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

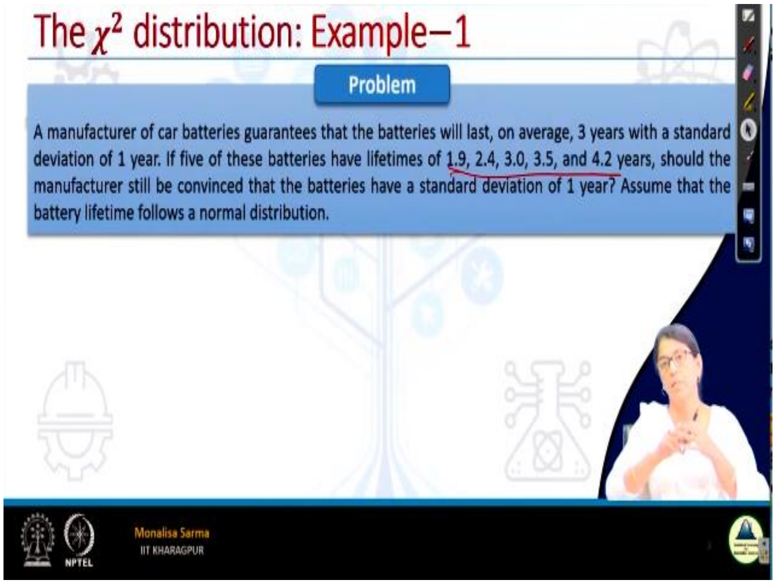
Lecture - 19
Sampling Distributions (Part 4)

Welcome back guys. So, in continuation of our discussion on sampling distribution last class in my last lecture, if you remember.

(Refer Slide Time: 00:37)

We discussed chi square distribution, and just we ran short of time in discussing a problem. The means related to chi square distribution first I will start with that and then I will go to t distribution.

(Refer Slide Time: 00:51)



So, I think you guys can remember what is chi square distribution? Chi square distribution, we use for sampling distribution of variance, like sampling distribution of mean we have sampling distribution of variance, sampling distribution of variance, why it is used? We used to infer if we were to infer something about the population variance, usually it is more mostly used in food and beverages industry, where quality is very much a paramount importance.

And then we need to find out how much variance is it from the specified volume or the specified weight. In those cases, we need to find out the; infer the volume I mean the variance of the volume, for that, we need the sampling distribution of the variance. And for sampling distribution of the variance, we use chi square distribution. So we have discussed what is chi square distribution in my last class.

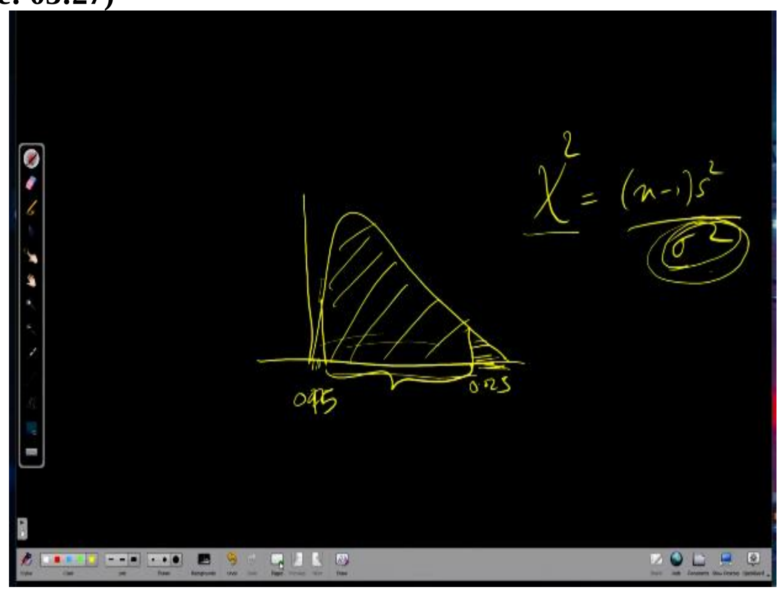
So today, we will be solving 1 problem, of course, I will be solving some other problem in my tutorial classes also, this is a problem so that you can understand the topic better, basically. So now, what is the question here? A manufacturer of a car batteries, guarantees that a battery will last on average, 3 years with a standard deviation of 1 year. So, manufacture it guarantees that on an average battery lasts 3 years with a standard deviation of 1 year.

That is the manufacturer's what to say his claim. So now, if 5 of these batteries have lifetime of 1.9, 2.4, 3.0, 3.5 and 4.2 years, so what happened a person he bought this, that reorder might be you can take in this way, like he wanted to do a bulk purchase of batteries. Before that he is might want to check it whether what the manufacturer is claiming is correct or not, maybe that is the reason or maybe he has bought it for his own need.

And he found that lifetime of this battery is this 1.9, 2.4, 3.0, 3.5 and 4.2 years. So, the manufacturer still be convinced that the batteries are the standard division of 1 year. So, with this, we are not interested in finding them average lifetime, just what we are interested? So, the manufacturer still be convinced that the batteries have a standard division of 1 year that means the variance among them is 1 year or not.

