Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

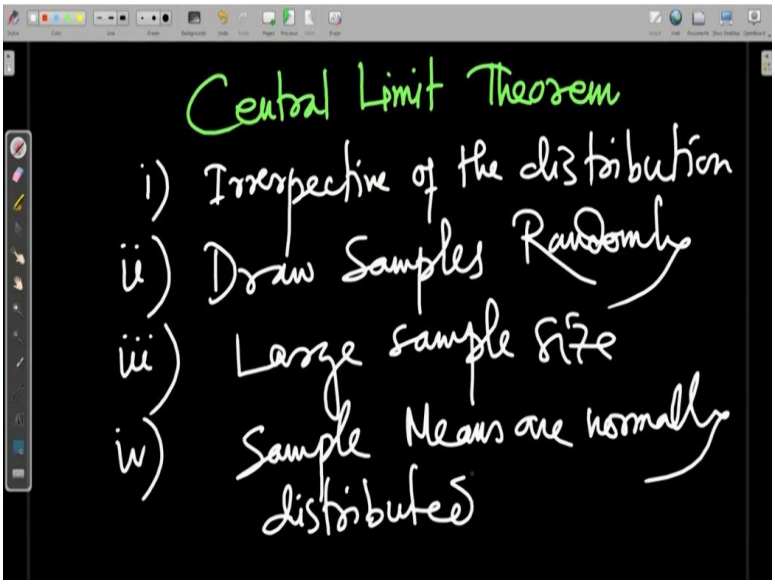
Module - 3
Lecture - 25
Central Limit Theorem

Hello and welcome back to the lecture on Applied Econometrics. So, we have been talking about probability distribution. In the previous lecture, we spoke about sampling. Now, in this lecture, we are going to talk about a very important natural law; we will see that this law is going to be a foundation of many theorems that we are going to deal with; and this law is central limit theorem.

This is essentially a natural law, and because it is a natural law, it is in all pervading application, be it in physics, be it in chemistry, mathematics, statistics of course, economics; everywhere we will see the application of this law. Now, what exactly is this law? So, I will just tell you few keywords which will help you to remember this law; and these keywords are, you can draw samples from any distribution; any distribution in this world, you draw samples; if you draw the samples randomly; the first condition is you have to draw the samples randomly; and second, the sample size has to be large.

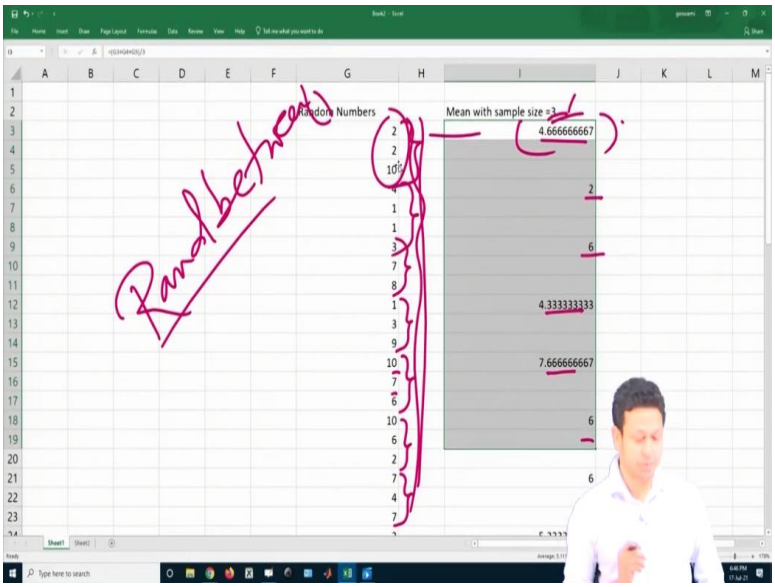
So, from any distribution, whatever be the nature of the distribution, you draw samples; you draw them randomly; and you have sufficiently large sample size. If you do that, and then you take sample means, what you will get is that, the sample means you will see are normally distributed. So, I will write down the key points.

(Refer Slide Time: 01:52)



So, first is that; let me use a different colour; any distribution; so, irrespective of the distribution; you can draw samples from any distributions. Then you draw samples; you draw them randomly; you have to have a large sample size. If you do that, the sample means that you get in this exercise, you will see, the sample means are normally distributed. So, let me actually show you what exactly; whatever just I have said, I will show you with some numbers. So, I actually have created an Excel sheet for you.

