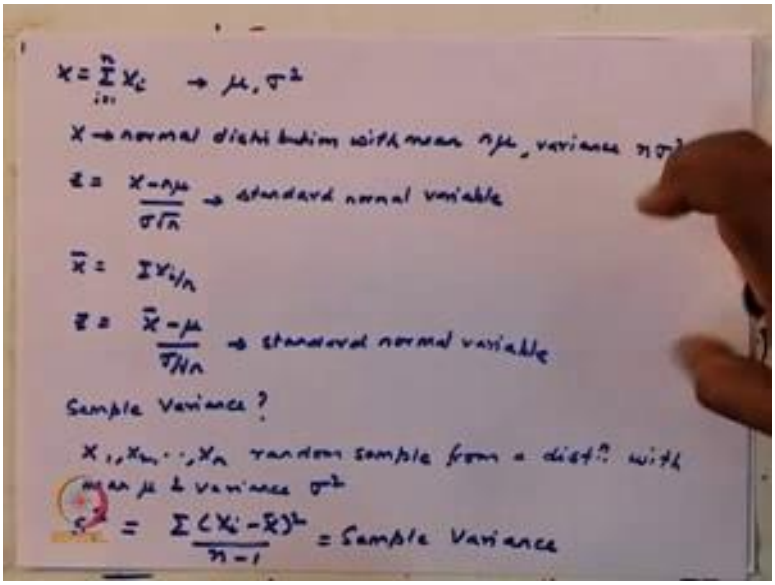
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 29
Central limit theorem - IV and Confidence intervals

Hello and welcome to today's lecture. In the last two classes we had discussed about the central limit theorem. So, it is one of the most important theorems which establish a link between the theory of probability and statistical interference. So, essentially in central limit theorem you are the most powerful statement is even for non normal population's the statistics from sampling distributions of essential statistics like mean or sum of random variables are normal or follow normal distribution if the sampling sizes is large.

(Refer Slide Time: 01:04)



So, two things we had in the last class we had discussed that if I define x as summation of x i. Then an each of x i has mean mu and variance sigma square then summation of x x equal to summation of x i will follow. So, x will follow normal distribution with mean n times mu. So, for summation x i for n random variables; n mu, and variance n sigma square.

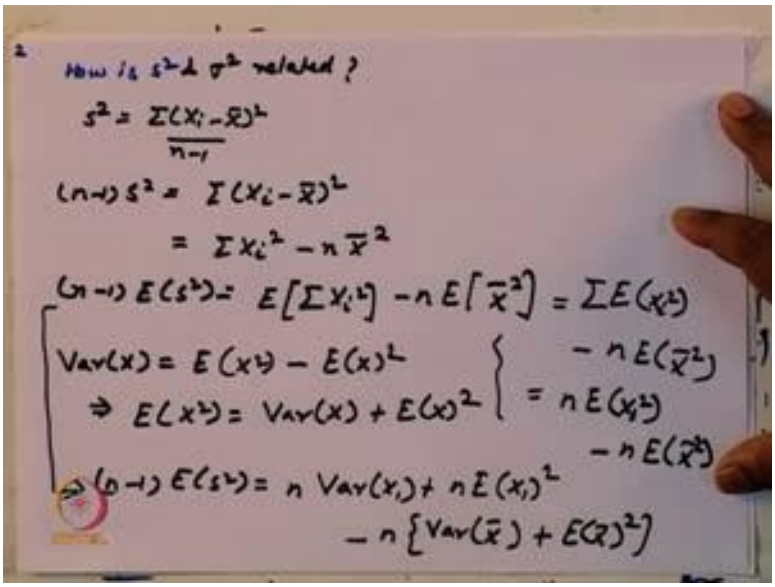
So, we can also convert this normal variable to; we can define this variable Z as x minus n mu by sigma root n. So, Z gives us a standard normal variable. So, Z gives us a standard normal variable which means its mean is 0. So, mean of Z is 0 and its variance

is 1. Similarly, if I define x bar as the mean simple arithmetic mean is summation x i by n then we have found that Z is equal to defined by x bar minus mu by sigma by root n is a standard normal variable.

Now, what can we say about the sample variance; what can we say about the sample variance as another metric. So, how do we define the sample variance? So, the sample variance s square, so if you have x 1, x 2 comma x n as a random sample from a distribution with mean mu and variance sigma square. So, I will define my sample variance as s square (Refer Time: 03:39) s square is equal to summation of x i minus x bar whole square by n minus 1. So, this is the sample variance.