That is the manufacturer whatever his claim, is it correct? That is what he wants to find out. So, remember, when the while discussing chi square distribution.

(Refer Slide Time: 03:27)



We have seen that let me take this chi square distribution is something of some skewed sort of distribution, remember, so it was something this portion at both the position is same, but my figure is not very good. So it is like 0.9975 and this is 0.025, this area is 0.025 and this whole area is 0.975. So, if the probability lies in this range, yesterday we have discussed then whatever its claim, our claim is correct that is what we were.

And that is a reasonable value the probability basically, if it lies within this range that is a reasonable values of probability, we have seen yesterday there in last class basically, if the probability lies in this range which is very less probability under what situation the probability will lie in this range. Remember what was the chi square value? Chi square value is n - 1 S^2 / σ^2 probability will lie in this range.

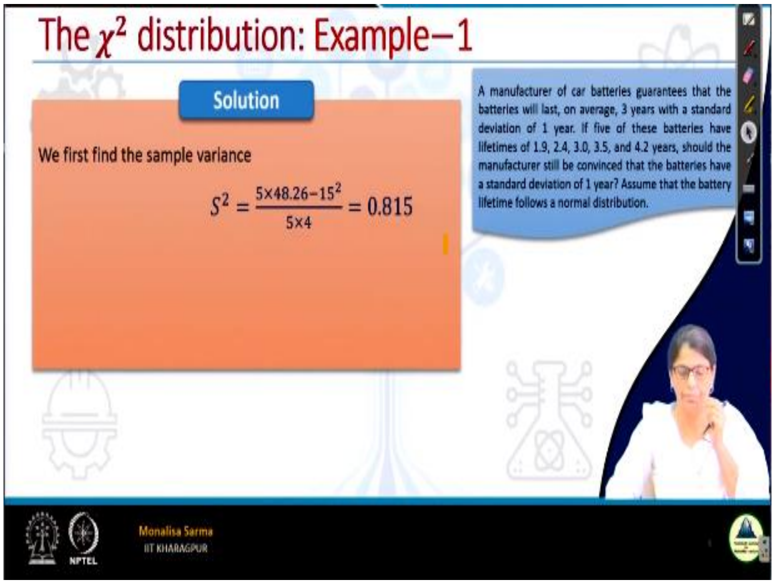
When chi square value is quite high, under what situation chi square values will be quite high, when this value will be very less when sigma square value is very less than we will get a very high chi square value so σ^2 value does a variance value very less variance value that is very much unlikely it is not that it cannot be but it is very much unlikely that it will have a very less chi square value.

At the same time, we will have a chi square value is very lesser chi square value if we have something in this region a chi square value will vary less, is not it, because it is starting from 0. So, under what situation we will have when chi square value is when σ^2 is very large. So, that is also quite unlikely why when the σ^2 is very large that means, this process is not stable when the process is not stable, why at all will go for doing this statistical inference is not it?

It is an unstable process. So, that is the reason that the value lie in this range is very, very unlikely, but it is not that it may not happen it may happen, but if it happens in this range, it might probability lies in this range then it is. Now, this is the claim is correct. So, now, essentially this problem we need to find out whether our; what to say probability lies in that acceptable range are not.

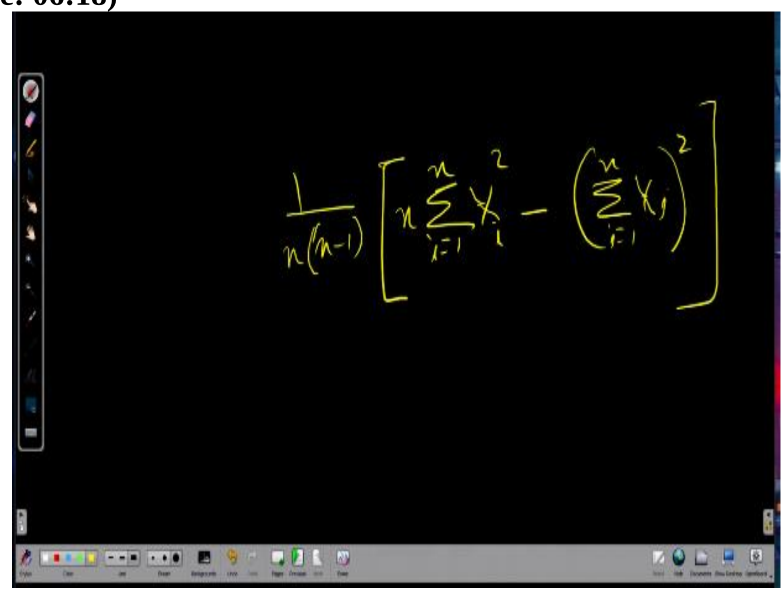
So, how do we solve this question first is that we need to find out the chi square value for chi square value, we need to find out the S^2 that is the variance of the sample.

