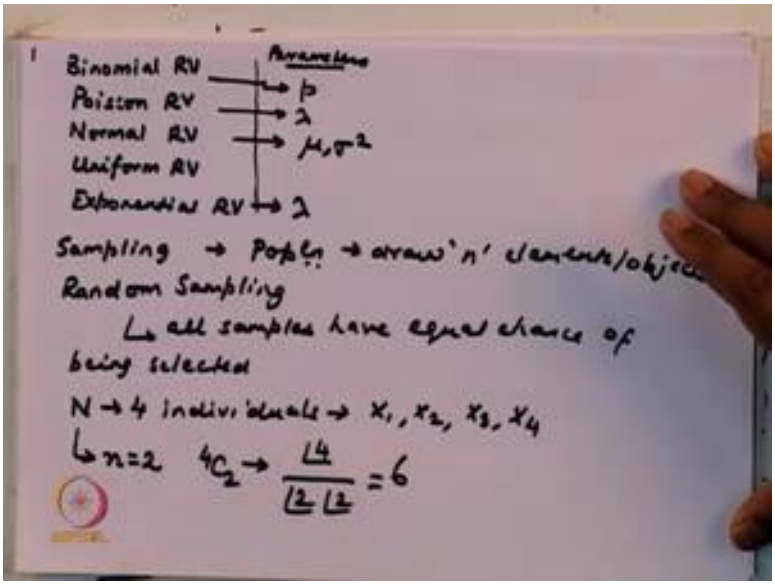
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience & Bioengineering
Indian Institute of Technology, Bombay

Lecture - 26
Sampling distributions and Central limit theorem Part-I

Hello and welcome to today's lecture. So, in the last few lectures we have discussed about the various kind of special random variables. So, these would include the binomial random variable.

(Refer Slide Time: 00:29)

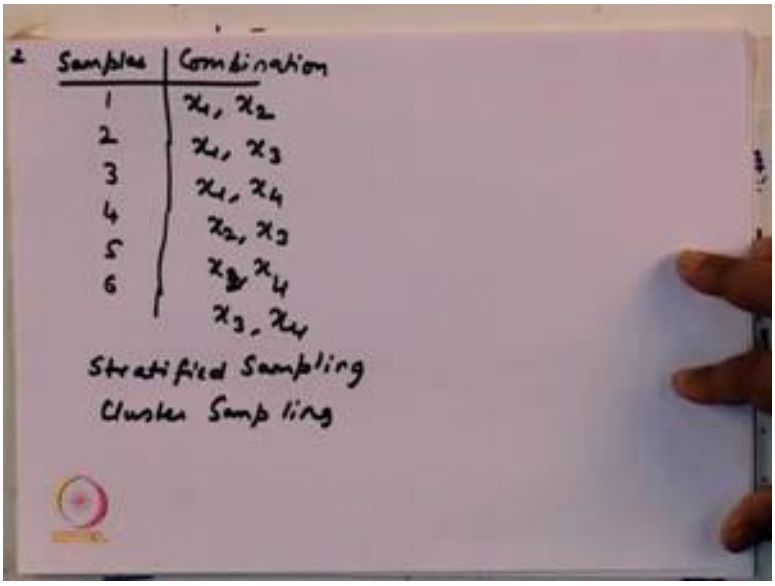


The Poisson random variable discussed about the normal random variable the uniform random variable and the exponential random variable. So, for each of these there are some parameters associated with it, for example, for the binomial random variable the parameter is the p probability of success, for the Poisson random variable you have lambda, for the normal distribution you have mu comma sigma square for the exponential r v also you have lambda.

So, in all of these cases we had solved the problems assuming that we know these parameters; however, in the real world scenario your job is to actually estimate these parameters. So, how do you do that you actually use sampling to get insight into what these parameters are going to be. So, what is sampling? So, sampling is from the population from your population you draw let us say n elements or objects.