So, can I say anything about the link between the sample variance and the population level variance?

(Refer Slide Time: 04:11)

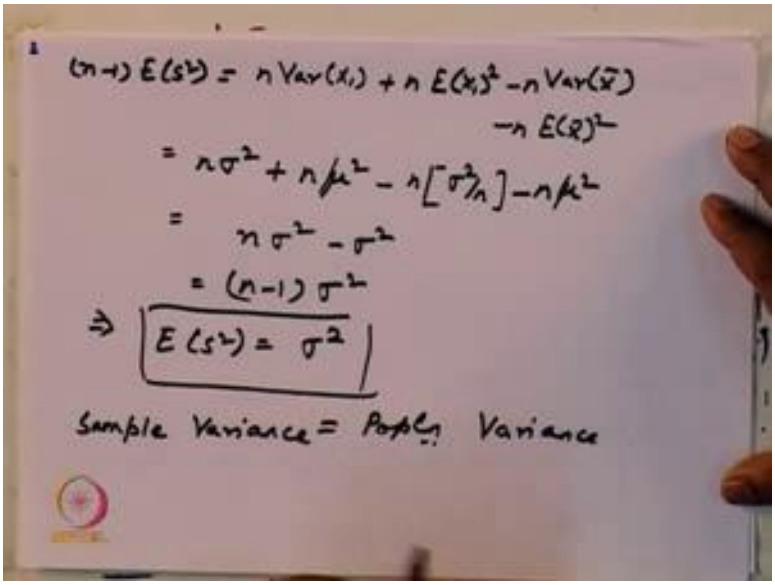


So, essentially the question is how is s square (Refer Time: 04:17), how is s square and sigma square which is the population variance related? Since, s square is defined as summation of x i minus x bar whole square by n minus 1, so I can write n minus 1 times s square is equal to summation of x i minus x bar whole square. Now this I can expand and as we have done before it can be shown that this will come out to be summation x i square minus n x bar square. So, I can take the expectation on both sides of this equation. So, I can write n minus 1 expectation of s square is expectation of summation x i square minus expectation of x bar square.

Now, we know that variance of x is defined as expectation of x square minus expectation of x whole square. So, from this equation I can write expectation of x square is equal to variance of x plus E of x whole square. From this equation I can then simplify; now because each of these excise are independent I can write this equation as summation E of x i square minus n times E of x bar square. So, summation E of x i square can be written as; so I can write n minus 1 times expectation of s square is equal to. So, since each of these E x i is mean n, so I can simplify it as n expectation of x 1 square let us say minus n expectation of x bar square.

So, this becomes n times variance of x 1 plus n times expectation of x 1 whole square. And this part minus n is variance of x bar and plus E of x bar whole square.

(Refer Slide Time: 07:16)

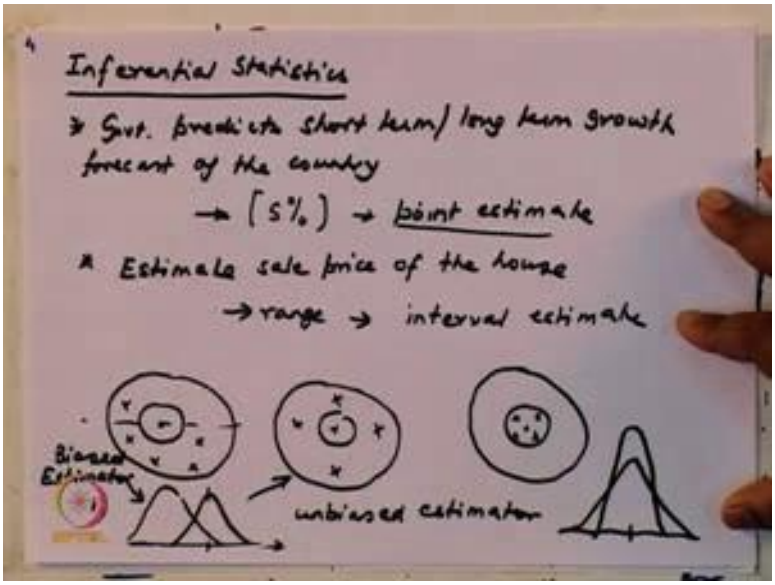


So, I can then again simplify this equation to write n minus i E of s square it is shown as; so n times variance of x 1. Let me write you have n times variance of x 1 plus n times E of x 1 square minus n times variance of x bar minus n times E of x bar whole square. So, this I can simplify as n variance of x 1 is sigma square n times E of x 1 is mu so you have n mu square. Variance of x bar is, so x bar has variance of sigma square by n which we derived. And n a E of x bar is simply equal to mu, so minus n of mu square.

