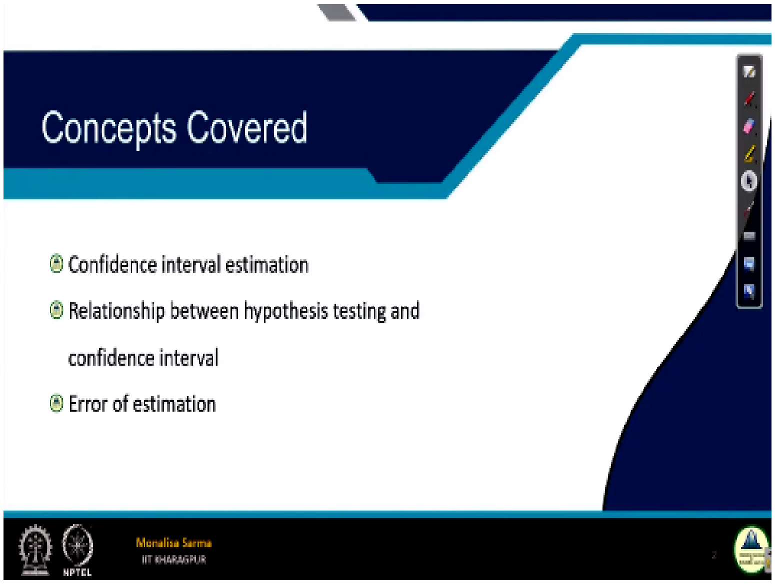
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology – Kharagpur

Lecture – 26
Statistical Inference (Part - 4)

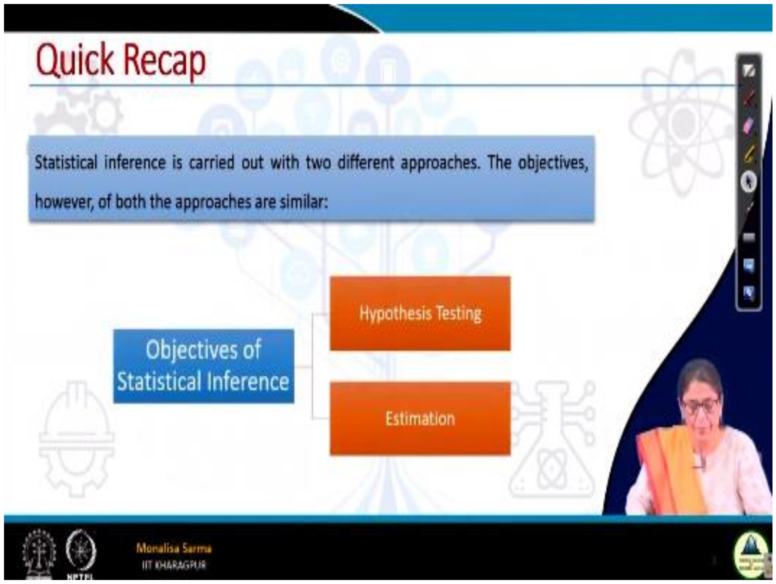
Hello guys. So, we have done one tutorial in my last lecture. So, now, we will start with our next continuing with our discussion on statistical inference, as I told you, there are lots of things to learn in statistical inferences.

(Refer Slide Time: 00:40)



So, in today's lecture, what we will see? What is confidence interval estimation then what is the relationship between hypothesis testing and confidence interval and also we will see what is an error of estimation.

(Refer Slide Time: 00:52)



Now, confidence interval estimation I think you remember when we initially started to talk on the statistical inferences, then I have already mentioned statistical inference we can carry out between 2 different approaches one approach is hypothetical testing and another approach is confidence interval estimation. We have discussed hypothesis testing for last 3 lectures and in today's lecture we will discuss the other one that is the confidence interval estimation.

(Refer Slide Time: 01:24)

So, before going to confidence interval because this is we are studying with a different approach let us just have a quick recap of what we have learned till now. So, we have of course, learn what is a statistical hypothesis if the hypothesis is done with the parameters of this population and we call it a statistical hypothesis I think you will remember what is null hypothesis, what is alternate hypothesis, null hypothesis is always we try to keep that hypothesis as a null which maintains the status quo we always try to give that hypothesis is null hypothesis.

Which has type one error may be quite significant accordingly to; based on the cost of the type one error we will consider the significance level. And moreover the alternate hypothesis is something which we want to test it that we may keep it as an alternate hypothesis, then what is rejection region or critical region that we have seen based on the significance level whatever significance level is mentioned based on the significance level we find out the rejection region.

Like from the table we can find out the z value corresponding to the probability that is the α, α is nothing but probability of type 2 error, probability is nothing but the area under the curve is not it? So, when the z curve so, then the corresponding z value corresponding to α is nothing but the rejection region like if I take if I remember the action which we have taken the medicine company that it should fill in the 8ml tube if it is 8ml tube so, initially what I was discussing how can we pick the rejection reason.

Suppose, find $\mu = 8$ that is okay but then there can be some changes we can accept. So, that suggests we have decided it to be it if it is from 7.92 8.1 we can accept that. So, what is the 7.9 and 8.1 this is the starting of the rejection region or the starting of the critical region and the probability so, the area corresponding to 7.9, 8.1 it is nothing but α. So, what in a hypothesis testing rejection region is not given α is given from α we find out the rejection region.

Then what is type 1 error, type 2 error if remember type 1 error is true null hypothesis we are what to say rejecting a true null hypothesis that is type 1 error, type 2 error is we are accepting a false null hypothesis then choosing between α and β that also we have seen how to choose α, how to choose β that we have seen some example based on that example we have seen how to do that.

(Refer Slide Time: 04:01)

Then on different ideas what we have covered is the 5 steps of hypothesis testing has 5 steps. Remember then what is the significance level of a HT hypothesis testing that is α, statistically significant results do you remember what is statistically significant results when we call a result a statistically significant essentially a results that is statistically significant when we reject the null hypothesis why it is say? Means the value what we get from the sample and whatever value we have hypothesis this both the values are statistically different.

This result is very different whatever we have hypothesis based on the sample we got the result, this result is very much different from whatever we hypothesis so, we call this result as a significant result. So, that is why it is called statistically significant results. So, easily statistically significant result when we get statistically significant results; that means we reject a hypothesis or we get some value like it. If you remember we have taken an example of constructing a test.

