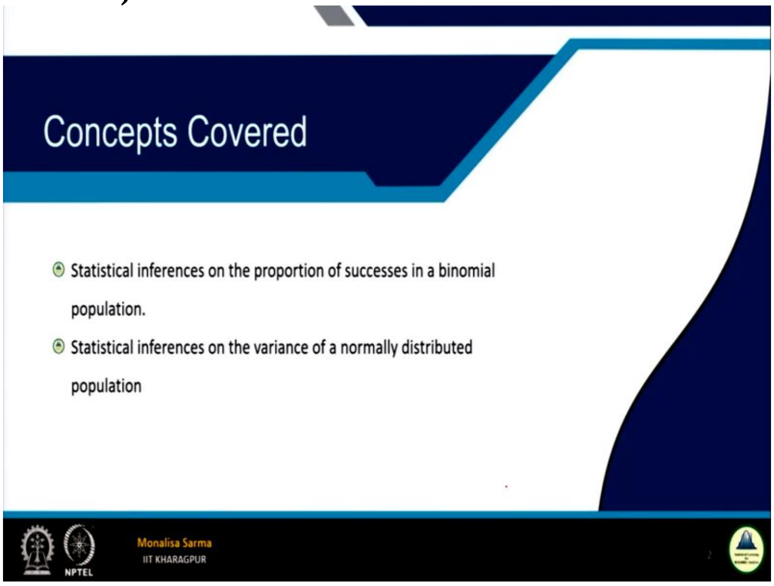
σμ√α∞ β χStatistical Learning for Reliability Analysis

Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology, Kharagpur

Lecture - 27
Statistical Inference (Part - 5)

Hello, everyone so in continuation of our earlier discussion on statistical inferences.

(Refer Slide Time: 00:34)



Today again we will be learning few more topics on that. Today what we will see is statistical inferences on the proportion of success in a binomial population. So, we have seen statistical inferences on mean, we have also seen how to find out a confidence interval as well. So, now we will be seeing statistical inferences on the proportion of success. And when we talk about proportion, I think you guys can remember.

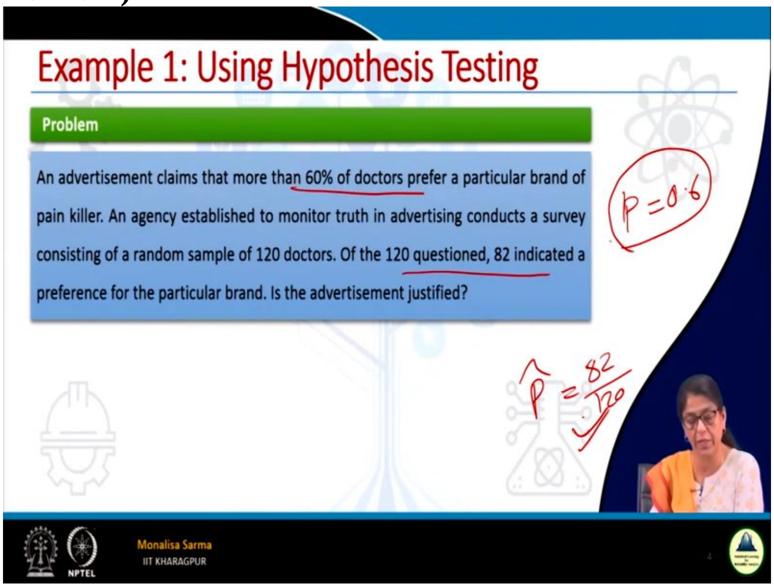
So, when we want to infer something about the proportion of our populations, their populations and then we could find out that it is basically it has the characteristics of a binomial populations random variable basically, it has the characteristics of a binomial distribution and that is why we can consider the population as a binomial population. And moreover, the sampling distribution also for a sampling distribution of the proportion we have seen that we can use normal distribution if the sample size is larger.

So, that as I told you again and again sampling distribution is basically the backbone of any inferences. So, when we will be seeing the proportion of success statistical inferences on a proportion of success, we will actually consider the sampling distribution of the proportion.

Similarly, we will also see the statistical inferences on the variance of a normally distributed population.

When we have to infer about the variance rather I should say when we you consider when we use chi square distribution or t distribution, always our population has to be a normal population, if not normal, at least not too much away for a normal populations, otherwise, the data that we will get will not be a precise result. So, that chi square distribution, F distribution and t distribution is very much sensitive to normality assumptions that we have seen again and again. So, now, first inference on a proportion, we will start with an example.

(Refer Slide Time: 02:32)



So, this example is a good way to understand that concept, is not it? So, like what in this example, what we have here? An advertisement claims more than 60% of doctors prefer a particular brand of painkiller more than 60% that is basically it is talking about a population. So, it is talking about the whole population as a whole about the whole lot it is talking that more than 60% of the doctors prefer a particular brand.

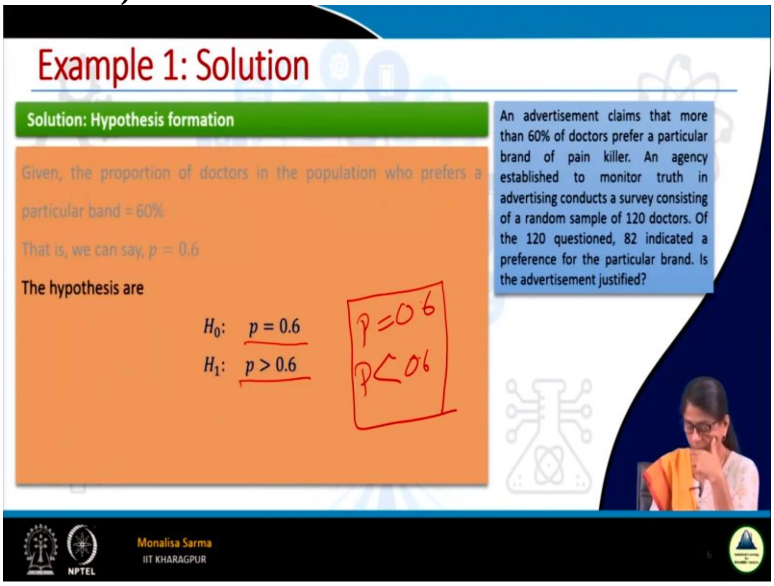
That means I can say p = 0.6 that this is the caution an agency established to monitor truth and agency advertisement is claiming that more than 60% of the doctors prefer this particular brand, and agency third party basically third party to monitor the truth in advertising conducts a survey consisting of a random sample of 120 doctors. So and a third party basically wants to see whether it is really correct.

