and 35, 36 and 40 and so on. We can do that very easily using a histogram option in excel. So if you go to data, data analysis and choose histogram, excel will ask you for the input range. This is our input range which is all these sample averages and it will ask us for a bin range. The bin range is the bucket values for which we want to do the counts.

In excel, if you specify this bucket values 30, and 35 essentially what excel will do? It will count the number of sample averages below 30, then between 31 to 35, 36 to 40, and 41 to 45 and so on. So excel has calculated this sample averages for us.

If I want to look at a frequency distribution simply inserting a chart, what will that look like and this is where the central limit theorem comes in. The central limit theorem says that if you take multiple random samples from an underlying population and you look at the frequency distribution of the sample averages. That distribution will be normally distributed and in fact it will be normally distributed even if the underlying population data is not normal.

Let’s put this into distribution chart visualization you can see this is approximately normally distributed. Of course we have very wide bins, if we had change the bin sizes into narrower, we will see a much nicer normal distribution. The more samples you have, the larger the samples you will see more it will approximate a normal distribution. This is what the central limit theorem says.

>>>

If you take a multiple random samples from an underlying population and look at a distribution of sample averages, the sample averages will be distributed normally even if the underlying population is not normal. In fact mathematically the distribution of sample averages will be distributed normally with the mean equal to the population mean and a standard deviation equal to the standard deviation of the population/square root of sample size as long as sample sizes are at least 30.

To recap, the central limit theorem says irrespective of the underlying population distribution. When you pick multiple random samples from an underlying population with the sample size of at least 30. The distribution of sample averages will be normal even if the underlying population is not normal.

Remember in a normal distribution, you need to specify two parameters to correctly setup the normal distribution. One is the mean and the other is the standard deviation. This distribution that we are looking at is not a distribution of the population. It is a distribution of sample averages. What you think will be the average of the sample averages.

In other words each of these cells has the individual sample averages where every sample is one row. But if I took an average of these numbers, what would I expect. I would expect to see 50. Does that happen? Let's take a look. If I take an average of these numbers, these 100 sample averages I should ideally expect to see the number that is very close to 50 and you can see that.

In fact the population average will also be 50. You can see that the population average and the sample average are almost exactly the same. In other words the distribution of sample averages will have an average equal to the population mean. Remember it is the average of sample averages and that distribution of sample averages its mean will be equal to the population mean. What about the standard deviation?

If I were to calculate the standard deviation of the population and that is easy to do. We just take all the numbers and include it in the standard deviation calculation. We get a standard deviation of the population itself to be 28.8. Do you think the standard deviation of the sample means will be the same as the standard deviation of the population?

Again we can check this. It turns out that the standard deviation of the sample averages will be a lot lower than the standard deviation of the population. Why is that? Let's look at it intuitively. In this example, we were picking samples of size 500 from its population of 10,000.

Now supposing that you were picking samples of size 500 and someone picking at samples of size 40. In which of these two set of samples would you expect to see a greater variation in the sample averages. In samples of size 500 or in samples of size 40? Most of us intuitively understand that the variation of sample averages that you will see will be a lot higher in the samples of size 40 relative to the samples of size 500.

Therefore it has to be the case that the standard deviation of the averages of samples has to be inversionally proportional to sample size. The larger the sample, the more likely that you will see less variation in sample averages. The smaller the sample, the more likely you will see greater variation in sample averages and therefore the standard deviation of the sample averages will be equal to standard deviation of the population/square root of sample size.

That's why the standard deviation of the sample mean averages that you can see here is a lot lower than the standard deviation of the population. This will be standard deviation of the population and divided by the square root of sample size.

In our example sample size is 75. Every sample contains 75 data points. So if we take the population standard deviation and divide that by the square root of sample size, we should get the number that is close to the sample standard size that we just calculated. Let me try that and see. I have divided the standard deviation of the population by the square root of sample size.

You can see that this number standard deviation of sample averages is a lot close to the calculated standard deviation of the population divided by the square root of sample size. The reason that it is important to understand this is because remember we are dealing with the distribution of sample averages and we are saying that the distribution of sample averages is normal and if you have a normal distribution you need to know what is the mean of that normal distribution and what is the standard deviation of that normal distribution.

How does this help us with hypothesis testing? It turns out in hypothesis testing, we always deals with samples. We look at a sample outcome and we want to check, what is the probability of observing the sample outcome simply because of random chance? In order to calculate a random chance probability we need a sample distribution and because we are dealing with random samples it turns out that the averages of these samples will be distributed normally.

So even if you don't have access to the population you have no idea what the population distribution is like. You can still go ahead and use a normal distribution calculation for calculating probability of observing a sample average as extreme or more extreme than what is observed. In order to calculate that probability using a normal distribution you know what the mean and the standard deviations will be.

The mean of this distribution of sample averages will be the same as the population average. The standard deviation of this distribution of sample averages will be equal to population standard deviation/square root of sample size.

That is why in hypothesis testing very often people end up using normal distributions to calculate probability of observed outcomes. Because the central limit theorem allows us to use normal distribution even if the underlying population distributions are not normal or unknown.

>>>