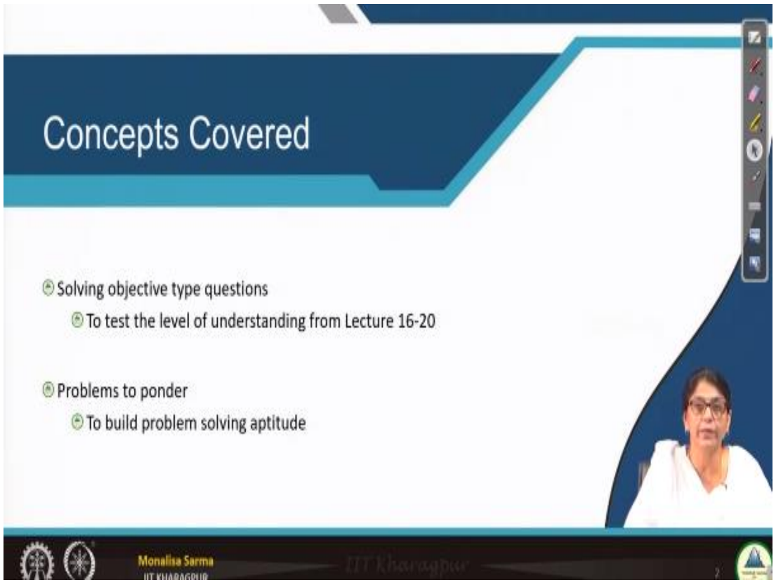
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology - Kharagpur

Lecture - 21
Tutorial on Sampling Distributions

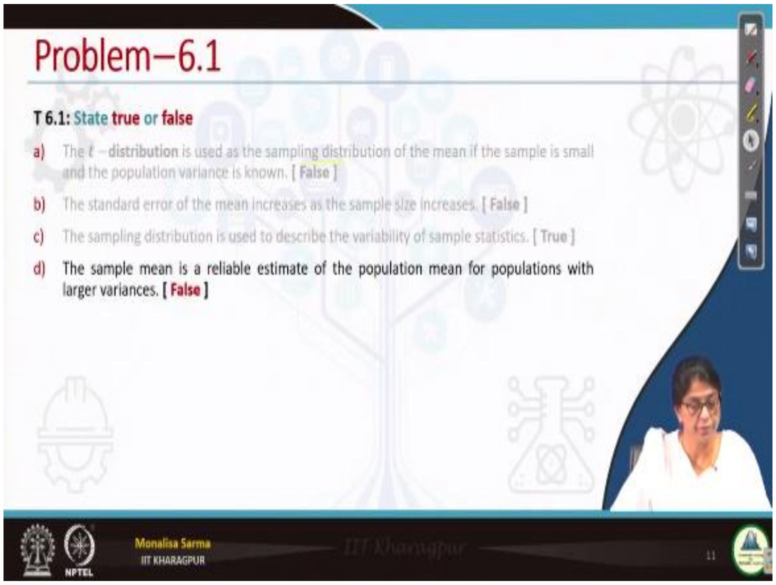
Hello everyone. So, today we are going to end the topic on sampling distribution like last lecture, I already told that we have covered almost everything of sampling distribution, I should not say everything I have covered, but then whatever is necessary at this point. So, I have covered those topics and today I will end those classes basically with a tutorial that is a tutorial on sampling distribution. And after that, I will be starting a new topic.

(Refer Slide Time: 00:54)



So, as usual in a tutorial, I always try to keep few objective questions, which will basically help you to take a quick recap of whatever we have learned. So, and then we will be doing few problems as well.

(Refer Slide Time: 01:12)



So, coming to the objective type questions, it is simple I do not have much objective question in this tutorial was there some simple true or false statement? So, first question here is the t distribution is known as the sampling distribution of the mean if the sample is small, and the population variance is known, so, what it is given? It is given that, so, let me take the pen, so, it is given that sampling distribution of the mean.

So, for sampling distribution of a mean we use 2 distributions if you can remember that one is t distribution and one is z distribution, when we use t distribution and when we use z distribution? We use z distribution when it is when the population variance is known, and we have to estimate about the mean of the population and population variance is known and we also have an estimated value of the population mean then we use z distribution on the contrary when the population variance is not known, which is actually more practical.

Knowing the population variance is not very much possible they actually so, when the population standard deviation whether population variance is not known, then we use t distribution. So, this statement is false. Second statement the standard error of the mean decreases as the sample size increases. So, while in a sampling and what is the central limit theorem do you remember when we have done central limit theorem? So, what was that central limit theorem what does it say?

It says the mean of the sampling distribution is the mean of the population and the variance of the sampling distribution is given by $\sigma^2 / \sqrt{n}$ right this was the variance of the sampling distribution of mean. So, what does this variance indicate? Variance indicates if this is the variance sorry it is variance is σ^2 / n not $\sqrt{n}$. So, if I this is whole square is variance and if I removed a square and this is the standard deviation, what the standard deviation means?

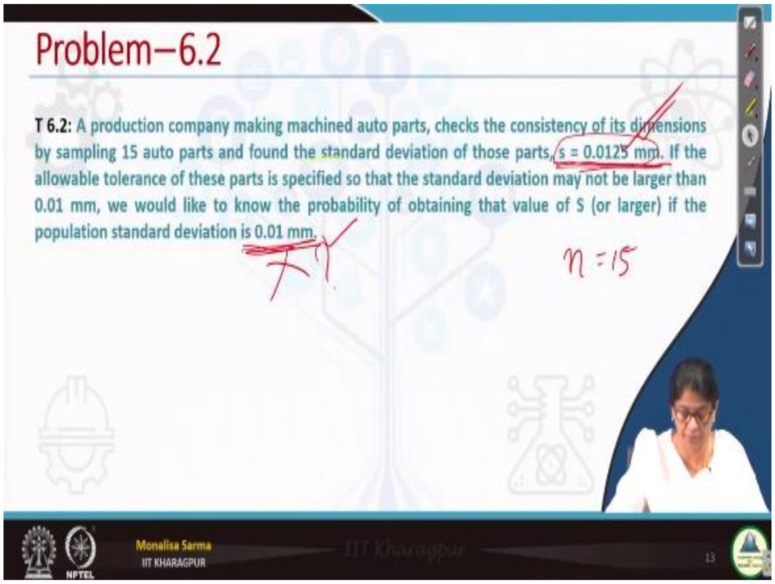
Here in this case, standard deviation is basically the standard error of the mean, means how much the mean varies among themselves if you take mean of different samples, how much the mean of different samples varies among themselves that is, we call it a standard error among the means. So, the standard error of the mean increases the sample size increases. So, what happens this is the standard error. So, if the sample size increases that means, if n increases then what happens the denominator is higher than the result will be lower.