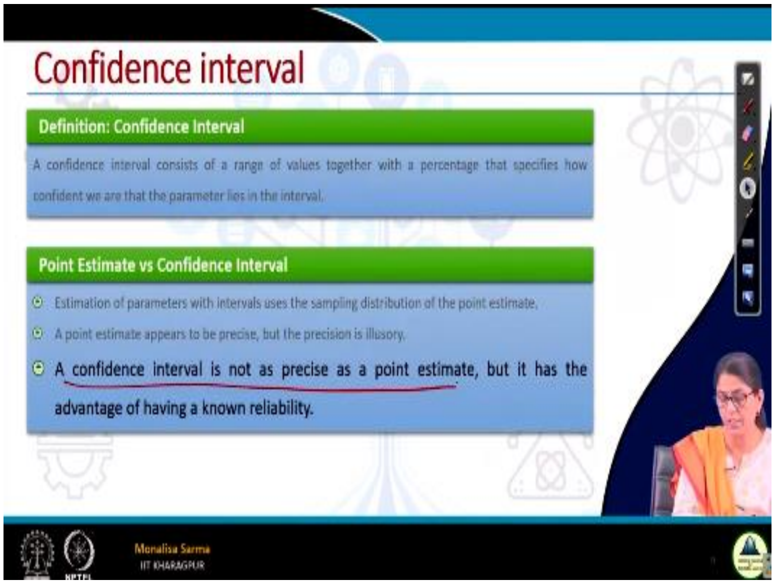
So, in the last tutorial we have seen in constructing a test, so, we got the value very close very, very less that means, but then we still we could not reject the null hypothesis that value also we can consider it as a statistically significant result. When the value is very, very different from whatever we have hypothesis value of below what is indicates? This indicates this probability of this occurrence if this is true what is the property of disappearance, is not it. So, now, why do we focus on type 1 error I have already mentioned.

Then what is p value? p value reporting is necessary is necessary why because sometimes it is very difficult to just give a reject a hypothesis or do not reject a hypothesis given just a yes no result it becomes too much because what I have explained it in some examples for similar type of results with slight change we see sometimes it gets accepted sometimes it gets rejected for very slight changes even for difference slight change of significance level also may get a different result.

And moreover the person who is carrying out the hypothesis testing that is the statistician and the person who is the actual decision maker maybe the 2 different ways in person. So, we should leave the decision to the decision maker whether we will give the result now, the decision maker will take a decision whether to reject or not to reject. So, this is call p value reporting basically based on what is the calculated test statistic whatever value we get the probability of test statistic is the p value.

Operating characteristic curve it is nothing but β versus μ. So, in power of test is? Power of a test is 1 - β versus μ power of test means rejecting a false null hypothesis that is the power of a test.

(Refer Slide Time: 06:51)



So, now coming to the confidence interval so, before looking at a slide first let us take an example suppose a retail chain retail chain wants to open a big retail what to say store basically in one area where the so, before opening such a counter he needs to know the earning capacity of the people in that locality. So, if because why he why it is necessary, because if he can find out what is the earning capacity of the people accordingly based on that he will keep the things in his counter in his outlet.

So, if the I mean the spending capacity not on it the spending capacity is more definite he can keep some high end products in the spending capacity is very, very less. So, high end product will just be will just left into outlet and it will not be sold. So, he will have to go for some less expensive products specialty some expensive product. Now it is a totally new area. So, how to find out the spending capacity of people it is a huge area with diverse populations and totally know what to say pass knowledge from which you can have a hypothesis value.

In such a case there are many such examples again the another one example which I can say is that so some what to say in a chemical industry the amount of what to say byproduct it gives until and unless you do the experiment, you really cannot say what will be the amount of mean of those byproducts. So, in this when you cannot basically when you cannot hypothesis about the parameter of a population, then the only option left in front of you is there then you will have to take a sample.

From the sample you will have to find out the value of the statistics and based on the statistics we will infer about the population like for here is the person who wants to open a retail outlet here. So, he will since he has no idea about the spending capacity of the people he will select he will take a non bias sample he will take a set of people and from the set of people he will try to find out what is the spending capacity of the people from with that spending capacity of the people from the sample then he can infer about the whole population.

So, basically here is hypothesis value is not there. So, he has to estimate based on the sample. Now, this estimation there are 2 different ways the first one is whatever he got from the sample suppose he has taken his interview around I should not say interview he has he tried to find out from around 20 person what is the spending capacity by spending capacity means how much a person arms in annually. So, from there suppose we found out on an average person in that industry 20 around among these 20 people on an average a person are not around 8 lakhs per annum.

So, 8 lakhs per annum is the mean salary of this 20 people what he has interviewed. Now, based on this, if he directly tells what you said the mean of earning mean of the spending capacity of this people or the earning of these people in this area is 8 lakhs per annum. So, that is basically that is called point estimation we have just we found out the mean and we are just telling this mean only we are telling that it is the population mean if we take this mean of the sample as the population mean.

If we infer the mean of the population mean of the sample as the population mean this estimate is called point estimate of course, and secondly, this estimate the probability that it is correct is very very less because, as I told you before this samples this statistics vary if I took a particular sample of 20 people I found as many as 8 lakhs per annum, I took a different sample maybe I found it 7 lakhs per annum I took in different sample maybe I got some other value. So, if I take a point estimation reliability of this estimate is very less.

So, what we can do is that instead of just giving a point estimate we can just hedge a bit what we do we will give an interval that means the earning capacity earning of this people in this locality is around from 5 lakhs to 10 lakhs we have given an interval directly without specifying the earning of this people is 8 lakhs per annum, we are given telling us a 5 lakhs to 10 lakhs. So, we have given an interval so, we are estimating within an interval in this is called interval estimation.

Now comes the question of what is confidence? Confidence is the how much reliability of this estimation that is called as confidence interval estimation. Now, how we can give the statement about the reliability of that this is again very much dependent on the significance level what a significance level remember significance level is the probability of type 1 error here also if the significance level is α that means my confidence coefficient is 95%. So, accordingly I will find out the confidence interval. So, let us see in the slide.

