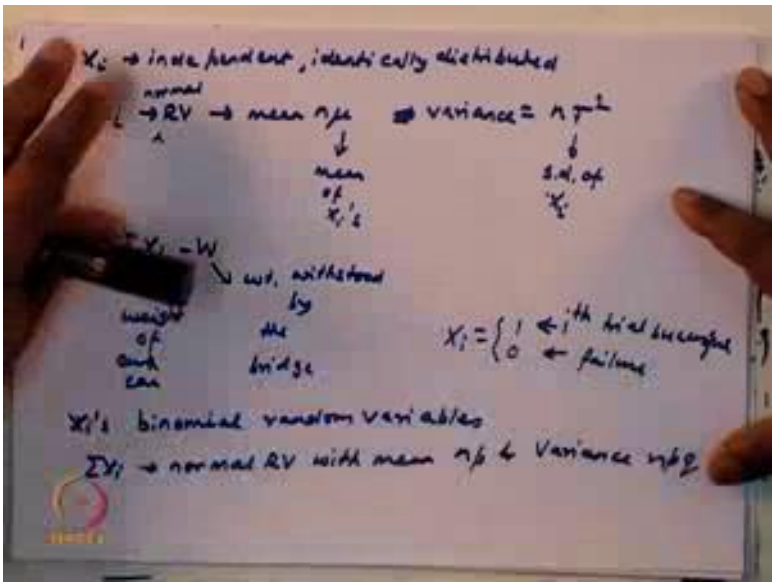
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 28
Central limit theorem Part-III and Sampling distributions of sample mean

Hello, and welcome to today's lecture. We will continue where we had left of in last class discussing the Central Limit theorem.

(Refer Slide Time: 00:30)



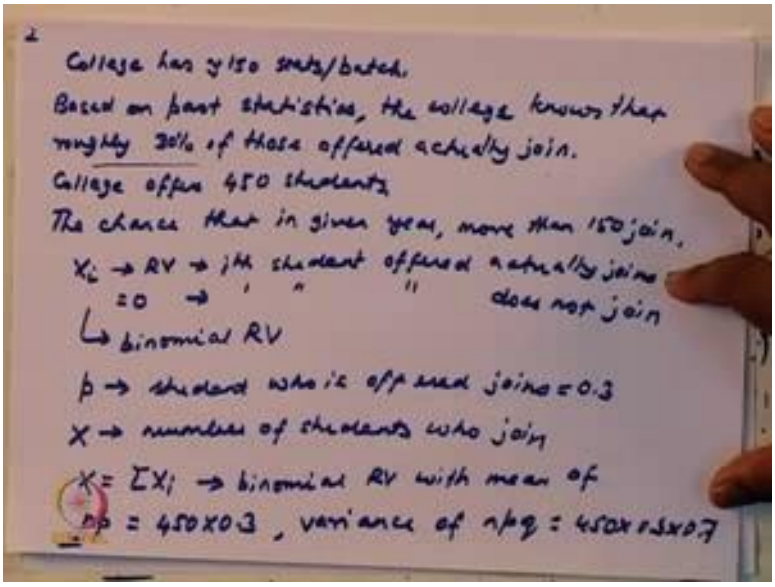
Just a brief recap, what we had covered in last class was, if you had random variables x i which are independent, but identically distributed. Then we have summation of x i is a random variable or normal random variable with mean equal to n mu, where mu is the mean of each of these excise; mean of excise. And standard deviation or we can say variance equal to n sigma square, where i is from one to n; n sigma square where this is the standard deviation of our variance of each r be x i.

And we had seen using the example of the cars on a bridge to see whether at what point the bridge will become unstable. In that case we had defined this random variable y as summation of x i minus w where x i was the weight of each car. And w is the weight that can be with stood by the bridge. And we found how we can calculate for this problem, for this we can find out its mean and standard deviation for this random variable and then translate it to using the normal distribution to estimate the probability of damage.

So, one thing we are discussed in last class was for binomial; if f i's for a binomial random variables then in this case my summation x i is a normal random variable with mean n times p and variance npq, so where p is the probability of success for i-th trial. So, we can define in this case x i is equal to 1 if i-th trial is successful, and 0 if it is a failure.

So let us see an example where we can make use of this concept of binomial random variable.