So, the standard of the error of the mean it will decrease when the sample size will increase. So, this is again a false statement. Third, the sampling distribution is used to describe the variability of sample statistics. That is, of course, that is the reason why we use the sampling distribution. So, represent the variability of the sample statistics different sample will have if we take from the same population if we take different sample all the sample will have the same value for the statistic it is quite unlikely.

So, the sampling distribution is used to describe the variability of the sample statistics this is a correct statement. Then, the sample mean is a reliable estimate of the population mean for population with larger variances. Same similar to question number b, the sample mean is a reliable estimate of the population mean for population with larger variance. So, for population with again same thing is what will be the standard error? The standard error will be $\sigma / \sqrt{n}$.

So, if this factor is more, definitely the standard error will be more if the standard error is more than on an every sample mean is equals to the population mean that is not a very much a reliable estimate. So, sample mean is a reliable estimate of the population mean for population has larger variance that statement is false. So, that is all from the objective side. Then we will be solving few problems.

(Refer Slide Time: 05:41)



So, first problem you will see a production company making machine auto parts, checks the consistency of its dimensions by sampling 15 auto parts and found a standard deviation of those parts. So, standard deviation of those part is given, how many parts we have taken a sample sizes that means my n = 15, sample size I have taken 15 and of this 15 the standard deviation is given when it is as that means it is talking about the sample.

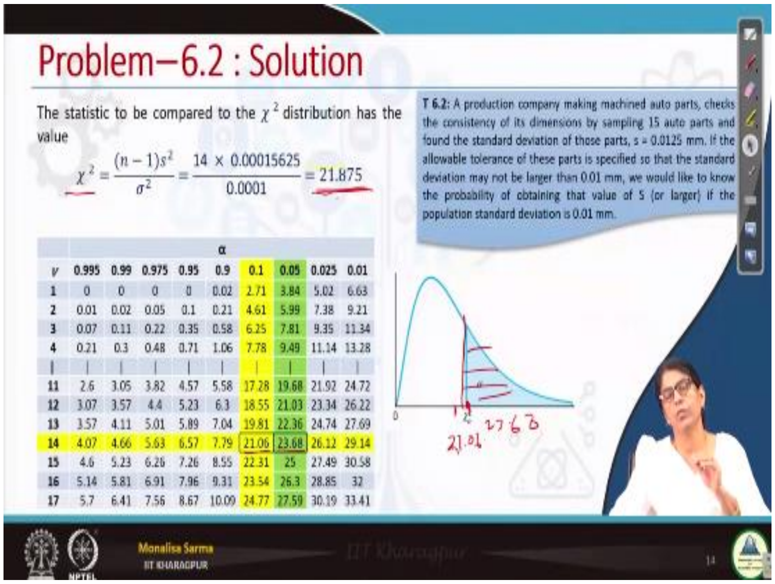
If it is population, I would have written as σ. So, then the standard deviation of the sample is given this value 0.0125mm if the allowable tolerance of this part is specified, so, that the standard deviation may not be larger than 0.01mm. So, what we want is that our requirement is that the standard division should not be larger than 0.01mm we would like to know the probability of obtaining that value of S or larger if the population standard division is 0.01mm.

So, given if the population standard deviation is 0.01mm, if this is the population standard deviation, what is the probability that if we take a sample from the sample standard deviation will be this value that is what we need to find out. If you go through the question, you will understand it is basically this asking that if the population standard deviation is this what is the probability from a sample we will get this much standard division. So, it is talking about variance.

So, that means, we have to infer about the population variance, because this is something what we got from the sample this cannot be wrong, we have collected the sample we have calculated it and we have got this value. So, this value cannot be wrong, but we are not sure about this value. This value may be wrong this value may be correct. We do not know about this, we have just estimated this. Now, from this value, we will find out if that means what we will assume that this is correct, if this is correct, what is the probability of getting this.

If the probability of getting this in the probability of getting this is very less that means our assumption is wrong means we are trying to prove it the other way basically. So, for this remembers what we have done, this is the inference of a population variance that means, we will be using chi square distribution. So, what was the statistic? n - 1 s^2 / σ^2.

(Refer Slide Time: 08:08)



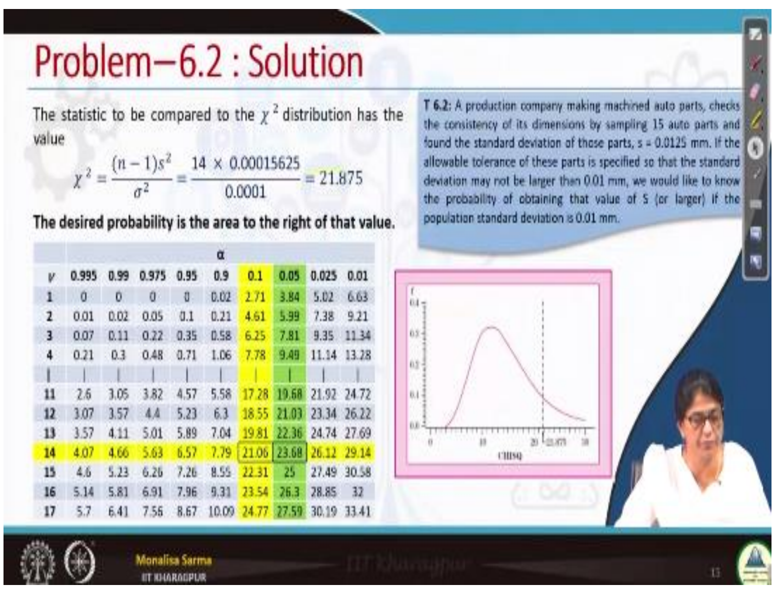
So, the statistics will compute is chi square n - 1 s^2 σ^2. So, this is n – 1 14 σ s^2 s^2 value is given σ^2 value so, we got is 21.875. Now, we will find out what is the probability of getting this value 21.875 remember chi square always the value is towards the unlike this normal distribution what we get? We get cumulative distributed value, but here we get is the area towards the right.

So, if suppose, this is my 21.875 this value this portion corresponds to this 21.875 then what is this area basically, and what is the degrees of freedom in chi square distribution parameters is degrees of freedom. So, what is my degree of freedom? Degree of freedom is 14. So, corresponding to 14 I am trying to find out for what probability I will be getting this value probability means this area here this areas are given for degrees of freedom from 14 for what probability I will be getting this area?