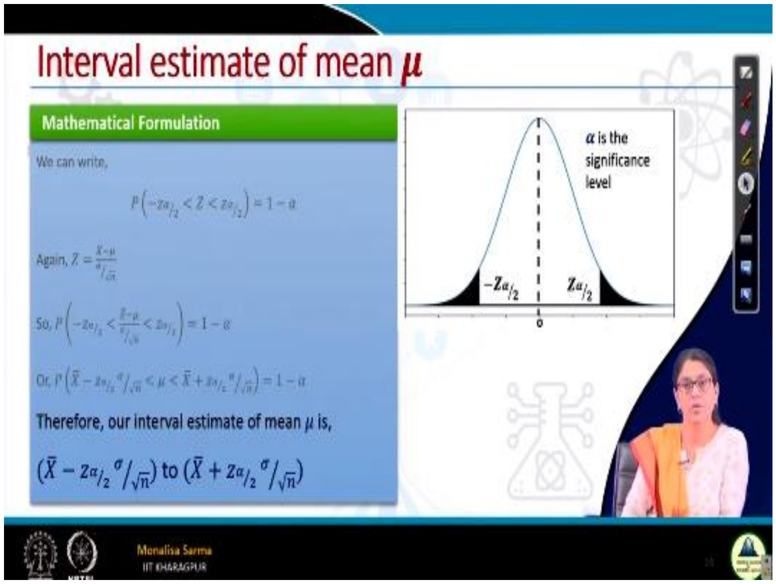
So, first say what we have a confidence interval consists of a range of values together with a percentage that specifies how confident we are that the parameter lies in the interval. Interval consists of a range of values together with a percentage that specifies how confident we are. So, it is 5 lakhs to 10 lakhs is a range now along with that we have to specify a percentage that we will have to give a percentage that specifies how confident we are that the parameter lies in that interval, that is the confidence interval estimation.

So, estimation of parameters with interval it also again like hypothesis testing hypothesis testing my main backbone is the sampling distribution my CPU is the sampling distribution is not it? here also for confidence interval estimation also my CPU was the sampling distribution estimation parameters with interval uses the sampling distribution of the point estimates we use sampling distribution based on the point estimation whatever mean I got from the sample I have collected one sample from the sample whatever I got mean.

Mean or variance or anything whatever we want to infer. So, this I will find the sampling distribution of that particular point estimate like how we do in hypothesis testing similarly. A point estimate appears to be precise, but the precision is illusory. The point estimate if I tell the mean income of the population of this area is 8 lakhs that is a very precise statement, rather than if I say may mean income of this area is from 5 lakhs to 10 lakhs this is not very precise statement, I am giving a huge range.

But if I am specifying just one value that is a very precise statement, but this precision is illusory as I told you why. A confidence interval is not as precise as point estimation, it is not as precise, but it has the advantage of having a known reliability that is the confidence level.

(Refer Slide Time: 14:12)

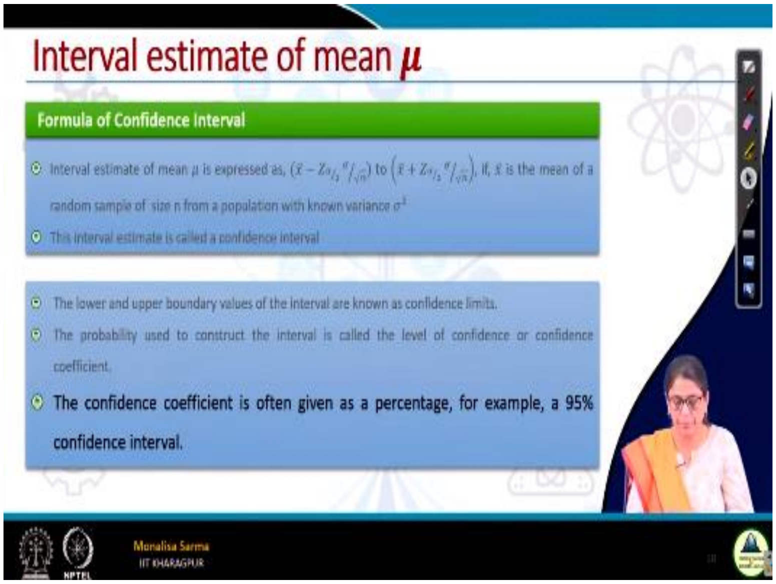


So, we will see how it is. So, what we do is that? We have seen till now, this concept of significant level that critical region or the rejection region when I specify α as my significance level, so for two tailed, when I am not specifying then I will always consider 2k. So, it is α is a significance level. So, this value z α / 2 corresponding to value z α / 2, this is my rejection region, is not it? This is my rejection region. So, if my value falls within this region, then I accept it if my value falls within this region.

Now, if this area put together as α, what is this area? This area is 1 - α, this area and this area total is α then this area is 1 - α is not it? So, probability meant that my z z will lie within this range is what is 1 - α. So, now what is z? z is a X bar - μ σ / √ n. So, I will substitute z with that now, I will do a bit of simplifying what I get I will be getting this sort of an expression μ this value is less than μ, μ less than is this value.

And the quality of this is 1 - α where α is the significance level. So, this is my interval the my μ value will lie between these to this my μ will lie between this to this, this is my interval this is my interval assist estimation and what is the confidence of this is 1 - α probability that my μ value will lie between this is 1 - α. So, there are interval estimate a mean μ is this where x bar is the mean of the sample.

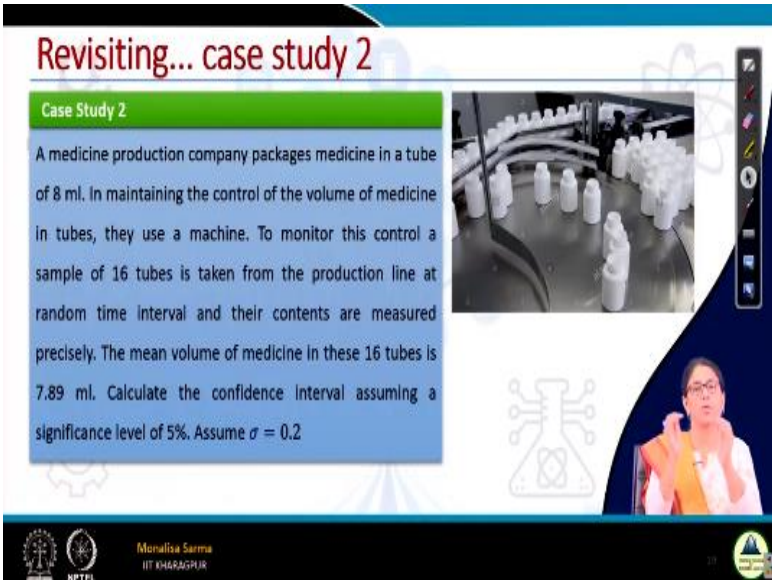
(Refer Slide Time: 16:16)



This interval estimate is called a confidence interval the lower and upper boundary values of the interval are known as confidence limit the lower confidence limit upper confidence limit and here this is my lower confidence limit, and this is my upper confidence limit and what is the confidence the percentage the probability is called a confidence coefficient. So, what is my confidence coefficient here 1 - α so, if α is 5 a confidence coefficient is 95%.

The probability used to construct the level is called the level of confidence or the confidence coefficient the probability used to construct the interval I call it as level of confidence or the confidence coefficient. So, if α is my significance level, what is my confidence coefficient or level of confidence it is 95%.

