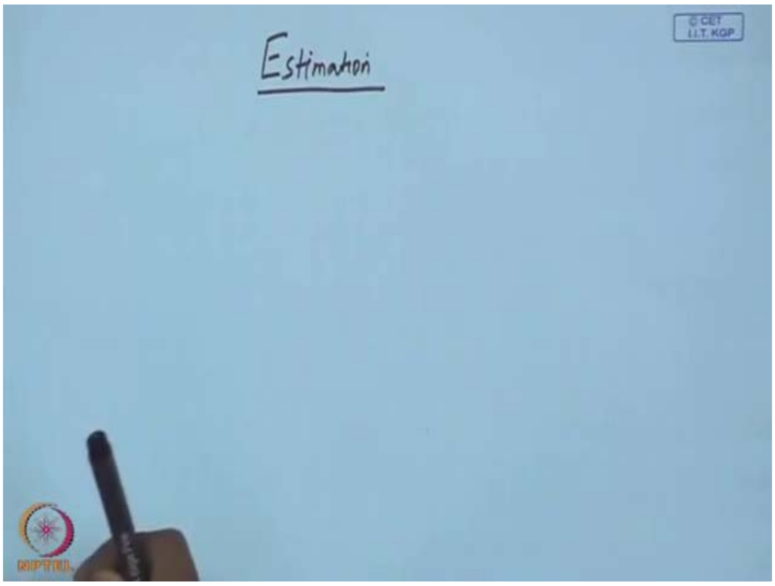
Applied Multivariate Statistical Modelling
Prof. J. Maiti
Department of Industrial Engineering and Management
Indian institute of Technology, Kharagpur

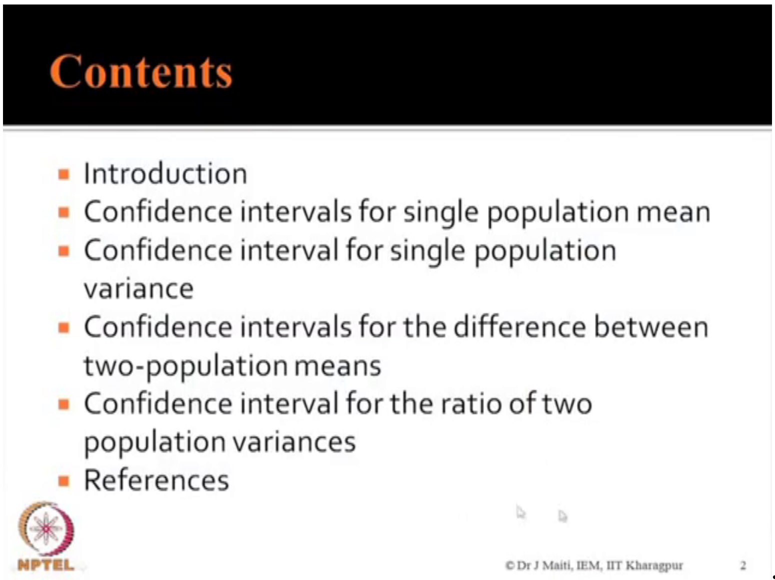
Lecture - 5
Estimation

(Refer Slide Time: 00:28)



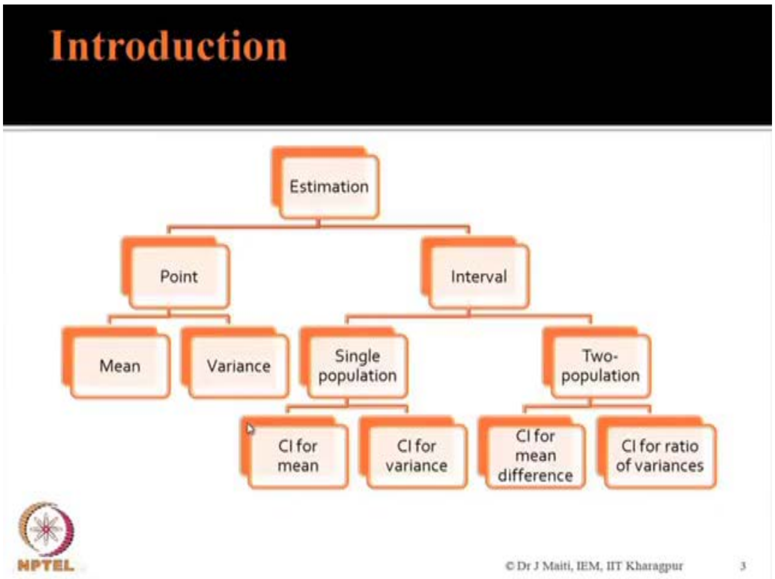
Hello, good morning. Today we will discuss univariate statistical topic estimation. Estimation comes under univariate statistics.

(Refer Slide Time: 00:43)



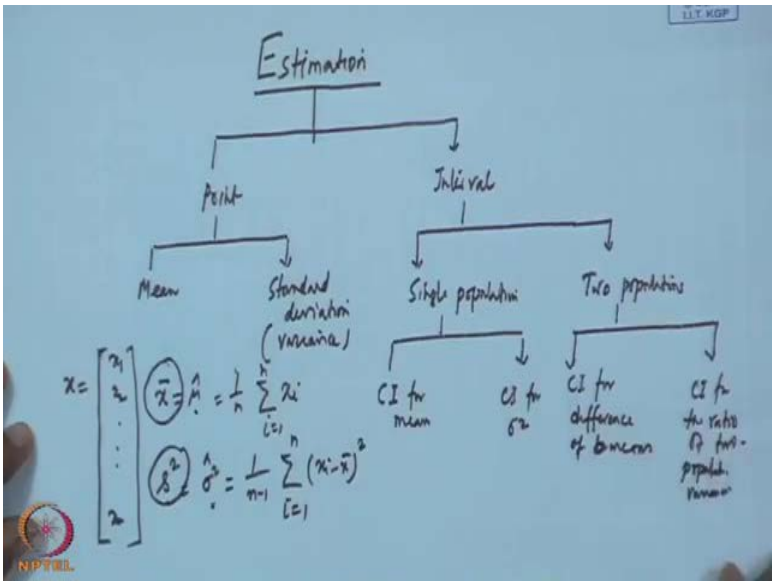
Today's content is we will start with what is estimation? Then I will tell you the different types of estimation like for single population mean for single population, variance, confidence intervals for the difference between two population, means confidence interval for the ratio of two population variances followed by references.

(Refer Slide Time: 01:14)



Now, if you see this slide estimation has two parts.