(Refer Slide Time: 06:06)



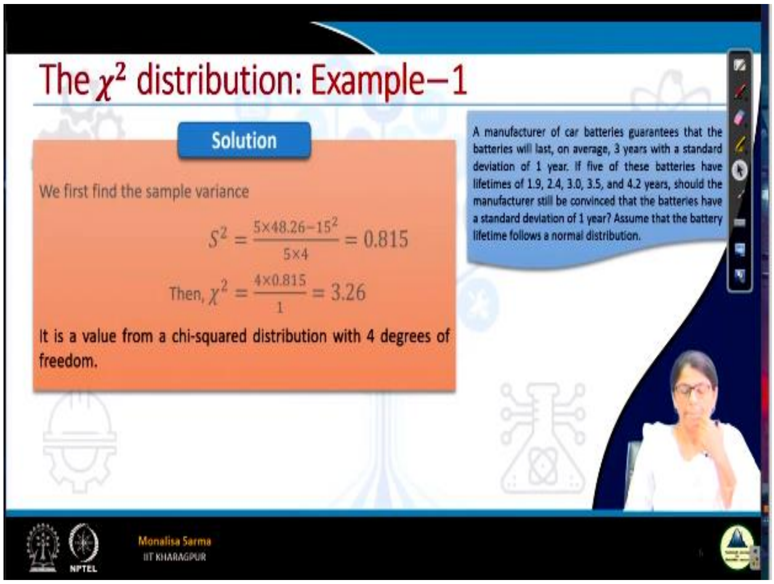
So, variance of the sample this is variance you know that formula for variance or we can use one more, there is one more formula for variance which are.

(Refer Slide Time: 06:18)



Let me give you there is 1 more formula for variance like it is 1 / n x n - 1 this is an n x n - 1 than sum n into ∑i = 1 to n y i ^{2} - i = 1 to n y i sorry it is not y i always x i have x i whole square. So, this is also 1 formula for let me rub it. So, this is also one way of calculating variance. Variance formula I think you know the one the standard formula which you know you can calculate it using that formula also or you can calculate it using this formula also.

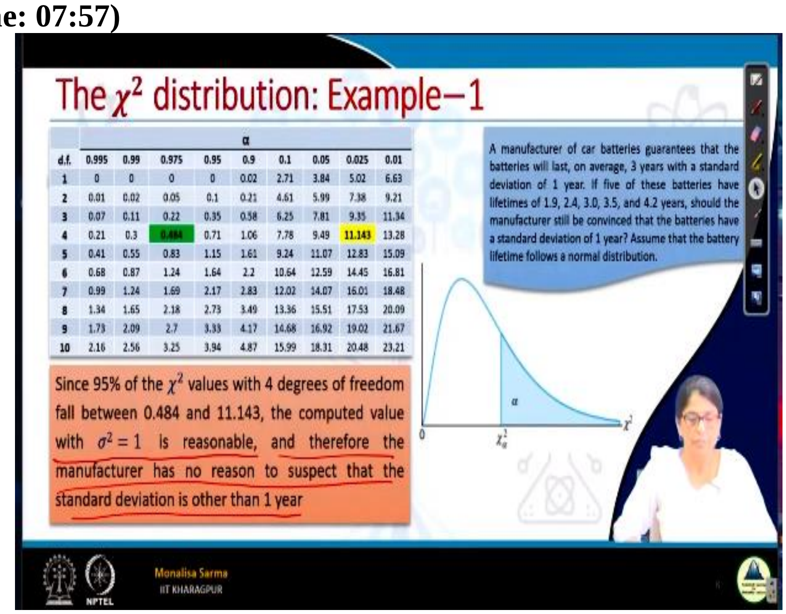
(Refer Slide Time: 07:15)



Here, I have used the other formula which I have just now written. So, we calculate the variance from the sample. Sample variance I already told you we specify it by S square. So, we found S square sigma square is given we know what is n, n is 5 so, it is n - 1. So, we found the chi square value, chi square value we got is 3.26. Now, we need to find out for what is the degree of freedom here? Degree of freedom is n - 1 that is 4.

So, we need to find out for 4 degrees of freedom what is the; acceptable range that we will for that we will consult a chi square table. So, I told you we have a chi square table so, as to consult the value.

(Refer Slide Time: 07:57)