(Refer Slide Time: 03:35)



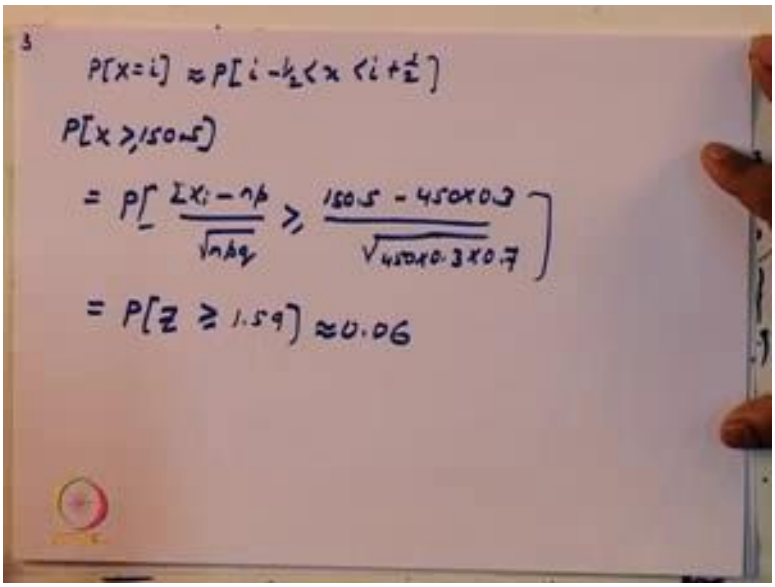
So, imagine you have a college has roughly 150 seats per batch. So there are 150 seats per batch. However based on statistics on past statistics the college knows that roughly 30 percent of those offered actually join. So, given this the college offers. So, to get around this problem that less number of students joins; the college offers 450 students. So, we want to calculate the chance that in a given year more than 150 join. So what you see is, in this case you can make use of binomial random variable.

So, what you can define is x i is a random variable, it is a random variable such that which says associated with if i-th student offered actually joins. What you can see is if i-th student offer this (Refer Time: 05:42) I can assign x i is equal to 1. And x i is equal to 0 if i-th student is offered, but does not join. So, x i in this case are binomial random variable. And what is the probability of success? So how do we define p? P is student

who is offered joins. So, this probability we know is given as roughly 30 percent. So, my p will equal to roughly 0.3 is equal to 0.3.

So, let x be the number of students who join. So, what will be x? X is nothing but x I can write as summation of x i because each x i is one if the student joins. So, if i add all of them then x is the random variable of the number of students who join. Since each of x i is a binomial random variable x will also be a binomial random variable with mean of n times p which is equal to, so n is the total number of students offered which is 450 into 0.3 is p; with mean of this and variance of npq 450 into 0.3 into 0.7.

(Refer Slide Time: 07:47)

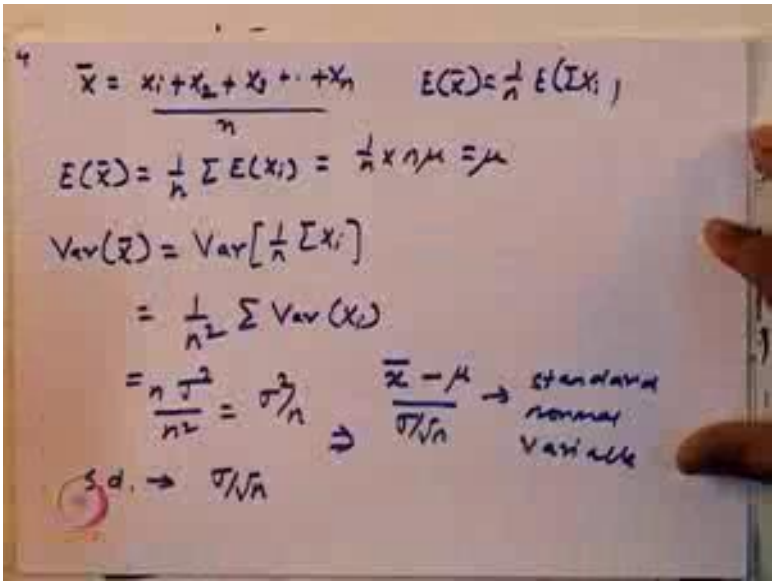


So, if you are asked the question what we want to know for a binomial random variable. So binomial x equal to i, you can approximate this because it is a discrete random variable. You can approximate this as probability of half less then x less then i less than half. So, this is approximation from discrete to continuous. So, what we have to find out is probability of x greater than 150.5. So, we need to calculate the probability that more than 150 join. So, 150.5 is a consequence of this continuous approximation.