(Refer Slide Time: 01:20)

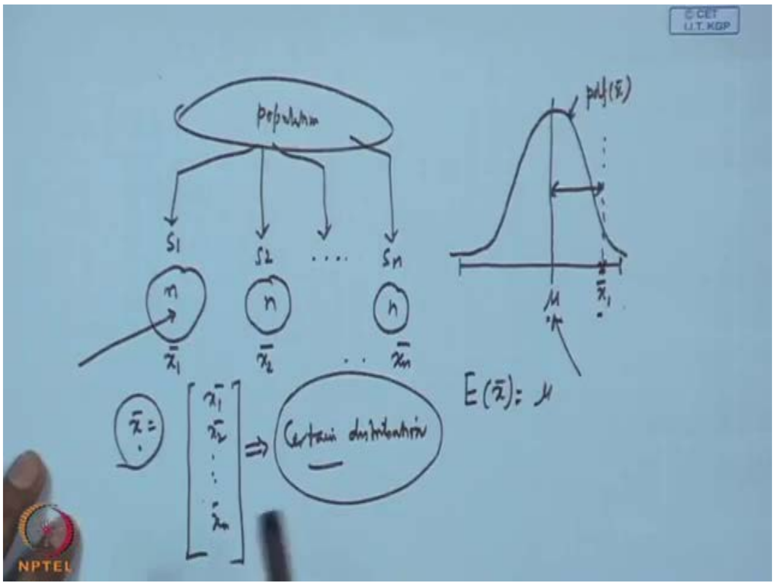


One is point estimation, another one is interval estimation. So, under point estimation we will be discussing about that point estimation of mean, and point estimation of standard deviation or we can say that point estimation of variance that is the square of standard variation, that we will be discussing. And under interval estimation here we will be discussing for first for single population, that is confidence interval for mean confidence interval for mean and confidence interval for variance. We will also discuss today that interval estimation for difference of means between two population, two populations here also difference CI confidence interval for differences, difference of means between two population, and ratio of confident interval for the ratio of two population variances, two population variances, okay?

Essentially what will happen here? Ultimately you will find out when you talk about the interval estimation, the intel logic, the logic remain same whether we will go for the single population or two population and the difference you will find in the little bit in the computation. So, if you have n observations x, x 1, x 2 like this x n, n observation all of you know that the mean, the estimate of mean is the average of the sample data. So, if I say x bar is an estimate of mu, then that you all know that this is i equal to 1 to n x i. So, what do we say that x bar is the estimate of population mean.

Similarly, we will calculate variance, sample variance which we say that the estimate of population variance which will be n minus 1 sum total of i equal to 1 to n x i minus x bar square. When you compute like this that you collect a sample and compute x bar s square from the sample, this is your point estimate. So, x bar is the point estimate of mu s square is the point estimate of sigma square. Now, as we have discussed in last class that what will happen when I go for several samples collected from a population.

(Refer Slide Time: 05:02)



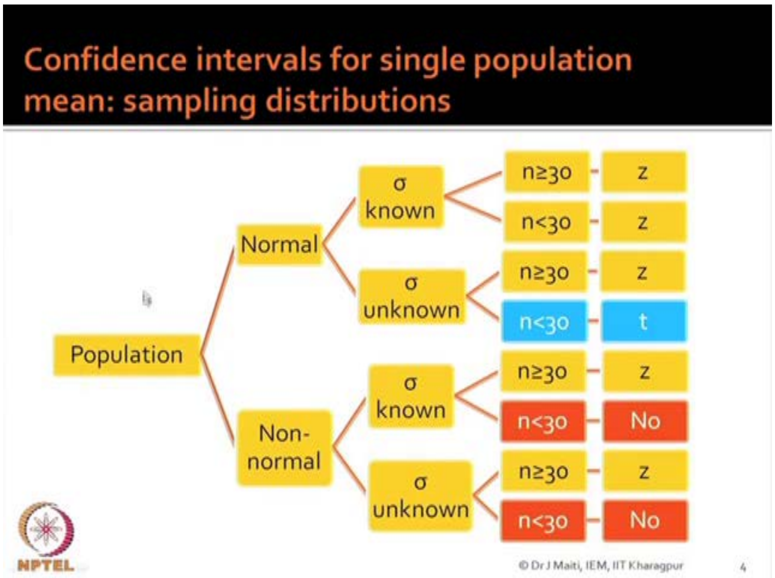
I told you in the last class this is my population and if I go for several sample collected so for example, you have collected sample 1, sample 2, then sample n, all of the samples with size n one, size n equal sample size n and here also n and if you calculate the point estimate that will be x 1 bar, x 2 bar like this x n bar and we have seen that this x bar if I write in this vector form that x 1 bar, x 2 bar, let x n bar. So, this follow certain distribution, last class we have seen this follow certain distribution.

So, if x bar follow certain distribution. Now, you are collecting one sample and computing x bar, what is the guarantee that the computed x bar will be representing the population mean? So, we want to have certain amount of confidence in our estimate. So, that confidence is known as confidence interval, by confidence interval what do we mean? We mean that suppose the distribution of x bar is like this, this is my pdf of x bar and all of us know now that expected value of x bar will be mu because this is the property of unbiased estimation. So, then your mu is coming here.

Now, let us talk about the sample 1 and you have computed x bar using S 1 and it is falling here, the value of x 1 bar is falling here. So, what is our interest here using confidence? We want to know that whether this x 1 bar or the x bar collected using sample 1 is representative of mu or not. It all depends on the distance between this two.

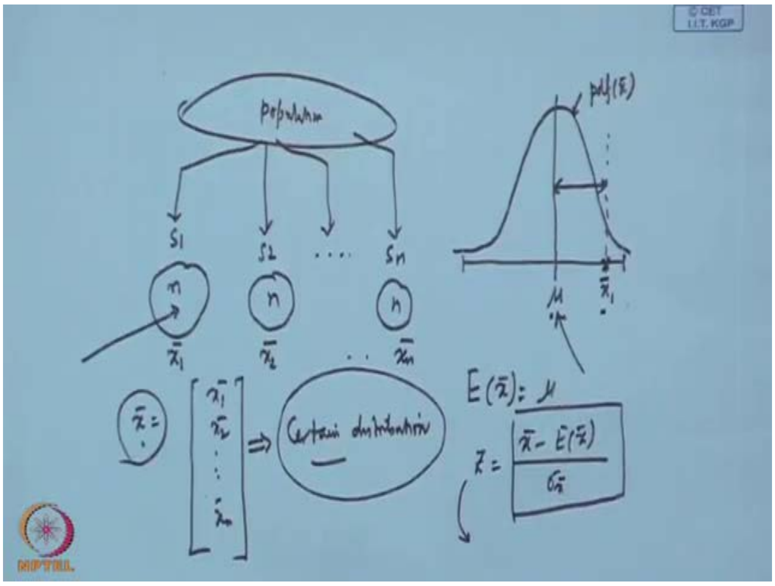
If x bar is far away from mu then it can it will not be a representative 1. So, in order to know whether the x bar contains the interval of x bar contains mu or not, we will go for first identifying the what is the distribution sampling distribution that is applicable. And using this sampling distribution, we will generate a interval and we want to find interval for mu, that is a population mean and we want to find out that whether that interval contains mu or not.

(Refer Slide Time: 08:19)



Now, come back to the slide here. What will happen ultimately? Population can be normal and non normal. Now, of you sample from normal population with sigma known and sample size whether small or large that is not a problem, not a question issue at all. Then you you will fall a z distribution, then what quantity will follow z distribution? I will discuss, but you please remember that if your population is normal and sigma is known irrespective of the sample size. The statistics will generate for x bar, which is which is basically z equal to x bar minus expected value of x bar by sigma x bar.