(Refer Slide Time: 17:20)



Now we will see with the help of an example, we have seen this example again that is why I have given a 3 digit the case study 2 we are considered when we have just tried to learn what is hypothesis testing? So, same example and bringing it out here. Here we have done using this we have used this example to find out hypothesis testing. Now, we will use this example to find the confidence intervals. Now, there are 2 things first is that when we use confidence interval first there are 2 cases.

Once we use confidence interval when we cannot hypothesis a value then we try to estimate the value of the parameter within an interval with a known confidence coefficient. When we cannot hypothesis if we can hypothesis the value then we can do the hypothesis testing and then we can find out whether it is accepted or rejected. If we cannot do that, we can find out the confidence interval and we can tell that our parameter value lies within this range. This is one reason why we do confidence interval estimation.

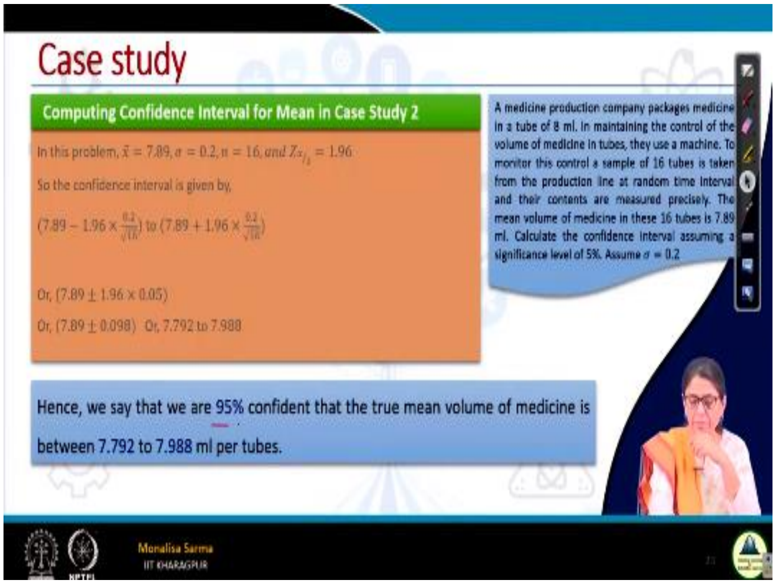
There is another one reason another one reason is that like when we do an experiment we have done an hypothesis testing like any other example what we have taken suppose we have taken $\mu = 8$ is null hypothesis μ not equal to. So, it is the alternate hypothesis. Now, suppose on the hypothesis testing whatever we have done and we found that null hypothesis is rejected that means μ is not equal to. Now μ is not equal state find that but μ is how much μ is not equals to it is okay.

But μ is how much that how we will get we can calculate it in using this confidence interval of course, we will not go to specific values of μ but we can tell my μ falls within this interval. So, these are the 2 reasons why we do confidence interval estimation. So, now this question what it is given we know all this but still going through medicine production, company packaged medicine in a tube of 8ml. That means I am I required 8ml in maintaining the control of the volume of medicine in tubes they use a machine.

To monitor this control a sample of 16 tubes is taken from the production line at random time interval and the contents are measured precisely. That is required is 8ml is required and whether it is really filling 8ml what we have done we have taken a; what to say random sample random sample of 16 tubes. And we have measured it precisely but we found the mean volume of the medicine in this 16 tubes is 7.89 mm. Calculate the confidence interval assuming a significance level of 5% as the mean $\sigma = 0.2$.

Now, it is asking us to calculate the confidence interval only, we are not asked to do the; what to say hypothesis testing. Remember this problem when we have done hypothesis testing for same data, we had rejected the hypothesis null hypothesis, we have the null hypothesis was $\mu = 8$. Alternative hypothesis is μ not equals to 8 and we have rejected the null hypothesis we found that μ not equal to for the same question now, let us see what we get in a confidence interval.

(Refer Slide Time: 20:24)



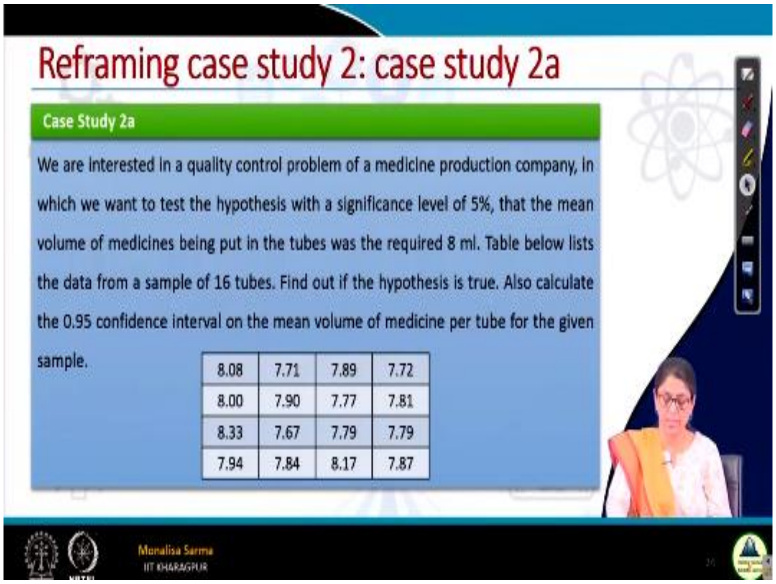
So, from this what it is given from the question it is x bar is given σ of the population is given n is 16. So, and significance level is 5% it is two tailed means α / 2. So, z α / 2 it is value is 1.96 you can find it from z table it is 1.96. So, confidence interval will be this is a confidence interval this is the formula X - z α / 2 σ y √ n X bar X bar - z α / 2 z α / 2 means this is the relative value of z corresponding to α / 2 it means if α is 0.05 value of z corresponding to 0.025.

That is that α / 2 this is also a value of z corresponding to 0.025 means this area is 0.025. So, what is this value of z? z that I am specifying as z α / 2. So, confidence interval I am putting all the values here and I got a confidence interval this 0.7 this is my confidence interval my μ lies in this range. Now, see here in hypothesis testing, I have rejected this null hypothesis if you can remember if you cannot remember you can go to the lecture again I think it is in the second lecture on statistical inference.

So, we have rejected this null hypothesis that μ is not equal to now, what is confidence interval value what it gives? confidence interval value also gave us that μ lies in this range 7.79 to 7.988 it means our hypothesis testing result is so, true means we have rejected that it is not equals to 8, but actually it is true you see in from the confidence interval also we found our μ lies in 7.79 to 7.98 8 is not included in this range.