(Refer Slide Time: 02:59)

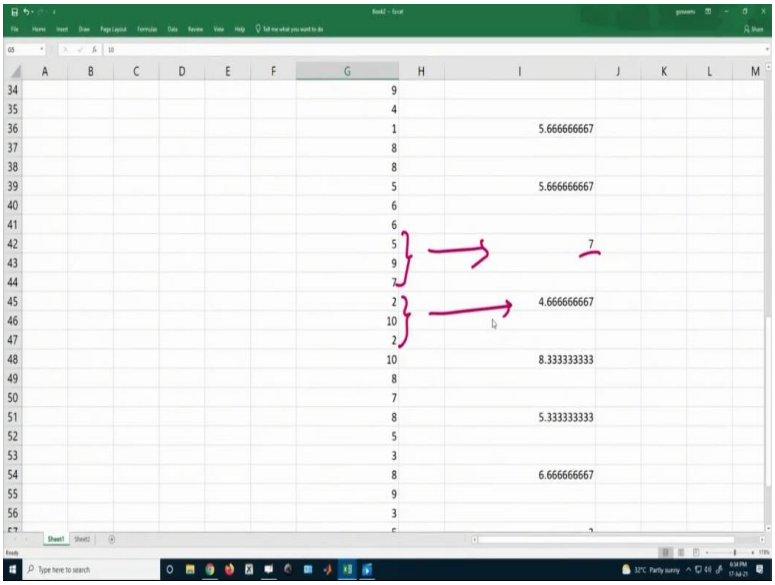


So, I used, if you remember in the previous lecture, we used some function control called RANDBETWEEN. So, if you give this RANDBETWEEN command, so, you can create, you can generate these random numbers. So, I have actually created these numbers somewhere else and I copied it, pasted it here. So, I used the function PASTE SPECIAL just to make sure that it does not change every time I click somewhere else.

So, essentially, I got all this random numbers. Here is, all these numbers you see here are random numbers generated using the RANDBETWEEN function. Now, once I have done that, I have taken 3. So, it is just randomly done, not arranged in ascending or descending order; just as it is. Then, what I have done is, I have taken a sample size of 3, like each 3 number I have taken 2 to 4; or 4, 1, 1; then 3, 7, 8; 1, 3, 9; 10, 8, 6; 10, 6, 2; 7, 4, 7; and so forth.

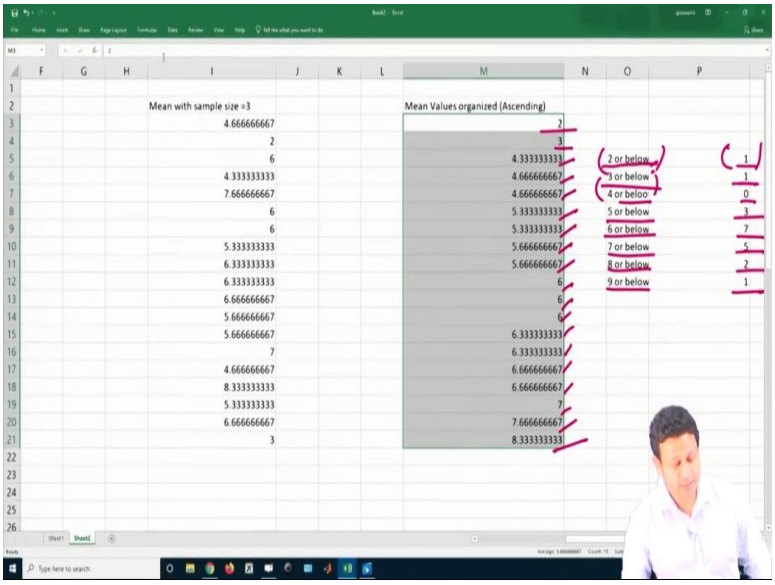
And then I have taken their mean. So, that means, 14 by 3, which is 4.66; then 4 to 6 by 3 is 2. Then 3, 7; 10; 18 by 3 is 6. Then 3 + 1, 4; 13 by 3 is 4.33. 10, 7; 23 by 3 is 7.66 and so forth. So, all these means that I obtain, they are done using these 3 numbers.

(Refer Slide Time: 04:28)



We got this means using these 3 numbers. So, we have got the sample mean. So, here the sample size is 3. You can actually increase the sample size to get a better result. Now, once I get the sample size, what I have done is, I have used a separate sheet, I have brought them together.

(Refer Slide Time: 04:53)



So, here we have the mean values, the same value from the previous sheet. Now, once I get the mean values, now I have done some ascending or descending order, whichever order I wanted to organise the numbers; just because this will help me to count the frequencies. Now, I have created several beans just because I need to count the frequencies. So, I have taken 2 or below; 3 or below, so, means 3 or 3 to 2, basically up to 2; 4 or below, up to 3; 5 or below, up to 4, 5 to 4; 6 or below, basically 6 to 5; 7 or below, 7 to 6; 8 or below, 8 to 7; and 9 or below, 9 to 8.