So, what that says that it takes a random sample of 120 doctors, so out of 120 questions 82 indicated a preference for a particular brand. So, what is my proportion is 82 / 120 is not it? My p cap is 82 / 120 proportion of the sample which has preferred the particular brand. So, it is the advertisement justified. Now, the question is with this proportion when assuming that because advertisement is claiming that this is true.

Advertisement is claiming that p = 0.6 and from the sample what result I got this proportion of the sample which actually prefers this brand is 82 / 100. So, this is of course, this is correct, because this we have tested it this is not something which we have tested, this is we have tested it. So assuming this correct is this possible, that is what we will check right. So, is the advertisement justified so, if we get a quite acceptable probability for this, Then we can say yes, acceptance, advertisement is justified.

Where actually I want to bring to your notice is that first, definitely, it is a question of hypothesis testing, we will do hypothesis testing, now how to frame the hypothesis?

(Refer Slide Time: 04:54)



See here, if we frame now here who is doing the research a third party. So, when the third party is conducting the research what the third party wants to prove as I told you the alternate hypothesis is always something what you wants to prove. So, the third party wants to prove whether it is greater. It has doubt in the advertising. So, it wants to prove that whether it is better, so, definitely alternative hypothesis is p greater than 0.6.

And null hypothesis is p = 0.6. Why I am discussing this question, because I want to bring to your notice one more thing. Suppose this same example the same thing if instead of the third party even if what to say the particular manufacturer, the manufacturer whoever advertise is the manufacture wants to do this experiment wants to check then how he would have formed a hypothesis.

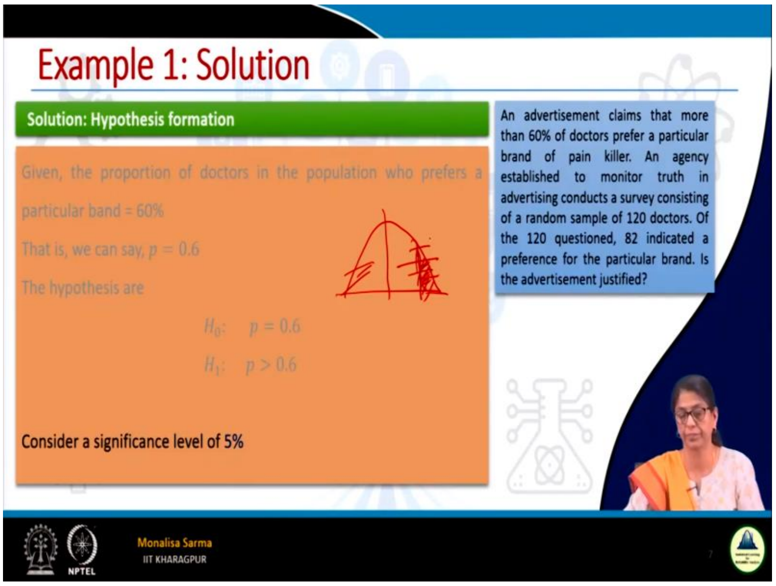
Because the manufacturer he believes that 60% of the doctor prefers this brand, he believes that. So, for him that is the status quote that is greater than 0.6. So, he wants to check whether it is less than, he does not want that less than to happen, he wants because type 1 is something which we want, what is the null hypothesis something which you want that to happen. That is why we have type 1 error very less amount, the type 1 error we have the significance level that is the significance level is a very less value we keep is not it?

Because we always put that in null hypothesis, which we want it to happen because which it maintains the status score. So, if the particular manufacturers may have wanted to test this, then for them, they would have framed a hypothesis, they would not have framed this hypothesis instead what do you have different hypothesis? The alternative hypothesis will be p less than 0.6 they would have want to put is it p less than 0.6.

They just for checking purpose they are sure that it is greater. So, for them step this is it is greater that means greater or equal. So, as I told you null hypothesis, we always specify with equality. So, if it is less other, because again one more thing I have mentioned, remember the both the hypotheses are exhaustive, is not it? It should cover all the things so if it is less than definitely will be other one will be greater and equal.

So, has to cover all the things all the values. So, if the firm would have manufacturing, firm would have frame the hypothesis, they would have framed in this way, but now it is a third party they want to prove that it is really greater than 0.6. So, they are framing the hypothesis as p greater than 0.6. So, fine we have this.

