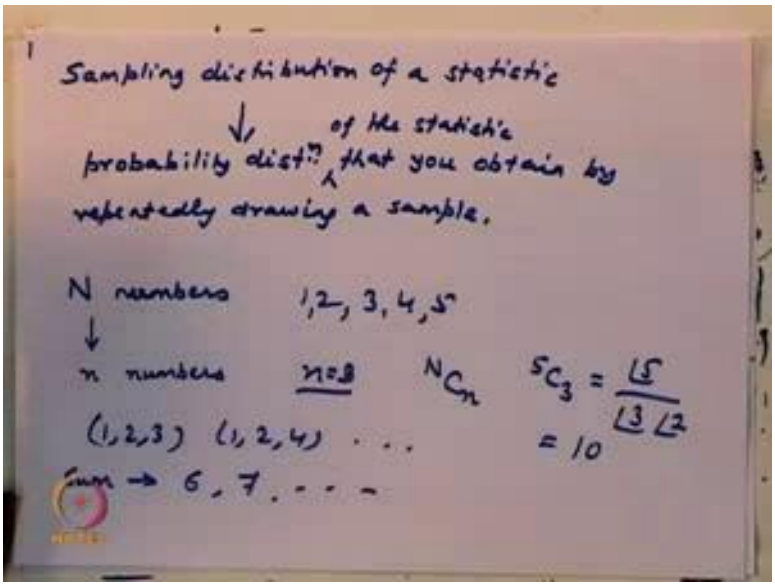
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 27
Sampling distributions and Central limit theorem Part-II

Hello and welcome to today's lecture. So, in the last class we had briefly touched upon the idea of central limit theorem right. One of the most important theories in the field of statistics, so before going to central limit theorem, I will begin by discussing about again sampling distribution right.

(Refer Slide Time: 00:36)



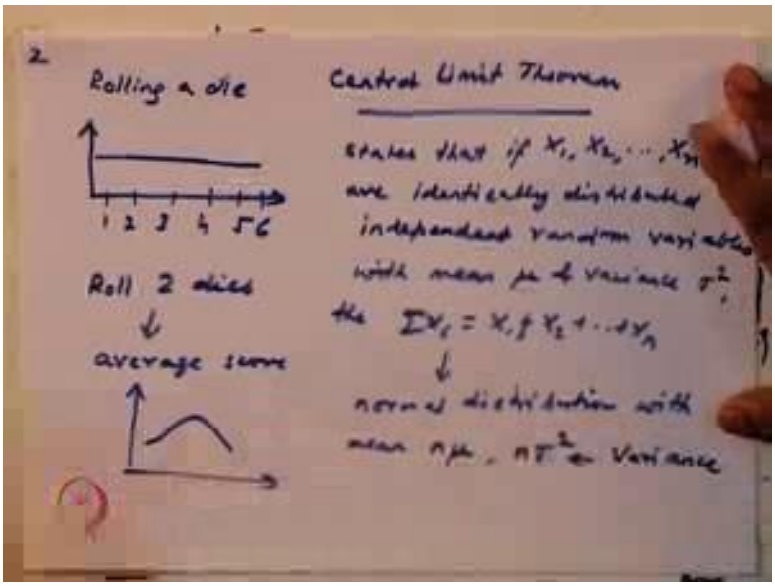
So, what is the sampling distribution and you can find the sampling distribution of a statistic. So, by this I mean you can compute the sampling distribution of a mean of a median or any other measure that you define.

So, what is the sampling distribution of the statistics? So, a sampling distribution of the statistic is the probability distribution that you obtain. So, probability distribution of the statistic that you obtain by repeatedly drawing a sample, in other words imagine you have a population of n numbers. Let us say n numbers and from this you draw a sample of n numbers. So, based on this for example, if my 5 numbers are 1 2 3 4 5 and let us say my n is equal to 3 right. So, in how many ways can I draw 3 numbers from n? So, it can be n c n, the different ways in which I can draw 3 numbers from 5 right. So, this is 5 c 3

equal to factorial 5 by factorial 3 factorial 2 is equal to 10. So, there are 10 ways of drawing a sample of 3 and one example is 1 2 3 or one 2 4. So on and so forth.

So, from these you can let us say compute the sum of these numbers in that case this will be 6 this will be 7 and so on and so forth. So, you can ask the question what is the probability distribution of the sum that you obtain. And this is the sampling distribution. So, ideally the concept of sampling distribution is very important, because you do not know the; you want to find out some information about the population based on the sample that you draw. So, central limit theorem is that theorem which allows us to correlate the sample mean or median or any other metric with the population metric.