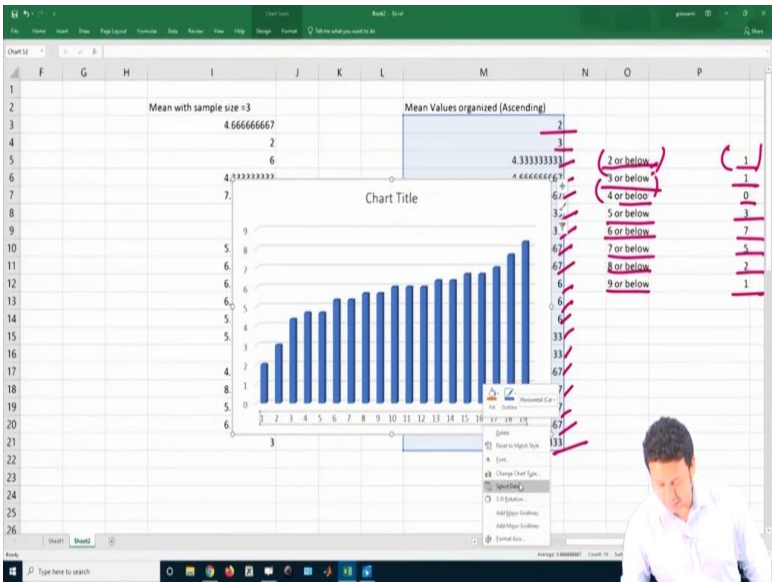
So, this way, we have counted the frequencies. So, let us see from this table, how we can actually count the frequencies. So, 2 or below, we have only 1 number 2, right? 3 or below, we have only 1 number which is 3. 4 or below, there is no number of 4 or below; so, between 4 and 3, there is no number. Now, 5 or below, we have this one, this one, this one; there are 3 observations.

6 or below, we have this one, this one, this one, this one and this one. So, when you are seeing 6 or below, we are counting 6 in; so, essentially, that means, we have 7 observations. Now, then we count 7 or below. So, we can take 7 or below. So, it is like 1, 2, 3, 4, 5; we have 5. 8 or below, we have 7 and 7.66, which are 2, and then this 8. So, these are the frequencies we have got.

Now, with these frequencies, if I now plot the; so, these are basically the mean values that we obtain from the different samples, and their frequencies. Basically, these are the mean values,

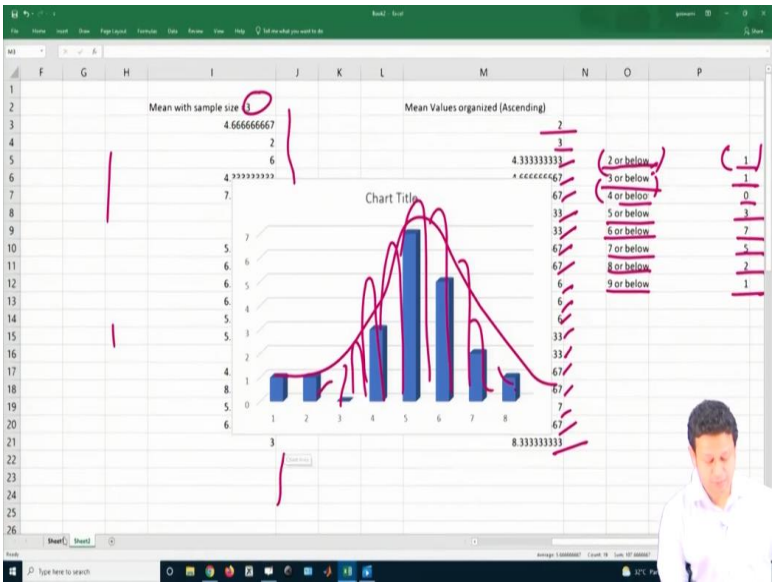
and these are their frequencies. Now, if I plot these in a, just a bar chart, if we just plot; simple bar chart.

(Refer Slide Time: 06:55)



And if I take this, the data if I change; select data, and I select data from here, and I okay it.

(Refer Slide Time: 07:12)