(Refer Slide Time: 07:40)

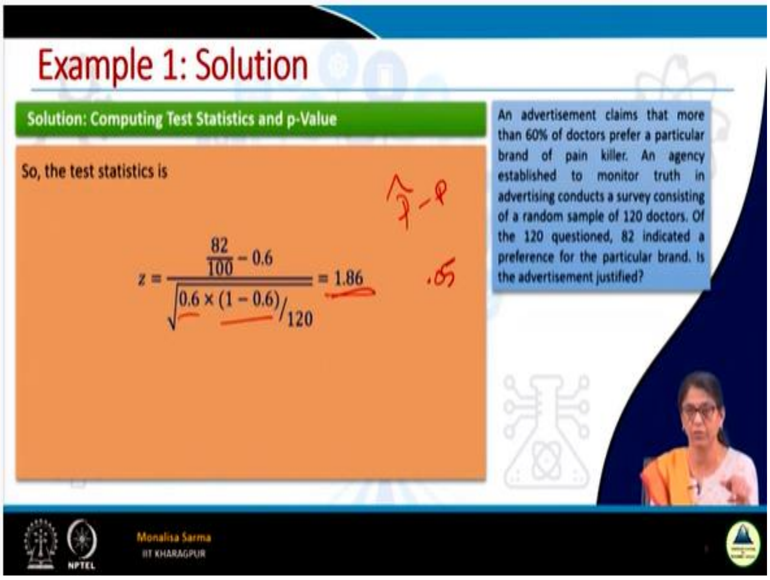


Now, how to test it? It is given a significance level of 5%. So, it is a one thing, so, when this is a one tailed will have only one rejection region. So, for greater than it is sender, greater than 0.6 that means, our rejection region will be in the upper tail, when we are talking of this is the normal distribution, this is the upper tail, this portion is the upper tail, this portion is the lower tail.

So when it is, when we are looking for whether it is better, so, our rejection region will be in the upper tail, because our null hypothesis is less than equals to is not it? A null hypothesis if it is greater what is a null hypothesis? The null hypothesis basically less than equals to so, if we get a value which is much greater that means, in the upper tail then we can say that your null hypothesis is rejected.

So, this is the case p greater than 0.6 so, it is given significance level of 5% so, we have to find out the rejection region in this region in the upper tail.

(Refer Slide Time: 08:50)

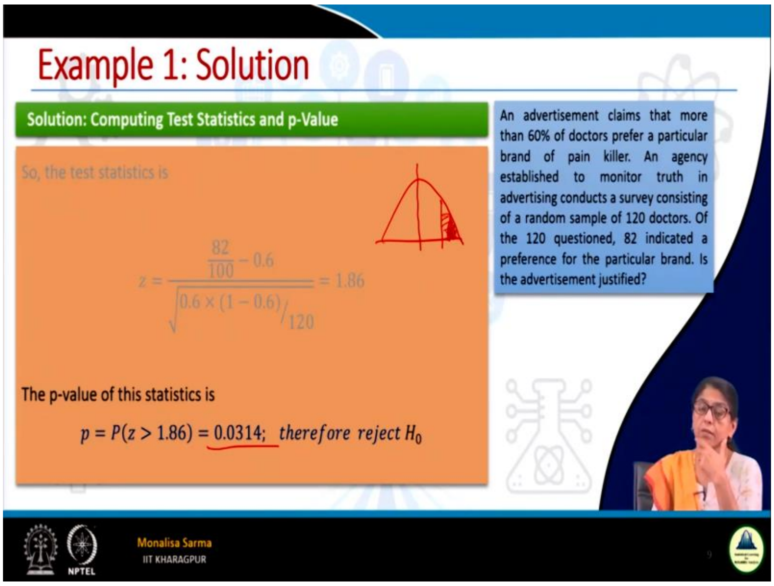


So, we will definitely will be using the z statistics here, z is the z distribution what is the value for this that is peak p bar and p cap p cap - p so this p cap - p then it is the σ / √ n, what is the standard deviation from the binomial population that is we got a p x 1 / 1 - p. So, this is p x 1 - 2 we have done it while discussing sampling distribution here the standard deviation is p x 1 - p that is, so, this is p 1 - p / 120.

So, we got 1.86 value so, corresponding to 1.86 there is 2 way of doing it, either you find true corresponding to this 5% significance, what is the critical region that you find if the value what we get from the z if this value is greater than the critical region, then we reject the hypothesis. Another way is that we find the p value of this and critical significance level is given 5% that means our 5% means it is 0.05.

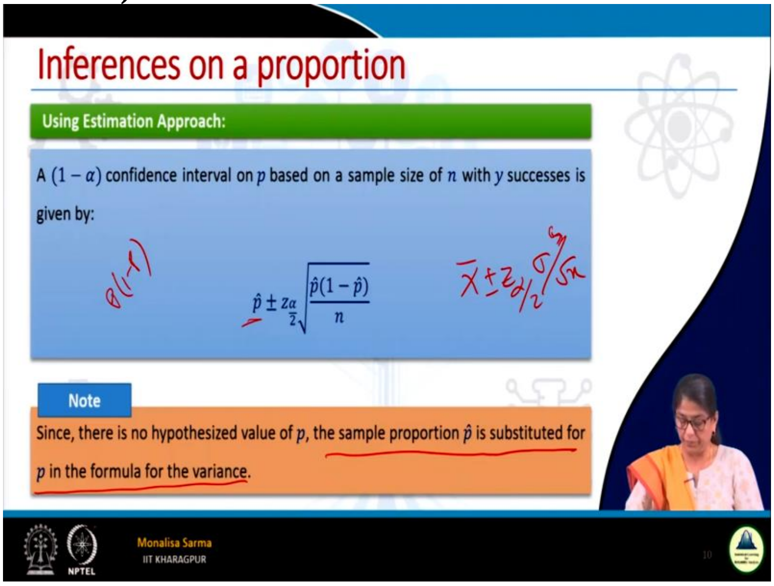
Area is 0.05, so, if it is the area corresponding to this is lesser than 0.05 then we will reject a hypothesis. So, let us do it by computing p value, you can you can do the other one that whatever it is.

(Refer Slide Time: 10:06)



So, p value for this is we will get it from the table it is one tail. So, you do not have to plus it twice. So, it is 0.03 and 0.03 less than 0.05. So, if this portion is 0.05 so, 0.03 will be somewhere set this portion. So, if fall in the critical region is so, therefore, reject H 0.

(Refer Slide Time: 10:40)



Now, this is using hypothesis testing, we have also learned the other one approach that is the confidence interval estimation. So, confidence interval estimation I have already discussed under what situation we use? The 2 situation first one is, number one is that when we do not have any hypothesis value we cannot guess we cannot make a guess what will be the hypothesis value, in that case, we just have to estimate how we estimate?