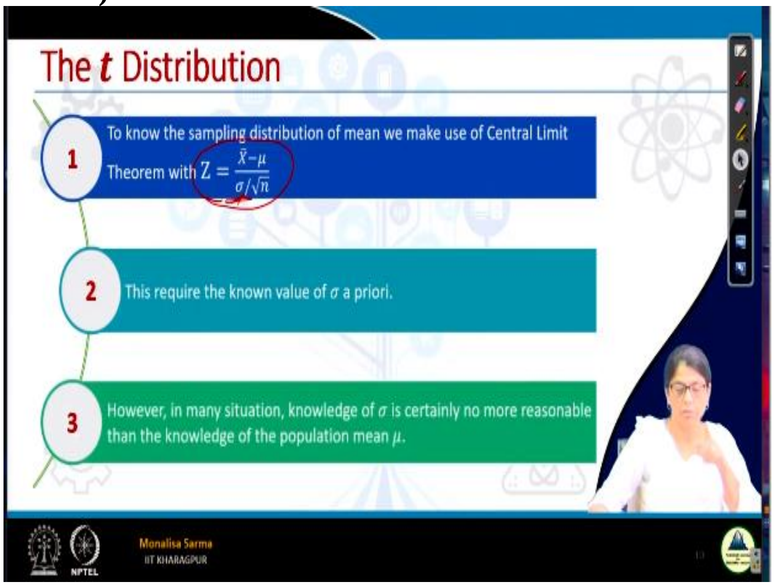


So, basically you see this is the chi square table and chi square table you will see 4 degrees of freedom. This is the area corresponding to this is the value of chi square corresponding the value area of 0.975. So, this maybe this area sorry this area if this area is 0.975 this value is 0.484. This is for 4 degree of freedom and then it is 0.025 maybe this region 0.025. This and the value of this maybe 11.143.

So, if my value lies between this and this than it is acceptable. So, what value I got definitely it lies within this range. So, what value I got? I got 3.26, chi square value 3.26. So, 3.26 lies very much within this range within this 4.484 to 11.41, it lies within this range. So, that is why since, you have written it since 95% of the chi square values with 4 degrees of freedom fall between 0.484 and 11.143, the computed value with $\sigma^2 = 1$ is reasonable.

It is a reasonable value because it follows an acceptable is not it is which is falls it in 95% of the values then therefore, the manufacturer has no reason to suspect that a standard deviation is harder than 1 year. So, what the manufacturer claim is it we can say that it is correct the he has no reason to suspect it, but always all these things when you talk about probability always all these things is a probability concept problem is the uncertainty it is already always there. It is not a never it is a definitive answer. So, that is the chi square distribution.

(Refer Slide Time: 09:57)



Now, coming to t distribution like chi square distribution and we use for sampling distribution of variance. For we have seen for sampling distribution of mean we have seen that we use Z distribution. So, symbol for sampling distribution of mean we also used t distribution. Now, under what situation we use t distribution that we will see first you see, when we have used sampling distribution of mean when we have used Z distribution, say this was our value.

This is the value, where we use to find out the Z value. Z is X bar what is X bar? X bar is the mean that we get from the sample X bar – μ σ / √ n here you see when we compute Z while explaining sampling distribution of mean I have told again and again this μ value what is our objective? From sampling distribution of mean our objective is to infer about the population mean.

So, you may think our we have to infer about the population mean, then we already have mu and what we have to infer it is already given no it is not given it is I told you it is not it is estimated it is or I can say it is an educated guess it is in maybe the producer or manufacturer or whoever the person maybe he is just claiming from the sample we have to find out whether what he is claiming is true or not, we have just taken it as estimated value or predicted value.

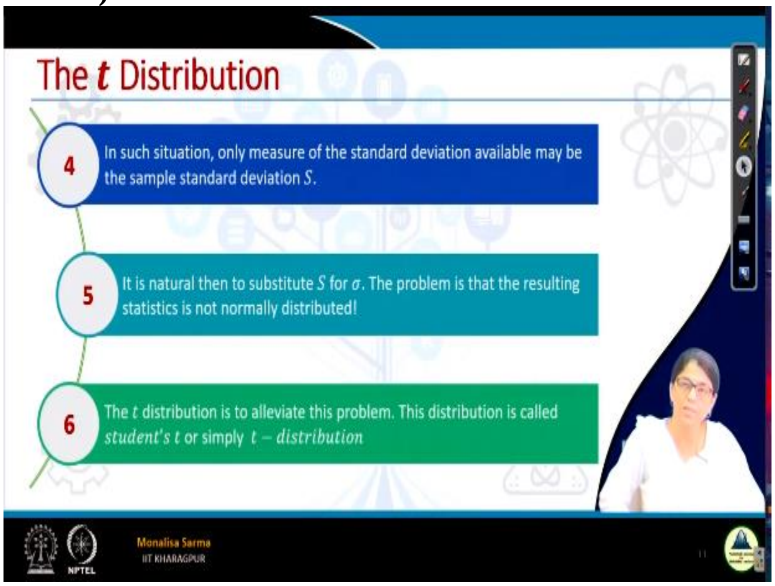
Now, the question is when we do not know the mean of the population having an idea what the variance of the population is very, very remote, we do not have mean only, how can we know the variance of the population isn't? If the population is something very known population or from past experience, then we can say variance of the population maybe so, and so, the easily in most of the cases when mean is not known having a knowledge about the variance is difficult.

In such cases when we do not know the variance of the population μ 2 have guessed it or we have predicted it that we will find it out using the sampling distribution of the mean whatever we have predicted is true or not that will find it out, but the variance what we have predicted variance what we have predicted that is usually it is very remote that you can predict a variance. In that case, when you do not want a variance we cannot use Z distribution because Z distribution the formula says directly says that we needed.