So, this is nothing but probability of Z, so Z is given by summation x i minus n p by root of npq and greater than greater than equal to let us say 150.5 minus 450 into 0.3 which is your n p by root of. So, this is probability of Z greater equal to good, 1.59. And this turns out to be roughly 0.06. There is still a 6 percent chance that more than 150 students will join.

So, this shows you how you can make use of binomial random variable to and central limit theorem to ask or answer this kind of questions.

(Refer Slide Time: 09:45)

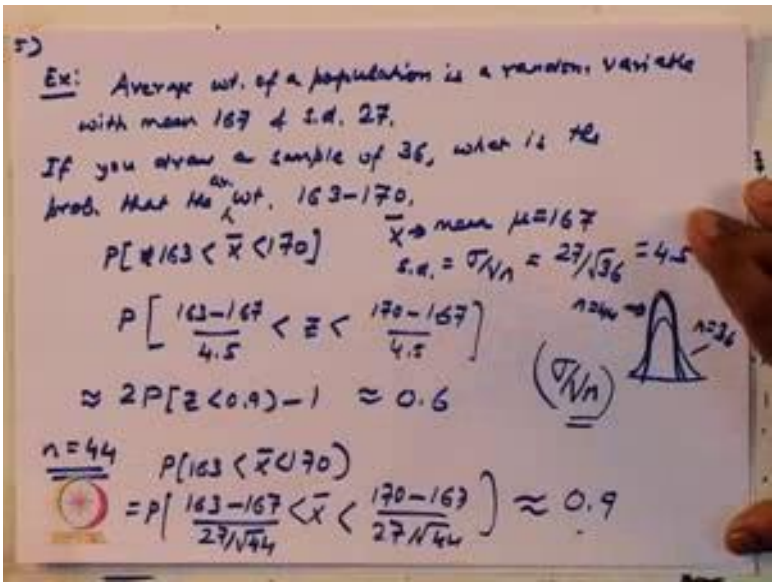


Till now we were discussing about the sum of all values. So x was summation x i, what will happen to the sample mean. So, for doing that; if I define x bar as x 1 plus x 2 plus x 3 x n by n, then E x bar 1, can be given a 1 by n summation E x i and this summation it is actually. So, E x bar is 1 by n E summation x i, since x is are independent I can make this approximation. So, I get 1 by n into n times mu equal to mu. So, the total sum is a normal random variable with mean of n mu, but the mean is a normal random variable with mean of mu. And what about variance of x bar? So, variance of x bar is variance of 1 by n into summation x i. This 1 by n is a constant. So, if you take it out you scale it by 1 by n square as we have derived earlier, and because these x is are independent I can simply again write summation variance of x i.

So, you have sigma square into n by n square equal to sigma square by n. So, this would mean that standard deviation of the sample mean x bar is given by sigma by root n. This would mean by the sample mean x bar minus mu by sigma by root n is a standard normal variable. So, this is two major differences. For summation x i your mean is n mu and variance is sigma root n, but for x bar the mean is mu and the variance and the standard deviation is sigma by root n. In one case you have sigma root n in the other case you have sigma by root n.

So, let us try doing some few examples of sample mean.

(Refer Slide Time: 12:13)



So let us say, let us take an example. Average weight of a population is a random variable with mean 167 and standard deviation 27. If you draw a sample of 36, if your sample size is 36 what is the probability that the average weight will lie between 163 and 170. So, we want to calculate the average weight lying between 163 and 170. We want to calculate 163, so this is average weight; this is an x bar less than 170. So, what do I do? I use central limit theorem again. I convert it from x bar we convert it into Z. And we know x bar has mean of 167 and x bar has. So, x bar will be have random variable with mean of mu equal to 167 and standard deviation of sigma by root n. So, sigma is 27 by root of 36. This will come to 4.5.

So, I can then convert this into a standard normal variable. So, this is nothing but 163 minus mu which is 167 by 4.5 less then Z, you are standard normal variable less than 170 minus 167 by 4.5. So, this turns out to be. So, you can make this simplification and this is roughly 2 probability of P, Z less than 0.9 minus 1 is roughly equal 0.6. So, for your sample size of 36 the probability that the average would lie between 163 and 170 is 0.6.

