INDIAN INSTITUTE
OF
TECHNOLOGY
KHARAGPUR

NPTEL
National Programme
on
Technology Enhanced Learning

Applied Multivariate Statistical Modeling

Prof. J. Maiti
Department of Industrial Engineering and Management
IIT Kharagpur

Lecture – 05

Topic

Estimation
(Contd.)

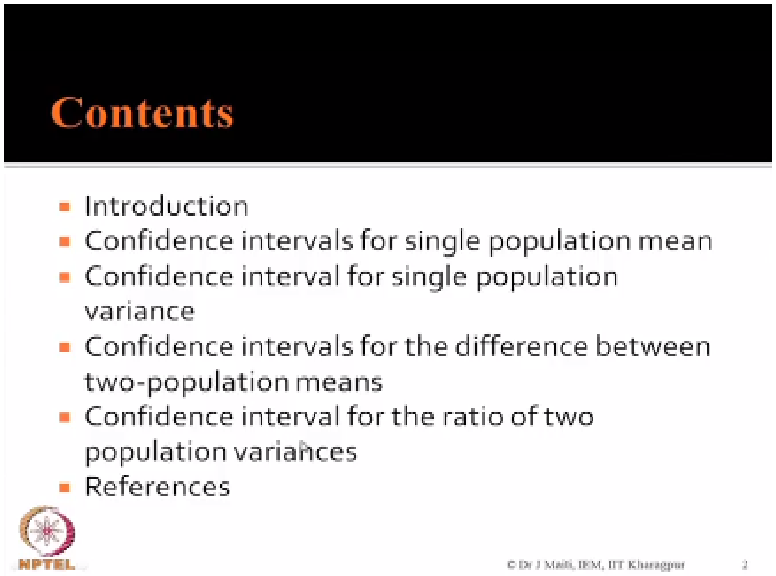
Hello good morning today we will discuss univariate statistical topic estimation.

(Refer Slide Time: 00:28)

(Refer Slide Time: 00:30)

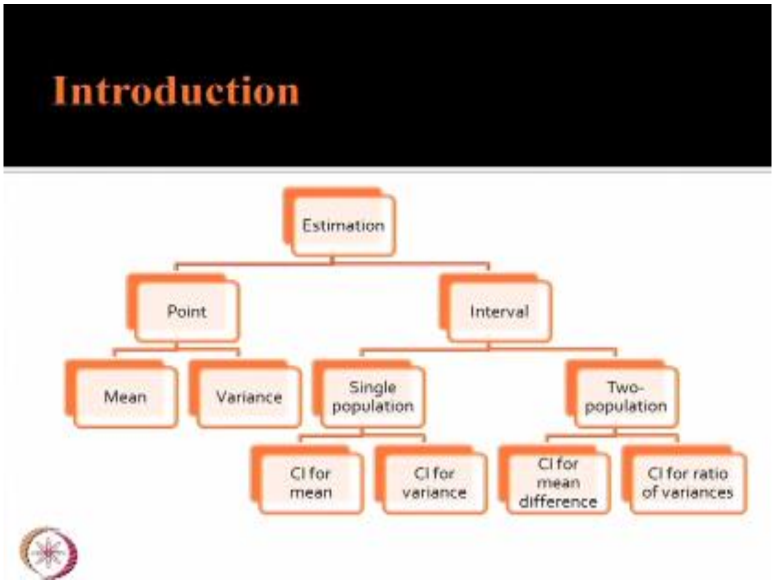
Estimation comes under univariate statistics today's content is.

(Refer Slide Time: 00:43)



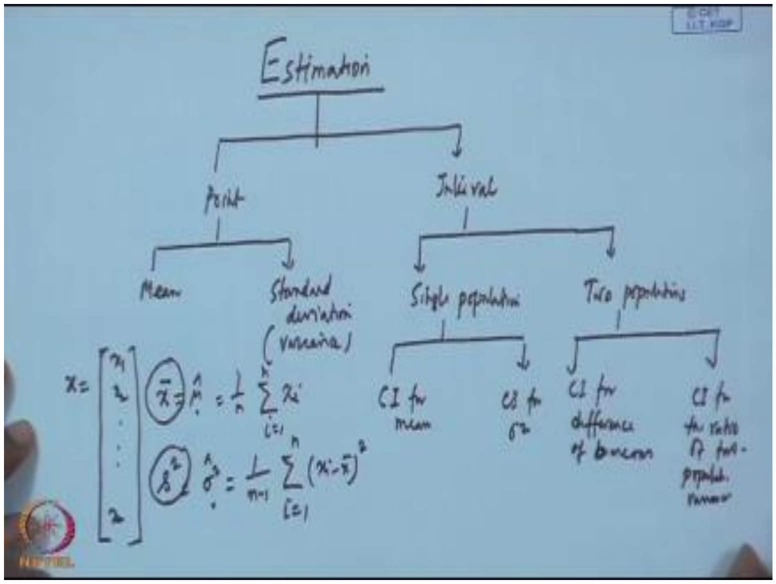
We will start with what is estimation then I will tell you the different types of estimation like for single population mean for single population variance, confidence intervals for the difference between two population means confidence interval for the ratio of two population variances followed by references.