The sigma here directly says then how can we use Z distribution. So, instead of σ we will need to use S that is the standard deviation of the sample or rather we will have to use the variance of the sample. So, when we use instead of σ when we use S then it becomes then it no longer fits to a normal distribution. So the distribution that we use here is the t distribution. So, that is what, so, whatever I told here this required a known values of σ a priori.

However, in many situation knowledge of σ is certainly no more reasonable than the knowledge of the population mean μ.

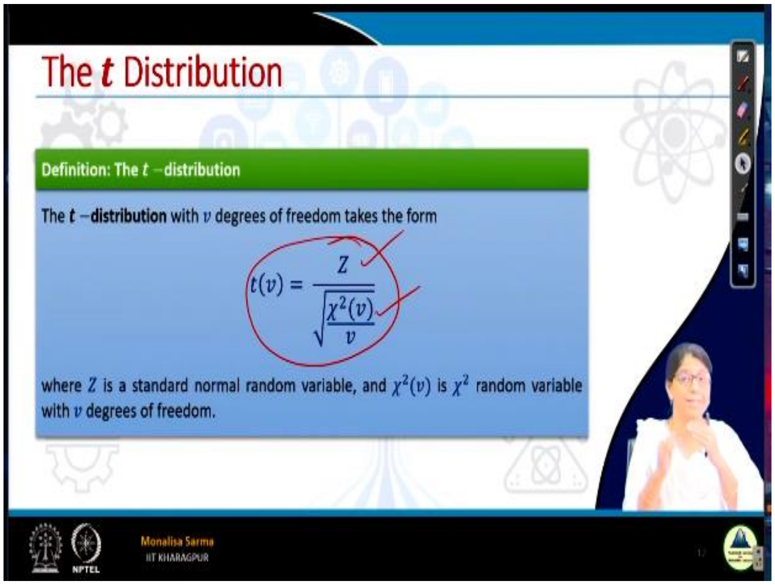
(Refer Slide Time: 12:56)



So, in such situation only measures are the standard deviation available, maybe the sample standard deviation S. So, if we substitute σ / S, the problem is that a resulting statistic, so, what is the statistics now the sample statistic is a t value. So, sample statistic is now not normally distributed, earlier my sample statistic was Z value for sampling distribution of men that was normally distributed in sampling distribution of variance my sample strategy statistics was the chi square value.

So, that chi square value was chi square distributed. So, now, this distributed empirically it is found that it fits a distribution which is called a t distribution, it is called student's t distribution or simply we can say this t distribution.

(Refer Slide Time: 13:49)

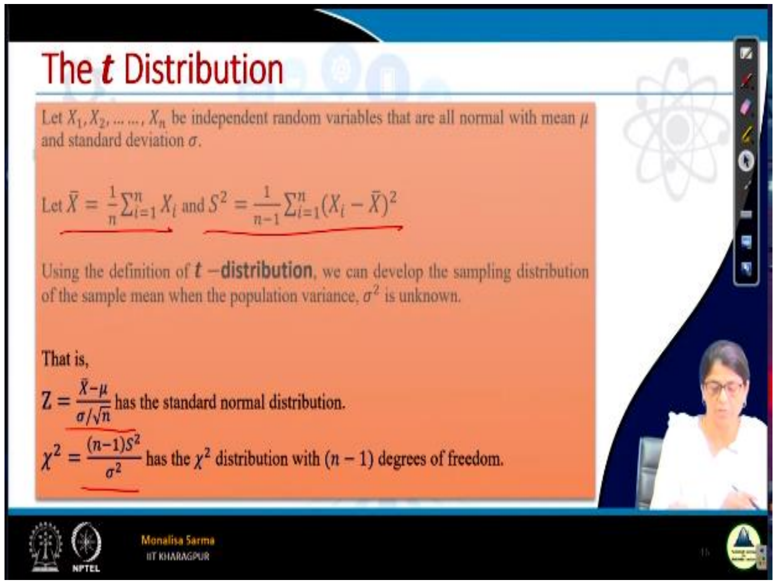


Now, what is t distribution? Any distribution means, we need to know its PDF probability distribution function, if it is continuous then call the density function whatever is we need to know that. So, t distribution, this is the PDF of t distribution. Here also there is only 1 parameter that is a degrees of freedom, t distribution with v degrees of freedom it takes Z by chi square with v degrees of freedom divided by the degrees of freedom that is a PDF of t distribution.

When if you just Google t distribution you may find a different expression, but one simplified form this is a simplified form of t distribution. Similarly, if you Google chi square distribution you may find a different form whatever form I have given that is again a simplified form different simplification different way of simplifying because chi square we use in different application in different application we will use that its different representations.