So, this gives us equal to n sigma square; so this has this term cancel each other right out. So, we have n sigma square minus sigma square equal to n minus 1 into sigma square.

Implying expectation of s square is simply equal to sigma square. Thus, the sample variance is equal to population variance. This is another important equation.

(Refer Slide Time: 08:58)



So, coming back to the central limit theorem, where can we make use of central limit theorem. So, we come for the idea of inferential statistics. Why is inferential statistics important? Some examples let us say, government often predicts the short term or long term growth rate right or growth forecast of the country. So, this might be let us say- at the country the GDP will grow at 5 percent at 7 percent so on and so forth. So, here you have what is called a point estimate; that means we are predicting or the statistics is used to predict a single value. Versus, let us say you can estimate. So, you have a house you want to sell it, you want to estimate the sale price of the house. So, this can be arranged. What is the minimum you can expect to get, what is the maximum you can expect to get?

So, in this case what you come up with is something called interval estimate. Now consider the point distribute; let us take out the first example which is a point estimate. Like imagine you have a dart porter, and this is the true value of what you want. So, you want to hit the bullseye or the center point. But when you get the sample the data or you hit it in the board many times let us say you are repeatedly throwing it and this is how you are getting your points. So, these are estimates of what you want of the bullseye.

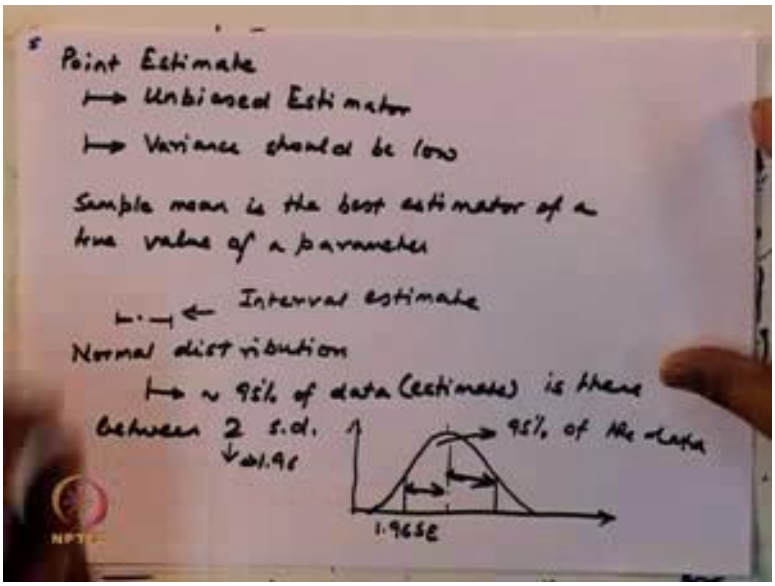
So, this is one example. So, where you see that these values are mostly below this axis. Or you can have a situation where you have points which are all over the place. So, the

difference between this equation here or this equation here; or let us take another case the points are here. So, what is, how do we discriminate between these three cases? So, what is chosen when you want to come up with a point estimate, what should be your yardsticks?

So, number 1: if this is your true value of a parameter that you are trying to estimate. You want, so let us say this is a true value, so this is some axis. This is a true value of the parameter you want to estimate. You want an estimator which is unbiased. In other words it has equal chance of predicting slightly higher or slightly lower values. So, this would be closer to this one and this kind of an estimator is an unbiased estimator; you have an unbiased estimator. Versus, in this case let us say the example we have drawn here its mostly for example you can draw it us like this.

So, this is the true value, so most of the times your values are either underneath. So, this is an example of a biased estimator. Also to compare between this and this, for both of them whatever is the true value is this, one has this as the representation and the other one has this as the representation. So, you can clearly see that in this case it is better because here the estimate, thus variance of this estimate is lesser.