So as one would imagine depending on how you draw these n elements or objects, the results of your distribution will drastically vary. So, this is why there are enough theories is to how one should go about deriving a sample and what should be the sample size, but let us discuss a very simple case as to what is random sampling. So, in random sampling we derive these objects. We derive these objects such that all samples have equal chance of being selected. So, for example, let us say you have a population n of 4 individuals. 4 individuals, and let us say these you can label these individuals as x 1, x 2, x 3, x 4. And from this population you want to draw a sample of size 2. This is a sample of size 2. And you want to know what are the possible combinations.

(Refer Slide Time: 03:48)

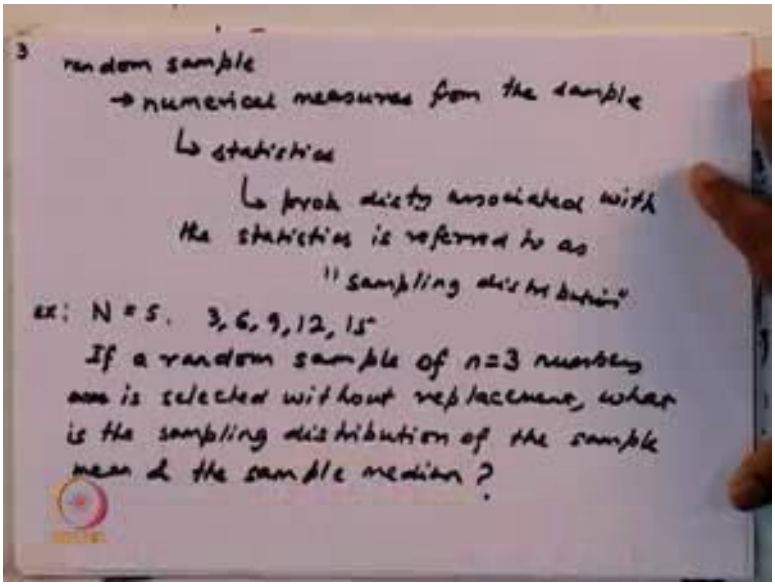


So you can clearly see that 4 c 2 are the number of possible combinations, which is factorial 4 by factorial 2 factorial 2 is equal to 6. So, what are my samples? My samples, I can put down the samples as 1 2. So, I know that there are 6 samples. So, I can have and this is my combination. And I am taking 2 right. So, I can take x 1 and x 2, I can take x 1 and x 3, x 1 and x 4, x 2 and x 3, x 2 and x 4, x 3 and x 2, x 2 and x 4, and x 3 and x 4. So, these are the different samples that you can draw. So, this would be called a random sampling, if the probability of drawing sample 1 is equal to the probability of drawing sample 2 so on and so forth.

But apart from random sampling, there are other types of sampling for example, stratified sampling or cluster sampling. So, in each of these cases, for example, in

stratified sampling; if you know your population is composed of multiple sub populations. You want to have a way such that all these individual sub populations get represented in your sampler. So, this is a special type of sampling. So, there are various other types of sampling. We would not go into details about the sampling, but we will start discussing about sampling distributions. So, when we draw a random sample. So, as we had drawn from the previous case. So, in this is the previous case different types of sample that we can draw.

(Refer Slide Time: 05:36)