We take a sample from the sample we calculate a particular statistics whichever we are interested and based on that statistics that statistic we call it as a point estimate based on the statistics we form the sampling distribution of that statistic, and then we find out the confidence interval that is one way, another way is that not another way that is the one reason why we go for confidence interval another reason is it supposing the hypothesis is they have rejected.

So, like we have done an example $\mu = 8$ alternate hypothesis is μ naught = 8 so, if μ naught $\mu = 8$ is rejected that means, μ is not equals to 8. Now, μ not equals 8 it means what is the value of μ? So, in that case also we can find the confidence intervals to know actually μ falls in what range. So, now, similarly for proportion also we can find a confidence interval.

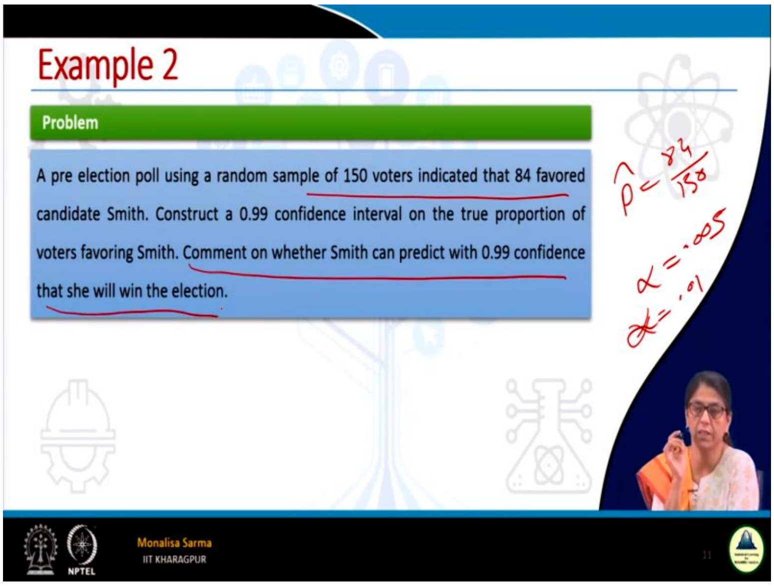
So, saying whatever formula we use to find out the confidence interval for inferences and mean same formula to find out what to say confidence interval estimation for a mean what is the formula if you remember it was x bar + - z α / 2 σ / $\sqrt{}$ n is not it? x bar + - z α / 2 σ / $\sqrt{}$ n. So, here instead of x bar, x bar is the mean of the sample.

So, we have proportion that is the p cap and we have σ what is the σ distance placing the see note it, the σ is the population standard deviation remember σ is the population standard deviation. So, and when we use when the population standard deviation is not known to us, then we use t distribution, in case of t distribution instead of σ we had s is not it? s that is the standard deviation of the sample.

Here similarly, here when we consider the standard deviation, so, what is the standard deviation of the population? Standard deviation of the population is equals to p x 1 - p, but if we do not know the hypothesise value what is the p of the population then what we will use like for t distribution we use s similarly here if the p is not known to us, we the sample proportion p bar is p cap is substituted for p in the formula for the variance.

We substitute p cap when we consider, when we calculate the estimation, it is mostly because we do not have the value of p we do not have the proportion value and proportion of the population. So, we substitute p with p cap so, my standard deviation is p cap x 1 - p cap.

(Refer Slide Time: 13:50)



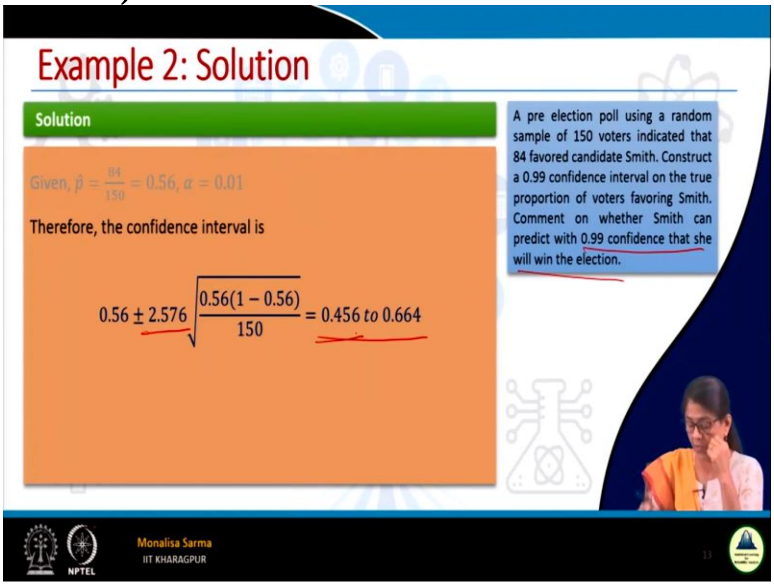
So, you see this example here, if pre-election polls using a random sample of 150 voters indicated that 84 favoured candidates Smith. So that means my p cap this is not p cap that is not talking about the whole population, it has taken a random sample of 150. So, what is that p cap? p cap is 84 / 150. Construct a 0.99 confidence interval we have to construct a 0.99 confidence interval. In the 0.99 confidence interval what is the significance level that means? A, α is 0.01.

So, if it is a two tailed if you have to check for two tailed 0.01 means it will be both side is α will be is equal to 0.005 and then if it is two tailed, but if it is single tailed we will just consider α = 0.01, now see this situation is what two tailed or single tailed whatever what it is asking, construct a 0.99 confidence interval on the true proportion of the voters favouring Smith. So, it is not talking about a single population one tailed or two tailed it is not talking about anything.