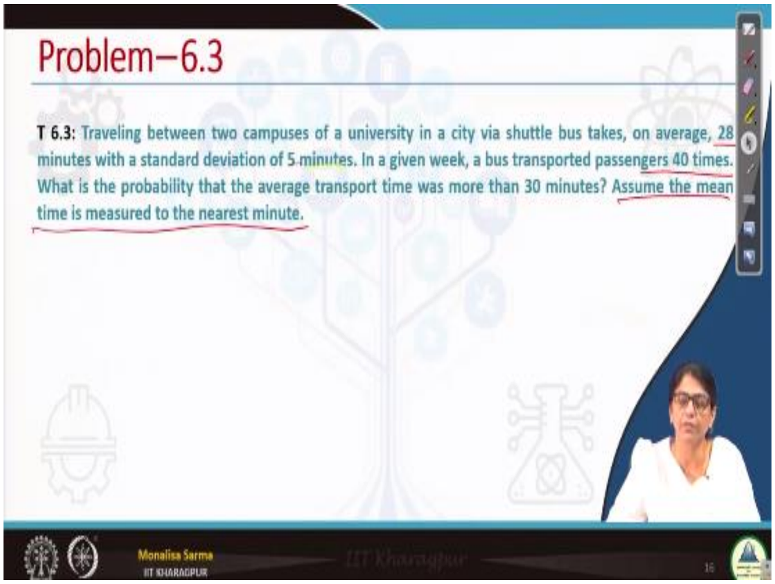
But that area is not available in the table what is available here? You see here I am getting a value of 21.06 and after that, it is 23.68 that value is not there. That means from here we can find it out if this is 21.06 that means this one let me first proceed and also it properly. So, if this value is suppose this value is 21.06 and this value is suppose 23.68. So, my value will be somewhere here, is not it? Somewhere in between this from 21 to 23.

So, that means, corresponding to 23 what is my area what is my probability is 0.05 corresponding to 21.06 my property is 0.1 that means, my probability will be when between 0.1 and 0.05. So, that is the probability it is smaller than 0.01 but greater than 0.05 that is my probability.

(Refer Slide Time: 10:25)



(Refer Slide Time: 10:28)



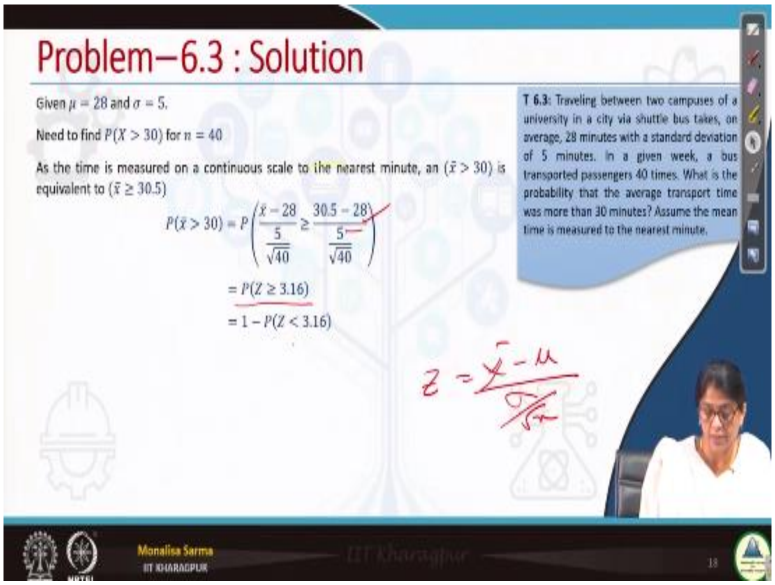
So, next question travelling between 2 campus of a university in the city via shuttle bus takes on average 28 minutes with a standard deviation of 5 minutes, this is something it is this estimated or we can say from our previous knowledge from a past knowledge we can say travelling between 2 campus it takes one and a average 28 minutes with a standard deviation of 5 minutes. In a given week, the bus transported passengers in 40 times that means my n is 40 here. What is the probability with that average transport time was more than 30 minutes.

In a week, the bus transported passenger, 40 times bus as shuttle it to and from and then average transport time was more than 30 minutes. As in the mean time is measured to the nearest minute. This say assume the mean time is measured to the nearest minute, that means measure to the nearest minute means when I have means specified is more than 30 minutes means if any of my journey, if I got is 30.1 then I will write it 30 only 30.2 I will still at 30, 30.3 I will still write it 30 till 30.5 I will write it 30.

Above 30.5 I will write it 31 minutes. So, it is nearest to the nearest minute means still 30.5 I will write it 30 minutes. So, essentially, when it is asking what is the probability that every transport time has more than 30 minutes. So, essentially I need to find out what is the average transport time what is the probability that the average transport time is more than 30.5? Because 30.5 also we will write it as 30 minutes because we are measuring it to the nearest minute.

Now what we have to find out here? Let us move this to the nearest minute the surrounding and follows the different and simple issue. Now coming to the main question what we need to find out here, but it is given so an average is 28 minutes standard deviation is 5 minutes. So, it is we have to infer about the mean 28 is the mean and 5 is the standard deviation. So, we have taken a sample, sample is doing to and from 40 times and from there we got it on an average 30 minutes.

(Refer Slide Time: 12:50)



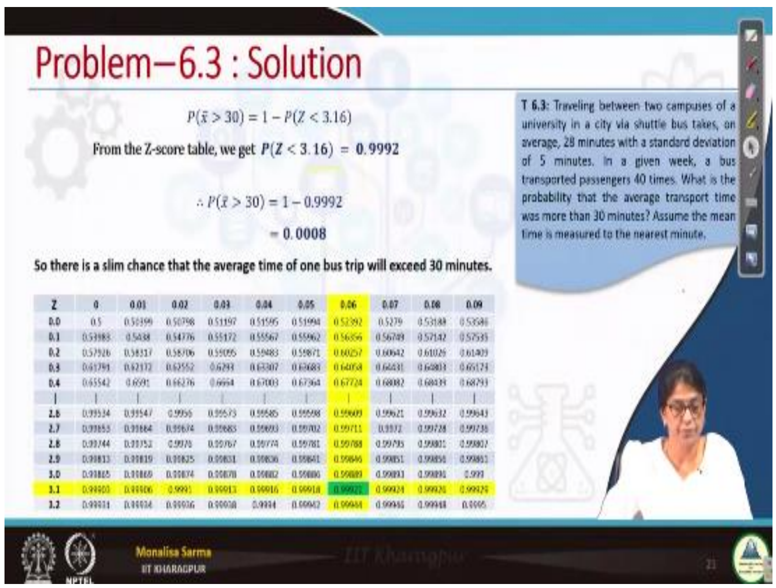
So, what we have to find? μ is 28 given σ is 5 given means this is estimated this is if this is known only why at all we are doing all this is not it? These values are not known this or we are just what to say predicting it or estimating it. I do not want to use the word hypothesize. Basically we are hypothesizing that term hypothesis this value, I will be using this word after this lecture on what basically. So, this is we are just let me use the term estimate. Now we are just estimating or we are just guessing it of that estimation.

Now from the sample whatever we got, I am repeating this again and again from the sample what we get that is the actual value. So, from the sample basically, we need to find out assuming this is correct, whatever we got from the sample, what is the probability of that? If that probability is very less that means our assumption is wrong, same thing in all the questions. So now, since here, we have to find out the mean, we have to infer about the population mean, and our standard deviation of the population is given.