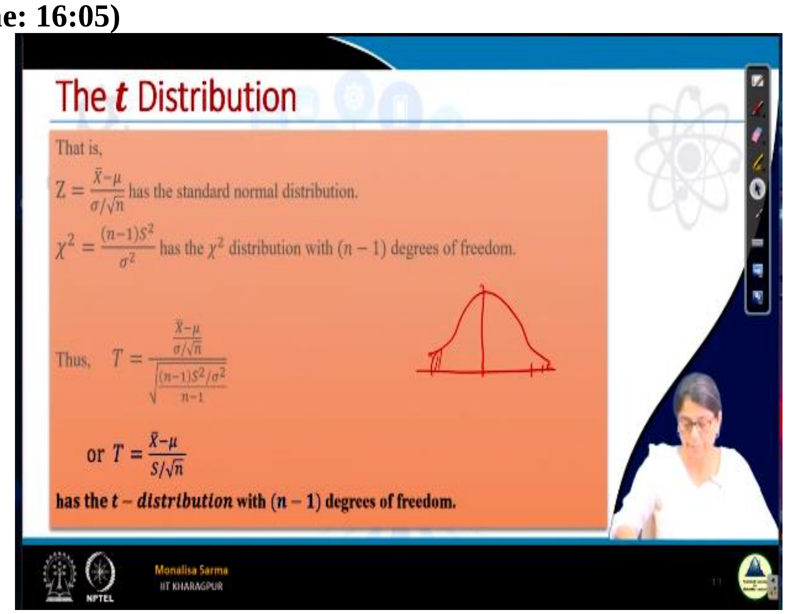
So, similarly, t distribution we will be using this representation. So, what is this Z by square root of chi square with v degrees of freedom divided by the v degrees of freedom. Now, we know what is that well, we know what is chi square with v degrees of freedom what is that if we substitute this then maybe we will get it whatever we need we need something in terms of sample parameter, a sample statistics. For in case of sample I should never talk as parameter, parameter is always for population.

(Refer Slide Time: 15:28)



So, now, see here now isn't here? So, we have Z is equals to this, chi squares this, we have seen chi square is this we have seen Z is equals to this and from the sample if X 1, X 2 are the independent random variable and with the mean mu and standard deviation, then I can find x bar is this way, and S square is just that is the sample variance this already I know. So, now, in this t distribution formula if I replace substitute the value of Z and chi square, what do I get?

(Refer Slide Time: 16:05)



I am replacing this Z by chi square value is just a simplification after upon simplification, I will get this as my this is the PDF takes this form X bar – μ/S / √n, S is the sample variance. Now, I do not need to know the population variance, even if I do not know population variance, I can still make a guess on that I can still make an inference on the population mean, but of course, I will have to first take an educated guess or I have to take some estimated value of the population mean or I have to predict some population mean.

So, this is the thing so, this is the statistics, this t value is the statistics that we will have to find out from the sample given a sample I like to find this t value what is how do I find the t value? t value is X bar – μ / S / √n and then accordingly now, the way we do for Z distribution the way we do for chi distribution, chi square distribution. Similarly, for t distribution, we have also lookup table from the lookup table we can get the value.

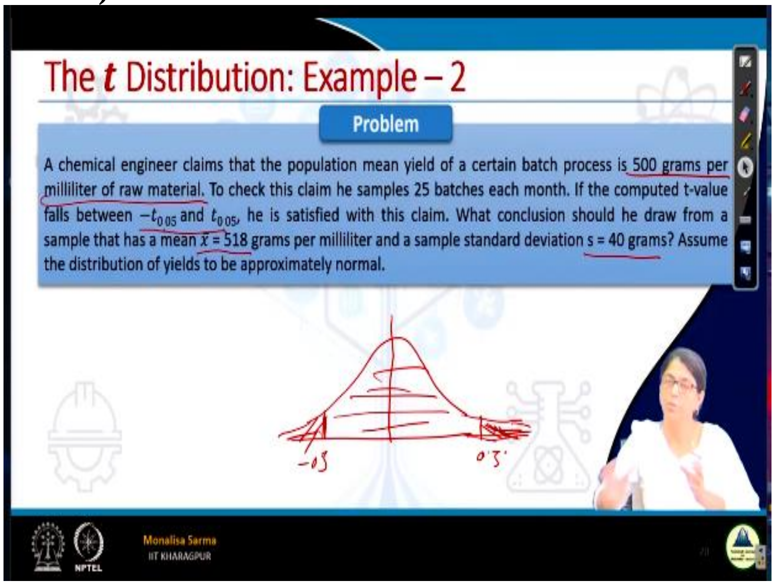
But t distribution is also one of the characteristics of t distribution this is also this distribution also very much symmetric like a normal distribution. So, since the symmetric so on the table, we have only the values for the upper tail, but right tail in the lower tail values are not given in the distribution because if we know the right tail values, we can go get the values for the low left as well because it is symmetric, is not it, like for them distribution, the whatever value we get for Z - 2, the same value will be getting for Z equals to + 2, is not it?

So because it is symmetric, similarly for t also first so that is why in fact, in Z we have seen we have given values for all the upper tail and a lower tail, but t distribution in a table only in most of the table only the upper tail is given, but we can definitely find it out from the lower from the upper tail we can find out the lower tail. So, now t distribution, some more characteristics of t distribution t distribution is like if I draw the figure for t distribution, like it is very much similar to normal distribution.

But it says a fatter tail, this tail portion is fatter compared to the normal distribution, fatter means probability here, the probability of occurrence random variable is quite more compared to the normal distribution. And one more thing, when my sample size is big, then this difference does not make this if I take whether I am taking S square or whether I am taking σ^2 it is the difference does it make have a much impact and that is why for greater sample size instead of t distribution.