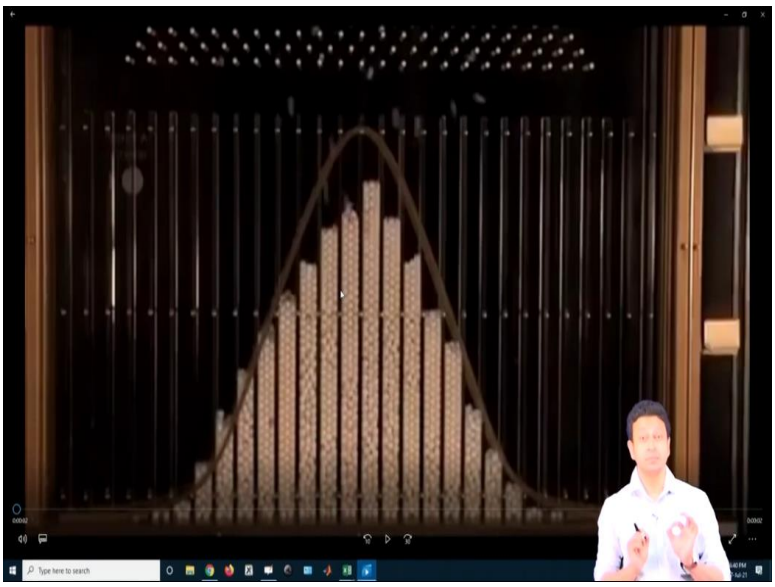
So, then I will see something like this, some sort of, this kind of curve. So, this essentially slows that it has a bell shaped curve nature. You can actually increase the sample size, and you can actually take larger number of observations and see; this will actually gradually come together and it will give you a beta sort of distribution. So, that we can actually do it, we can probably do in some assignment.

So, this is essentially the essence of central limit theorem. So, irrespective of whatever distribution; this is a random number, so, it is not a prominent distribution, this is a random number. This one, random number; these numbers we got, just a random number; but you can actually have a distribution, you can actually have a logarithmic or exponential or whichever distribution you think you want to take, you can take it.

And you take the numbers from there, and you can actually do the experiment yourself. So, that is something that you should do. So, essentially, the point is that we get this normal curve. Now, I will explain why do we get something like this. And this requires some bit of understanding. And whatever background, whatever knowledge we have gained so far in this course, will be sufficient for us to explain why this central limit theorem is actually observed in nature; and I will explain that; but before that, I will share some stories with you; and how the central limit theorem actually came into prominence.

I think I told you about Francis Galton. And Francis Galton, in his time, he was a prodigy. So, another claim of fame is that, he is a cousin brother of Charles Darwin. As a matter of fact, Francis Galton had much more fame at his time than Charles Darwin. Charles Darwin, at that time, did not discover his master creations. So, Francis Galton, he was; so, this is a story; and he actually demonstrated central limit theorem in Royal Society, it was based in England. So, one evening, he actually took something called Galton Board. And the Galton Board is something like; you probably have seen it.

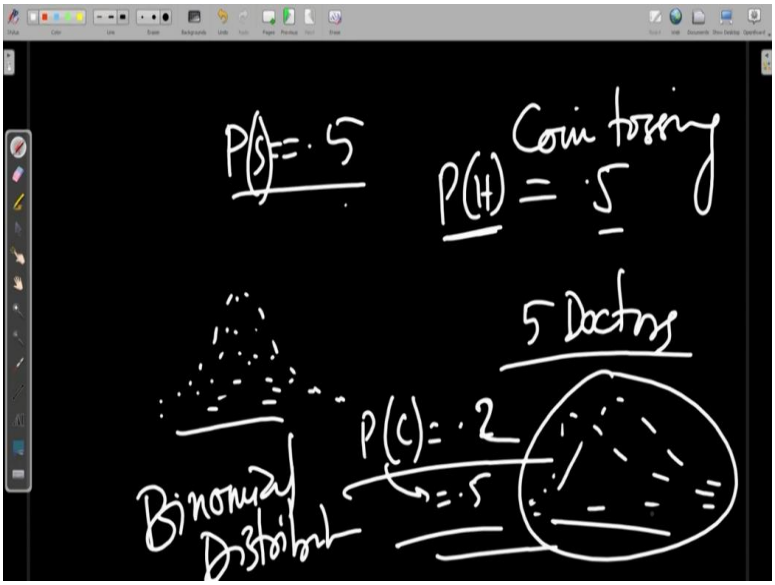
(Refer Slide Time: 09:28)



So, I will just show you. This is a Galton Board. So, this, I have taken it from internet. So, this is something like this; it is sort of a bean board, and you can drop beans or some sort of metal balls inside it and see how the distribution looks like. So, what he did is, he actually dropped these metal balls; you can see the metal balls on the top. And he actually dropped the metal balls inside the box.

And what happened then is like this; **(Video Starts: 09:57)** something like this. **(Video Ends: 10:04)** Now, essentially, what you see is, all the balls when falling from the top, they have an equal probability, probability 0.5 of going towards left or right.

(Refer Slide Time: 10:25)



Now, when the balls, when they have a probability 0.5, it is actually very similar to the coin tossing experiment that we do when we actually explained the Bernoulli trial or binomial distribution or Bernoulli distribution. So, it is just like that; probability of success is 0.5, here also probability of success is 0.5; so, probability, the ball goes left or right is 0.5. Now, so, it can go either way. Now, why suddenly it has this kind of distribution?

Why suddenly it is spiking up here? So, that is something we need to further see. And if you remember, when we did one example where we actually did the normal approximation to binomial distribution; so, when we get the example of doctors in a hospital, if we have 5 doctors in a hospital and we wanted to see the probability of all the doctors getting infected, when the probability of getting COVID is 0.2.