And we do not know the value of p as well we just whatever information we have just from the sample like the example what I have given, so, the person wants to open a retail outlet and he does not know the spending capacity of the people, so, he has this taken a sample and try to find out what is the annual income of the people, so, he does not have any idea so, he just took out the value from the sample.

So, similarly, we just have the sample value, we do not have any idea what maybe the population proportion we do not know p. So, p, we found out p cap here. So, comment on whether Smith can predict with 0.99 competence that she will win the election.

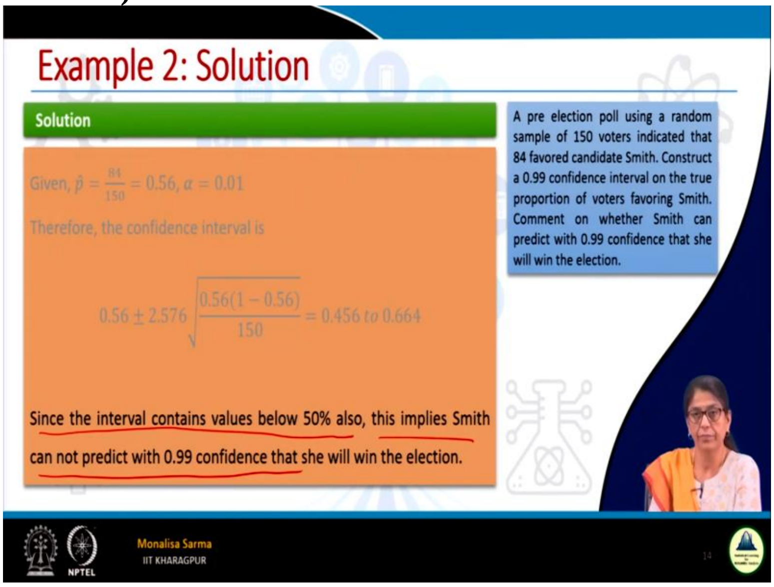
(Refer Slide Time: 15:49)



So, now here we have p cap is this α = 0.01. So now, what is the confidence interval here the confidence interval same value whatever we used here this is the confidence interval p cap + - z α / 2 p cap x 1 – p cap / n √ n. So, just putting the value this is z of α / 2 values corresponding to that is 2.576 you can see it from the z table. So, we got this is the value we got.

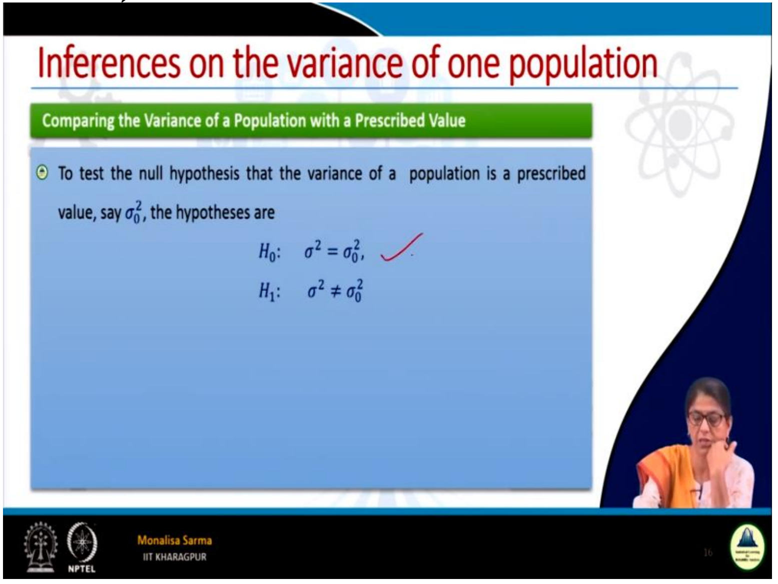
This is the confidence interval from 0.45 to 0.66 that means what is the last question it is asking comment on whether Smith can predict with 0.99 confidence that you can win the election for winning the election at least more than 50% so, both. So, here the person who are voting is we have values below 0.5 also of course, we have values above 0.5 but we have below values be below 0.5 as well.

(Refer Slide Time: 16:52)



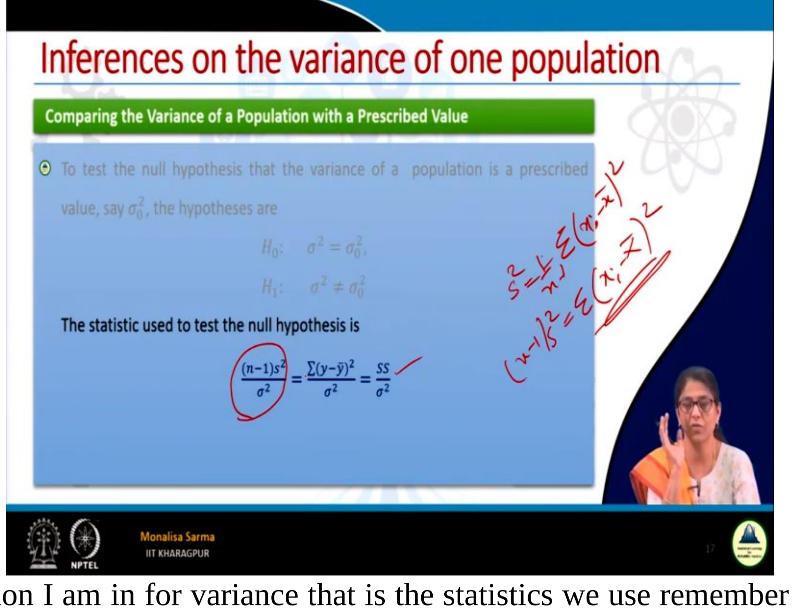
So, since the interval contains values below 50% also this implies Smith cannot predict with 0.99 confidence that she will win the election. So, next is; we will see how we infer on the variance of one population.