(Refer Slide Time: 03:30)

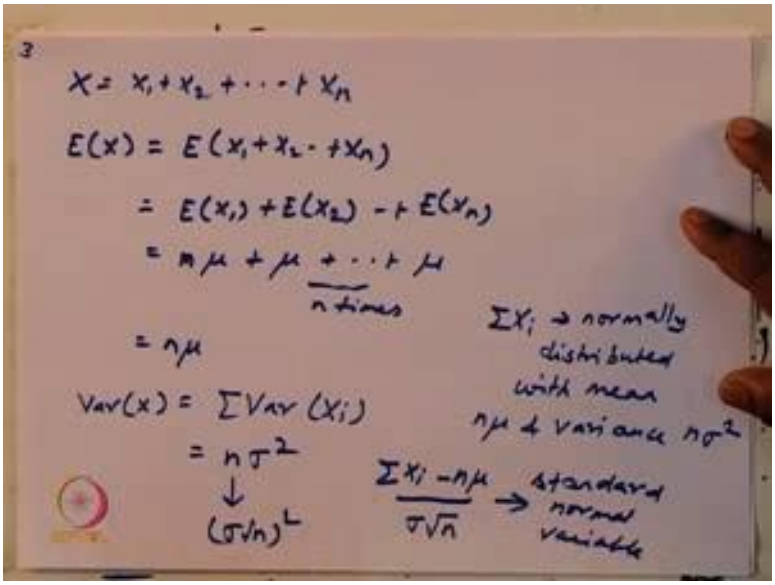


So, what does central limit theorem state. So, I had discussed this brief example in last class right, where if you are interested in rolling a die and checking out it is value that you get right. So, the distribution would be fully flat. So, each of these values is 1 2 3 4 5 or 6 this distribution is flat, but if you roll 2 dice together and you compute the average score that you get. So, you will see begin to see that slowly your distribution will look closer and closer like a normal distribution. And with increase in the averaging process, and this is what central limit theorem says. So, what does the central limit theorem state?

So your central limit theorem states that if x 1 x 2 x n are identically distributed, independent random variables, with mean mu and variance sigma square. Then summation of x i that is x 1 plus x 2 plus x n, this is also a random variable, it will follow

a normal distribution with mean n times mu and variance n sigma square. So, this is your variance. And this is very easy to prove.

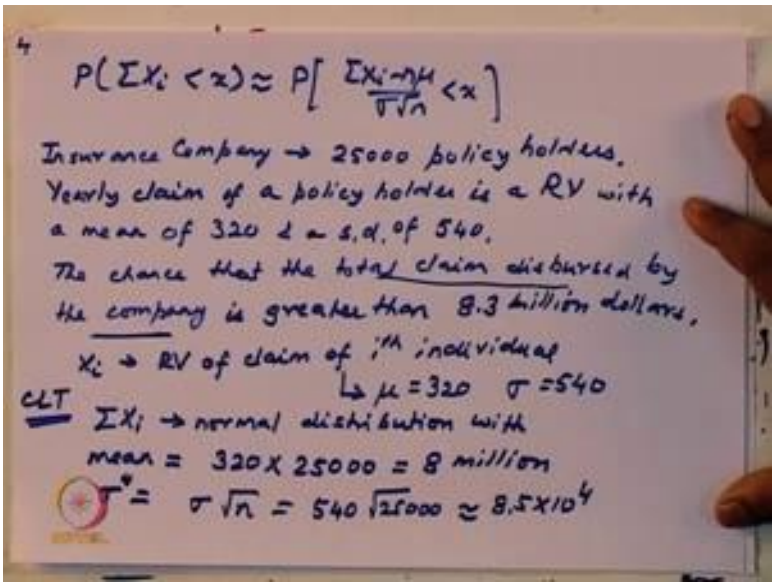
(Refer Slide Time: 06:00)



So, I can write x as x 1 plus x 2 plus x n. So, since each of these variables are independent, the expectation of x is e of x 1 plus x 2 x n, because each of these random variables are independent. We can break it down into e of x 1 plus e of x 2 So on and so forth, plus e of x n. So, this is going to be n. So, mu plus mu plus mu, this is n times. So, that is why this will give you the expectation of this x is simply equal to n mu. And by the same token I can write variance of x is equal to summation of variance of x i by again because these random variables are independent and so, I will get n times sigma square. So, this variance is n times sigma square

So, what in other words which would mean that if you have, I can say that summation x i is normally distributed with mean n mu and variance n sigma square. So, then I can say that sigma x i minus n mu by sigma root of n. So, this is nothing, but sigma root of n whole square right. So, this is approximately equal to a standard normal variable.