(Refer Slide Time: 01:13)



Now if you see this slide estimation has two parts.

(Refer Slide Time: 01:21)



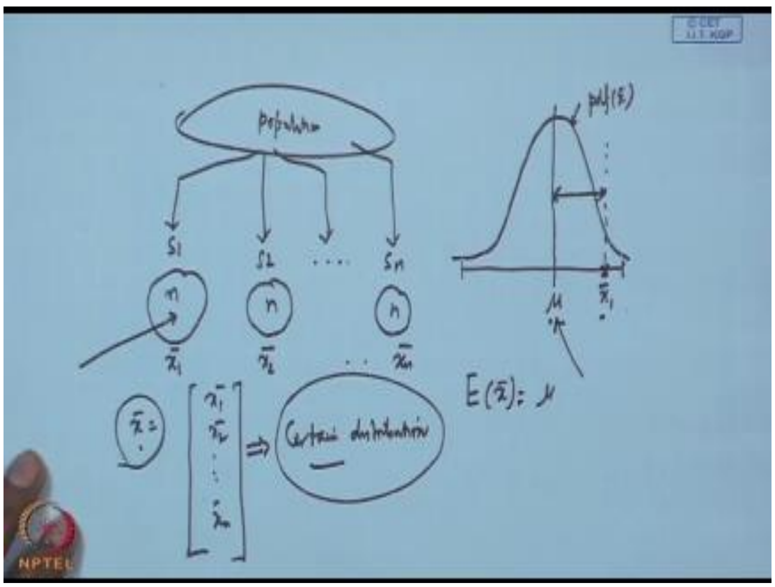
One is point estimation another one is interval estimation so under point estimation we will be discussing about that point estimation of mean and point estimation of standard deviation or we can say that point estimation of variance that is the square of standard variation that we will be discussing 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.

And 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 $\bar{x}$ is an estimate of μ then that you all know that this is i = 1 to n x_i so what do we say that $\overline{x}$ 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 -1 sum total of i equal to 1 to n xi - $\overline{x}^2$ when you compute like this that you collect a sample and compute $\overline{x}$ as well S square from the sample.

This is your point estimate so $\overline{x}$ is the point estimate of μ s square is the point estimate of σ2 now we have discussed in last class that what will happen when I go for several samples collected from a population.

(Refer Slide Time: 05:01)

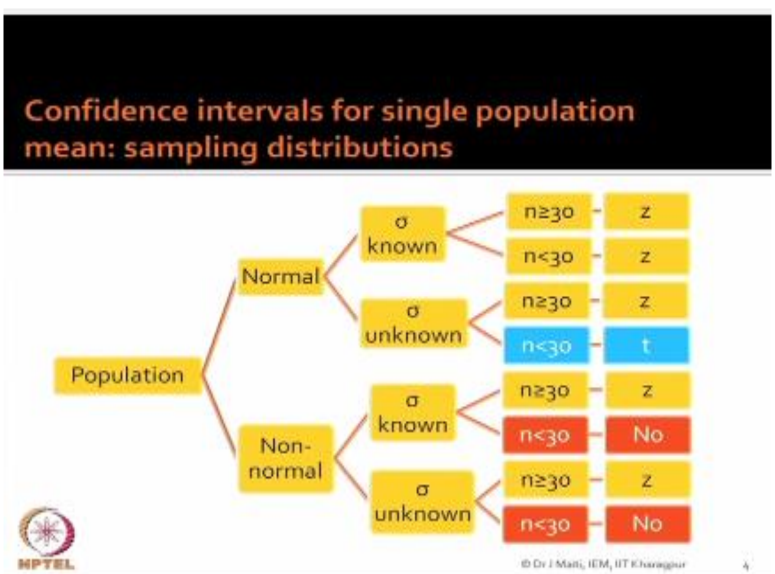


I told you in the last class this is my population and if I go for several sample collected so for example you have collected sample1 sample 2 then sample n all of the samples with size n one size n equal sample size n and here also n and if you calculate the point estimate that will be $\overline{x}_1$ $\overline{x}_2$ like this $\overline{x}_n$ and we have seen that this $\overline{x}$ if I write in this vector form that $\overline{x}_1$ $\overline{x}_2$ 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 $\overline{x}$ what is the guarantee that the computed $\overline{x}$ 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 $\bar{x}$ is like this, this is my pdf of x bar and all of us know now that expected value of x bar will be μ because this is the property of unbiased estimation so then your μ is coming here now let us talk about the sample 1 and you have computed x bar using S1 and it is falling here the value of $\bar{x}1$ is falling here so what is our interest here using confidence we want to know that whether this $\bar{x}$ 1 or the x bar collected using sample1 is representative of μ or not.