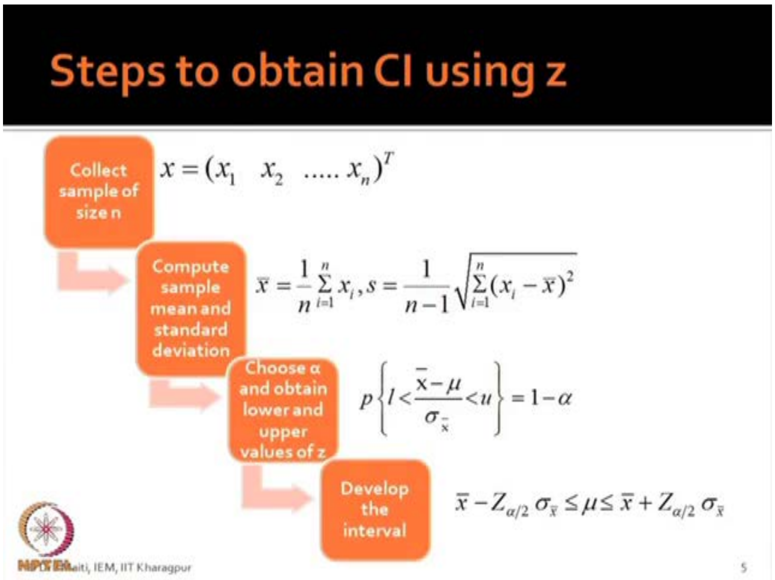
(Refer Slide Time: 09:06)



These quantity follows that this x bar minus expected value of x bar by sigma x bar follows z distribution if you sample from normal population, irrespective of the sample size. But sigma must be known, if sigma is unknown your sample size is large then this quantity follows again z distribution, but if sigma that sample size is small and sigma is unknown then this quantity follows t distribution.

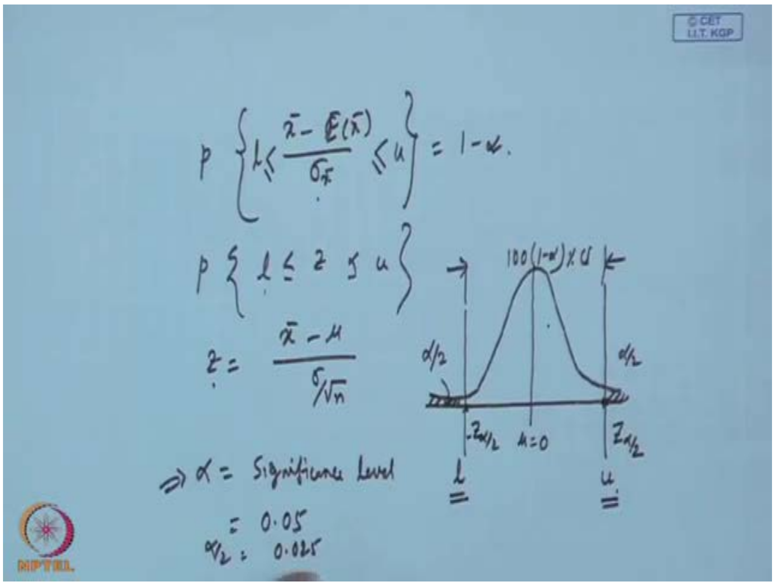
Now, if you sample from non normal population when sigma is known and your sample size is large then this quantity again the same quantity, this quantity follows z distribution even if sigma is unknown is also, but sample size is large that is again z. But other two cases when sample size is less than 30 that is the small sample size, then no parametric distribution possible, okay? So, whether you will use t or z distribution, the mathematics and the procedures remain same. Only you have to use z table or t table.

(Refer Slide Time: 10:35)



Let us see here what is what way you can calculate the confidence interval, you see what we have said, first we collect data. Then find out the statistic what you want to compute, that is the estimate of mean population, mean is x bar, estimate of population standard deviation is s. Then you choose a particular alpha value and we know that probability that this quantity what the statistic what we generated.

(Refer Slide Time: 11:10)



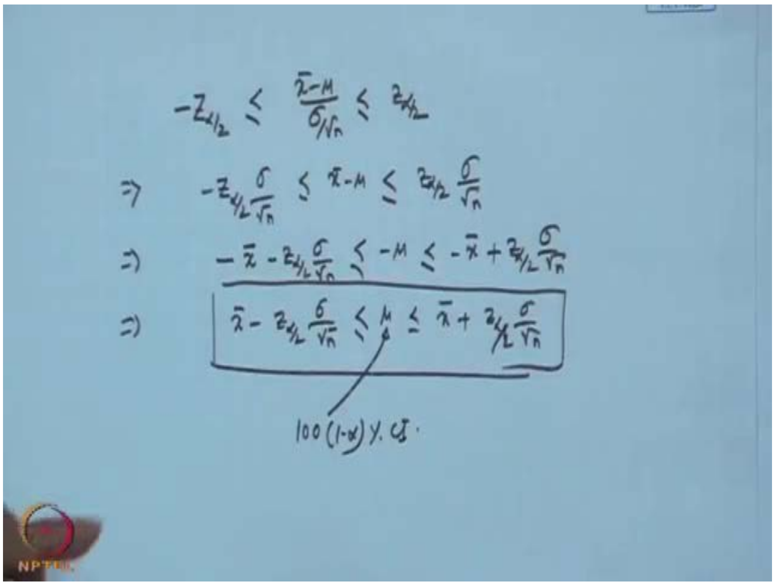
That x bar minus expected value of x bar divided by sigma x bar, this one probability that this value will be greater than l and less than or greater than equal to l and less than equal to u, that will be 1 minus alpha, okay? So, if your distribution this one is nothing but z that l less than equal to z less than equal to u, when some the condition satisfied like from normal population irrespective sample s sigma is known then what we will write basically? We will write z equal to x bar minus mu by sigma by root n, the sigma x bar is sigma by root n. We have seen in the last class and expected value x bar is this.

So, essentially what you can write that this is normally distributed unit, normal distribution, this is my unit normal distribution and there is one value which is the lower value. We are expecting considering, another one is the upper value you are considering and what we are saying the probability, that this z value lies in between l and u and that is the confidence interval which is 100 into 1 minus alpha percent CI.

What is this alpha? How do you determine this alpha? Alpha is known as significance level, this alpha when our, this our this x bar minus mu sigma by root n that is z distributed, it is a two tailed case. So, left hand side and right hand side will be that probability value will be equally divided. So, the portion here is alpha by 2, here is alpha by 2. So, significant level of significance or significance level alpha what it indicates? It indicates that if I consider that l to u that is the confidence interval then what is the error you are consuming? That error is alpha, what is error the probability that that true mean lies in this portion that is basically alpha percent probability, okay?

So, now how do we get this value? This l and u what will be l and u? It all depends on what will be your alpha. If we consider alpha equal to 0.05, then alpha by 2 is 0.025, then what do you require to know? Now, that you have to find this z alpha by 2, this is your z alpha by 2 and this left hand side this value will be minus z alpha by 2. So, you see the table and find out z alpha by 2 value and accordingly you compute.