(Refer Slide Time: 07:57)



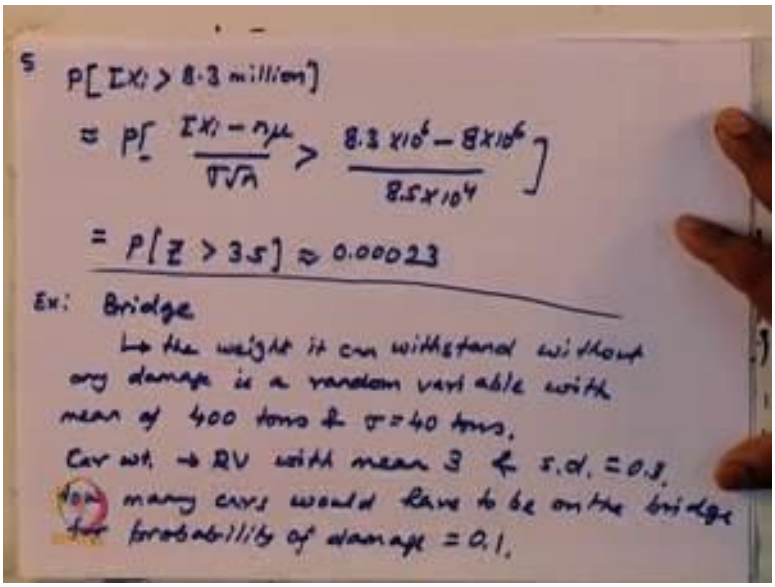
So, in other words in other words probability of summation x i, let us say less equal less then x you can approximate is the probability of summation x i minus n mu by sigma root n less than x this is the approximation.

So, let us first try out few examples, how this theory can be used. Imagine you have a insurance company which has 25000 policy holders. So, the yearly claim of a policyholder is a random variable with a mean of 320 right and standard deviation of 540. So, this is the background information. So, we want to compute the chance that the total claim disbursed by the company disbursed by the company is greater than 8.3 million dollars.

So, this is what we want to compute the chance that the total claim is greater than 8.3 billion dollars. So, if you have to make use of the central limit theorem. So, what we are given is. So, each x i is a random variable of claim of ith individual. So, we want to. So, the total. So, the total claim disbursed by the company is going to be summation of x i. So, each x i has a mean equal to 320 and sigma equal to 540 dollars. So, my summation x i must follow a normal distribution. So, this is by central limit theorem. What I can say is summation x i follows a normal distribution with mean equal to 320 into the total number of policyholders. So, for each x i you have mean as 320. So, for summation x i the mean has to be 320 into 25000.

So, if you plug in the numbers this will come to be 8 million and standard deviation will be equal to I can say sigma star equal to sigma into root of n. So, which is 540 into root of 25000 so, this will approximately come to be 8.5 into 10 to the power 4.

(Refer Slide Time: 12:07)



So, we are asked to calculate probability that summation x i will be greater than 8.3 million. So, we can approximate this as the probability summation x i minus n mu by sigma root n is greater than 8.3 into 10 to the power 6 for million minus. So, n mu is going to be 8 into 10 to power 6 that we found and sigma root n is going to be 8.5 into 10 to the power 4. So, this is equal to probability of. So, this is your standard normal variable now. So, this is z greater than. So, if you plug in the values this will give you a value of 3.5 and this comes to approximately 0.00023.

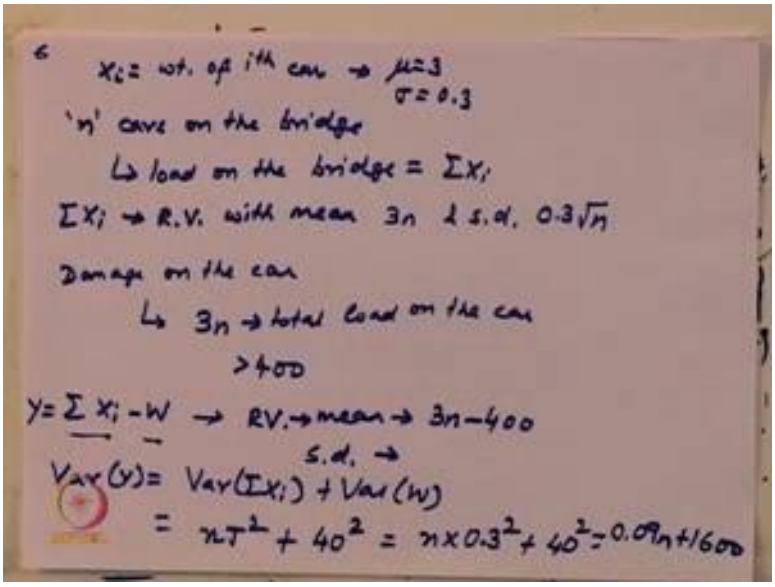
So, as you can see that where there is very little chance, that the claim is going to be more than 8 .5 million. So, this is how you make use of central limit theorem, to gain an idea of or to make a prediction of calculating the chance that this particular insurance company will have to disperse this once. And these are calculations which the company does at it is end to figure out how much the insurance amount should charge. So, that it makes a certain amount of profit.