It all depends on the distance between this two if is $\bar{x}$ far away from μ then it can it will not be a representative okay so in order to know whether the $\bar{x}$ contains the interval of $\bar{x}$ contains μ 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 μ that is a population mean and we want to find out that whether that interval contains μ or not now come back to the slide here. What will happen ultimately?

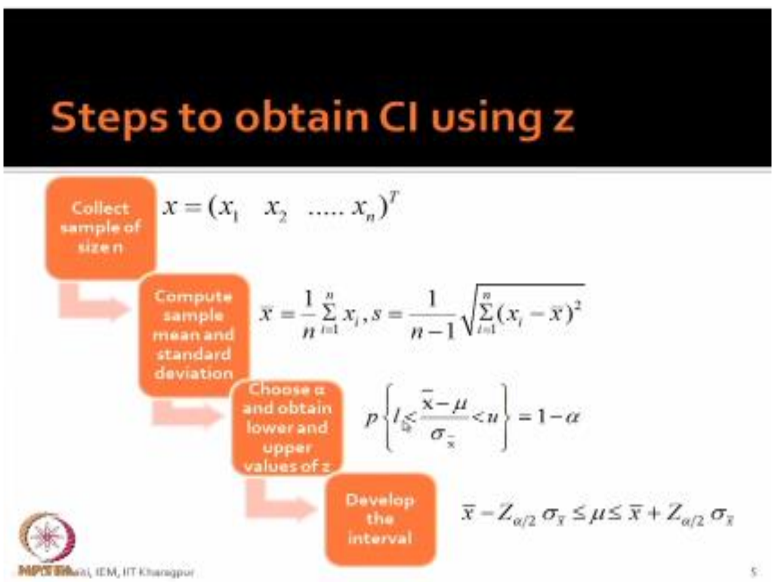
(Refer Slide Time: 08:19)



Population can be normal and non normal now of you sample from normal population with σ known and sample size whether small or large that is not a problem not a question issue at all then 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 σ is known irrespective of the sample size the statistics will generate for x bar, which is which is basically z equal to $\bar{x}$ - expected value of $\bar{x}$ by σ $\bar{x}$ these quantity follows that this $\bar{x}$ - expected value of $\bar{x}$ by σ $\bar{x}$ follows z distribution if you sample from normal population irrespective of the sample size but σ must be known, if σ is unknown your sample size is large.

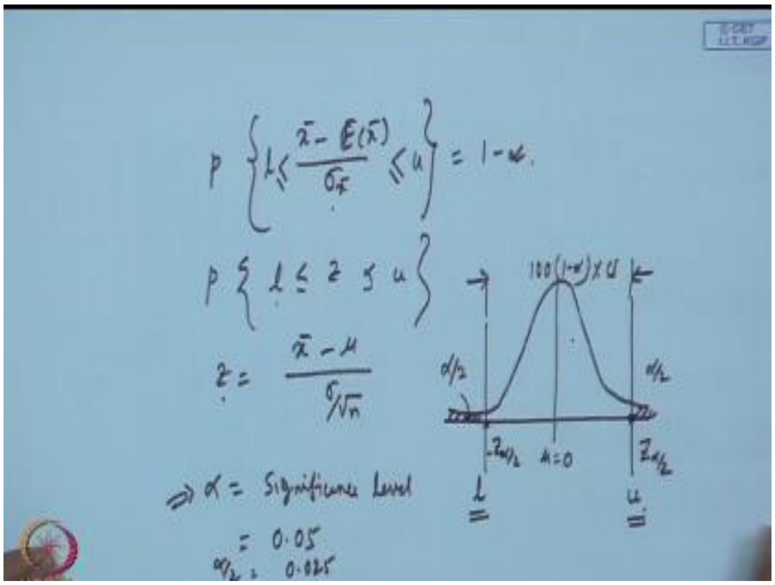
Then this quantity follows again z distribution but if σ that sample size is small and σ is unknown then this quantity follows t distribution now if you sample from non normal population when sigma is known and your sample size is large then this quantity again the same quantity. This quantity follows z distribution even if σ 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 tor z distribution the mathematics and the procedures remain same only you have to use z table or t table let us see here.

(Refer Slide Time: 10:35)



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 $\bar{x}$ 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 that $\bar{x}$.

(Refer Slide Time: 11:10)