(Refer Slide Time: 13:29)

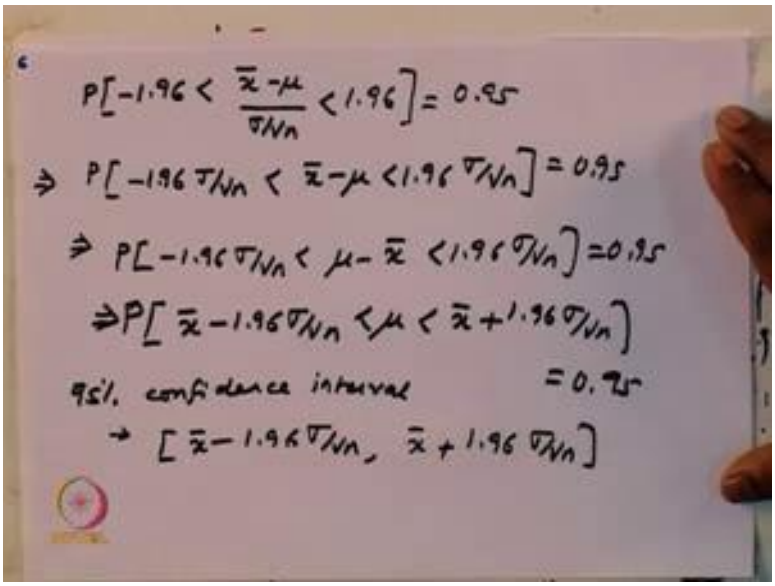


So, that brings us to the idea that for point estimate; write it down, when we want to come up with a point estimate there are two rules: you want an unbiased estimator and second the variance should be low.

So, clearly in our case what we found was the mean the best or one of the best estimator of true value of a true value of a parameter. Why, because we showed that the sampling mean follows a normal distribution and then depending if your sample size is larger, then your variance can also come down. Because the variance for sample mean is sigma square by n. But in the general case as opposed to predicting a small value it is better to come up with a range, it come up with a range. Or, and this kind of a range is called an interval estimate. For a normal distribution we know, so I can write; so we know that between 2 standard deviations 95 percent of the data is there. So, roughly 95 percent of data or is that is estimate is there between 2 standard deviations; so instead of 2 it s actually roughly 1.96 in to be exact.

So, if I were to draw this, this is your distribution that you obtain. Let us say this is your sample mean, this is 1.96 times s E both this and this is 1.96 times s E. And this total area this covers 95 percent of the data, which means that the probability that any estimate falls within this range is 95 percent.

(Refer Slide Time: 16:36)



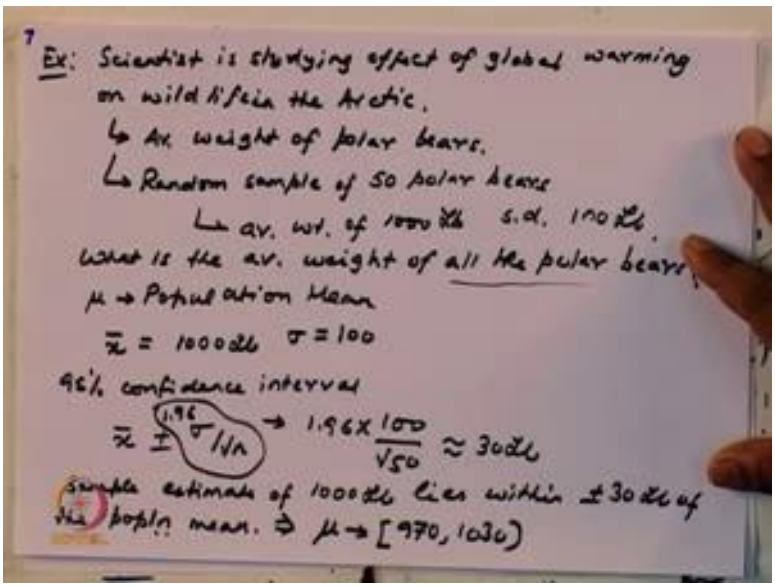
So, I can write down this statement that the probability, so if you are doing with a sample mean we can write down the probability that between minus 1.96 and 1.96. So, the probability that your standard normal variable lies between minus 1.96 and 1.96 is 0.95. So, this is what it means for 95 percent chance that your value estimate is going to be

within 2 standard deviations. So, I can rewrite this equation. I can multiply by sigma by root n, I can write x bar minus mu plus n 1.96 sigma by root n; and this is 0.95.

So, I can again simplify it I can multiply by a negative sign. So, I can write minus 1.96 sigma by root n less than mu minus x bar less than 1.96 sigma by root n. So, this can be simplified further implying, if I add x bar to both of them x bar minus 1.96 sigma by root n less than mu less than x bar plus 1.96 sigma by root n 0.95.