So, let us go to another example. So, think of a bridge. So, each bridge has a certain capacity of how much weight it can bear. So, imagine for a given bridge, so the weight it can bear. The weight it can withstand without any damage is a random variable with

mean of with mean of 400 tons and sigma equal to 40 tons. So, cars fly on the bridge on an average. So, the car waits imagine, it is a random variable with mean 3 and with standard deviation 0.3.

So, the question is how many cars would have to be on the bridge for probability of damage equal to 0.1. So, what this question tries to understand is how many cars is the you know can the bridge bear. So, there is damage of 0.1. So, how do we go about this problem? So, what you have to calculate what we have been asked to calculate is.

(Refer Slide Time: 16:20)

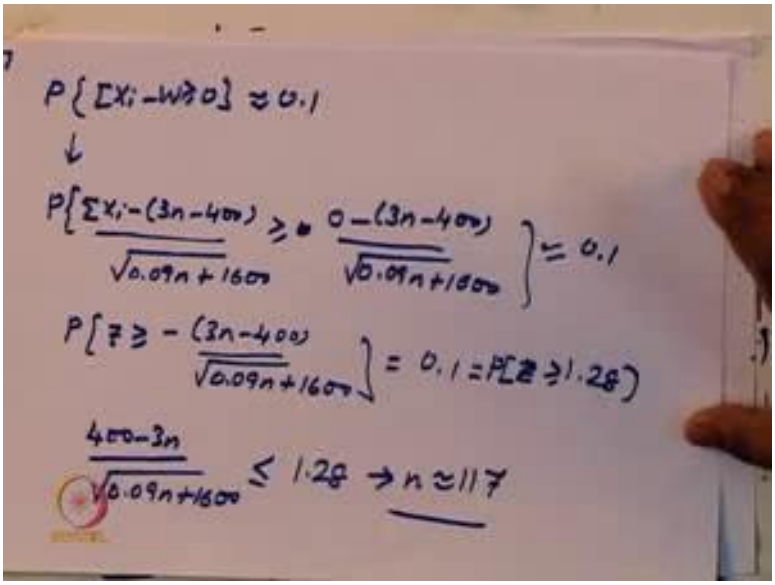


So, let us say x i is weight of ith car. So, if there are n cars on the bridge. So, the load on the bridge is equal to. So, since x, so the load on the bridge is equal to summation of x i right. So, what can I say about summation of x i.

So, each of these x i is a random variable with mu is equal to 3 and sigma is equal to 0.3. So, I can say summation is sigma x i is a random variable with mean 3 times n. If you have n cars 3 is the mean for each car. So, mean will be 3 n and standard deviation will be 0.3 into root n 0.3 into root n. So, when I want to calculate the damage on the car. So, from the damage on the car we are thinking of a situation that 3 n which is the total load on the car. This is greater than 400. 400 is the weight which can be withstood by the bridge. So, summation x i minus w, so this is the effective weight mismatch between the weight of that cars on the bridge and the weight that can be withstood is a random variable, with mean this has a mean of 3 n minus 400 and a standard deviation.

So, how do we calculate the standard deviation here you have 2 random variables. So, you define a random variable. Let us say y as summation of x i minus w mean is this and you know. So, since these 2 are independent of each other. So, your variance of y will be variance of summation x i plus variance of w and variance of summation x i is nothing, but n times sigma square and this is whatever has been given to you which is 40 tons. So, this is 40 standard deviations. So, 40 square, so this is n into 0.3 square plus 40 square equal to 0.09 n plus 1600. So, variance is this. So, y which is defined as the difference between the net weight from the cars on the bridge to the weight that can be withstood is this. So, y is the random variable, which has a mean of 3 n minus 400 and a variance of 0.09 plus 1600.