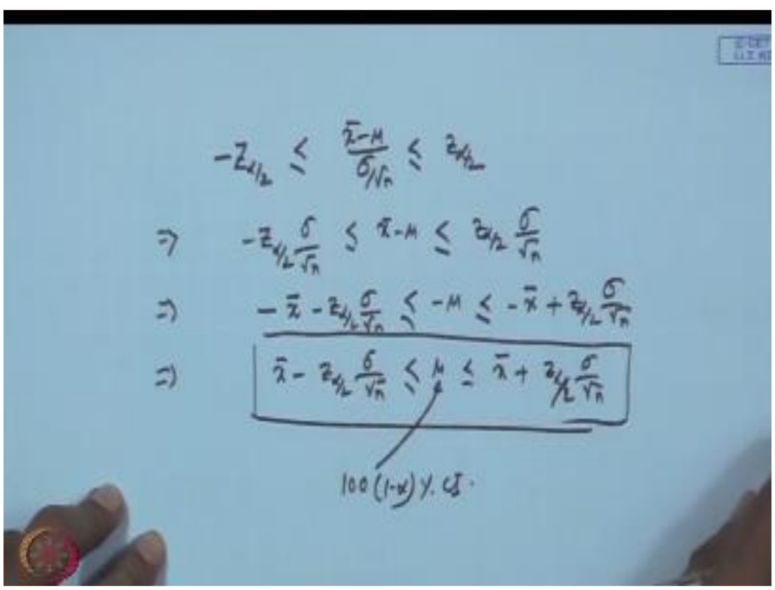
Expected value of $\bar{x}$ divided by σ x bar this one probability that this value will be greater than l and less than or greater than equal to l and less than equal to u that will be 1 - α okay so if your distribution this one is nothing but z that l less than equal to z less than equal to u when some the condition satisfied like from normal population irrespective samples σ is known then what we will write basically we will write z = $\bar{x}$ - μ by s σ by √n the σ $\bar{x}$ is σ 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 x 1 – α percent CI what is this α how do you determine this α, α is known as significance level.

Okay this α when our this our this $\bar{x}$ - μ σ by $\sqrt{}$ 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 α by 2 here is α by 2 so significant level of significance or significance level α 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 α what is error the probability that that true mean lies in this portion that is basically α 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 α if we consider $\alpha = 0.05$.

Then α by 2 is 0.025 then what do you require to know now that you have to find this z α by 2this is your z α by 2 and this left hand side this value will be minus z α by 2 so you see the table and find out z α by 2 value and accordingly you compute so then mathematically what is happening here.

(Refer Slide Time: 14:39)

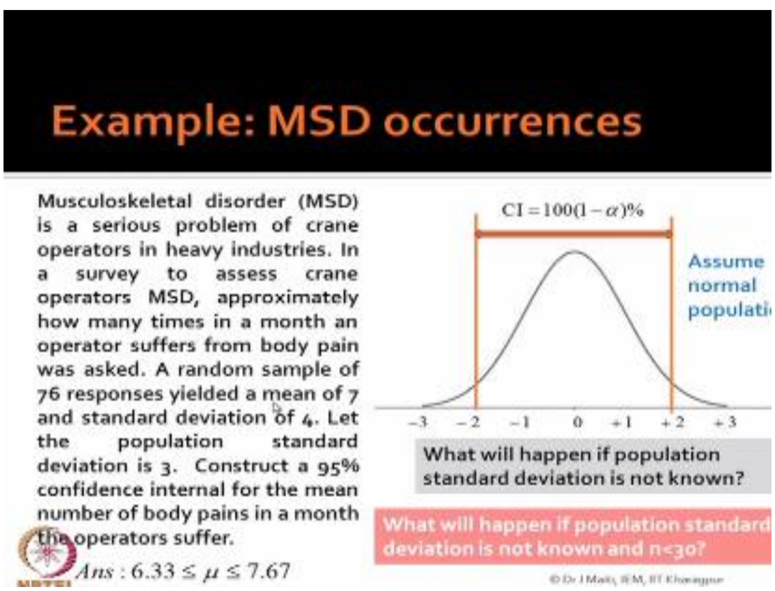


Mathematically is my l is - zα by 2 and this will be less than equal to $\bar{x}$ - μ by σ by $\sqrt{}$ n that equal to + zα by 2 now if you rearrange this one what you were getting z α by 2 σ by $\sqrt{}$ n less than equal to $\bar{x}$ - μ less than equal to z α by 2σ by $\sqrt{}$ n then again we will manipulate this so I will just bring that because we are looking for confidence interval of μ so what do you do we basically separate we will take $\bar{x}$ from this, the middle portion then if I write again I will write like this - $\bar{x}$ - z α by 2 σ by $\sqrt{}$ n this less than equal to - μ less than equal to - $\bar{x}$ + zα by 2 σ by $\sqrt{}$ n correct.

Very simple manipulations you are just now taking out $\bar{x}$ from the middle portion putting to the left hand right hand side then if you little modify now we do not want - μ we want+ μ so you are multiplying it by - 1 now it will just the reverse will take place what will happen $\bar{x}$ - zα by 2 σ by$\sqrt{}$ n less than equal to μ less than equal to $\bar{x}$ + zα by 2 σ by $\sqrt{n}$ so this formula is applicable when you this is the confidence interval for μ 100 x 1- α percent CI for μ 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 μ 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 i think let us go for a problem first.