So, it is a straightforward question, we will just have to use z distribution. So, probability of this is x bar greater than 30, mean is x bar greater than 30.5. When I am telling actually greater than 30, that means actually, I am actually implying x bar greater than 30.5, because I am nearing it to the nearest minute. I am measuring it to the nearest minute. So, what is this? What is the formula for z bar for sorry, so z is equals to basically x bar – μ / σ √ n.

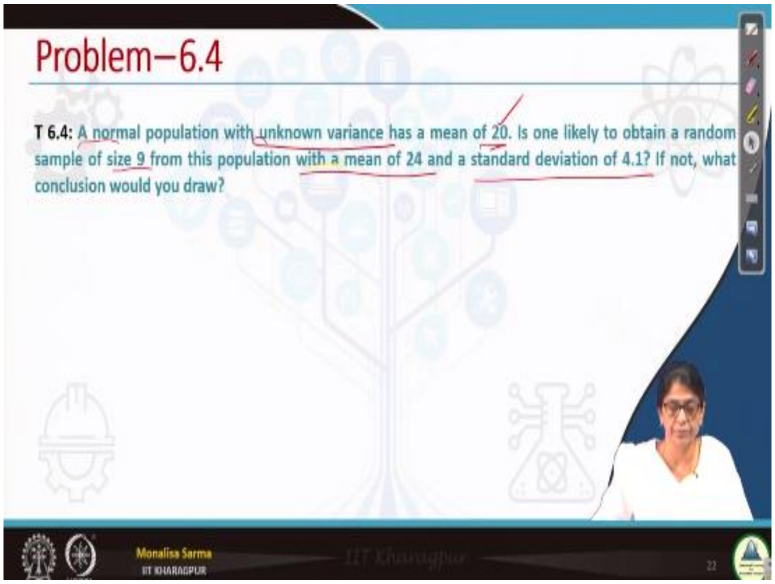
So, my x bar is 30.5, μ is 28. What is my σ √ n 5 / 40 the root of a 40. So, this is what probability of Z greater than equals to 3.16. z distribution always I get it from - ∞ to that value. Whereas for chi distribution, t distribution we get the opposite side from that value, value to the infinite. So here, z equals to, so I will find it 1 - P Z at least, probability that Z is less than 3.16. So this value we can get it from the table this I have not shown here the table I have shown it many times.

(Refer Slide Time: 15:08)



So, from the value if we can calculate it I have here I have given a table I forgot it. So, my value compared to 3.16 see how do we remember 3.1 is this then 6 is this so 3.16 this is the value this is the probability the Z less than 3.16, how is the z distribution remember the figure if I draw it this way, so, this is 0 the side is minus this side is plus so, it is 3.16 maybe somewhere in here 3.16. So, this whole area this whole area is this is 0.9992 and what I want is greater than this, I do not want less than this, I want greater than this that means, I want this portion. So, what is this portion? 1 minus of this will be given this portion.

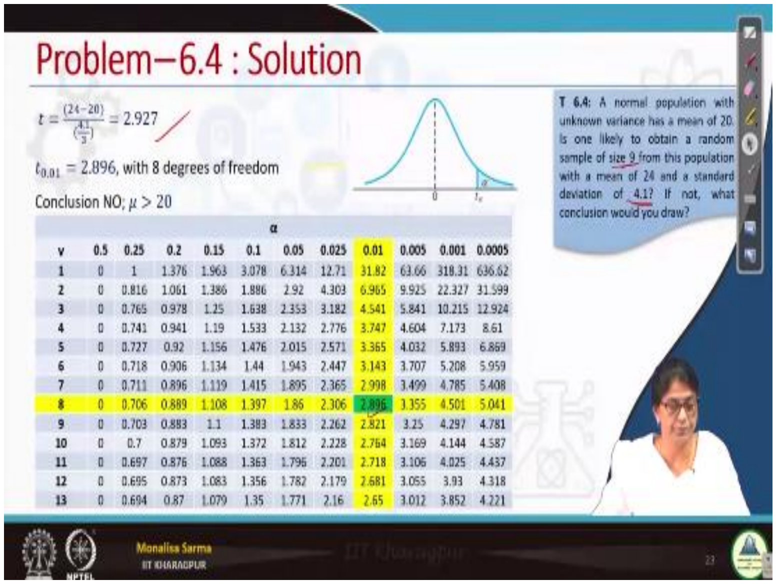
(Refer Slide Time: 16:01)



So, next question in normal population with unknown variants has a mean of 20. Is one likely to obtain a random sample of size 9 from this population with a mean of 24 and a standard deviation of 4.1. If not, what conclusion would you draw? So, see here it is given unknown variance has a mean of 20. If this is the case, it is a population is normal. Always remember I am repeating this again chi square population, a chi square distribution and t distribution and f distribution we can use only if the parent population is normal.

So, it is mentioned here the normal population with unknown variance, variance is not known. That means somehow we can we are guessing the value mean is 20. If says this is the case, is one likely to obtain a random sample of size 9 with a mean of 24 standard deviation of 4.1. If not, what conclusion would you draw? Same we will use here, we will use t distribution because variance is not known and we have to infer about the population mean is not it? Because it is asking is it likely to obtain a random so, for with a mean of 24.

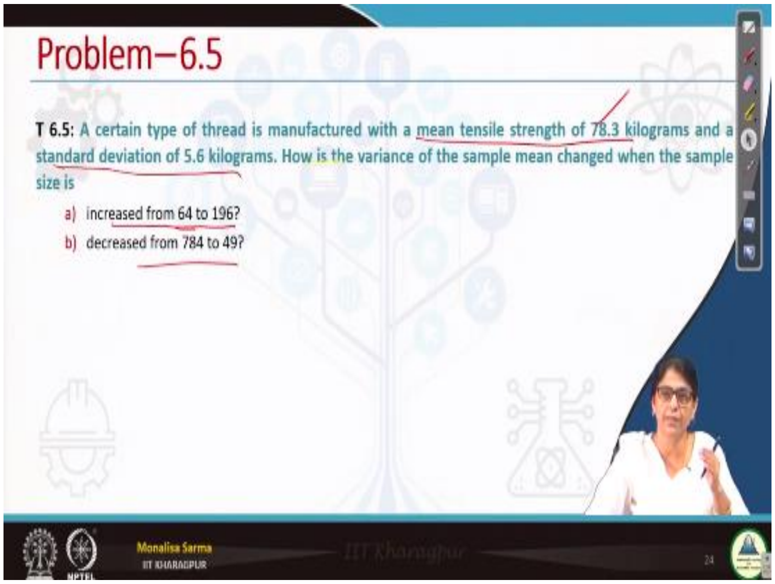
(Refer Slide Time: 17:17)



So, we are using t distribution. So, using the value for t distribution, what is the thing representation, I mean the formula for t distribution what the t statistics what this gives t = x bar - μ / s / √ n. So, what is my s here? s is 4.1 √ n is what is the size is 9? So, √ of 9 is 3 so, 4.1 / 3 24 - 20 this is the value 2.927 so, 2.927 with 8 degrees of freedom because it is sample size is 9, 8 degrees of freedom 2.927, 8 is not there in a table, so, 2.896 closer value is there. So, it will be above this.