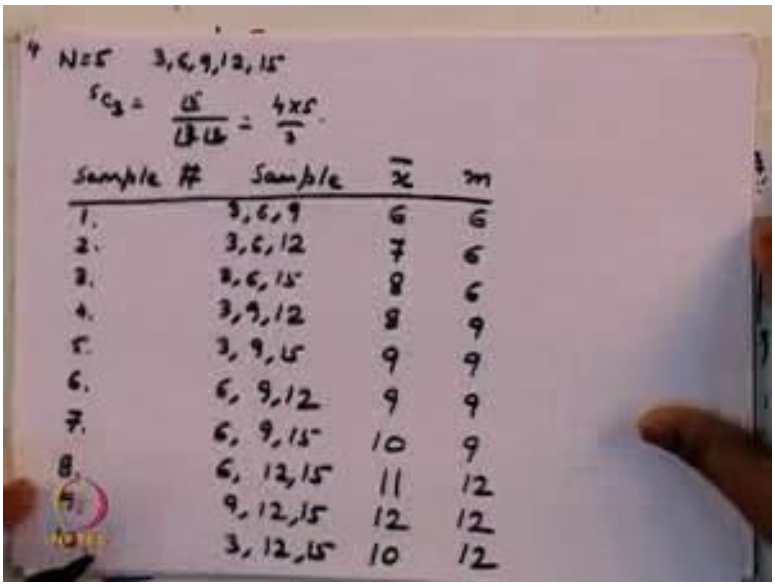
So, imagine these variables x denotes the height of the individual. So, for each of the samples I can somehow come up with metrics to quantify the popular or the sample. So, I can think of average height. So, these metrics are called statistics. So, there are you can come up with numerical measures, from the sample. And these are called statistics. And the probability distributions associated with these statistics. So, the probability distributions associated with these statistics with the statistics is referred to as sampling distribution.

So let us consider a simple example. Imagine you have an n of 5 numbers. So, n equal to 5 the numbers are 3, 6, 9, 12 and 15, if a random sample of n equal to 3 numbers are drawn is selected without replacement. What is the sampling distribution of the sample mean and the sample median? So, what is made clear is you have to select a random sample of 3 numbers without placement. So, this would mean if I choose one number, let

us say 3 for the first time, then I do not put this number back in apparently the box where these numbers are stored.

So, the next time you can only draw one of the 4 numbers. Similarly, once you have drawn let us say 9 you cannot draw 9 again for the third time. So, you have to draw from one of the other numbers. So, this means that in your random sample, none of the same number will be repeated more than once. So, let us see how will we determine the sampling distribution of the mean and the median. So, once again your n equal to 5 your numbers are 3, 6, 9, 12 and 15. So, in how many different ways can you draw 3 numbers from 5 numbers?

(Refer Slide Time: 09:05)

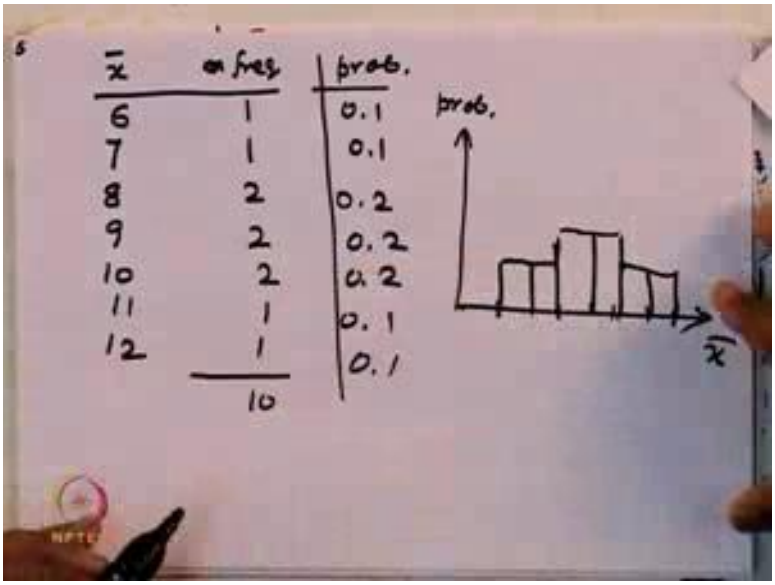


So I can draw it for in 5 c 3 different ways. 5 c 3 different ways which is factorial 5 factorials 3 factorial 2 is equal to 4 into 5 by 2; yeah it should be given 10. Let us list down the numbers; I will make the sample number. I will write down what is my sample, and for each of these I will calculate what is the x bar is the sample mean. And let us say m is a sample median. I can draw 3, 6, 9 for these cases my x bar is nothing, but 6 my median is 6. My sample number 2 is 3 6 12. So, in this case 2, 15, 21, x bar is 7 mean median is 6. 3, 6, 15 x bar is 15 18 and 6 into 4. So, 8 mean median is 6. 4th 3, 6, 3, 6, 9, so 3, 9, 12; so 12 and 3, 21 this is 8, median of 9. I can do 3, 9, 15, 7.