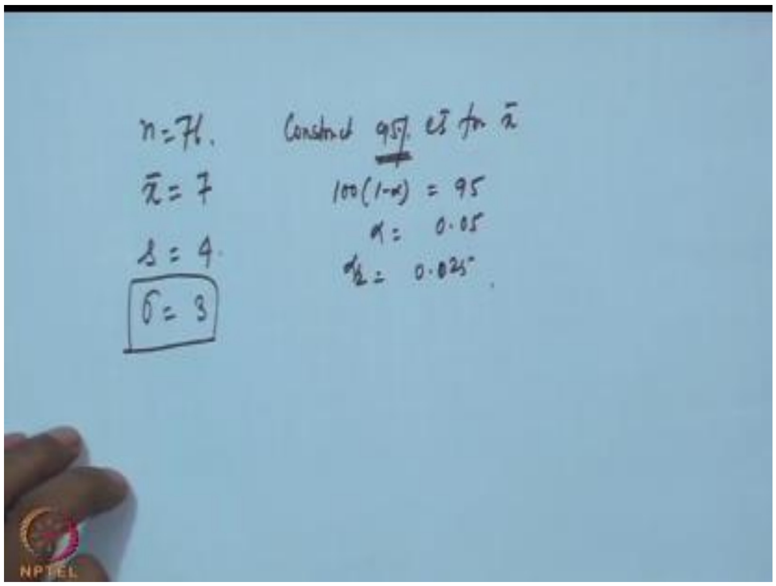
(Refer Slide Time: 17:19)



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 = 76.

(Refer Slide Time: 18:00)



You have computed $\bar{x}$ which is 7 and standard deviation that S = 4 let the population standard deviation is given 3 that is $\sigma = 3$ constructs 90% confidence interval for $\bar{x}$ your work is construct 95 % CI for $\bar{x}$ this is your work now see when I say 95% that means we are saying that 100 into 1 - α = 95 so you are getting α = 0.05 so what is my α by 2 0.025 now here it is clearly given that σ is3 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.

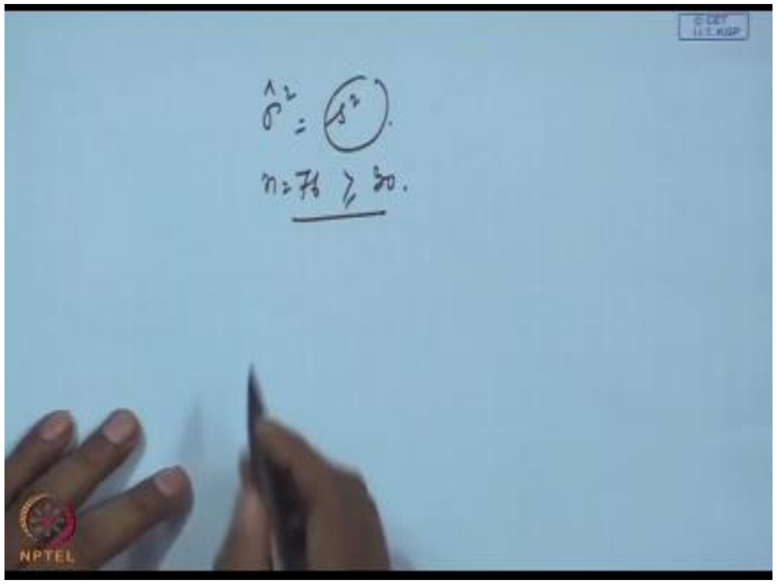
So you see this is normal assume normal distribution then what we will use we will use z distribution and accordingly our interval will be $\bar{x}$ - z α by 2 σ by $\sqrt{n}$ less than equal to μ less than equal to $\bar{x}$ + z α by 2 σ by $\sqrt{n}$ every values are known to you so your computation is be $\bar{x}$ will be 7you have to know z 0.025 into σ population σ is 3 and your n is 76 so less than equal to μ less than equal to 7 + z 0.025into 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 - 1.963 by root over 76 less than equal to μ less than equal to 7+ 1.963 by root 76 and the answer will be 6.33 less than equal to μ less than equal to 7.67 this is the confidence interval for μ.

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 μ and interval estimate of μ what is how to know that whether these estimate this one this one contain the μ true μ now see when we convert the statistic that $\overline{x}$ 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 - 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 μ term what will happen ultimately see you are getting positive left hand that 6.33 is also positive and 7.76 is also positive okay 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 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 that σ2 will a square population standard deviation is known.

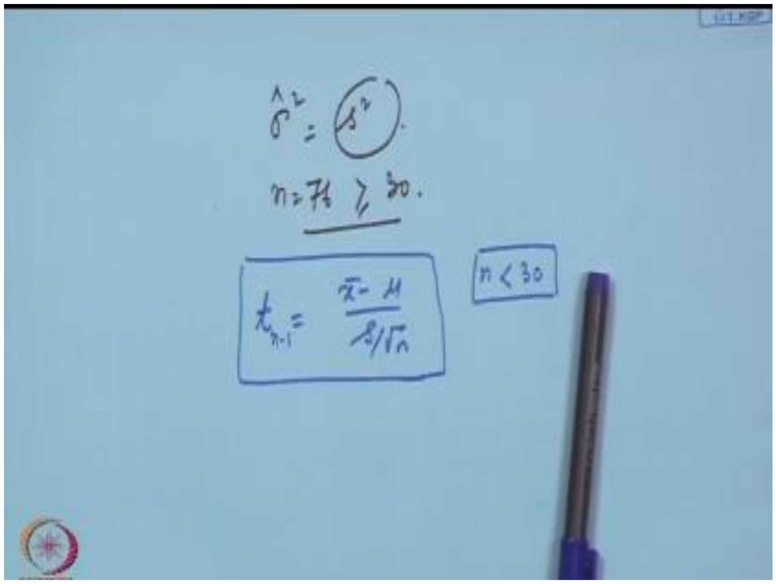
(Refer Slide Time: 23:33)