So, the probability of this will be lesser than one person less than 0.01, it will be between 0.01 and 0.005 less than 0.01. So, that means it is a very less probability. So, that means the whatever we have estimated that is this our estimation this mean of 20. So this our estimation is not correct. So, conclusion no μ is greater than 20 μ has to be greater than 20 it cannot be less than equal to 20.

(Refer Slide Time: 18:41)

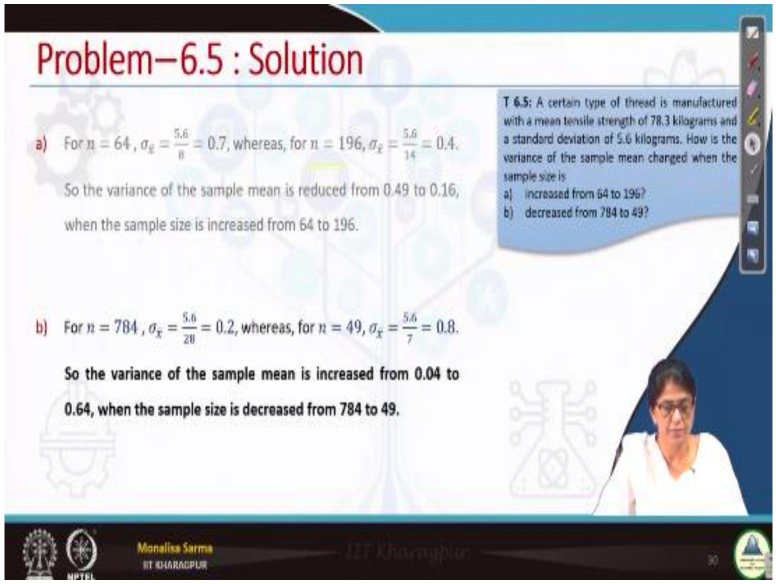


So, next question is certain type of thread is manufactured with a mean tensile strength of 78.3 kilograms and a standard deviation of 5.6 kilograms, mean is given this is the mean and this is the standard division how is the variance of the sample mean changed when the sample size is increased from 64 to 196 or decrease from this I will just one second one will also be the same. So, it is a very simple question what it is given a certain type of thread is manufacture mean tensile strength is 78.3 and a standard deviation is 5.6 kilogram

So, that means, for this for sampling distribution of mean if we draw a sampling distribution of mean what will have? Mean of the sampling distribution of mean will be this 78.3 and what will be this variance? Variance will be 5.6 divided by the sample size, is not it? So, what it is given? So, how is the variance of the sample mean when a sample sizes so, sample size initially sample size is 64 then we have seen the sample size to 196. If we do that, how the variance will change.

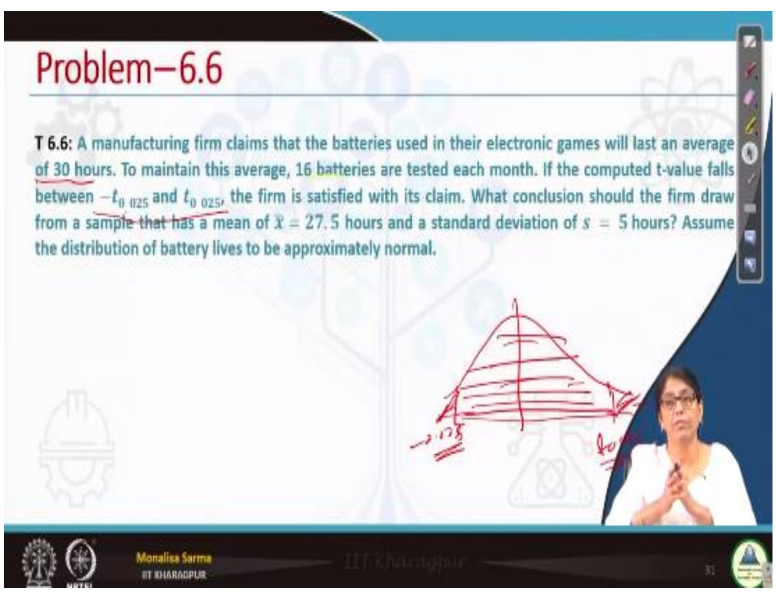
Remember we have already done in the objective type my second question is the same thing actually, when my sample size is increased, whatever as my variance decrease my sample sizes decrease my variance increase. So, here initially sample size is this then so, see what happens.

(Refer Slide Time: 20:09)



So, this is how we find out the standard deviation that is the standard error just the standard deviation of the population divided by the what is a sample size root of our sample size that is root of 64 is 8 I got 8 is 0.7 whereas, if n = 198 if I got variance I got 0.4. See, when the sample size increased my variance our standard error decreased that is it obviously, but we have seen in objective type question similarly, for the next question decrease from 784 to 49. So, variance will increase. So, I am not discuss this it is a solution is given there you can see.

(Refer Slide Time: 20:50)



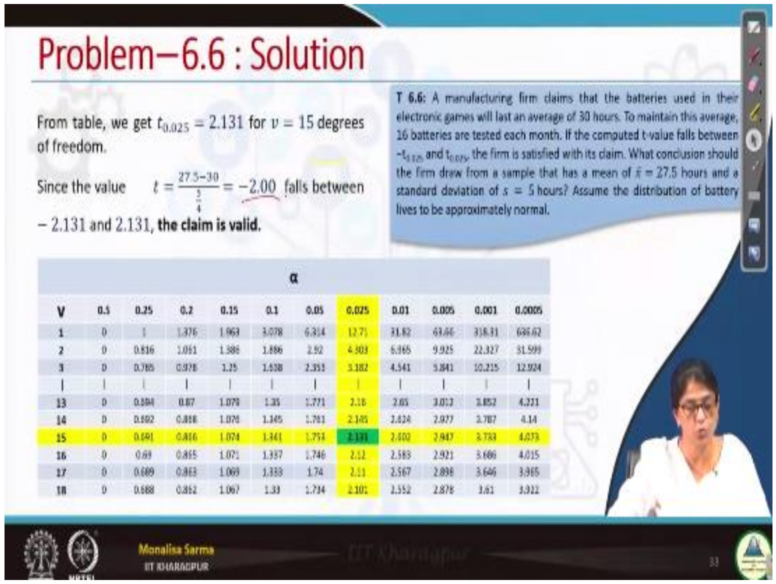
So, next question a manufacturing firm claims that the batteries used in the electronic games will last an average of 30 hours. So, this is a claim. So, they are claiming that their mean lasting time that the mean time that the game will last is 30 hours to maintain this average 16 that is tested each month. So, whether this average is maintained at a manufacturing firm it is claiming there, but before it goes to the market, it wants to check again and again. So, what is so, it tries to check it taking 16 batteries each month.