My x bar is 9. Median is 9, 6, 9, 6, 9, 6, 12, 6, 15, 9 12, 9, 15. So, then I can do 6, 9, 12. So, mean is 9 medians is 9, 6, 9, 15. So, that would mean 15 and so, x bar is 10 medians

is 9, 8, 9, 15 we have 6, 9, 12, 6, 9, 15, 6, 12, 15. So, in this case 21 and 12 and 11 and 12, 3, 6, 9, 3, 6, 12, 9, 12, 15; this is 12 what is the 10th number. So, I should have. So, I missed out the 3, 12, 15. I missed out the 3, 12, 15, in this case this is 10 this is 12. So, you have 10 possibilities here. And these are get all your possibilities have we missed out any other case.

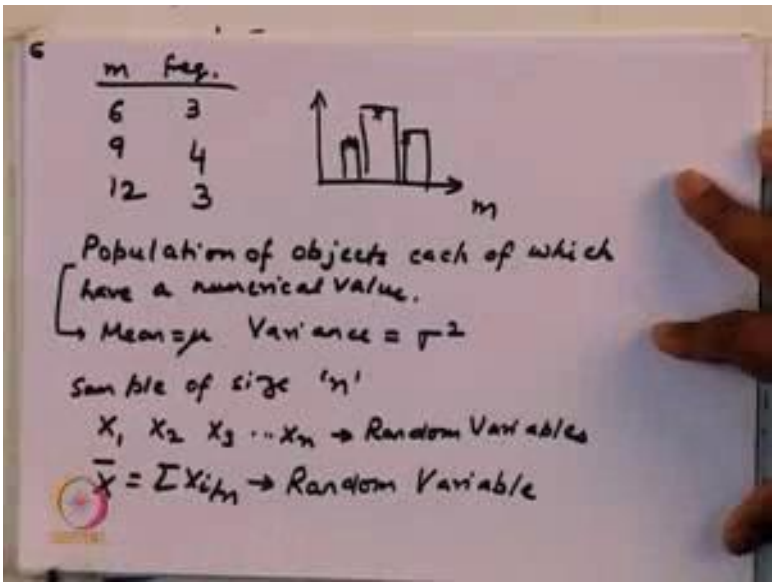
(Refer Slide Time: 13:22)



So from this list, we can find out that say what are the different x bar values and what are the m values where we can do this x bar and frequency. So, the values of x bar we get our 6, 7, 8, 9, 10, 11, 12. The frequency of 6 frequencies of 6 is only 1. Frequency of 7 is only 1. Frequency of 8 is 2, frequency of 9 is 2. Frequency of 10 is 2, frequency of 11 is 1, frequency of 12 is 1 and total frequency here is 2 plus 2, 6, 8, 10. So, I can convert this into probability. It is 0.1 0.1 0.2 0.2 0.2 0.1 0.1.

So, if you plot this distribution, x bar and probability. You have 6, 7, 8, 9, 10, 11, 12. So, these 2 will be. So, this is what your frequency will look like, for the sampling mean, similarly you can do the same thing for the median.

(Refer Slide Time: 15:05)



So, for median m, median you have obtained values of 6, 9 and 12. 6 has been obtained 3 times, 9 has been obtained 4 times, and 12 have been obtained 10 times. So, what you see. So, your median will have these 3 values. This will be a sampling distribution of the median.