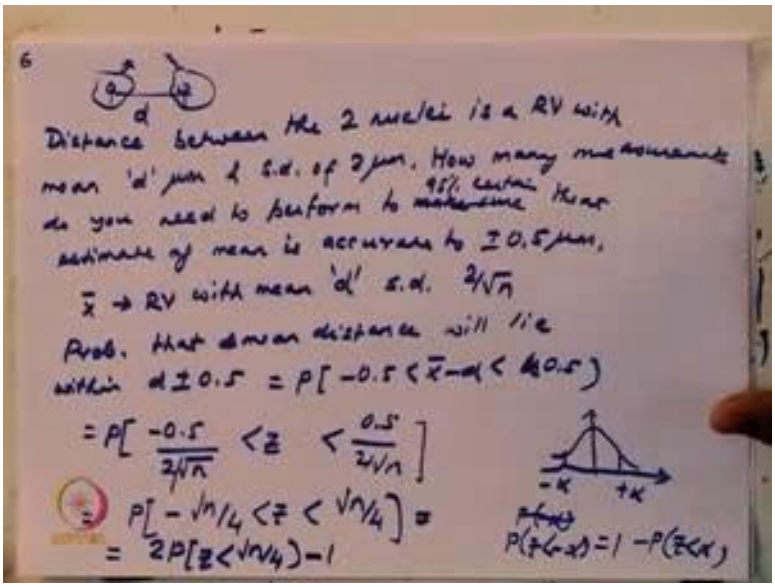
But if you use the sample size of 44 and you do the same calculation, so in that case my probability of 163 less than x bar less than 170 would be probability of 163 minus 167 and this is 27 by root of 44 by root of n less than x bar less than 170 minus 167 by 27 by root of 44. And this probability turns out to be 0.9. So, what you see is by increasing the

sample size merely from 36 to 44, your probability of this estimation increases drastically. From 60 percent you go as high up to 90 percent even for small increase in sample size.

So, this is very important to note. Your spread is inversely proportional. So, you are spread is inversely is the standard deviation of the distribution and that is sigma by root n. Because of this, because you have sigma by root n so with increase in n here the distribution becomes more and more peak. So, this is what we found. So, this was for n equal to 36 and this was for n equal to 44; actually I did not draw it nicely but it was much more peaker than for n equal to 36.

This shows that because of this reverse scaling when you increase the sample size you are much more confident of predicting the mean value.

(Refer Slide Time: 16:58)



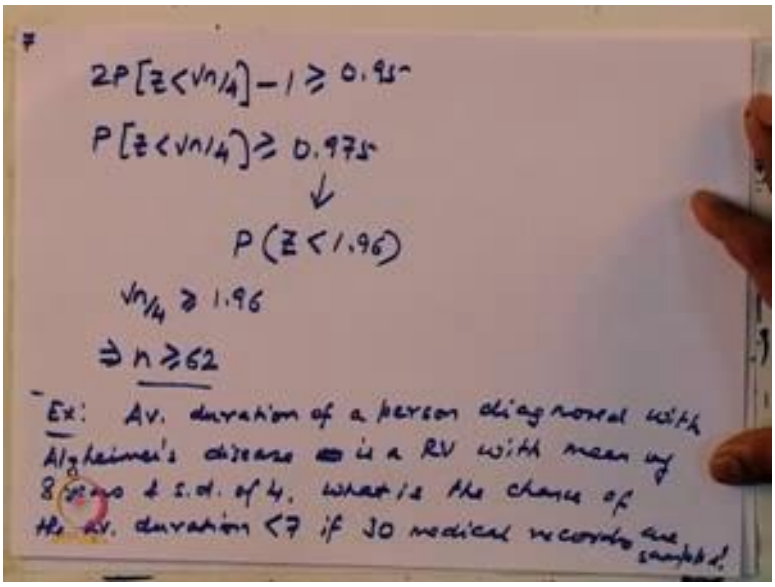
Let us to another example. So, let us say you are trying to do a biology experiment where you have two cells which are migrating, but maintaining some separation between the center of the nuclei; this is d. And what you find is, the distance between the two cells; between the two nuclei let say is a random variable with mean of ‘d’ microns and standard deviation of two microns. So, how many? So we want to know again, we saw that you have this dependence on n right so the question is as follows.

How many measurements do you need to perform to make sure that are to make sure to be 95 percent certain estimate of mu is accurate to plus minus 0.5 microns. So, the question is asked. So, you have this d is the average, you want to know that the estimate of mean is accurate to plus minus 0.5 microns. And given the scaling sigma by root n we know that if we increase the number of n increase the sample size and you would be much better measurement.

So, the question we are posed is as follows mathematically. So, x bar is the distance between the two cells. So, this is a random variable with mean equal to d and standard deviation is to be 2 by root of n. 2 is the standard deviation of each of these, so x bar will be 2 by root of n. So, we want to know. The chance to the probability that the mean distance will lie within d plus minus 0.5 is equal to is given by. If the mean minus d must lie between minus 0.5 and plus 0.5. And what I am also given; so I can convert this into the standard normal variable. So, this is same as probability of minus 0.5 by 2 by root of n; less than Z less than 0.5 by 2 by root n. And this is equal to probability of minus root n by 4 less than Z less than root n by 4.