So, how it does? If the computed t value falls between this t value is also similar to t distribution is also similar to normal distribution just to the fatter tail is not it? So, the computed t value falls between t value of 0.025 means 0.025 maybe says this is t value of 0.025 maybe this portion and t value of -0.025 maybe this portion. So, if my computed value falls in this range, that means falling in this range means, from the sample what I am getting assuming this is true from a sample what I am getting is be significant probability.

Probability is quite good, it is not a very less probability, because if it falls in this region, it falls in this region means very less probability. So, the manufacturer or the firm is satisfied if it falls within this range that means, if it is falls within this 95% of this area, it is satisfied with this claim. So, what we need to find out from the question, only it is very clear, that means we need to find out a t value first from the sample. And this our sample t values would fall within this range.

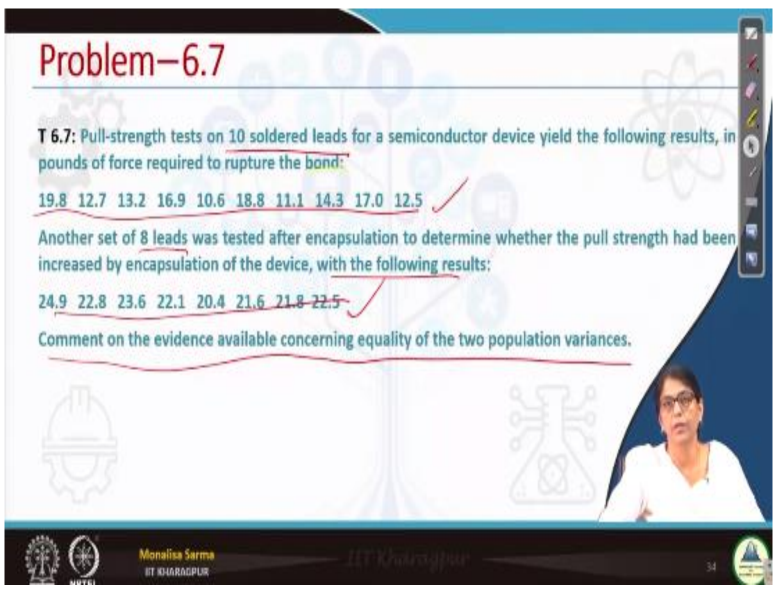
So, we will have to find out what is the t value corresponding to this area, we will have to find out what is the t value corresponding to this area, what is the t value corresponding to this area? How do we find out? This we will find out we will be able to find out from the table using what is the degrees of freedom 15 degrees of freedom. For 15 degrees of freedom what is the t value corresponding to 0.025 as I told you t distribution as symmetry and whatever we get it if we get value we have x then this will be –x is not it?

(Refer Slide Time: 23:17)



So, for 15 degrees of freedom t value of 0.025 is 2.131 that means from the sample statistics from the sample sorry from the sample the sample statistics that I will calculate if my calculated value falls within plus minus 2.131 then the manufacturing firm is satisfied. So, what is my t value t value formula you know how to calculate the t value it found is -2. So, it falls between the - 2.31 and 2.31. So, claim is valid.

(Refer Slide Time: 23:58)



So, we have seen question on the sampling distribution of mean, sampling distribution of mean we have seen using both z distribution and t distribution then we have seen the sampling distribution of variance that means to infer about the population variance we have seen some problems on that we have also seen problems on how the standard error varies with the as regard to the sample size. Now, what is remaining whatever we have learned of sampling distribution what is remaining?

This f distribution is remaining, f distribution and sampling distribution of proportion, sampling distribution of proportion. I do not think I have a problem here already we have discussed when we discuss 1 or 2 problems while we have discussed sampling distribution of proportion. So, now we will see f distribution. When we use f distribution, remember when we want to compare the variance of 2 different population as I had mentioned you earlier it is usually we used in the food and beverage industry.

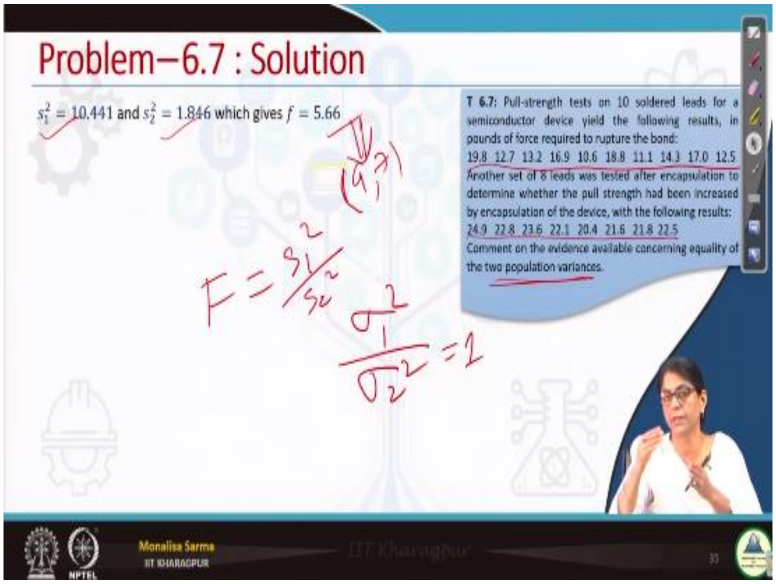
Where food and beverage industry where there are many players for the same type of product like cold drinks, there are many products for fruit juice for say mango juice, mango juice there are different players in the market. So, this is one of for this sort of products, usually this variance and comparing the variance of 2 different populations or different chemical factories comparing the variance of 2 different products, it is very important.

So, f distribution you will see a different example here, pull strength test on 10 soldered leads for a semiconductor device is the following results in pounds of force required to capture the bond, how much strength we have to give basically to rupture the bond. So, these are the values given for sample size is 10 here for f distribution, it is not required the sample size for both the population has to be same it is not required. So, as we have already seen that n 1 and n 2 we have used 2 different terms remember.

So, here another set of 8 leads was tested after encapsulation, we have encapsulated the thing with some material maybe and then to determine whether the pull strength had been increased by encapsulation of the device, we have encapsulate the soldier lead by some sort of material after that, do we need to what to say put more stress it is expected at we need to put more stress to rupture the bond. So, increased with the following results we got the following results. Comment on the evidence available concerning equality of the 2 population variances.