(Refer Slide Time: 17:12)



All this we are considering for one population. So, variance for one population similarly, the way we do it for mean we have done it for proportion. Similarly, this will be our null hypothesis this is an alternate hypothesis.

(Refer Slide Time: 17:27)

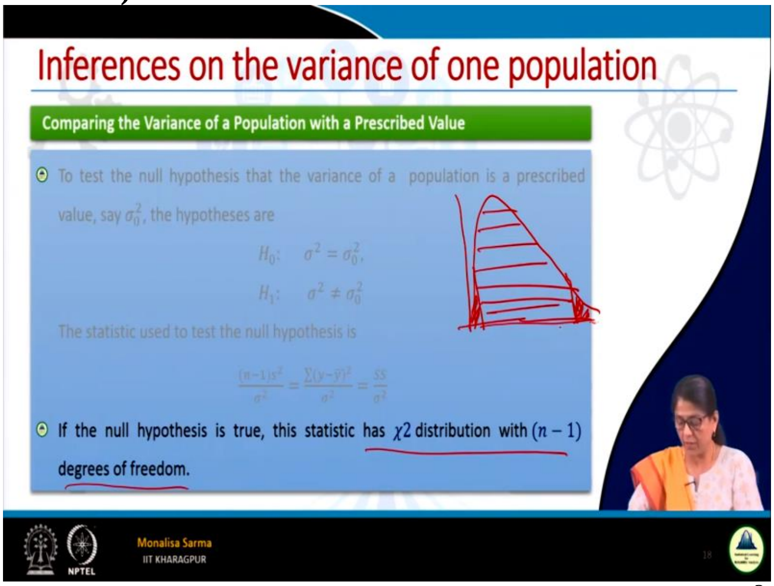


And for proportion I am in for variance that is the statistics we use remember for what to say to find out the sampling distribution of a variance we have used the statistics n - 1 s^2 / σ^2 this is the value and this n – 1 s^2 / σ^2 it it has a chi square distribution remember we have seen, it as a chi square distribution with degrees of freedom n - 1 is not it?

So, this is has a chi square distribution and what is this n - 1 s ^{2} you can also tell is a sum of SS is called sum of squares basically, I have shown you why it is called sum of squares, because if we find out the formula for s ^{2} what you get remember the formula for variance the whole portion is basically and then in it is 1 / n - 1 summation of x i - x bar ^{2} this is the formula for a square. So, if I bring n - 1 here n - 1 s ^{2} what remains is this summation of x i - x bar ^{2}.

So, this is a summation of square that is why when you say s ^{2} is also called summation of squares. So, there is nothing so, heard about that. So, you can use either this or you can use this. So, basically n - 1 s ^{2} / σ ^{2} is suite.

(Refer Slide Time: 18:45)



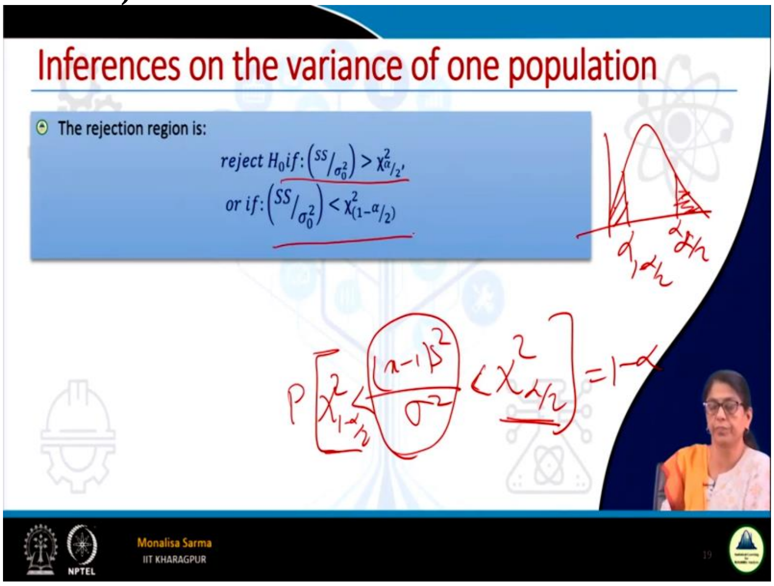
The statistics chi square is the null hypothesis is true this statistics n - 1 / s ^{2} / σ ^{2} this statistic says chi square distribution with n - 1 degrees of freedom, that we have already seen. So, now we will have to find a for if we want to infer on the variance we will have to find out the value of the statistics and then we will have to see whether this has a chi square distribution basically it meaning it has a chi square distribution means it has to have a significant probability not if it says very less probability.

We have seen it is 95% of the value falls within what range I am, if you remember have discussed that. So, if it is so, chi square distribution chi square distribution we have something of this chart. So, 95% of the value basically satisfies this, satisfy the chi square distribution of the for the hypothesised σ value. If our value falls in this region why will a value fall in this region?

If we assume variance to be very small which is very unlikely very small of course, it can happen it is not that is but it is very unlikely, if we assume a variance to be very small we will get in this range. Again if we assume our variance to be very big, very high then it will fall in this range that is also quite unlikely a very high variance, it is a very unstable process. So, that is also quite unlikely.

So, if our value falls within this 95% range, then we can say that it is it satisfies the chi square distribution. So, now here, so what is the rejection region corresponding to a particular significance level.