Hence, we can say that we are 95% confidence because a significant level is 5% that the true mean volume or dimension is between this.

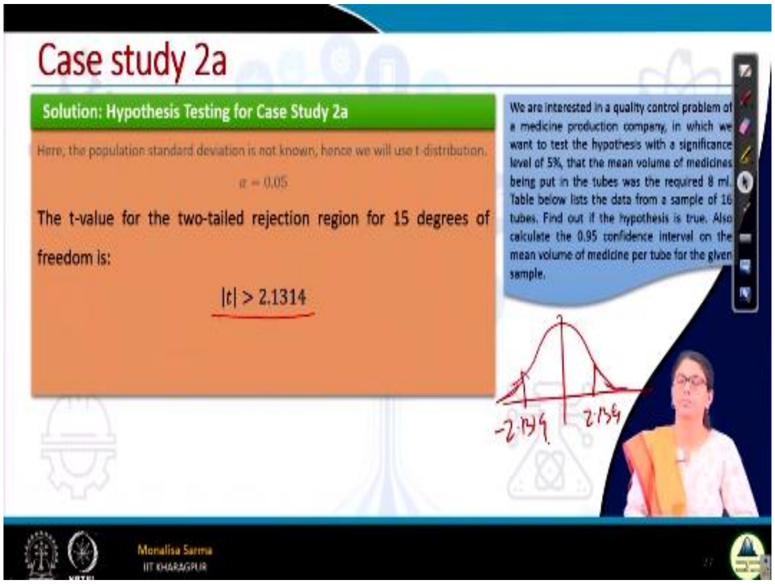
(Refer Slide Time: 22:23)



The same question whatever we have seen, but we have slightly changed up here whatever what we are doing is here our standard deviation in the first question our standard deviation of the mean was given here to see where I have mentioned here assume $\sigma = 0.2$, the standard deviation or mean is not given, then what we can do we cannot use the z distribution then we will have to use t distribution we will have to estimate the s^2 from the sample and then we will have to use t distribution.

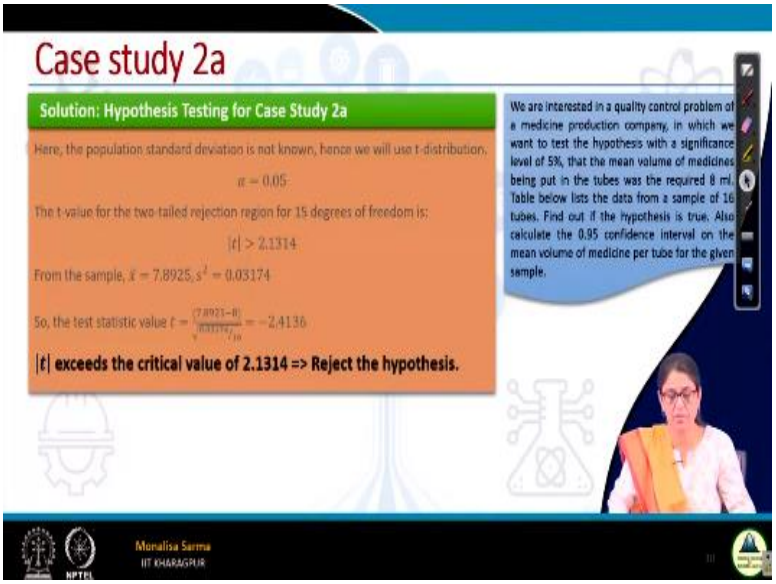
So, same question just a standard deviation is not given that means we will from the sample will estimate s^2 and use t distribution.

(Refer Slide Time: 23:05)



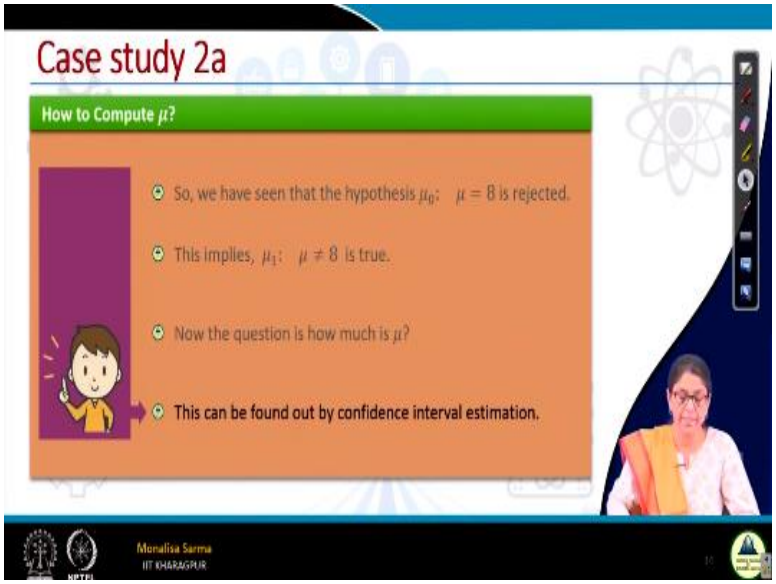
So, first we are solving it by hypothesis simple hypothesis testing how we have this we know already I am not going into details. So, $\alpha = 0.05$ and t value means degrees of freedom we have taken 16 samples so, degrees of freedom is 15 degrees of freedom 15 for $\alpha = 0.05$ that means α by 0.025 we found value of t should be absolute value of t should be greater than 2.13 means if my flatted tail t has flatted. So, the rejection region is this side is 2.1314 this is -2.1314 this is my rejection region.

(Refer Slide Time: 24:03)



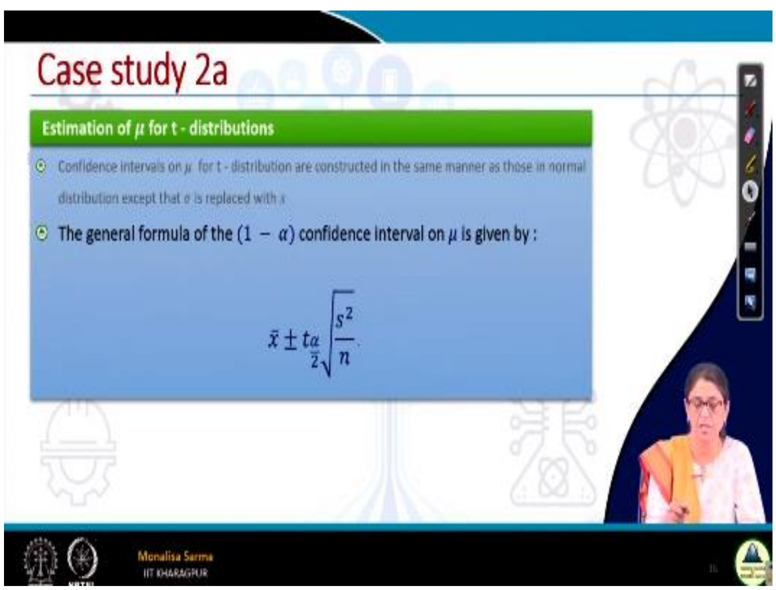
So, from the sample I have calculated x bar I have calculated s ^{2} from the sample data what is given and I found my calculated value is -2.4136. So, it is greater than rejection region so, that means my null hypothesis is rejected. Obviously, we got it for other case is also the null hypothesis is rejected.