I can very well use the normal distribution, the result will not be unreliable, it will be with good precision only instead of t distribution I can very well use the normal distribution for a bigger sample size.

(Refer Slide Time: 19:12)



So, now the question a chemical engineer claims that population mean yield or a certain batch is 500 grams per milliliter of raw material from per milliliter there is some raw materials are there from and that the mean yield of the population mean yield will be it is 500 grams. He claims that to check his claim, this person is suspicious of his own claim. So, he wants to check it whatever he is to check his claim he samples 25 batches each month. That means it the sample size is 25 now.

If the computed t value falls between this t of minus 0.05 to t of 0.05, he is satisfied with this claim. What conclusion should he draw from the sample that has a mean of X bar = 518 grams per milliliter and a sample standard deviation of S = 40 gram. see here population standard division he does not know he is just having a guess of mean, mean means he is guessing that it is around 500 grams, but he himself is not convinced.

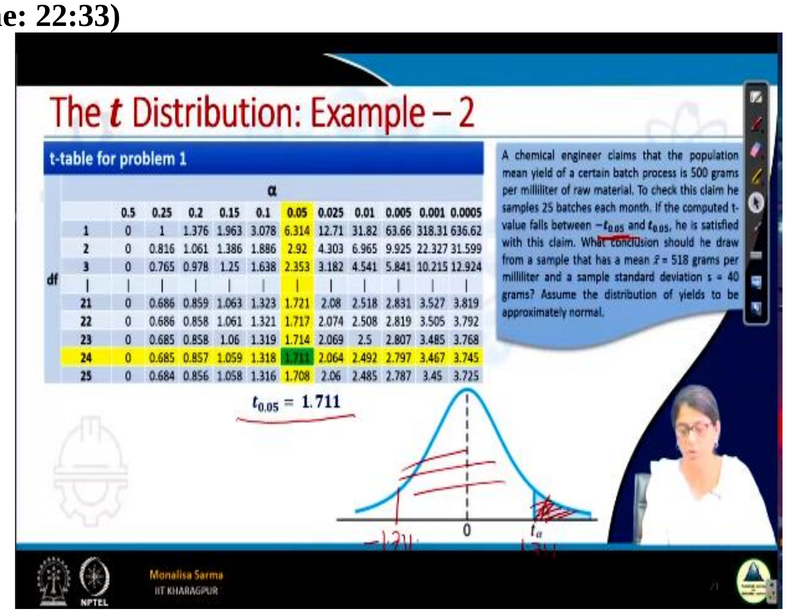
So, there so, he wants to check for more checking what he has done he has taken a sample batch of 25 that is a sample size and from there since he has to infer on the population mean, so, it will be sampling distribution of mean now sampling distribution mean we have learned to we can either use Z distribution, we can use t distribution. Now, here we cannot use Z distribution because we do not know the population standard deviation.

So, we will have to use t distribution because we do not know the population standard deviation and we have here we have the sample standard deviation definitely if it is not there also given the data we will be able to collect the sample standard deviation. So he will use calculate the t value now what he is claiming that if the computed t value falls within this range this range means what t of that let me draw the figure it will be something. So, this is I am very bad in drawing actually.

So, it goes this way, this way it goes. So, he is satisfied with this claim if it is falls in t of minus 0.05 means this portion plus 0.05 means this portion. So, he is satisfied, if it is falls within this range, that is quite reasonable that means it is falling within this range means probability of occurrence is quite high means, if this is true probability that I will be getting this 518 value is a high probability because this is very less probability is not it?

That is what his whatever explain he is taking quite justifiable experiment, because if this is true than from the sample, what he is getting probability of occurrence it should have a higher probability, if it has a very remote probability that means what he is claiming is not true. So, that is why he has taken this range.

(Refer Slide Time: 22:33)