And there we have actually showed that, depending on your probability of success, you have a distribution that could be skewed or that could actually become a normal distribution. So, we have seen that even in binomial trial, binomial distribution. We have seen that, if probability of success approximates 0.5, the distribution also approximates to normal. Now, the point is that, so, it appears that it is quite natural.

Everywhere we see that for a equal probability of things going left or right, we actually see something like a normal distribution. Why is that so? So, again, like I before, I am trying to create some suspense. So, again, I will show you another example, I am going to show you another example of something similar happening. And this is with an experiment of sugar. So, I have a jar of sugar and I have a plate.

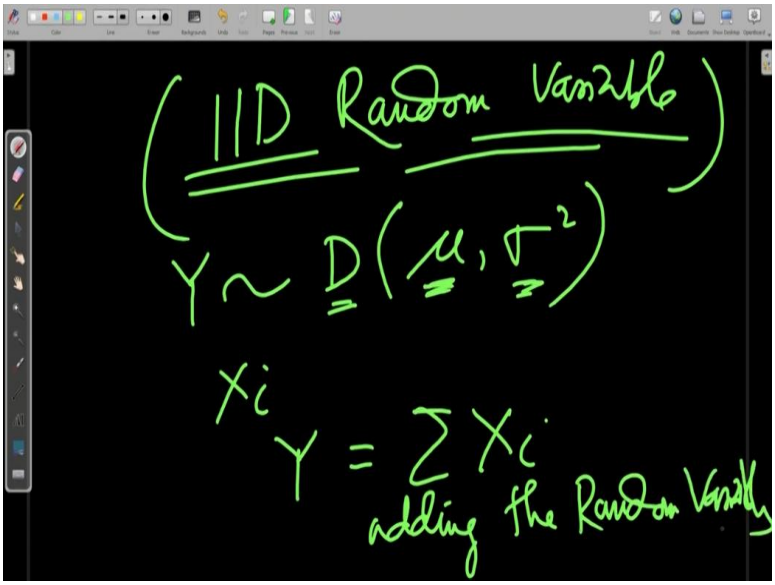
Now, I want to see what kind of distribution I will get if I was in the place of Galton. So, what I am going to do is, I am going to actually pour some sugar on the plate. And let us just do that. So, I am going to pour the sugar; so, it is an empty plate; no mystery magic here; it is just normal sugar. And let me just start pouring it. **(Video Starts: 13:01)** I will be careful. That is enough, I think.

So, I hope you can see the hump like distribution, the bell shaped distribution, the normal curve type distribution here. **(Video Ends: 13:27)** It is exactly the same thing that Galton has seen, that is exactly the same thing that you have seen when we were doing the normal approximation to binomial, when we were actually trying to see the infection rate of doctor in their probability of coming to office.

So, essentially, all these examples or the same thing we have seen with our random number experiment, so, all the experiments we are doing is basically to show you the phenomena what is happening here, that the normal curve is appearing from nowhere. Now, the point is, why the normal curve is appearing, and that is what we will see now. Now, note something. So, one is that, we have spoken; let us go back to the explanation of it.

So, we have said that, we need to draw the samples randomly; I will use a different colour. Now, what happens if we draw a samples randomly? So, when we draw a samples randomly, we have previously explained that we have IID.

(Refer Slide Time: 14:36)



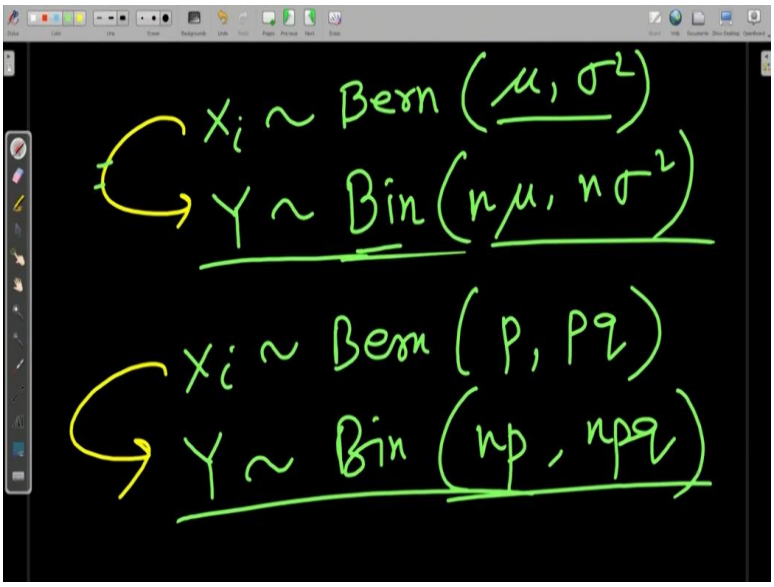
We actually use IID random variable when you are drawing samples randomly, we are getting IID random variable. And we will see why this criteria is so important here, this IID random variable, when we explain the central limit theorem. Now, if we have IID random variable, we know that for IID random variable; let us say, Y is a IID random variable and it follows some distribution, whatever distribution that is; it is a fixed mean, so, let us say mu; and fixed standard deviation or fixed variance, sigma square.