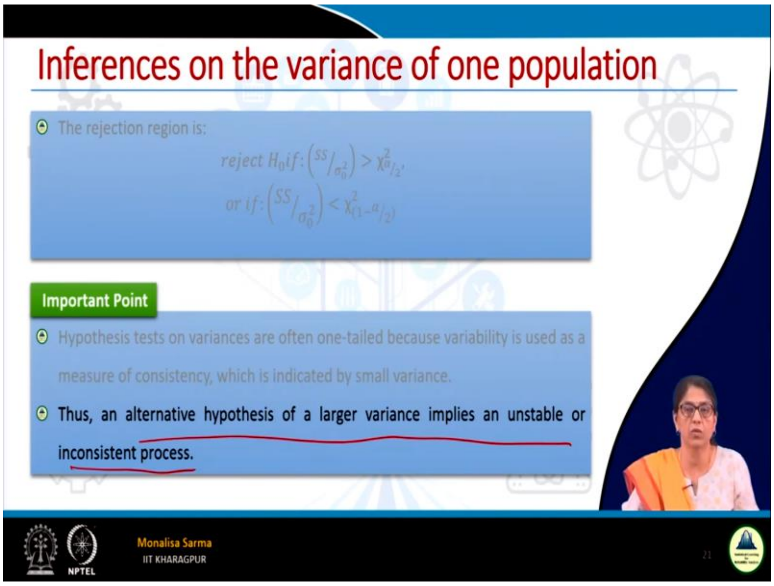
(Refer Slide Time: 20:29)



Remember how to find out a rejection region. So, it has like let me write it here only so, it is n - 1 s^2 / σ^2 basically from the confidence interval it will be easier to write it this way χ^2 1 - α / 2, this is less than χ^2 α / 2 this probability is equal to 1 - α is not it? My this value n - 1 s^2 / σ^2 it should fall between this χ^2 α / 2 or χ^2 1 - α / 2.

So, that means, what is my rejection region if my this statistics, if this statistics is if it is greater than χ^2 / 2, greater than χ^2 α / 2 means this region then we reject it or this value if it is less than this value χ^2 1 - α / 2 is value, χ^2 α / 2 is this portion. So, if it falls in this region or region then we will reject the hypothesis. So, reject H 0 if this is true or if this is true.

(Refer Slide Time: 21:46)

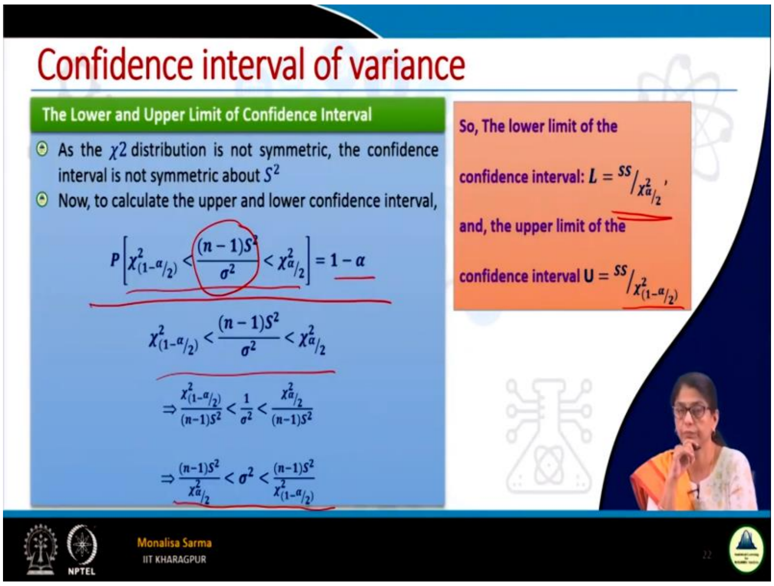


Coming back so now, if you see here hypothesis tests on variance are often one tailed because variability is is a measure of consistency hypothesis testing when we try to find out infer on the variance usually when we try to infer on the variance we never try to inference if variance is great what to say smaller than this value, because we need smaller variance what does smaller variance indicates? Smaller variance indicates consistency.

So, any product, any equipment, any material whatever it is we want variance to be as less as possible. So, smaller variance is a positive thing for us. So, definitely we will not try to check for usually we do not try to check for a smaller variance what we try to check for whether it is greater variable. So, mostly in hypothesis testing for inference and variance it usually for us it is usually that means we test for high and greater variances that way it is single tailed.

See alternative hypothesis of a larger variance implies an unstable or inconsistent process. So, in unstable process, no point carrying out hypothesis testing or confidence interval estimation.

(Refer Slide Time: 22:58)

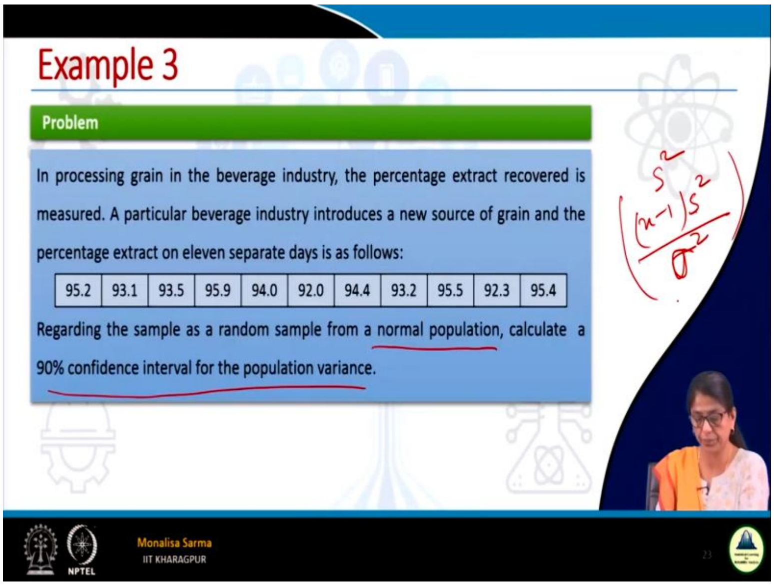


So, now, what is the confidence interval we have seen the using hypothesis testing how we do this now, we will see how the confidence interval for variance. So, this is what just now what I have mentioned if the confidence coefficient is 1 - α if the significance level is α, my confidence coefficient is 1 – α, if it is 5 confidence coefficient is 95%. So, probability that it falls within this range is 1 – α.