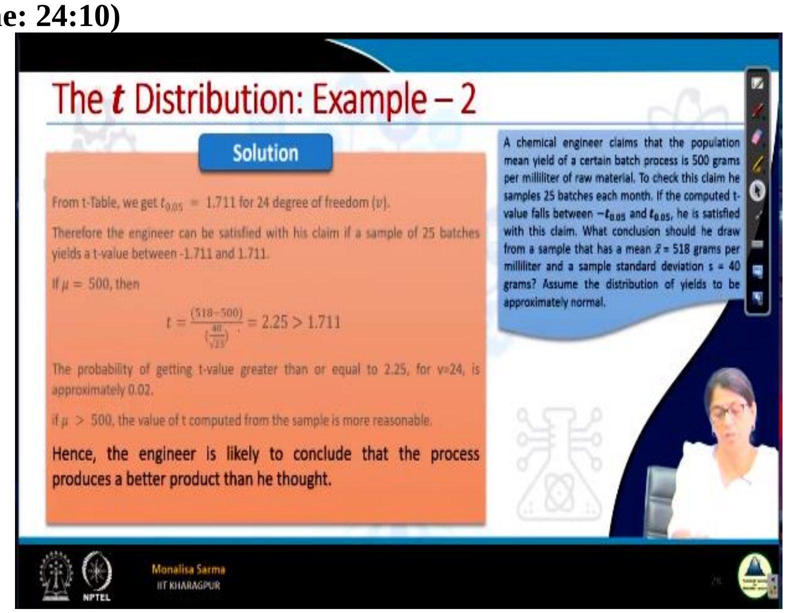


Now, let us see how he has done it. So, we have seen from the t table as I told you, in the t table only the upper tail values are given lower tail values are not given. So, in the t table, and one more difference like in Z distribution as I told you, we get the CDF cumulative distribution from minus infinity to x, but chi square is different chi square we get the value the right side of the value in fact 1 minus of the CDF that we get in chi square.

Similarly, in t distribution, t distribution alpha value is the area we got this is the area rather than this portion, we have this portion in a table. So, that is the reason t value and chi square is same we have the right side value. So, here or what is it how many degrees of freedom? Degree of freedom is 24. Because there is 25; sample size is 24 degrees of freedom for 0.05 what is the value? 1.711.

For 24 degrees of freedom t 0.05 is this value corresponding to this is 1.711. So, what will be the value corresponding to minus t of 0.5? It will be because it is symmetric it will be minus 1.711 this minus value is lower values are not given but it is symmetric. So, I can even if I know that I will be able to know this, is not it? So, if it falls it might calculate a t value falls within this region calculated t value falls within - 1.71 to + 1.71 that means, this is whatever his claim is justifiable. So, what did he get?

(Refer Slide Time: 24:10)



So, he calculated his t value, this is how we calculated his t value. Using the t value formula what is the formula for the t you remember, X bar - μS / √n this is the formula for the t value. So, he calculate the t statistics from the sample he got 518. This is the, here is the predicted this population mean 500. Then sample standard deviation 40 divided by root n that is 25 and what we got is 2.25.

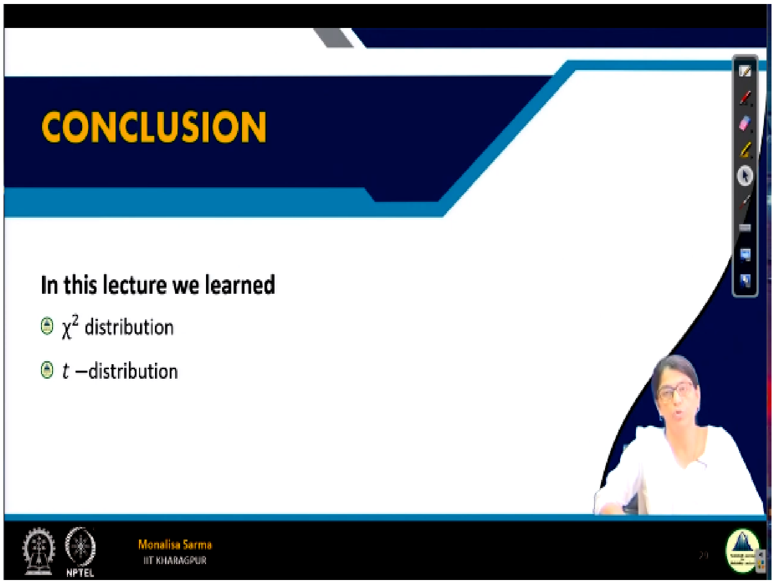
2.25 means if I draw the figure if this portion is say 1.711. And then he got it somewhere here that is 2.25, somewhere this portion this is minus 1.711. He did not get within this mean, what the value of what you got is beyond this that means probability of that occurrence is very less that is very remote probability that means, whatever he is claiming that mean yield 500 grams, that is not true.

If that is true we are from the sample not bigger sample is something which we have tested it cannot be wrong, it is we have this collected and tested this cannot be wrong, it is giving us a true picture. So, if that is true, then this should have been true, but what I got this is from this taking that is true, I got a very less probability that means what I have assume and whatever estimated the population mean 500 that is not correct.

So, now, you can see him that is not correct. So, what may be the probable values of 500? You see I got a value of 2.25 If I got if I would have got a lesser value than 1.711 lesser than under what situation I will get if this 500 would have been if this is instead of 500 if this value would have been more then I would have got a lesser value is not it? So, first thing the probability getting a t value greater than 2.254 days is approximately 0.02 very remote probability that is my first observation.

Second observation mu is better than 500 value from t sample is no more reasonable. Third, the engineer is likely to conclude that a process produces a better product than he thought actually, he was very pessimistic instead of doing this optimistic, he was telling that my thing mean yield is 500. But he did not know that his process yield is much more than that. That if he would have predicted his process yield is much more than 500, then his value would have been within this range.

(Refer Slide Time: 27:21)



So, that is all in this lecture so we have till now we have learned chi square distribution, we have learned t distribution and few more sampling distribution which will be learning and my next lecture.

(Refer Slide Time: 27:36)

Then, these are the references guys. Thank you.