So, this you can convert it as so, if you have a standard normal distribution, if this is minus alpha and this is plus alpha then probability of minus alpha or probability of Z less than minus alpha is minus 1 minus probability of Z less than alpha. You invoking this I can write this as 2 probability of Z less than root n by 4 minus 1.

(Refer Slide Time: 21:52)



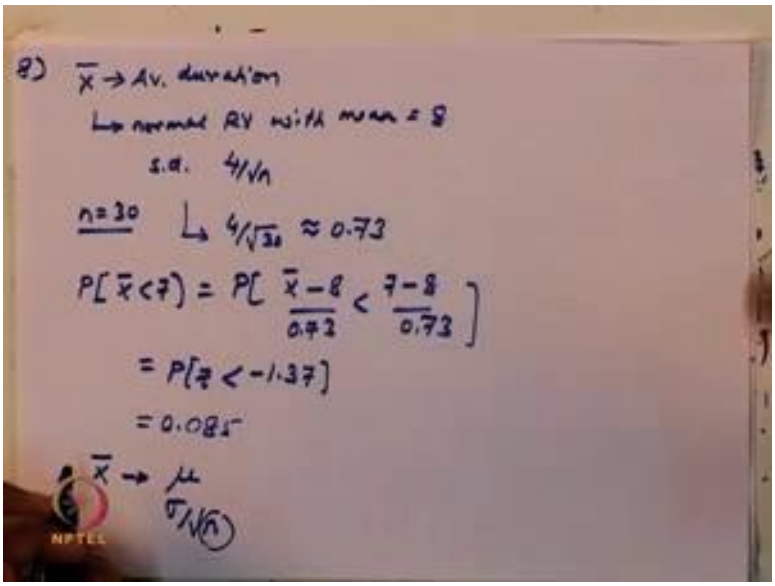
Now, this is the probability that the mean distance will lie within plus minus 0.5 of d. And I want to be certain 95 percent certain that this is so. So what I am given is this 2 P of Z less than root n by 4 minus 1 has to be at least 95 percent. So, this is the net probability and this is greater or equal to 0.95. So, this gives me probability of Z less than root n by 4 is greater or equal to 0.975. So, from now this you can look up the standard normal variable tables and this comes out to be Z is less than 1.96.

So, then what I would need is root n by 4 must be greater equal to 1.96 and this will give me n greater equal to 62. This would give me n greater equal to 62. So, this means that I need to measure 62 times and then average to be 100 percent sure that my error is less than 0.5 microns in terms of estimating the mean. This again shows you that it is very (Refer Time: 23:17) reporting mean you no need to cover adequate samples you need to make enough measurements just for particularly for biological experiments where there is lot of heterogeneity and lot of scope of error. It is important that you make enough number of measurements so that you can be confident of stating what the mean is.

We just give one more example of this; last example for this central limit theorem using for sample mean. So, imagine the average duration of a person diagnosed with Alzheimer's disease, so random variable with mean of 8 years and standard deviation of 4. So, what is the chance of the average duration? So, average duration being less than 7 if 30 medical records are sampled. So, if you sample 30 medical records what is the chance that the average duration will be less than 7?

Once again we need to invoke.

(Refer Slide Time: 25:25)



So, what we are given is x bar is the average duration; let us say x bar is the average duration. So, x bar will follow a normal RV with mean equal to 8 and standard deviation. So, in this case standard deviation is equal to 4 by root n and this in our case n equal to 30 because 30 medical records are chosen. For this your standard deviation comes to 4 by root of 30 is approximately equal to 0.73. So, we want to calculate probability of x bar less than 7 is same as probability x bar minus mu which is 8 by sigma by root n which is 0.73 is less than 7 minus 8 by 0.73. So, it is probability of Z less than minus 1 0.3 7 comes out. So, you put in the values and this comes out to be around 0.085; so just not too low which is around 10 percent close to 0 percent.

So, to conclude today's session we had discussed cases of how you can use x bar or the sample mean to find out to say things about the number of measurements you need. And because of the x bar dependence on the standard deviation of the sample mean is sigma by root of n. So, this is mu which is the same as the x a random variable mean, but the standard deviation is sigma by root n, because of the scaling when you increase the number of measurements you get a much better estimate of the sample mean.

With that I will stop here and we will continue in next class.