(Refer Slide Time: 20:01)

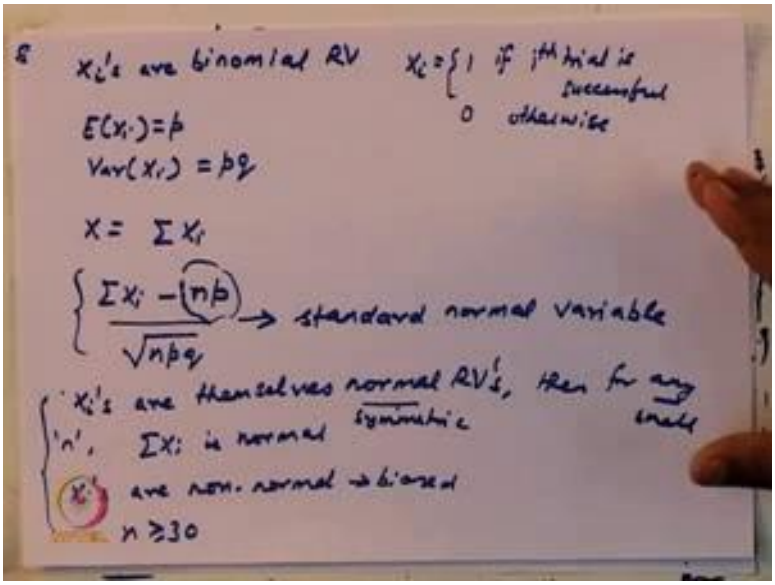


So we have been asked to calculate probability summation x i minus w greater equal to 0. Why? Because if summation x i is less than w there is no damage on the bridge. Only when summation x i begins to exceed w, then there any chance of the bridge. So, we have been asked to calculate this is roughly equal to 0.1. So, probability of summation x i minus w, this can be written as probability of z summation x i minus of 3 n minus 400. So, this is the net mean minus root of 0.09 n plus 1600. Because the variance was 0.09 n plus 1600 is greater equal to 0, p of z f greater or equal to 0. So, p of this is basically greater equal to 0 minus of 3 n minus 400 by root of 0.09 n plus 1600. And this we are given approximately equal to 0.1

So, probability of z greater equal to minus of 3 n minus 400 by root of 0.09 n plus 1600 is equal to 0.1. So, if you look up the table, this is roughly p of z greater equal to 1.28. So, this would mean that 400 minus 3 n by root of 0.09 n plus 1600 is less or equal to 1.28. And from here you can find out the value of n. So, this is one equation you take a square and determine the value of n. So, this n will be ordered 117. So, this allows if there are roughly 120 cars then there is a chance that 10 there is a 10 percent chance that the bridge will get damaged.

So, here we made use of central limit theorem, but we first defined our random variable y as the difference between the total weight of the cars on the bridge minus the weight that could be withstood because we are calculating the chance of damage.

(Refer Slide Time: 22:41)



So, imagine your x. So, x i s are binomial random variables, x can be of any distribution x can follow any distribution. Let us assume that x i s have binomial random variables. Then we know that e of x i. So, we can call x i one if ith trial is successful equal to 0 otherwise. So, we have previously demonstrated that e of x i is p and variance of x i is equal to p q. So, if we define a random variable x which is summation of x i. So, in that case again by you invoking central limit theorem, I can say summation x i minus. So, x is summation of x i. So, summation x i will have a mean of n p and a variance of n p q. So, so n p q is going to be the variance of this x and n p is going to be the mean of this x. So, this means that summation x i minus n p by root of n p q is a standard normal variable.

So, one thing is, there is a requirement on the value of n depending on what distribution these x i s follow. So, generally if x i s, if x i s are themselves normal random variables then for any n summation x i is going to be normal, but if x i s are non-normal, and maybe even biased, in that case you need to have n greater or equal to 30 for this distribution to be a normal distribution. So, for a binomial random variable, we have summation x i follows summation x i follows normal with mean of n p and variance of n p q. So, I will conclude my lecture today by saying that we would we can briefly remember about central limit theorem and how we can use central limit theorem to estimate certain things given the sample distribution.

So, as we saw that summation x i will follow a normal distribution with mean of n times mu and variance of sigma square n which means the standard deviation will be sigma root of n. And depending on the type of distributions, so if n if you are underlying x i s are themselves normal random variables or even symmetric, then for any or small values of n 2 or 3 you will have summation x i or x i or summation x i by n follow normal distribution, but if x i s are non-normal in nature, particularly they are an asymmetric then you need n greater equal to 30.

With that I conclude my lecture today. I look forward to meet you again in next class.