(Refer Slide Time: 24:26)



Now, if we want to so, we have seen that a hypothesis μ = 8 is rejected this implies that μ is not equals to 8 is true. Now, the question is how much is μ? This can be found by confidence interval estimation.

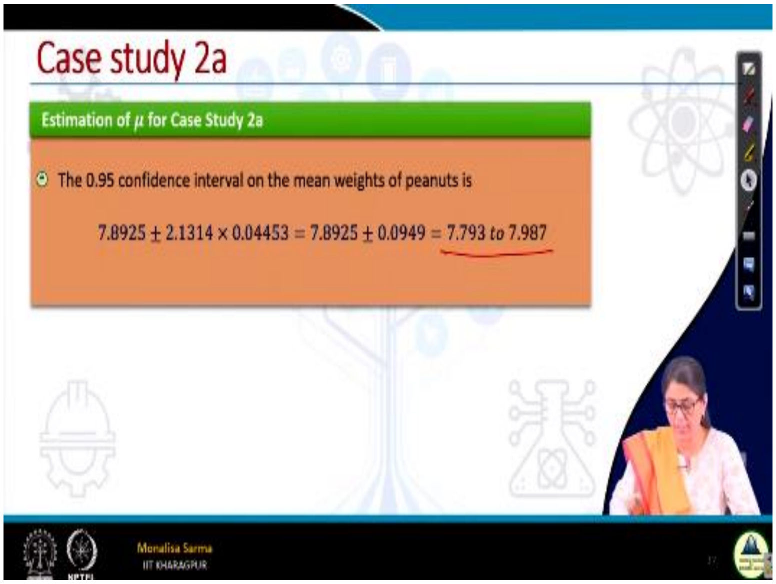
(Refer Slide Time: 24:43)



Now, how do we find out the confidence interval estimation here? Because here we do not know the σ so, we cannot use z distribution. So, in that case we will have to use t distribution. So, t distribution also confidence interval are constructed in the same manner as a z distribution but instead of z dot α / 2 we find out to take t of α / 2. So, same formula so, instead of σ / √ n, we have taken s / √ n or √ S^2 / n is the same thing is not it?

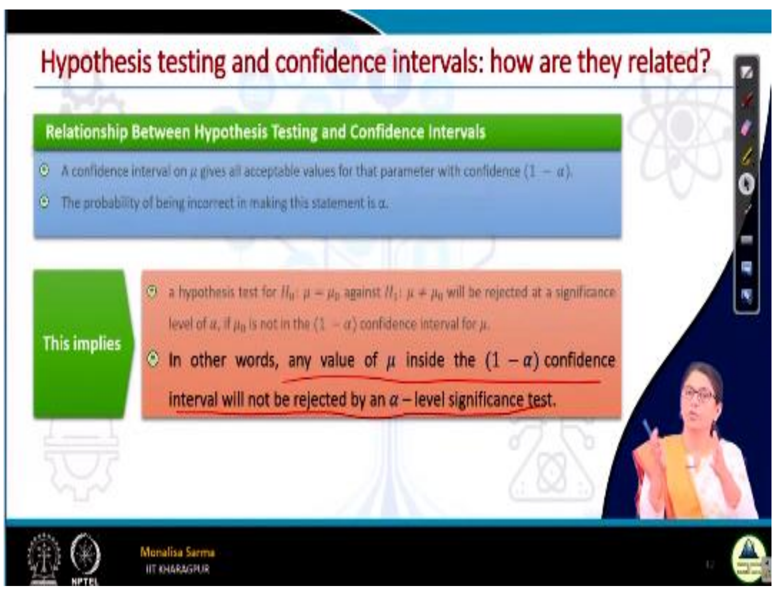
So, instead of z α / 2 I have taken t α / 2 everything remains same and instead of σ I have used s.

(Refer Slide Time: 25:20)



So, that is how I found a confidence interval in this range. So, it agrees with the hypothesis testing.

(Refer Slide Time: 25:30)



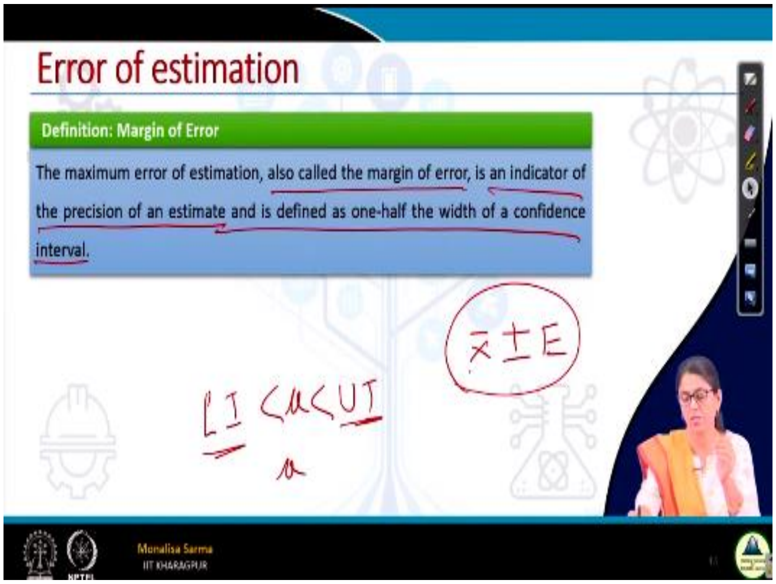
So, the relationship between the hypothesis testing and confidence interval, the confidence interval on μ so see a confidence interval and μ gives all acceptable values for that parameter we did with confidence 1 - α with 1 - α confidence I can see that I am 95% confidence my μ value lies in this range if my significance level is 5 that my μ value given a significance level of 5 I can say with 95% confidence that my μ will rise in this range say this is lower bound and this is upper bound.