And this mu and sigma square is going to be constant across the different random variables. Now, and it does not matter which distribution we are drawing it from. So, it does not have to be normal or anything, just for IID, these 2 are required to be constant across the different random variables. And one draw will not influence the other, so, they are basically independent.

Now, let us say, when you are drawing samples randomly, X i are all the different draws. And when we are drawing all the samples randomly, and what we are doing, when we are actually doing the average thing, first, before we get the means, we are actually summing them up, we first add them up, we are first adding these 3 numbers, we are getting the sum, and then dividing it by 3.

So, we have just done the whole exercise in one go, but we are actually summing them up. So, essentially, first I am getting, let us say a Y which is sum total of X i. So, first I am summing them up. Now, X i are the random variable. So, basically, I am adding the random variables. Now, when I add the random variables, what do we get?

(Refer Slide Time: 16:44)

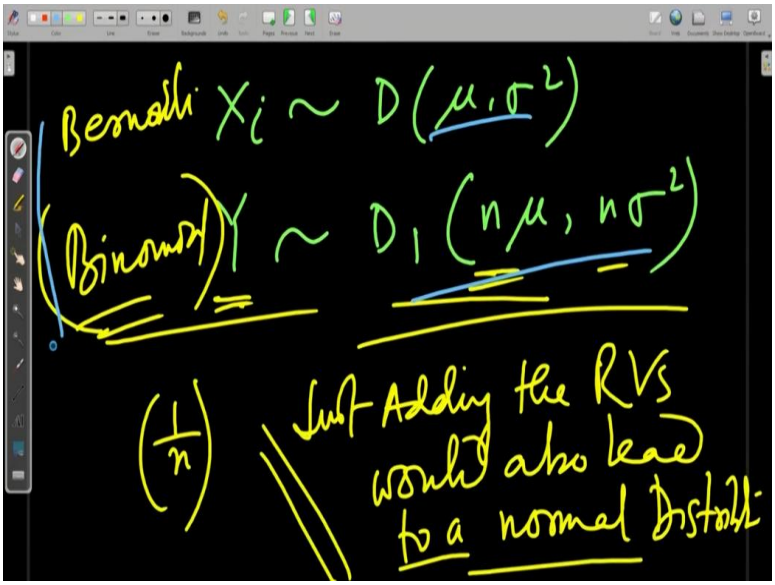


We have shown in our explanation from Bernoulli binomial, we have shown that, when we add the random variables, if my X i has a, let us say Bernoulli distribution with mu and sigma square, then, if I sum them up; so, usually we write; okay, let me write down, mu sigma square. Then my, for binomial distribution, we will have n mu and n sigma square. Now, to be very specific, the Bernoulli distribution mean is p; variance is pq.

And then, my Y will follow a binomial distribution with np, npq. So, this is something we have seen. Now, in a similar fashion, when we have this; so, all we have done from Bernoulli to binomial, we have basically added the random variable. So, here also, what we are doing is, we are actually adding the random variables. So, we have seen a binomial distribution could be derived from a Bernoulli distribution.

All you have to do, repeated Bernoulli trial would, when we sum them up, we will get the binomial distribution. Now, in the same fashion, here what we are doing is, when we are drawing these samples and we are getting the sample means; so, the first step of that is we are adding the samples, we are adding the random variables.

(Refer Slide Time: 18:27)



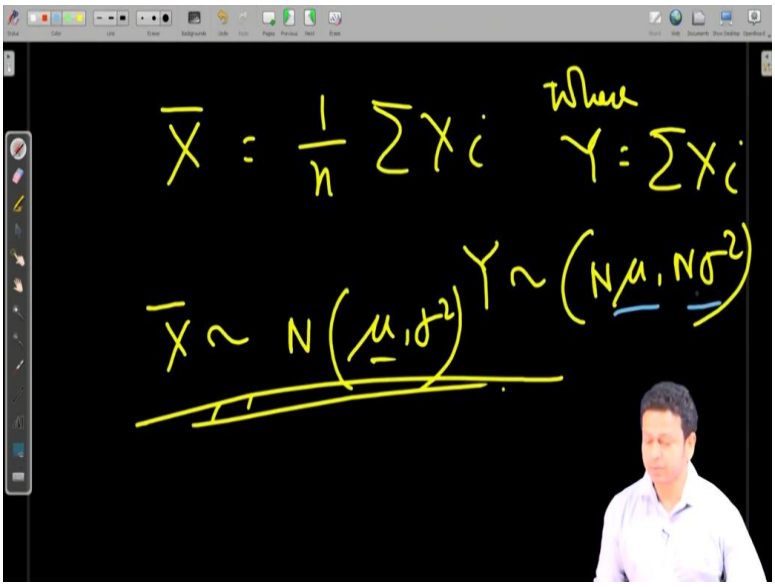
Now, when you are adding the random variables, let us say all the X i's are representing the random variables in a distribution mu sigma square, and we are adding them up, we are getting Y. Y is following, let us say, another distribution D 1 n mu n sigma square. So, it is exactly the same thing as the previous one, where we had this one as Bernoulli and this is a binomial. Now, how come it is becoming a normal?