(Refer Slide Time: 14:42)



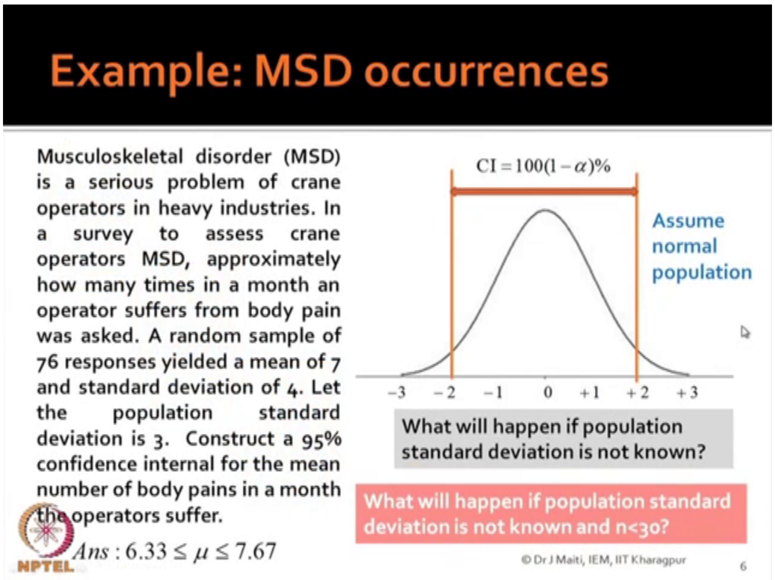
So, then mathematically what is happening here? Mathematically is my l is minus z alpha by 2 and this will be less than equal to x bar minus mu by sigma by root n, that equal to plus z alpha by 2. Now, if you rearrange this one what you were getting z alpha by 2 sigma by root n less than equal to x bar minus mu less than equal to z alpha by 2 sigma by root n.

Then again, we will manipulate this. So, I will just bring that because we are looking for confidence interval of mu. So, what do you do? We basically separate, we will take x bar from this, the middle portion then if I write again I will write like this, minus x bar minus z alpha by 2 sigma by root n. This is less than equal to minus mu less than equal to minus x bar plus z alpha by 2 sigma by root n, correct?

Very simple manipulation, you are just now taking out x bar from the middle portion putting to the left hand, right hand side then if you little modify. Now, we do not want minus mu, we want plus mu. So, you are multiplying it by minus 1. Now, it will just the reverse will take place what will happen x bar minus z alpha by 2 sigma by root n less than equal to mu, less than equal to x bar plus z alpha by 2 sigma by root n.

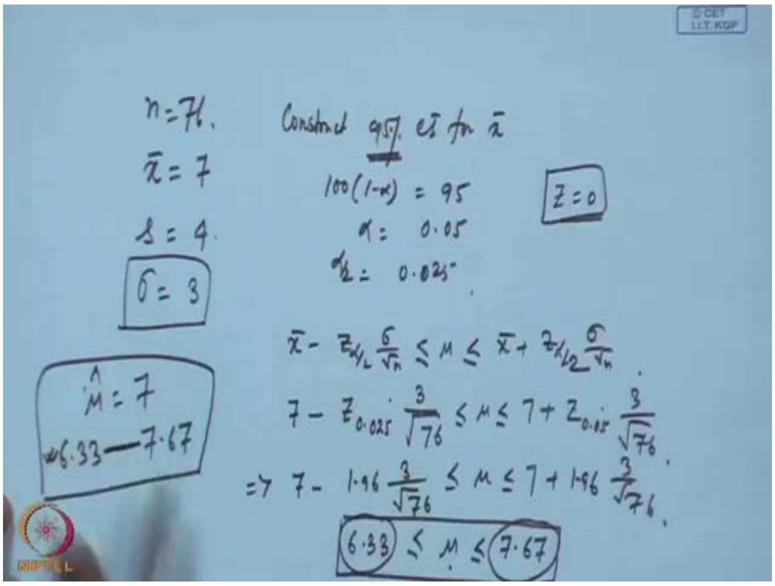
So, this formula is applicable when you this is the confidence interval for mu, 100 into 1 minus alpha percent CI for mu, correct? So, when you talk about confidence interval it is definitely for the population parameter, not for the simple statistic, getting me? Then once I know this one what will happen ultimately? If you know these then how do I know that whether this mu basically contain this within this, in this interval mean is contained or not that how do you know? Because if i it is basically the z value.

(Refer Slide Time: 17:20)



I think let us go for a problem first then suppose this is our problem, this is a problem, what is this? Musculoskeletal disorder is a serious problem of crane operators in heavy industries. MSD in a survey to assess crane operators MSD, approximately how many times in a month an operator suffer from body pain was asked. You are asked that this is the measure of one of the measure of MSD could measure of MSD that how many times in a month a random sample of 76 responses yielded a mean of 7. So, what is our problem here? We have taken collected n equal to 76.

(Refer Slide Time: 18:01)



You have computed x bar which is 7 and standard deviation that S equal to 4. Let the population standard deviation is given 3, that is sigma equal to 3 constructs 90 percent confidence interval for x bar, your work is construct 95 percent CI for x bar, this is your work.

Now, see when I say 95 percent that means we are saying that 100 into 1 minus alpha equal to 95. So, you are getting alpha equal to 0.05, so what is my alpha by 2? 0.025. Now, here it is clearly given that sigma is 3, that population standard deviation is known, and we are assuming that the sample has come from the normal distribution, population distribution is normal then this is normal. So, you see this is normal, assume normal distribution then what we will use? We will use z distribution and accordingly our interval will be x bar minus z alpha by 2 sigma by root n, less than equal to mu less than equal to x bar plus z alpha by 2 sigma by root n, every values are known to you.

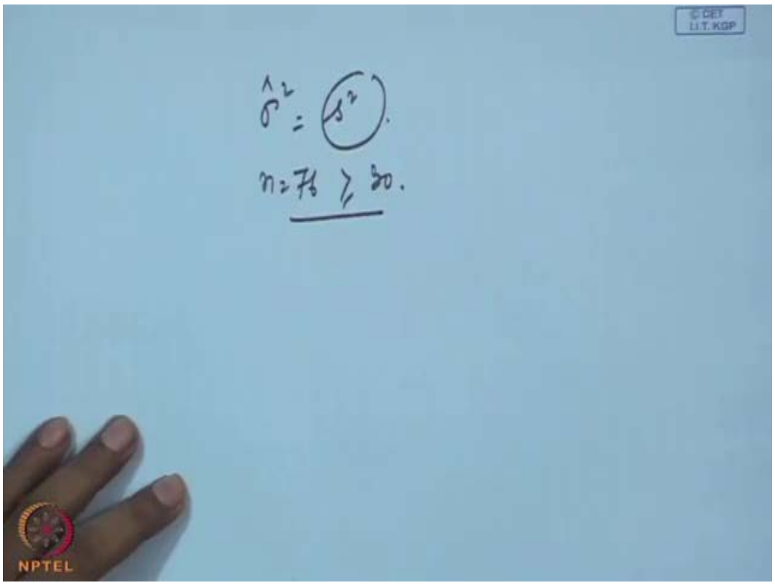
So, your computation is be x bar will be 7, you have to know z, 0.025 into sigma, population sigma is 3 and your n is 76. So, less than equal to mu less than equal to 7 plus z 0.025 into 3 by root over 76. All of you know that that z 0.025, if you see table it is 1.96. So, I can write further that our interval is 7 minus 1.963 by root over 76 less than equal to mu less than equal to 7 plus 1.963 by root 76 and the answer will be 6.33 less than equal to mu less than equal to 7.67, this is the confidence interval for mu.

So, that mean essentially what you are getting? You are not getting a point estimate only, here you are getting an interval estimate, point estimate says mu estimate is 7, interval estimate says it is not 7, it is in between 6.33 to 7.67. So, it is 6.33 to 7.67 that is the difference between your point estimation and interval estimation or point estimate of mu and interval estimate of mu.