But my this statistics falls is the less than the X and χ^2 α / 2 and greater than χ^2 1 - α / 2. So, if I simplify a bit this whole thing if I simplify a bit how I have simplified just divided both sides by n - 1 S^2 just simple simplification I will get this value. So, my σ^2 would fall within this range what range n - 1 S^2 / χ^2 α / 2 n - 1 S^2 / χ^2 1 – α / 2.

So, this is the confidence interval, confidence interval for confidence coefficient 1 - α. So, the lower interval lower limit which is a lower limit, lower limit is n - 1 S^2 / χ^2 α / 2 or I can tell SS / χ^2 α / 2 this is my lower limit what is my upper limit is assessed by χ^2 1 - α / 2 this is my upper limit.

(Refer Slide Time: 24:25)



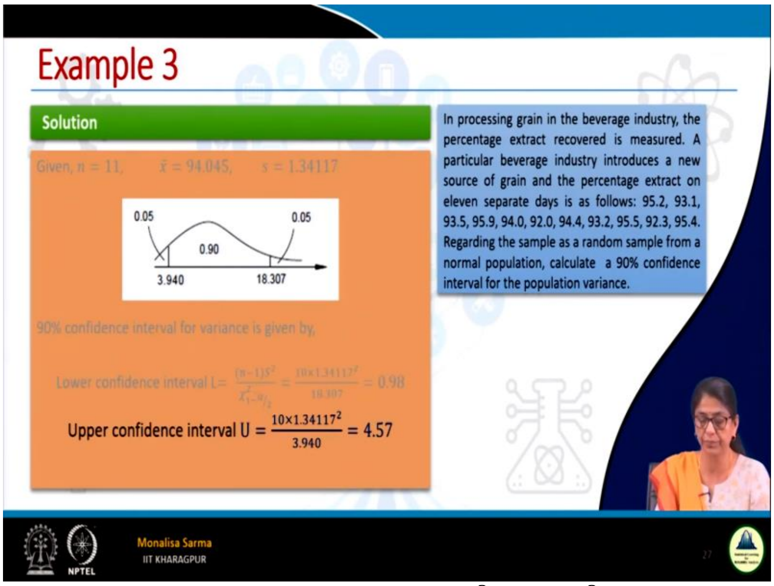
So, we will see a problem for that in processing grain in the beverage industry the person is extract recovered is measured, a particular beverage industry introduce a new source of grain and the percentage extract on 11 separate days is as follows. This is the sample what we have taken regarding the sample as a random sample from a normal population variance of course the population has to be normal.

Calculate the 90% confidence interval for the population variance we have to calculate the 90% confidence interval, here we do not know what is the population the variance is not given. We do not know what maybe the population variance or population standard deviation, we just may have taken a sample from the sample we got this data and from this data we will can calculate the 90% confidence interval.

So, basically from this sample we will try to find out the value of S^2 once you find out the value of S^2 that is the variance from the sample then n - 1 that is here total 11 days n - 1 is 10 n - 1 S^2 / σ^2. So, this is what to say this is the statistics we have to find out. So, what happened, what does the confidence interval say? See here to find out this, I do not have the value of σ^2, I do not know what is the value of σ^2 here, I know only S^2.

So, I cannot use sampling distribution based I cannot hypothesise the value because I do not have any idea. So, definitely I cannot use hypothesis testing here what I will do I will just go and find out the confidence interval. To find out the confidence interval what was this? This is the formula remember, this is the upper limit lower limit this is the upper limit.

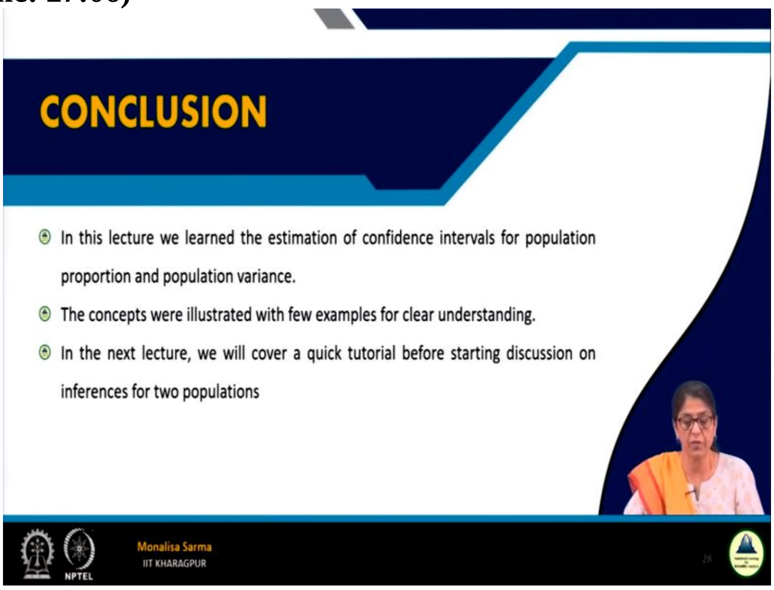
(Refer Slide Time: 26:16)



So, I found out x bar I found that s than what is the χ^2 α / 2 χ^2 α / 2 is this value that we will find it from the χ^2 table, what is α? A is given is 10%. So, α / 2 will be 5% that is 0.05. But this chi square value corresponding to 0.05 with 11, it is 10 degrees of freedom, 10 degrees of freedom we will get it from the chi square table. And similarly for this is 1 - α / 2 χ^2 of 1 - α / 2 for 10 degrees of freedom we will get this value.

So, now, when once we know this value, we can find out the interval low interval, upper interval and lower interval using the formula.

(Refer Slide Time: 27:06)



So, that is all in this class. So, what we have learned? We have learned the estimation of confidence interval for proportion and population variance. So, I have tried to illustrate the concept with some examples. And in the next lecture, we will cover a quick tutorial before starting the discussion on the inferences for 2 populations.

(Refer Slide Time: 27:27)

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