The probability of being incorrect in making this statement is α. So, this implies a hypothesis for a taste $\mu = \mu$ naught against μ is not equal to μ naught will be rejected at a significance of α if μ is not in a 1 - α confidence interval is not it? What does this mean a hypothesis test it will reject the null hypothesis that μ is known, if μ is not in the when it is at the high null hypothesis if it finds that μ naught is not in the confidence interval like here 8 was not in that interval in this example, but we got see 8 is not in this interval.

So, the null hypothesis rejected is not it? This imply in other words, any value of μ inside the 1 - α confidence interval will be rejected by α level significance it will not be rejected. In other words, any value of μ inside this confidence level will not be rejected by α level significance tests. So, hypothesis testing and confidence interval it is basically the 2 sides of the same point we can say if hypothesis testing it is rejected that means, it will not be in the confidence level.

If the value is not in the confidence level definitely it will be reflected in the hypothesis test as well. So, it is like 2 sides of the same coin.

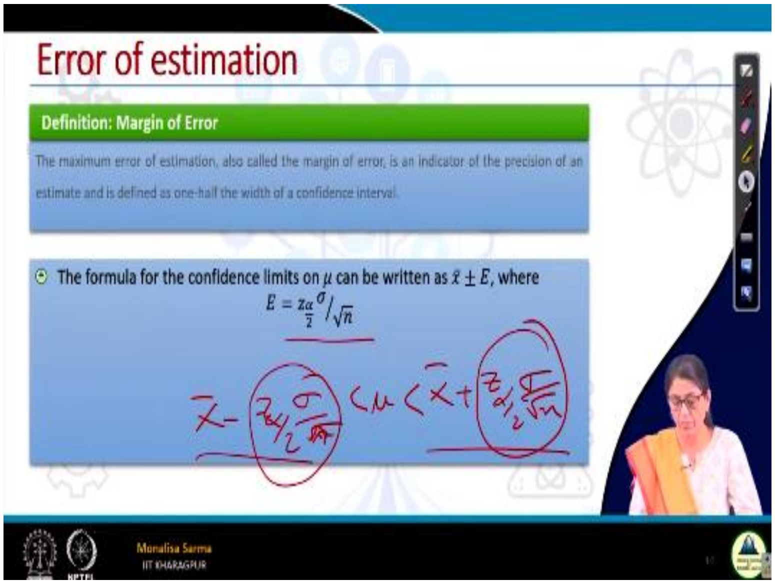
(Refer Slide Time: 27:30)



Now, here it is one more concept in confidence interval estimation that is the margin of error. So, x bar is the point estimation. So, the from x bar till how much we can go how much we can drill, as I told you we can hedge a bit So, how much we can hedge from x bar how much we can hedge that is my error of estimation. So, a maximum error of estimation is called a margin of error is an indicator of the precision of an estimate if point estimate is highly precise, is not it?

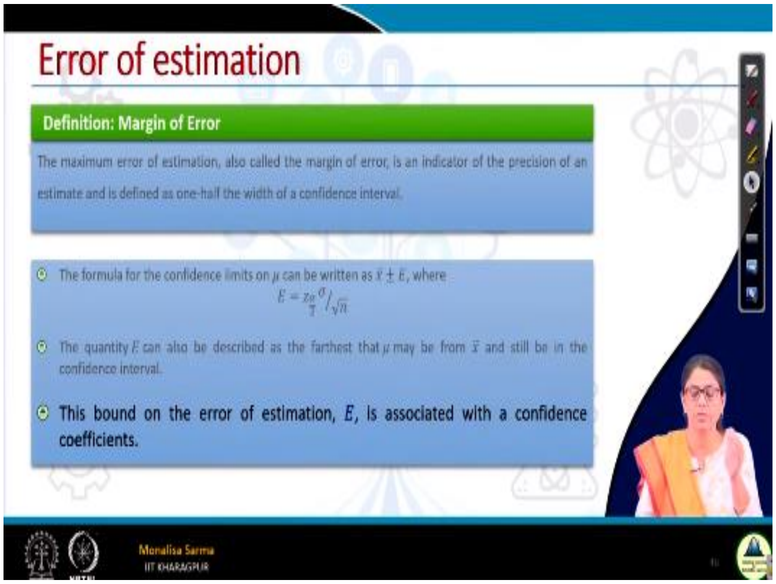
So, it is an precision of an estimate and is defined as one half of the width of the confidence level. So, it can tell that means some value this lower interval less than equals to μ less than equals to upper interval is not it? So, what is this μ I got basically this interval one half of this interval I call it other margin of error that is E x bar + - E because μ will lies in this range what is x bar + E. This is the upper interval x bar - E that is the lower interval.

(Refer Slide Time: 28:49)



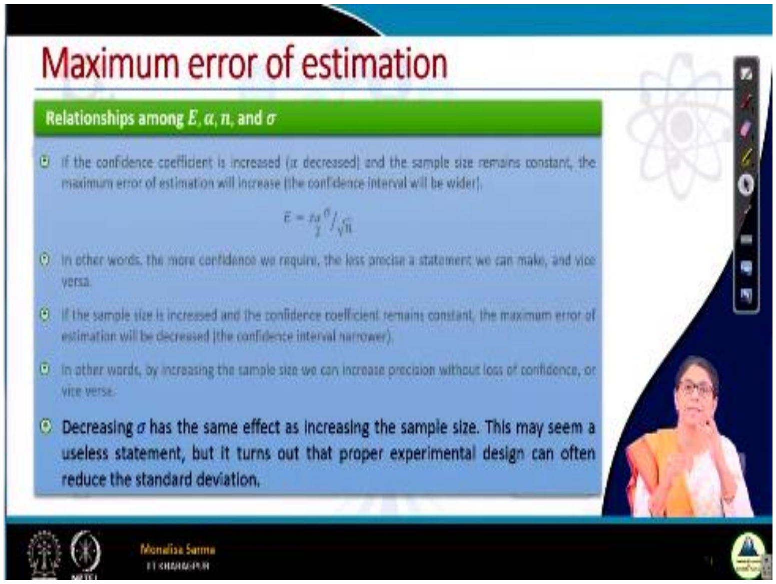
So, this is what is E? E is this value z α / 2 σ √ n because what we got X bar - z α / 2 σ / √ n this is less than equals to μ, μ is less than equals to x bar + z α / 2 σ / √ n this was our interval is not it? So, this is our lower interval this is our upper interval, this portion we call it as a error of estimation z α / 2 σ / √ n this is called the error of estimation. So, from this expression E is equals to z α / 2 σ / √ n and from this expression we can come to some observation.