What is how to know that whether these estimate this one, this one contain the mu true mu. Now, see when we convert the statistic that x bar to equivalent z by subtracting its mean and dividing it by standard deviation. So, what is the mean of z 0, it is 0. Now, here what is you are getting the interval? 6.3 minus 7.67. So, can you find out some meaning of that if I convert into z 0 and you are getting when you are again translating back to the mu term. What will happen ultimately? See you are getting positive left hand that 6.33 is also positive and 7.76 is also positive.

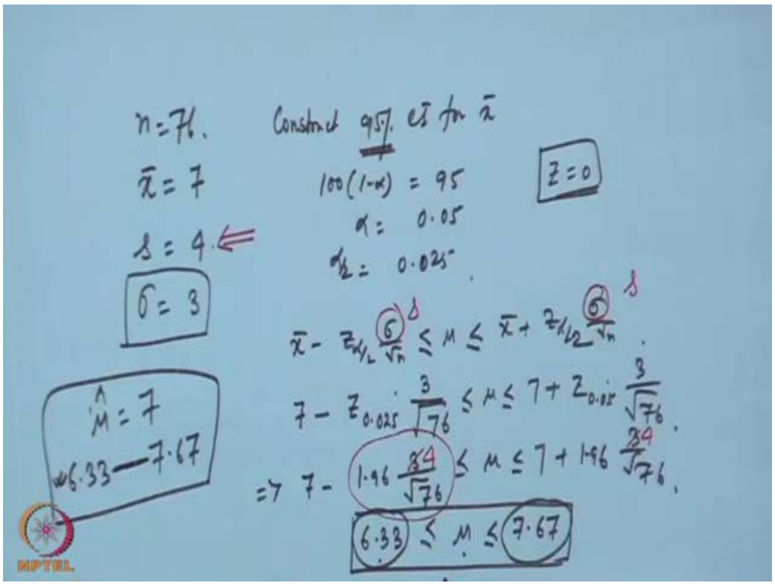
You think next class I will explain if I want to say that that how do I know my question to you that, how do I know that this interval contains the mean or not? Last class I also last but one population mean last but one I have given you 1 1 similar question also, but I have not asked this is one question we will discuss but you must remind me next class, because I am giving you in the belief that you will go through the book. Now, next question here is what will happen if population standard deviation is not known, you will use s.

(Refer Slide Time: 23:33)



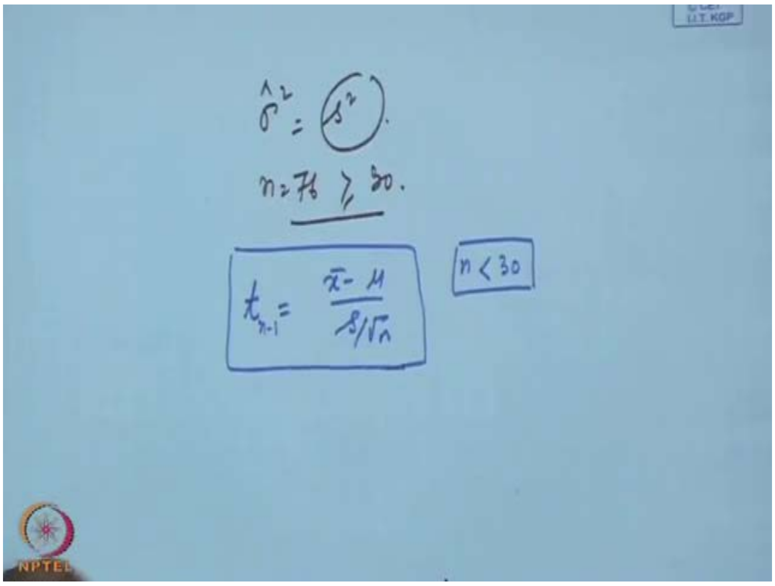
Yes, that sigma square will be s square, population standard deviation is not known, mean sigma square estimate that is what we are saying s square. So, you will be using s square. Now, here sample size is 76, it is a large sample because it is greater than 30. So, you can still use the z distribution but your change will be here.

(Refer Slide Time: 23:58)



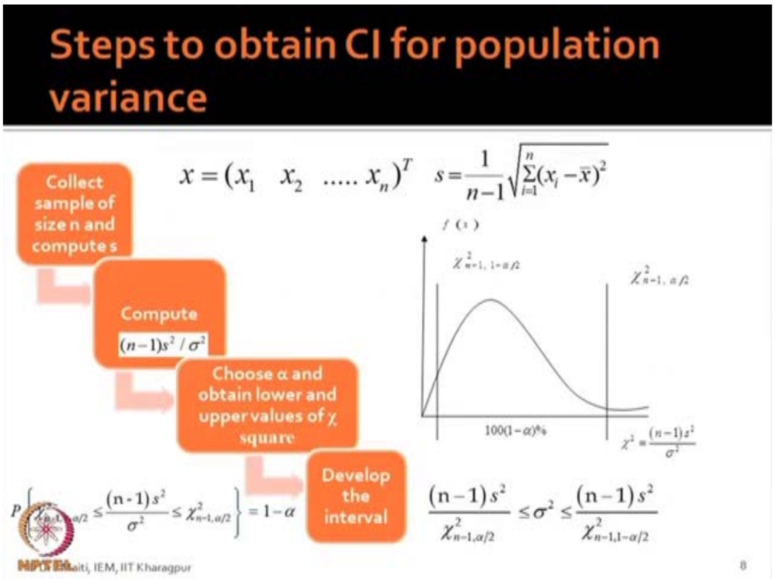
What will be your change? You will be using the same z distribution but instead of sigma, you are using s. You see in the given problem sigma and s are not same. So, s equal to 4 sigma, sigma equal to 3. So, you will replace everything, this will be by 4 by this and this will be 4 by this. So, ultimately what will happen? This resultant quantity will be bigger than the earlier one and the interval will increase. So, if you compute this you will be finding out interval will increase. Now, another question here is that what will happen if population standard deviation is not known and sample size in less than 30. So, you cannot use z distribution.

(Refer Slide Time: 24:58)



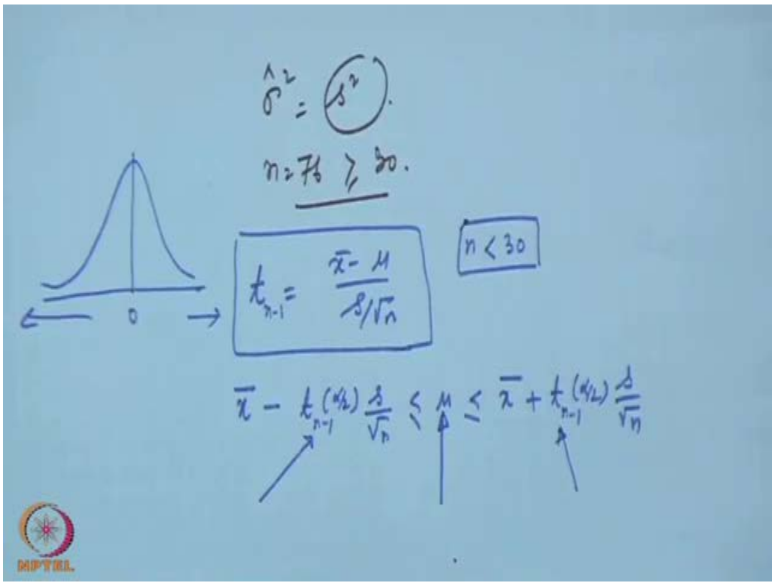
What is required? Now, your case is going like this. You will be using t distribution which is x bar minus mu by s by root n. You will be using distribution that is why the reason is n is less than 30. Please keep in mind when you sample from normal population with sigma population variance is known, irrespective of your sample space you will use z distribution. Population variance is not known but sample size is large, you will use z distribution. Population variance is not known, sample size is small but you are sampling from normal distribution, that is t you have to use. See rest of the things are same. Now, in case of t I told you earlier that there will be degrees of freedom for t distribution. In this case this will be n minus 1 degrees of freedom.