Now known mean σ2 estimate that is what we are sayings square so you will be using S^2 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 what will be your change you will be using the same z distribution but instead of σ you are using S you see in the given problem σ and s are not same so s = 4σ = 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.

Okay 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 what is required now your case is going like this you will be using t distribution.

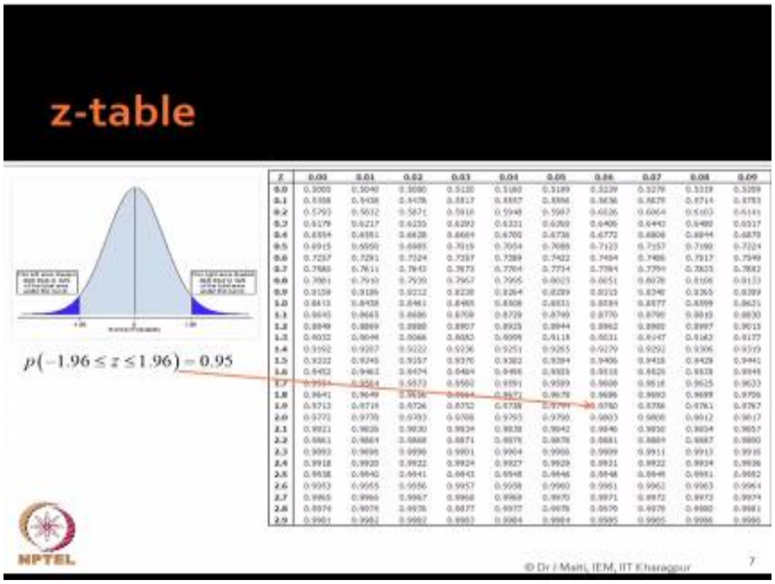
(Refer Slide Time: 25:00)



Which is $\overline{x}$ - μ by s by √ 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 σ 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 –1 degrees of freedom so let us see the t distribution.

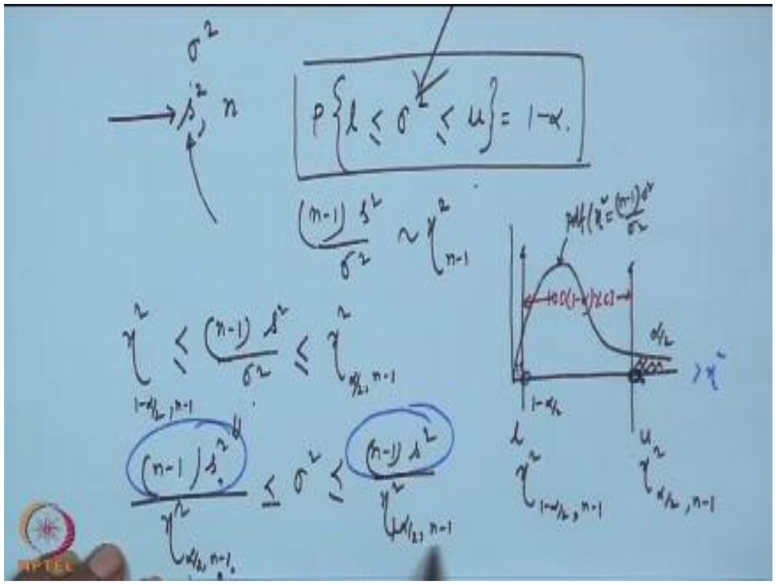
(Refer Slide Time: 26:08)



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 your interval will be like this $\bar{x}$ - t n - 1 α 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 $\sqrt{n}$ less than equal to μ less than equal to $\bar{x}$ + t n - 1 α by 2 s by $\sqrt{}$ 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 - 1 α 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 population variance means σ^2.

(Refer Slide Time: 28:02)



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- α 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^2 it is S^2 it is n - 1 S^2 by σ^2 follows with distribution chi square distribution what will be the degrees of freedom n - 1 so n - 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 - $1S^2/ \sigma^2$ and you want to find out a upper value.