Now, if you remember the binomial, any binomial distribution, we have shown it previously, could be approximated to normal when you have large number of observations, right? When we have large number of observation, it actually becomes a normal distribution, and the normality increases with with larger n. Now, if I actually look at it, so, it is basically the sum total of all the random variables, right?

So, when you are doing a mean, we are actually dividing it by n, the total number of observations, we are getting a mean. Now, essentially, the whole thing, even before we actually take a mean, even if we just sum them up, that too will follow a normal distribution. So, that is something we have to remember about central limit theorem. It is sort of a, rejoinder; that when we add them up, when we add the numbers, when we add the random variables, that also will lead to normality.

So, just adding the random variables would also lead to a normal distribution. So, you do not necessarily have to have the mean, like we claim the mean of the samples. We do not have to have the mean, even if you do not have the mean, you can actually get a normal distribution; But if you have it, what do you do? Essentially, you divide it by n.

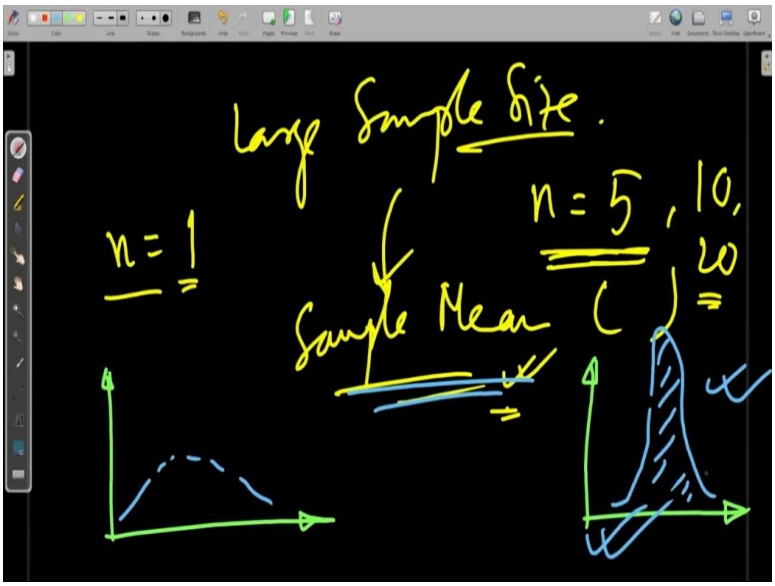
(Refer Slide Time: 20:24)



So, basically you get X bar is, let us say, 1 by n summation X i, where summation Y is equal to summation X i; and Y follows n mu, n sigma square. So, when you divide that by n, you get X bar follows a normal distribution with mu sigma square. So, essentially, what you have is that, you prove that your claim about central limit theorem is correct. So, the mean of the randomly drawn sample would follow a normal distribution.

Now, why we call a large sample size? Another criteria was large sample size. We have to have a large sample size. Why do we claim large sample size? So, I will just; you can intuitively understand that, the large sample size part.

(Refer Slide Time: 21:30)