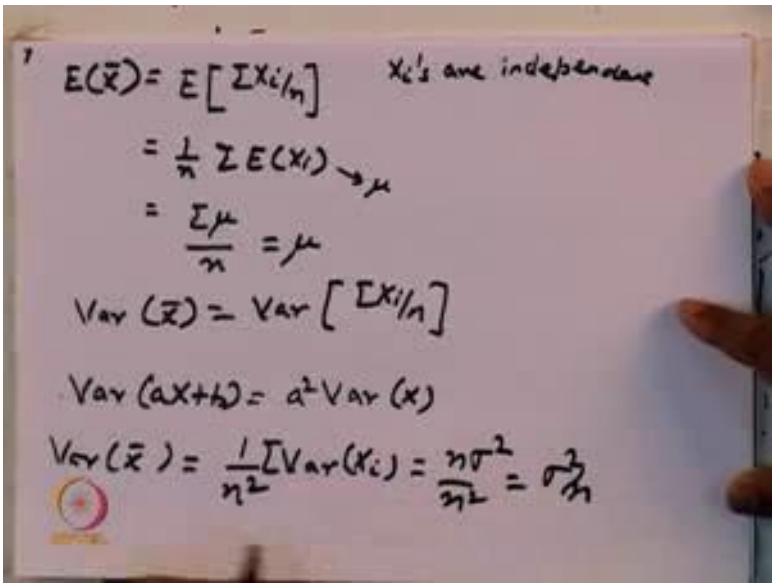
So, this gives you a clue that depending on how you draw your sample, but you will get a finite result, but what you also see is if we had changed the size of the sample then the distribution will change. So, as you know that when you have n equal to 4, there are many other ways in which you can do it. Or n equal to 2 you will get different result. So, you are sampling distribution is dependent on the size of the sample that you draw.

So in the let us consider another case. So, imagine you have a population. So, just above case population of objects, each of which have a numerical value and let us say mean, for the population you have a mean equal to mu and a variance equal to sigma square. Now let us say from this population, you draw a sample of size small n. And you label these numbers are x 1 x 2 x 3 x n. So, these each of these x is being random variables. So, each of them are random variables.

So, as for the previous case, you can see that each of these number is a random variable. So, first number is x 1 second is x 2 third is x 3. So, you can have each of them take different values. Hence the random variable, so this will also mean, that the mean, if I define x bar is equal to summation x i by n x bar will also be a random variable which

will have a given probability distribution. So, this was the probability distribution for the sample mean that we calculated when we do a sample of 3 from 5 numbers. So, similarly depending on your population size you will get x bar is x bar is itself a random variable.

(Refer Slide Time: 18:24)

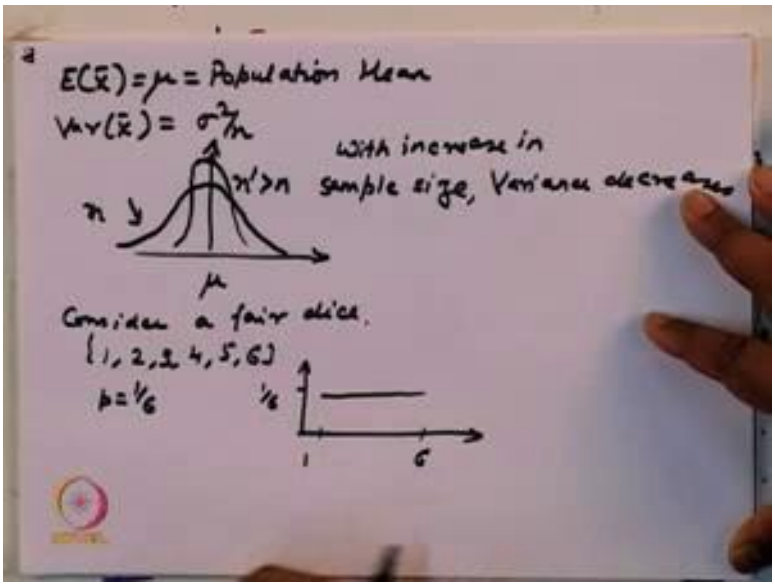


So, I can calculate expectation of x bar as expectation of x bar is expectation of summation x i by n summation x i y n.