But the manufacturer whatever they are claiming that both the population variances are same both the population variances are same. So, we have to comment on this whether the population variances same are not how we will come in? We will find out f value if the f value from the because these are the statistics. This is one sample, this is one sample from this 2 sample we will be able to calculate the f value and if the calculate the f value has a significant probability. Then equal the population variances are in the 2 population variances are equal, we will agree to that claim.

(Refer Slide Time: 27:23)



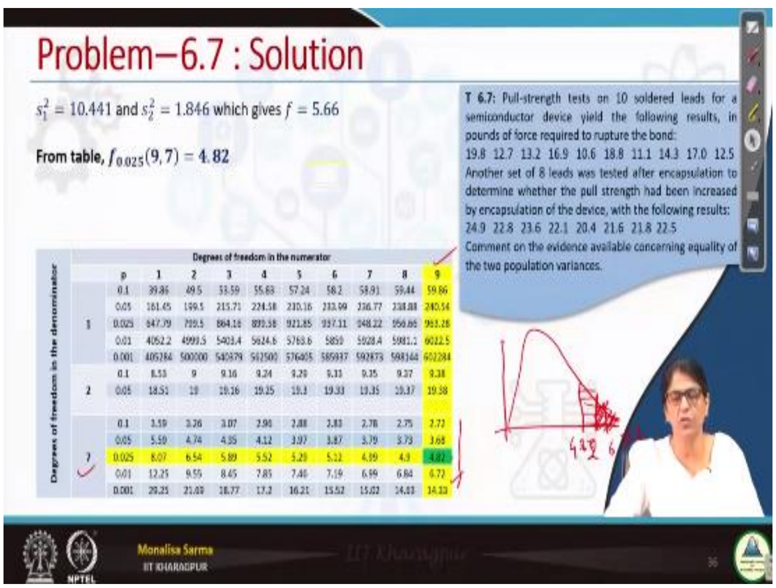
So, what we got so, from this set of data you will calculate the S ^{2} that is the standard variance I have not shown you how to you know how to calculate the variance with the same method calculating variance or we know that from class 9 onwards. So, calculated the variance from this value, calculate out the variance from this value, so one is this S 1 ^{2} S 2 ^{2}, then we will find out the f value. What it is mentioned?

We have to assume that the 2 population variances are equal if the 2 population variances are equal. So, what will be the f statistics value will be f statistic was S 1 ^{2} σ 2 ^{2} / S 2 ^{2} σ 1 ^{2} is not it? So, now σ 1 ^{2} σ 2 ^{2} becomes one in both are equal. So, if that is one, then σ 1 ^{2} / σ 2 ^{2} this is equals to 1 because if both the population variance are assumed to be equal, under this situation, what is my f? f turns out to be S 1 ^{2} / S 2 ^{2}.

Now, whether I should which one I will consider as S 1 which will consider S 2 as a general convention. The higher value I will consider S 1 is a general convention based on that also the tables are given in the standard textbooks. So, S 1 ^{2} S 2 ^{2}, which indeed I got this is my f value, f = 5.66. So, now, this is my f value what is the 2 degrees of freedom? One is the numerator degrees of freedom and the denominator degrees of freedom, what is the numerator degree of freedom, so, that is the total 10 size sample size is 10.

So, that is 9 and this is 8, that is 7. So, my degree of freedom is 9 and 7. For this f value, and this is my degrees of freedom, what is the probability corresponding to this that we will see, but again mind it for f value has very limited probability in that standard books. So, sometimes we will have to either we have to interpolate or we will have to find a value which like closely in this and then we can basically give a interval like we have done in t distribution or x chi square distribution in previous one for example.

(Refer Slide Time: 29:37)

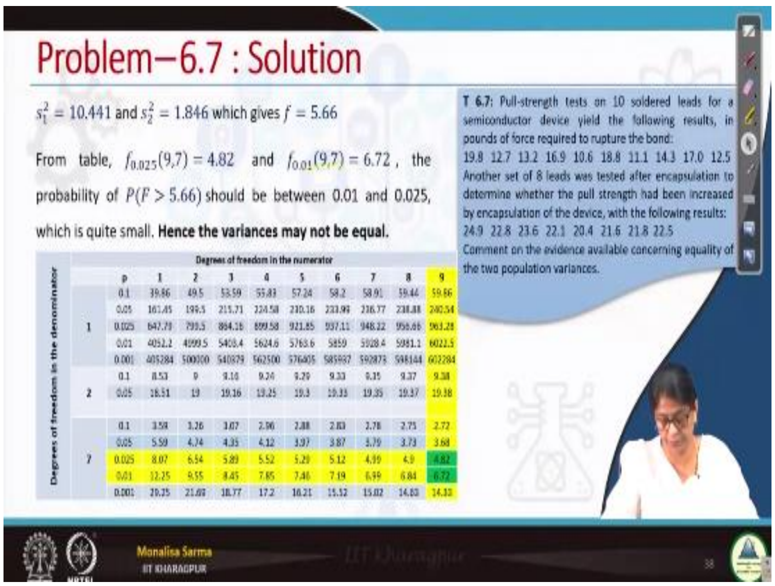


So, here what it is given, we got the value of 5.66 so, and degrees of freedom is 9 and 7. So, this is the numerator degrees of freedom here this is 9, this is 7, we in the table we have only this 5 probabilities given 0.1, 0.05, 0.025, 0.01, 0.001 and that means only these are the values which are given here which value closely resembles is 5 we have our value is 5.66. So, we do not have this value in the table what we have is 4.82 we have and 4.82 corresponding to area 0.025.

And we have 6.72 corresponding to area 0.01 that means, our value will lie within this range. So, if f distribution is something like that, so, this is one value this is another one value this is one value that is 4.82 and this is 6.72 it will so, 4.82 corresponding to 4.82 what is the area it is 0.025 that means, this area to the right and what is the area corresponding to 6.72 is this portion. So, my value is in between 4.82 and 6.72. So, my value may be somewhere in this portion. So, my area will be this.

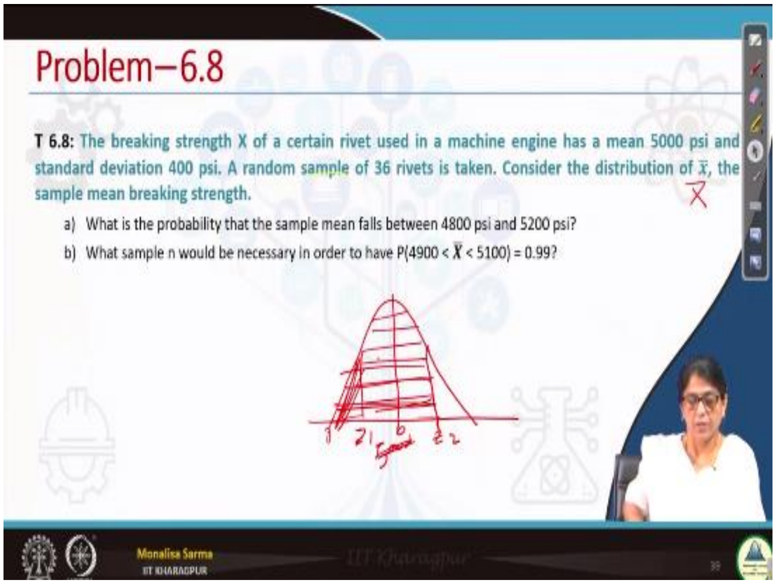
So, it will be in between 0.025 and 0.01 means it will be greater than 2% but lesser than 1% I mean sorry it will be lesser than 2% but greater than 1%.