(Refer Slide Time: 29:37)



So, first what is E the quantity E can also be described as the farthest that μ maybe from x bar and still be in the confidence interval. What does it E describe far does that μ can be from x bar but it is in the confidence interval. This bound on the error of estimation is associated with a confidence coefficient definitely it is associated with a z α / 2 is what it is associated to a confidence coefficient of α.

(Refer Slide Time: 30:06)



Confidence significance level of α confidence coefficient is 1 - α. So, in this if the confidence coefficient is increased that means, if α is decreased confidence coefficient means 1 / α is increase means α is decrease, if α is decreased, then what happens if α is decreased my value of z will be more is not it? In this this famous concept, so, if my α is decrease initially suppose my α is this much α was this much now I am reducing this portion that means.

What my z value is becoming more desert plus side as well as my saddle is becoming less it is going more decide. So, my Z value is increasing absolute value of Z is increasing when my absolute value of z is increasing that means, what my z α / 2 value is also increasing. So, this value is increasing that means, what this value E will be more in fact, if you can see it this way, when my rejection region become less my acceptance region becomes more when α is decrease my rejection region becomes very, very less.

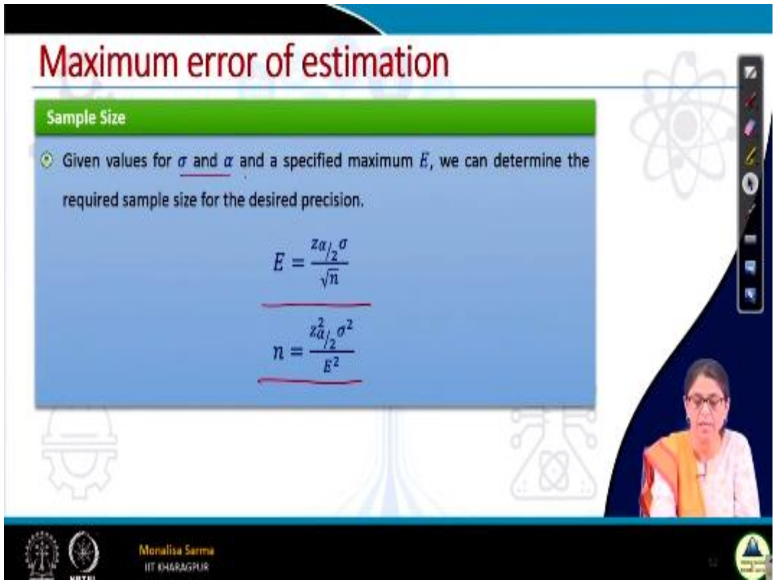
When my rejection region become less accepting region becomes more, is not it? So, what happens when my confidence coefficient is increased my precision decreases I am telling you value lies between 2 to 4 one statement another statement I am telling my value lies from 1 to 10 which is more precise 2 to 4 is more precise here 1 to 10 means have been reduced interval. So, my precision has gone down. So, similarly, when α is decreased, my confidence interval has increased that means, what my precision has decreased.

In other words, the more confidence we require, the less precise a statement we make our confidence has increased whatever statement precision has decreased if the sample size is decreased now, if we increase the sample size what happens if we increase the sample size here from the expression on only you can see if you increase the sample size my error estimation becomes less for the same of confidence coefficient.

So, it might increase the sample size without sending the confidence coefficient I can increase the precision because E value become less so, the sample size is increased and the confidence coefficient remain constant the maximum error estimation will be decreased the confidence interval becomes narrower. In other words, by increasing the sample size we can increase precision without loss of confidence or vice versa. And decreasing σ also has the same effect but decreasing σ is that something which we do not have any control is not it?

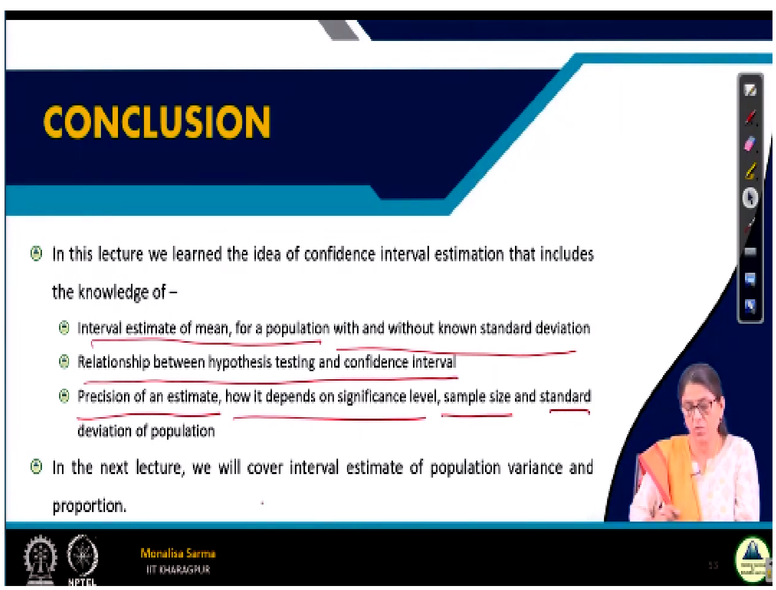
That is why it is in dismissing a useless statement, but when we how can we decrease σ how can we decrease the standard deviation or variance of a population if we have a proper experimental design while doing the well manufacturing the things are while doing the things only if I have my σ is reduced by good experimental design will have a very less σ. So, that is why it is decreasing σ it is not in the hands of a statistician it is totally on the hands on the people who are actually producing this stuff.

(Refer Slide Time: 33:31)



So, from this expression, this is the expression for you now, if we are interested in finding out what is the sample size required for a particular precision and particular significance level given the precision that is E and given the significance level what is the sample size required just from this expression we can find out what is the value of E given values for σ and α and a specified maximum E we can determine the required sample size for the desired precision.

(Refer Slide Time: 34:04)



So, in this lecture what we learned we learned about idea of confidence interval in estimation that includes the knowledge of interval estimate of mean of a poor population with and without known standard of deviation, if we know the standard deviation, if we do not know the standard division how we do then the relationship between the hypothesis testing and confidence interval we have seen that on the precision of an estimate what do you mean by precision of an estimate? How it depends on a significant level?

Precision of estimates how it depends on a significant level? It has significance level is decreases my precision decreases, is not it? I am sorry, my significance level is decreased my precision also decrease how sample size if my sample sizes increase my precision increased it my standard deviation is decreased my precision increased and the next lecture will also cover interval estimate or population variance and proportion here we have only seen interval estimation of population mean.

(Refer Slide Time: 35:02)

So, these are the reference. Thank you guys. Thank you.