Now this I can split it down, since x i are independent x i s are independent. So, I can write this. So, I can write this as 1 by n summation expectation of x i. And what is the value of x i is simply mu expectation of x i is mu. So, I have summation of mu n times by n, which will return you a value of mu. And what about variance of x bar. So, variance of x bar is variance of summation x i y n right.

So, you remember the expression we derived variance of a x plus b is a square variance of x and b does not come into the picture. So, this would mean that variance of x bar will be equal to 1. So, your pre factor is 1 by n. So, in this case a is 1 by n. So, variance of x bar will be 1 by n square. So, summation variance of x i because x i s are independent. So, each variance is sigma square. So, I will have n sigma square by n square is equal to sigma square by n. So, what this tells you, so your expectation of x bar is equal to mu which is same as the population mean. And variance of x bar is equal to sigma square by n.

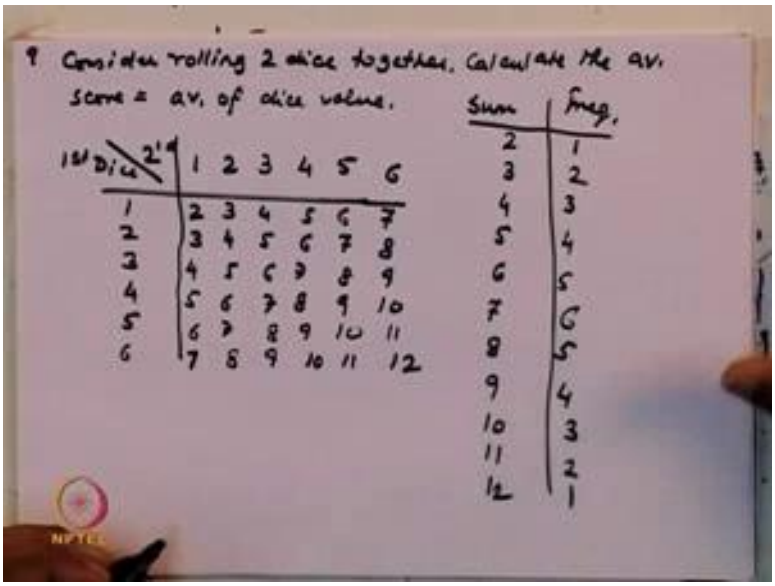
(Refer Slide Time: 20:16)



In other words, for a given distribution, so let us say this is my mu value for the sampling distribution. So, as I increase my n, with increase in sample size with increase, the variance decreases. In other words, your distributions will become more and more narrow. So, let us say this is n for sample size of n and this is for sample size n prime greater than n. So, you are mean would not change, but these distributions will get more and more piquant.

So this is what this proves. And let us test it with a simple example. So, you consider a fair dice. So, if you roll a single dice, you know your outcomes are 1, 2, 3, 4, 5 or 6. And for each of them probability is equal to 1 by 6. So, your distribution will simply be this from 1 to 6. And this value is 1 by 6, but say let us say I consider 2 dice together.

(Refer Slide Time: 22:18)