That is u and a lower value that is l for n - 1 square by σ^2 so again if I consider α that this one is basically this particular this side it is let it be that α by 2 and this one what will happen total is 1 1 – α by 2 you will be getting chi square 1 - α by 2 but please keep in mind chi square also having a degree of freedom that is α by 2 and degree of freedom is what n - 1 n - 1 getting me so then what you will write then you will write n - 1 S^2 by σ^2 it must be less than equal to chi square α by 2 n - 1 as well as it must be greater than equal to chi square 1 - α by2 n – 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 x 1 – α percent CI for the variance this is my but what do you want here in this equation what do you want.

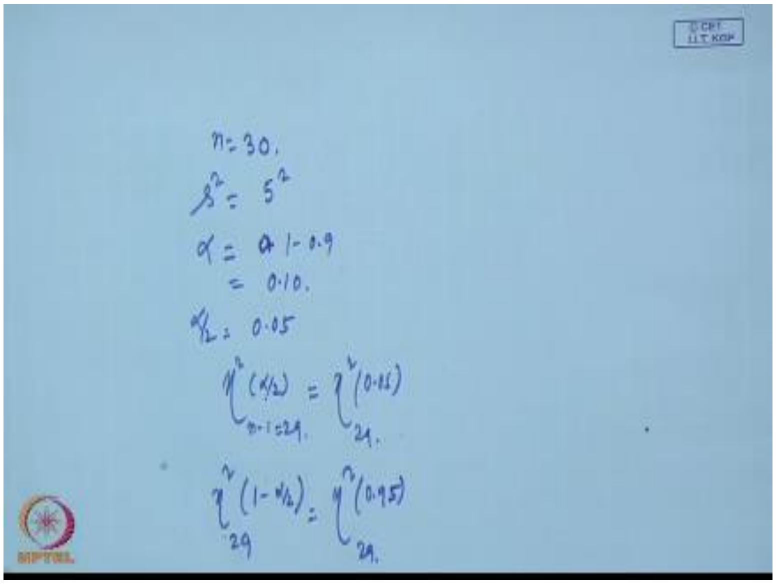
You want something like this, something like l less than equal to sigma square less than equal to u can you now find out that what will be can you not manipulate this you can easily manipulate so what will be the once you manipulate this that mean what will happen you will want to keep in between the less than equal to terms only the σ^2 so if you manipulate you will be getting like this n – 1S^2 by chi square α by 2 n - 1 you will get this s α by 2 n - 1 less than equal to σ^2 less than equal to n -1 square by chi square α by 2 n – 1 you see the here you see the denominator.

Here is n -1 x S^2 here also n - 1 x S^2 same quantity the difference is in the new sorry in the numerator both are like n -1 x S^2 but in the denominator that is s square chi square α by 2 n - 1 what value is 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 – S^2 by chi square α by 2 n - 1 is definitely less than n – 1s^2 by chi square α by 2 n - 1 and that is the interval okay now given data what you will do suppose this is the data.

(Refer Slide Time: 33:34)

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% confidence interval for the population σ what you will do you will all what is n value here n = 30 okay.

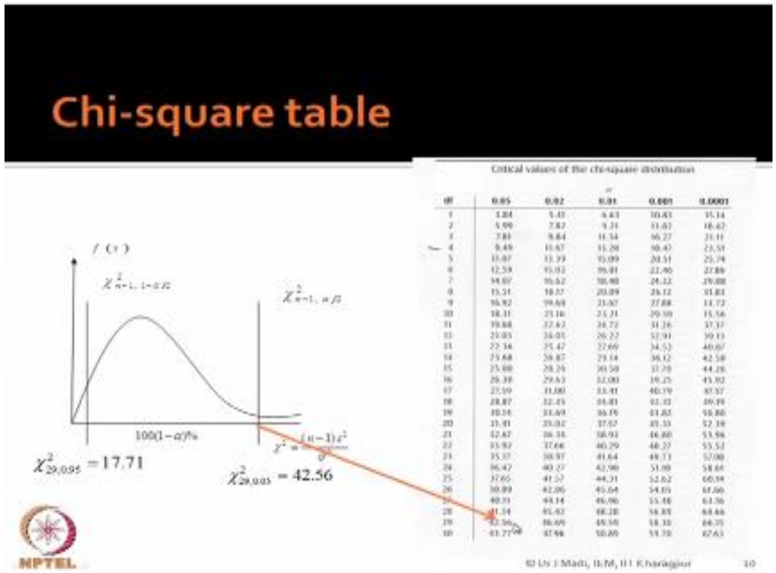
(Refer Slide Time: 34:24)



What is S^2 population that means you have collected a sample of 30 with standard deviation of 5 so that mean this is 5^2 correct then what more you want nothing only one thing you want to know that is what is α so we are saying 90% confidence interval so α = 0 1 that is 1 - 0.9 that means 0.10 so your alpha by 2 is 0.05 you want to calculate chi square 2 value chi square α by 2 n - 1 where n is30 that is 29 and α is your that α 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 – α by2 with again same n - 1 so that mean chi square 0.95 29 if you know these two value then put into this equation S^2 is known 25 n -1 is 29 chi square that is 0.05 29 chi square0.95 29 you have to find out the chi value see that table.