(Refer Slide Time: 26:10)



So, let us see the t distribution. So, z this one I have shown you, so t distribution I am not kept here. So, you have to use t distribution, you are getting me? What is required then here.

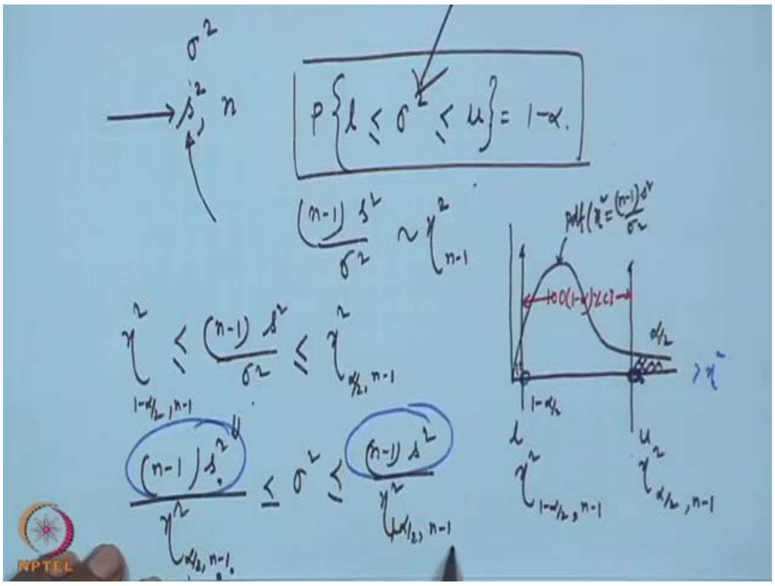
(Refer Slide Time: 26:22)



Your interval will be like this, x bar minus t n minus 1 alpha by 2 because t distribution also 2 del distribution. It is your min value is and this side and this side it is that two extremes negative to positive that minus infinite to plus infinite. Now, then s by root n less than equal to mu less than equal to x bar plus t n minus 1 alpha by 2 s by root n, the same problem if you collect observations which is less than 30 as well as your population variance is not known.

You use t distribution and then find out what is this n value and then find according n, what is alpha value and accordingly you find out the t n minus 1 alpha by 2 value and put into this formula. You will be getting the confidence interval for mean population, mean I repeat that always confidence interval for the population parameter, okay? So, same thing can be applied to population variance also but in case of population variance your distribution will be different. What is happening here in population variance?

(Refer Slide Time: 28:01)



Population variance means sigma square. What do you want? You want your, you have a sample is collected and s square is calculated, a sample standard variance is computed your n is the sample size. So, what do you require to know? You require to know a confidence interval based on this sample data. You want to know the confidence interval for this. Again what do you want? What will be the you got this s square value, but what is the lower value and what is the upper value? Similar manner you have to find out.

So, that means if I know the distribution. I can say this is what is this, this is 1 minus alpha, in true sense you want to do like this, that is what you want basically, but you have your this value, what is the computed statistic value is s square and we have seen under sampling distribution that s square. It is not s square, it is n minus 1 s square by sigma square follows with distribution, chi square distribution. What will be the degrees of freedom? n minus 1. So, n minus 1 degrees.

So, you know chi square distribution because we have seen chi square distribution depending on degrees of freedom. It will of different shape, one may be this is the chi square distribution, what you and this one is the PDF of chi square, which is basically in this case n minus 1 s square by sigma square. And you want to find out a upper value, that is u and a lower value, that is l for n minus 1 square by sigma square.

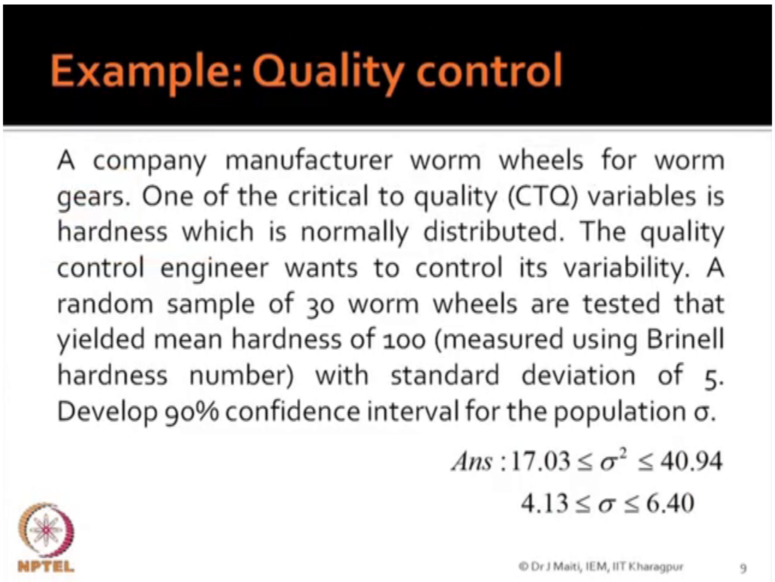
So, again if I consider alpha that this one is basically this particular, this side it is let it be that alpha by 2 and this one what will happen? Total is 1, 1 minus alpha by 2 you will be getting chi square 1 minus alpha by 2, but please keep in mind chi square also having a degree of freedom, that is alpha by 2 and degree of freedom is what n minus 1 n minus 1, getting me?

So, then what you will write? Then you will write n minus 1 s square by sigma square. It must be less than equal to chi square alpha by 2 n minus 1 as well as it must be greater than equal to chi square 1 minus alpha by 2 n minus 1. You see these two, this is left from a interval point of view and this total it is in between whatever it is there, that is what is our 100 into 1 minus alpha percent CI. For the variance this is my but what do you want here in this equation? What do you want? You want something like this, something like l less than equal to sigma square less than equal to u. Can you now find out that what will be can you not manipulate this? You can easily manipulate.

So, what will be the once you manipulate this that mean what will happen? You will want to keep in between the less than equal to terms, only the sigma square. So, if you manipulate you will be getting like this, n minus 1 s square by chi square alpha by 2 n minus 1. You will get this s alpha by 2 n minus 1 less than equal to sigma square less than equal to n minus 1 square by chi square alpha by 2 n minus 1.