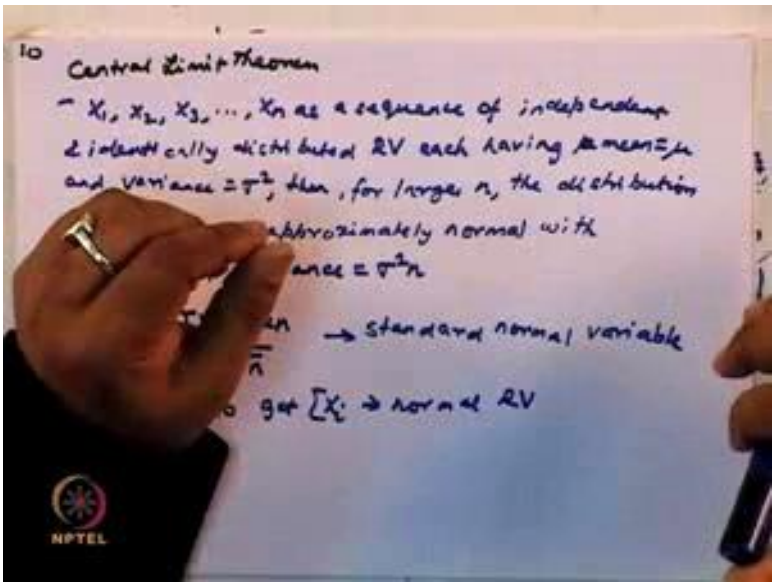


And from this I am calculating the average score that is equal to average of dice value. So, I can construct this table. Let us say first die what you would to return and second.

So, you can have for the second any values, from 1 to 6. For the first dice also you can have any values from 1 to 6. So, if I write down the average value, if I write down the sum be 2, 3, 4, 5, 6, 7. For 2 it is 3 4 5, 6, 8, 4, 5, 6, 7, 8, 9, 5, 6, 7, 8, 9, 10, 6, 7, 8, 9, 10, 11 and 7, 8, 9, 10, 11, 12. So, from this I can calculate what is the average. So, the sum, so I can make this plot, let me make the plot of some I will get any are 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. This is the sum and this is my frequency. For 2 it is 1, for 12 it is 1, for 3 it is 2, for 11 it is 2, for 10 it is 3, or 4 it is 3, 5 it is 4, 9 it is 4, 6 it is 5, 8 it is 5, 7 it is 6.

In other words, if we plot this data on this curve, so what you see? You will have a distribution which will look like this, it is symmetric. It will look like this and what you see by increasing. So, if I plot the average this will. So, some is 7 this is 3.5. It has the maximum value. So, what you also know note is the average of these numbers is 3.5. So, your E sample average is also giving use 3.5, which is same as the population average, but you transitioned from a flat distribution like this, to this just by doing this 2 times. So, with increase in number this gets on getting more and more distributed. So, if you increase your sample size, you will see that slowly it will go on transitioning into more and more piquant distribution.

(Refer Slide Time: 25:43)



So, this is the premise of the central limit theorem one of the most important theorems in probability. So, what does central limit theorem say? So, if you have x 1 x 2 x 3 x n as a sequence of independent and identically distributed random variables, each having mu having mean equal to mu and variance equal to sigma square. Then for large n the distribution x 1 plus x 2 plus dot dot x n is approximately normal with mean of n times mu and variance sigma square n. So, this would mean if I have the random variable z as summation x i minus mu n by sigma root n, then z would become a standard normal variable.

So if your individual x i r normal random variables by themselves, then this is always around normal random variable or random normal random variable independent of your choice of n. If x i s are from a symmetric distribution. As we did for the rolling of a die case for a symmetric distribution even for small values of n you will get a normal random variable and that is what we obtained right. So, in this case even from a for n equal to 2, we obtained a roughly normal distribution. So, increase in end will give you, but the event for a symmetric distribution you will get a normal distribution. Otherwise if the distribution is non normal for a population which is non normal you need n greater equal to it has been estimated n is greater or equal to 30 to get x bar or summation x i to be a normal random variable. So, that ends our discussion today.

So, we started with normal and sampling distributions. And we saw how one of the important properties of sampling distribution is by increase in the sample size. So, your sample mean is always same as the population mean, but by increasing the sample size your variance of the distribution keeps on decreasing. And we finally, discussed the central limit theorem, where if you have individual random variables which are identically distributed and independent. So, then the sum of these random variables will be a normal random variable with mean of mu n and variance of sigma square n. So, that is it for today and I look forward to next day's class.

Thank you.