(Refer Slide Time: 36:03)



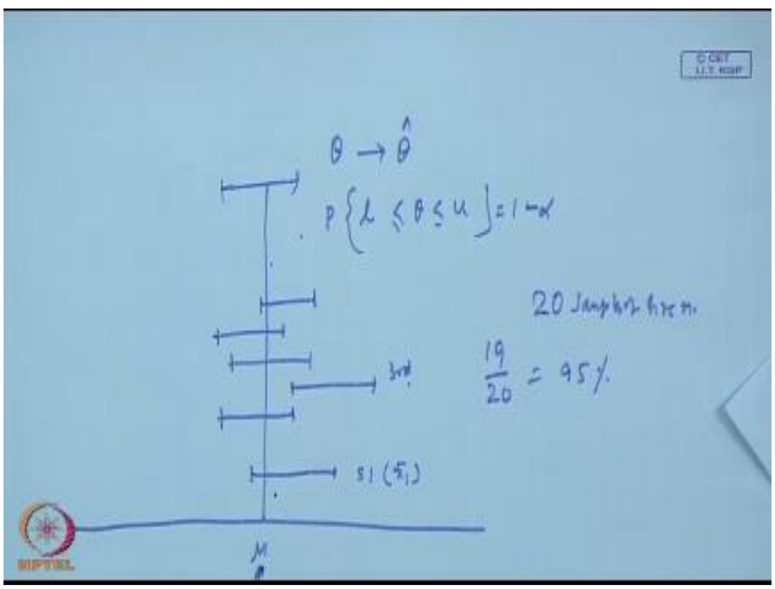
You see this table that in this table our α value a α by 2 is 0.05 so we want to first know that chi square our degree of freedom is 29 and what is this value that value is 42.50 so chi square our 29 0.05 which one is this chi square29 0.05 that is 42.56 similarly chi square that 99% that is 17.71 so once you know these two values m 17.71 and these values you know your computation becomes very simple n - 1 that is 29 x S^2 is 25 divided by chi square that one that mean 42.56 less than equal to σ square less than equal to 29 x 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 σ^2 less than equal to 40.94 and if you go by σ then square root of this 4.13 less than equal to σ 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 okay 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 $\overline{x}$ then you are converting into z or t if it is your σ^2 then you are converting into appropriate statistic n -1 S^2 byσ^2 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% 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% 90% of the cases this sigma value will lie basically 90% 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 θ and.

(Refer Slide Time: 39:50)



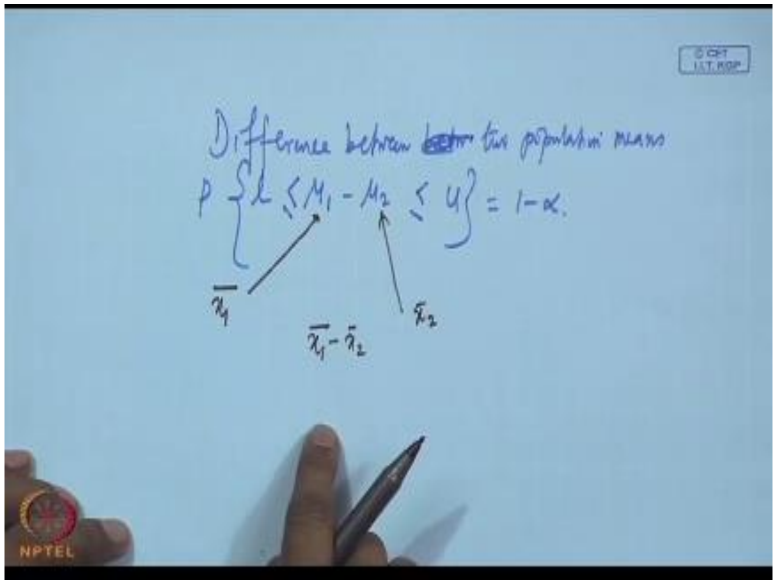
You have estimated θ cap using some sampling distribution now let the theta cap has a particular distribution then what you are doing you are basically doing like this less than equal to θ less than equal to u that you are doing which is 1 – α using the theta cap value that range you are getting theta cap value now let us concentrate consider the θ is μ that is the mean value there θ is μ suppose you have collected a sample with in size and you have computed the $\bar{x}$ and you found out that $\bar{x}$ interval is like this is your sample 1 and your $\bar{x}1$ its interval you found of because once you collect a sample you know $\bar{x}$.

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% 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% chance that it will contain the mean it may not contain that chance is 5% 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 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.

(Refer Slide Time: 42:49)



Here 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 μ_1 mean and population two mean is μ_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 μ_1 - μ_2 is 1 – α getting me so that is the issue so if this is the case then what is at your hand you have only $\overline{x}_1$ from the population one that mean1 and for this you have $\overline{x}_2$ and you have also the difference between $\overline{x}_1$ $\overline{x}$ 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 $\overline{x}_1$- $\overline{x}_2$ and standard deviation of x_1.