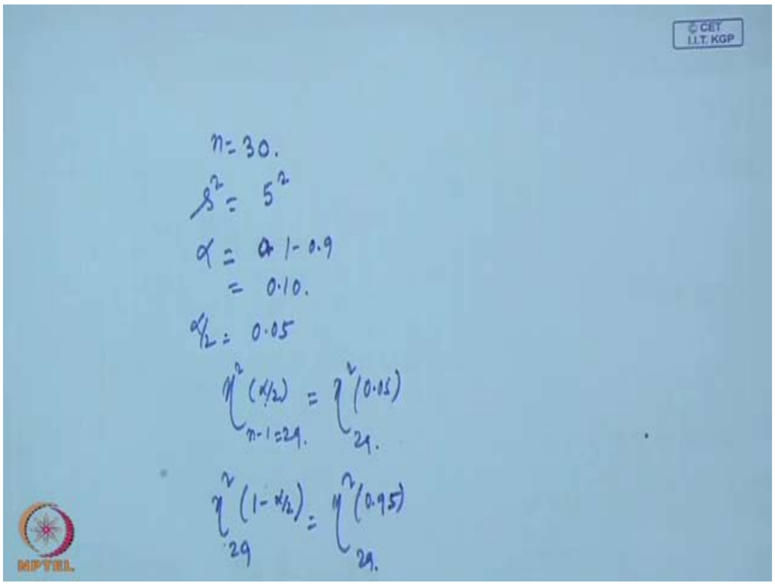
You see the here you see the denominator. Here is n minus 1 into s square, here also n minus 1 into s square, same quantity the difference is in the new, sorry in the numerator. Both are like n minus 1 into s square but in the denominator that is s square chi square alpha by 2 n minus 1. What value is this? This is this value and in the right hand side this value is other one and you can find out that this is the chi square axis. So, definitely this value is less than this value, so that mean n minus 1 s square by chi square alpha by 2 n minus 1 is definitely less than n minus 1 s square by chi square alpha by 2 n minus 1, and that is the interval, okay?

(Refer Slide Time: 33:35)



Now, given data what you will do? Suppose this is the data, can you not compute this a company manufactures worm wheels for worm gears, one of the critical to quality variable is hardness which is normally distributed. The quality control engineer wants to control its variability, a random sample of 30 worm wheels are tested that yielded mean, hardness of 100 which is measured using Brinell hardness number with standard deviation of 5, develop 90 percent confidence interval for the population sigma.

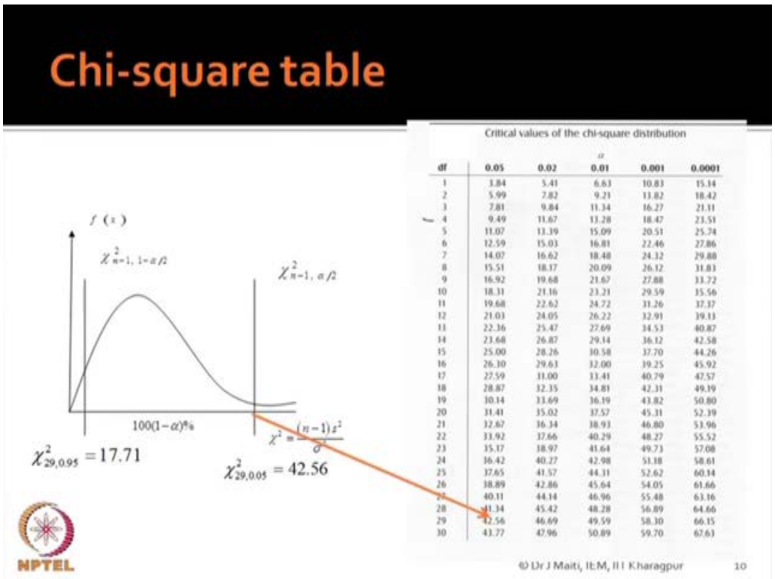
(Refer Slide Time: 34:24)



What you will do? You will all what is n value, here n equal to 30. What is s square population? That means you have collected a sample of 30 with standard deviation of 5. So, that mean this is 5 square, correct? Then what more you want? You want, you want nothing only one thing you want to know that is what is alpha. So, we are saying 90 percent confidence interval.

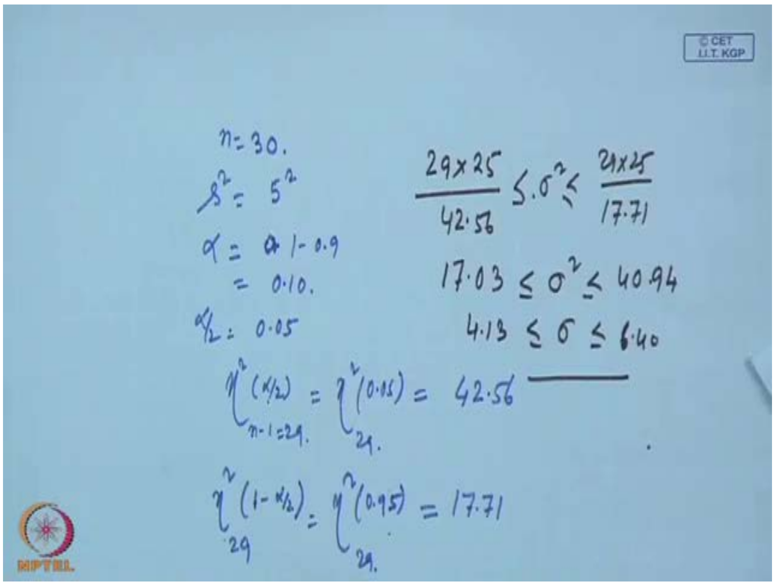
So, alpha equal to 0 1, that is 1 minus 0.9, that means 0.10. So, your alpha by 2 is 0.05, you want to calculate chi square 2 value, chi square alpha by 2 n minus 1, where n is 30 that is 29 and alpha is your that alpha by 2 is 0.05. That is chi square 0.05, that is chi square 0.05 29, you require to find and get as well as one more value you want to know that is chi square 1 minus alpha by 2 with again same n minus 1. So, that mean chi square 0.95 29.

(Refer Slide Time: 36:04)



If you know these two value then put into this equation, s square is known 25, n minus 1 is 29, chi square that is 0.05 29 chi square 0.95 29, you have to find out the chi value. See that table, you see this table that in this table, our alpha value a alpha by 2 is 0.05. So, we want to first know that chi square, our degree of freedom is 29 and what is this value? That value is 42.50. So, chi square our 29 0.05, which one is this, chi square 29 0.05 that is 42.56.

(Refer Slide Time: 36:35)



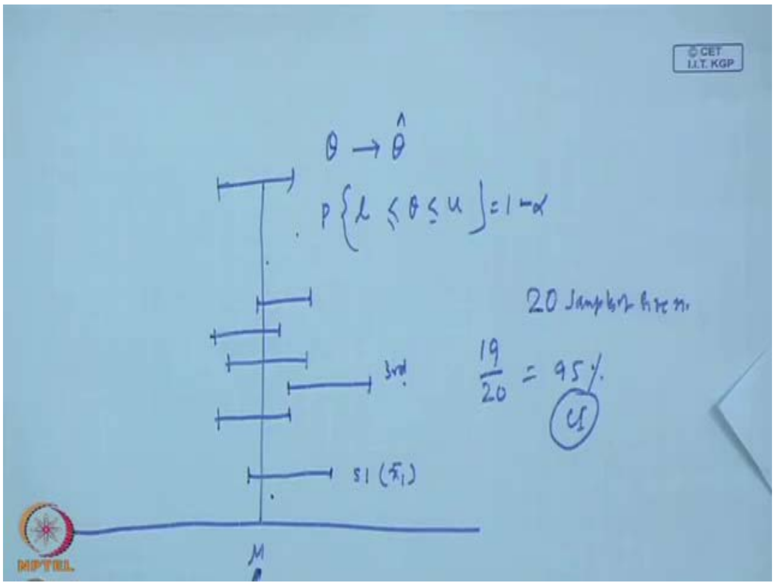
Similarly, chi square that 99 percent, that is 17.71. So, once you know these two values, 17.71 and these values you know your computation becomes very simple, n minus 1 that is 29 into s square is 25, divided by chi square. That one, that mean 42.56 less than equal to sigma square less than equal to 29 into 25 by 17.71 is in the formula, then what is the result? Ultimate result will be like this. So, you will be getting 17.03 less than equal to sigma square less than equal to 40.94 and if you go by sigma, then square root of this 4.13 less than equal to sigma, less than equal to 6.40, that is what is interval estimation.