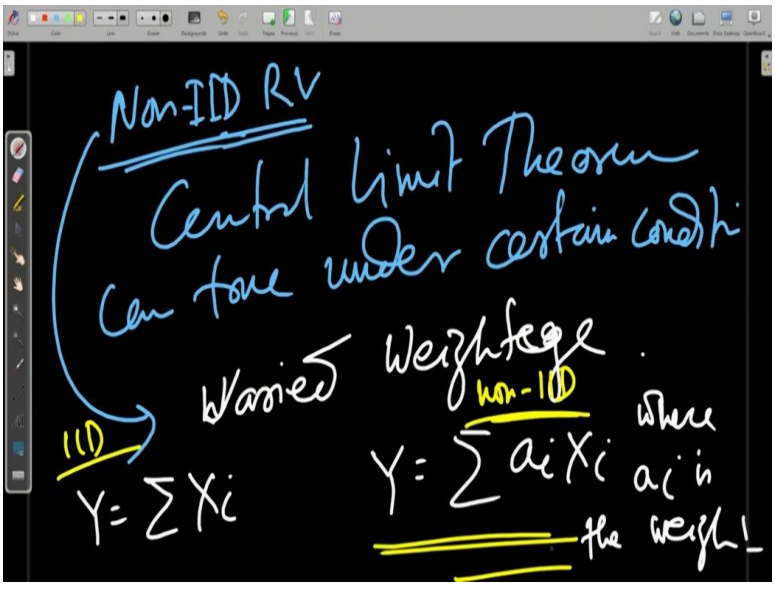
So, let us say you are drawing sample size with sample size is equal to 1; in another case, you are drawing sample with sample size 5, let us say. Now, what will happen? Let us say, if I draw with n is equal to 1, so, the mean of that value is going to be the value of the random variable itself. Whereas, if you get n = 5, you will get the mean as the mean of all those 5 random variables.

Now, getting n = 1, you essentially will not be able to cancel out errors as compared to if you have n equal to 10 or 5 or 10 or 20, whatever. So, if you have larger n, what will happen? The moment you take sample mean, in that sample mean, all these observations, they will cancel out their errors, and it will be more close to the population mean. So, as you increase the sample size and when you get the sample mean, you will get the number, you will get the value which is more close to the population mean.

And that is the reason, when we actually plot these two different distribution, you will see something like; here, if you get something like this; here, in this case, you will see something more prominent normal curve, more close to the population mean. So, that is the reason you prefer to have a larger sample size, because that will give you more sort of a better result. So, in our first example that we have given, we had sample size is equal to 3, but if we instead give a sample size of 10 or 15 or 20, we will see the result is actually improving.

So, it will look more normal. So, that is basically the idea of central limit theorem. Now, I said that the central limit theorem, the random variable has to be IID random variable.

(Refer Slide Time: 23:32)

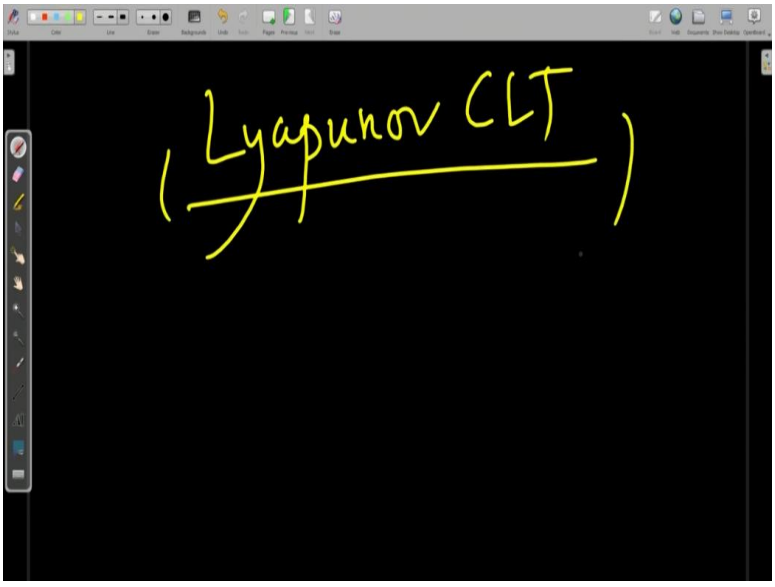


And why it has to be IID random variable? Because you have proved that, if you do not have IID random variable, this Bernoulli to binomial conversion is not going to be happening, because that is the fundamental of this conversion. So, that is why it has to be IID random variable. The mu and sigma square has to be constant, and they need to be independent. Now, what if a random variable is not IID?

So, even then, the central limit theorem can hold true. So, I will write for non-IID random variable, the central limit theorem can hold true under certain condition. There has to be some condition fulfilled. And what are the conditions? So, for IID random variable; so, what is the difference between non-IID and IID random variable? So, since in IID random variable, the mean and variance is constant, so, we have equal weightage.

Whereas in non-IID random variable, since the mean and variance is not constant, so, we have varied weightage. So, where for IID random variable we wrote Y equal to summation of X i; here it will be Y equal to summation of A i X i, where A i is the weight. So, this is for the IID random variable, and this is for the non-IID random variable. Now, if we have a very large number of observation, if we have a very large number of observation, it could be proved that this weight will not matter. So, essentially, it will behave as if it is a IID random variable. And I am not going to prove that. So, it is something called Lyapunov CLT, central limit theorem.

(Refer Slide Time: 25:32)



So, essentially, you can also have the central limit theorem holding true for non-IID random variable, if you have a very large number of observation or very large sample size. So, with

this we end the lecture on central limit theorem. In the next lecture, we are going to actually talk about another natural law which is law of large numbers. Thank you.