In other words, the difference between estimator, your domain the 95 percent confidence interval is the range given by x bar minus 1.96 sigma by root n to x bar plus 1.96 sigma by root n. This is called the 95 percent confidence interval.

(Refer Slide Time: 18:54)

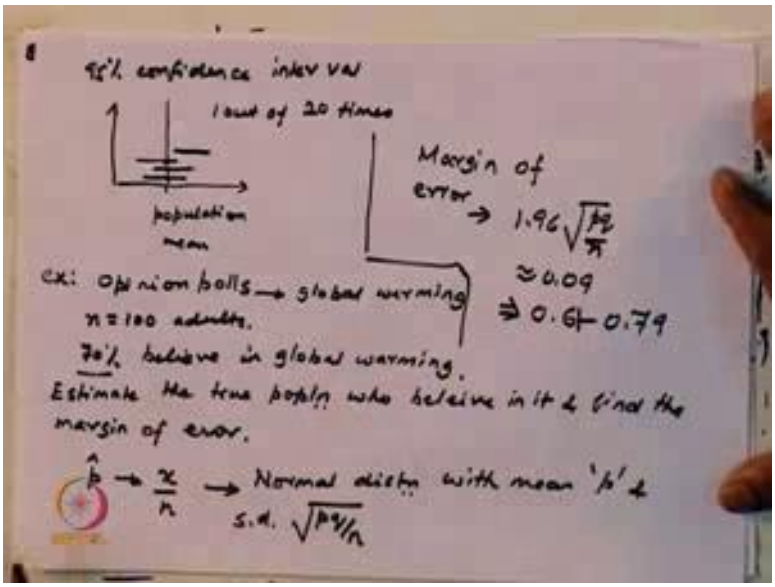


So, let us take up an example; let us do an example. Imagine a scientist is studying the effect of global warming on wildlife in the Arctic. So, as part of this the samples the average weight of polar bears. And what he founds from a random sample of 50 polar bears. So, he comes up with the average weight of 1000 Pounds, and a standard deviation of 100 Pounds. So, the question is based on this can we estimate; what is the average weight of all the polar bears. So, all the polar bears mean that you want to estimate mu which is a population mean. And what you have been given is a sample mean x bar which is equal to 1000 Pounds and the sample variance sigma is equal to 100.

So, for 95 percent; for creating the 95 percent confidence interval I want x bar plus minus 1.96 times sigma by root n. So, this would mean between, so this term will come out to be 1.96 into 100 by root of 50, this is if you calculate they will come to roughly around 30 Pounds

So, what you can say with certainty that the sample estimate of 1000 Pounds lies within plus minus 30 Pounds of the population mean. So, implying the population mean must lie between 970 and 1030 Pounds. So, with 95 percent confidence you can say that the mean is going to be lying between this and this.

(Refer Slide Time: 22:43)



So, what does exactly there is this 95 percent confidence interval mean? (Refer Time: 22:39). What do we mean by saying that 95 percent confidence? What it means? What it means is imagine this is your true mean; this is your true population mean. Let us say you sample it once and you found the sample range interval to be between this and this.

Similarly, another time you did it and you found a range which is somewhat like this, you get valleys between these two and so on and so forth. Only 95 percent means, only once out of 20 times. You will probably get an interval which is like this, where which does not contain the population mean. So, in all of these three cases the range contains the population mean, but this is an example where this range does not contain the population mean.

So, when we say 95 percent confidence interval this means that only in 1 out of 20 times you will have a scenario where the population mean does not lie in that interval. Let us take another example: this is considering about opinion polls. In opinion polls: so let us say you have taken a random sample of 100 adults and this is an opinion poll about global warming. So, an opinion poll and of 100 adults 70 percent believe in global warming; 70 percent believe in global warming.

So, we want to estimate the true population who believe in it and we want to find the margin, the margin of error. So, here we are talking about the proportion. Now for proportion this follows normal distribution. So, proportion is probably x the number of people out of a population who believe in it. So, p the proportion follows normal distribution with mean p which is given in our case to be equal to 70 percent and standard deviation root of pq by n, where n is your sample size.

So in our case, in this case the margin of error becomes 1.96 into root of pq by n. So, if you plug in the values they should be come out to be 0.09 implying the true population would lie between 0.6 to 0.79; roughly 0.61 to 0.79.

With that i would like to conclude our class for today. So, we saw how you can make use of the central limit theorem as a link between probability and statistical interference. And we make use of the idea of confidence intervals to gain a range within which the population mean should lie.

Thank you for your attention, I look forward to next discuss.