So, if you really want to know what, how to go about interval estimation, please keep in mind you must know the statistic. You are interested to know the interval estimation for a population parameter. First you know the population parameter, you must know the corresponding sample statistic, you also know that what will be the basically the statistics for which you want to develop the sampling distribution. If it is x bar then you are converting into z or t, if it is your sigma square then you are converting into appropriate statistic n minus 1 s square by sigma square which follows chi square distribution.

So, unless you do not know the distribution as well as the statistics required you cannot go about interval estimation, and what is the advantage of interval estimation? As I told you instead of a point estimate you are getting an interval, and what do you mean the 95 percent case, the mean value will lie within this interval. For example, in the this example variance case what we are saying the variance or standard deviation whatever you consider.

Suppose, variance is 95 percent, 90 percent of the cases this sigma value will lie, basically 90 percent we are confident that sigma value population sigma square value is in between this somewhere. It is there sir logically point estimation value will be inside the interval. Basically, if you want to understand interval that fast whether interval estimation, what is the meaning of this? Suppose, there is population parameter theta and you have estimated theta cap using some sampling distribution.

(Refer Slide Time: 39:49)



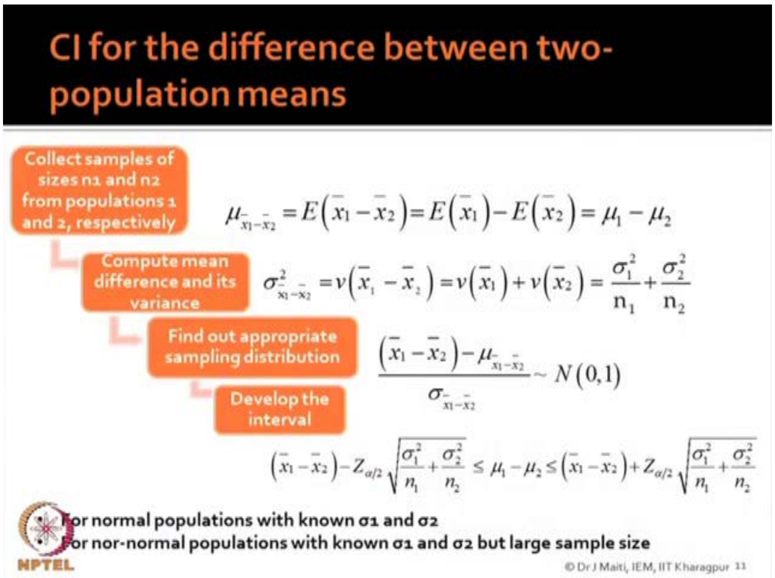
Now, let the theta cap has a particular distribution. Then what you are doing? You are basically doing like this, less than equal to theta less than equal to u, that you are doing which is 1 minus alpha, using the theta cap value that range you are getting theta cap value. Now, let us concentrate consider the theta is mu that is the mean value, there theta is mu.

Suppose, you have collected a sample with n size and you have computed the x bar and you found out that x bar interval is like this. This is your sample 1 and your x 1 bar its interval you found out, because once you collect a sample you know x bar. You can calculate the interval now, second one suppose like this, third one suppose like this, fourth one like this, suppose fifth one is like this, suppose sixth one is like this, seventh one, eighth one like this. Let all other suppose you have collected 20 samples, 20 samples of size n and you computed, you have found out that that this is the true mean. Hypothetically, we are assuming this is my true mean and 19 sample contain this mean because the interval contain this means, this is the constant value that mean.

Now, first sample the interval contain mean, second sample contain mean, but third one no. So, out of 29 contain mean that means it is 95 percent that is the message, that mean what we are saying when you collect the sample, it does not mean that that contain mean. But you were getting interval but that this in the total method is such that it says that there is 95 percent chance that it will contain the mean, it may not contain, that chance is

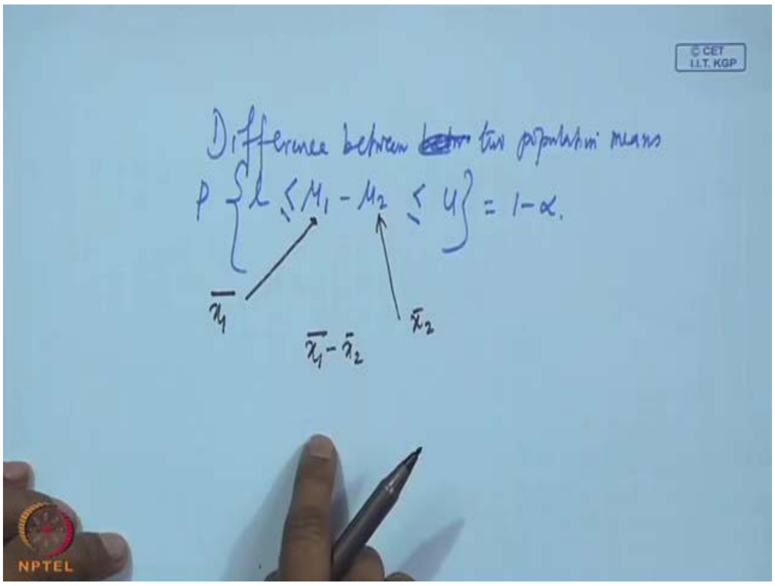
5 percent. That is why we are saying confidence interval. Actually we will not go for 20 samples, but then if you go, if you find like this that is the case then chi square, that we have seen this.

(Refer Slide Time: 42:21)



Then confidence interval for two population, difference between two population mean. What you will do here? See the entire procedure remain same, only you have to know what is the random variable, what is the statistic and what is the distribution sampling distribution. If you know then your work is over. For example, what we are now creating one variable here.

(Refer Slide Time: 42:50)



Suppose, you want the difference between two population means, this is my difference between two population, two population means. If your population one is having the mu 1 mean and population two mean is mu 2, you want to find a confidence interval for this. What will be the 1 and u value for which your probability that the interval contains mu 1 minus mu 2 is 1 minus alpha, getting me?

So, that is the issue. So, if this is the case then what is at your hand? You have only x 1 bar from the population one that mean 1, and for this you have x 2 bar and you have also the difference between x 1 bar and x 2 bar. We have seen earlier that x 1 in any statistics is a random variable. Now, difference between the statistic, also a random variable that mean you want to know. What is the mean value of x 1 bar minus x 2 bar and standard deviation of x.