(Refer Slide Time: 31:32)



So, probability of that f is greater than 5.66 so should be between this value. So, which is quite small this is a very less probability and so, it is less than 2%, less than 2% is a very less probability. So, hence the variation variance whatever it is assumed that the 2 population variances are equal that may not be true. In probability and statistics we cannot tell deterministic it is not true, this is correct, this is false, this is right nothing we can tell deterministic because it is a science of uncertainty probability. So, everyone see the statements where variance may not be equal.

(Refer Slide Time: 32:16)



There is the last question. The breaking strength X of a certain rivet used in a machine engine has a mean of 5000 psi and a standard deviation of 400 psi random sample of 36 rivets is taken consider the distribution of x bar the sample mean breaking strength, what is the probability that sample mean falls between 4800 psi and 5200 psi here it is one problem on error in the equation. If you can notice then it is very good if you till now if you did not notice I am mentioning it to you.

Because this is always it is a general convention random variable we always write it using capital letters. And a value that a random variable takes we write it is in small letters. So, when we talk about the random variable has a particular distribution that means we are talking about a random variable we are not talking about a value then always it will be what it will be capital letter. So, consider a distribution it is not small it is X bar. So, what is the probability that sample means falls between so it is asking that a sample between falls within this.

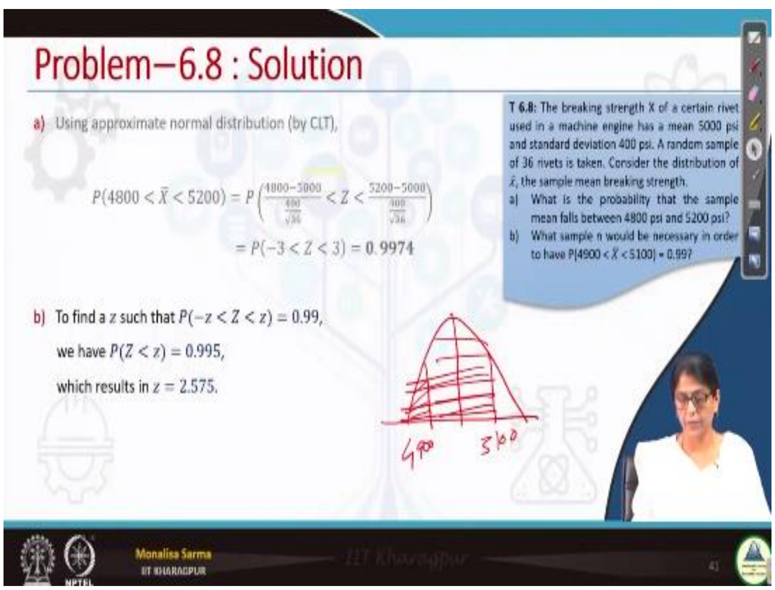
So, say what it is given let us let us just draw it when we draw things becomes very easy, X has a mean of this 500 psi this is the mean value. So, this is something like that and a standard deviation of 400 psi. So, consider the distribution of X bar sample mean distribution of X bar has a sample mean breaking straight what is the probability that sample means fall between 400 psi and 5200 psi? So, for 1400 will be somewhere maybe here and 52 will be somewhere maybe here I need this probability.

So, how do I find out? First this is something which we this type of problem we have done when we have discussed normal distribution. This is basically a question of that only not a question of sampling distribution basically. So, how do I find this value first is that this is in normal distribution, I will have to convert it to z distribution that means, which has mean 0 and standard division 1, that means each value I will have to convert it to z value.

And then from that z value I will be able to whatever some I will get some value say this is z 1, this is z 2 and this mean will be 0 basically, when it is converting it to z distribution and I have to find out this area how do I find this area? This area will be this whole area minus this portion because normal distribution I get cumulative from - ∞ to this point, is not it? So, I got this whole area and from here I got this area. So, this minus this area, I will get this area.

So, it is this repetition of that type of question which we have done many questions on while discussing normal distribution.

(Refer Slide Time: 35:18)

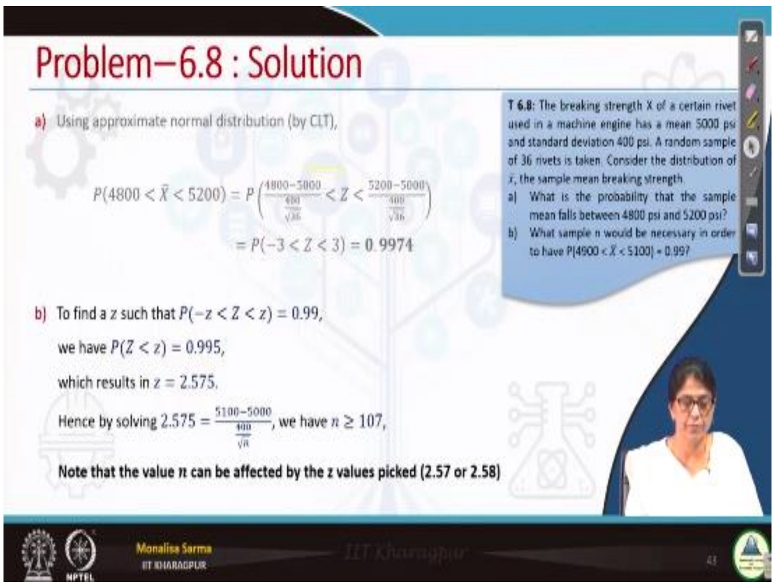


So, I will not be discussing in details here you can just see for calculated the Z value, so, this P of Z value falls between this. So, you can from the table you will be able to do this. So, I am not discussing that next also what sample n would be necessary in order to have probability of this is 0.99 here it is given that again, if you solve it yourself, it is given probability 4900. So, this is

4900 because mean is 5% 49 will be definitely this side and 5100 it is given that probability of this is that means this portion is given 0.99.

If this portion is 0.99 that means this to put together is 1% means this will be 0.05, this will be 0.05 is not it? So first, we will have to convert it to z. So that means this portion what will be the area of this portion? This portion will be 0.995 is not it? This portion will be 0.995 and what it will be sorry and this portion will be 0.05. So, subtracting from this portion to this portion corresponding to z value of that and we will get the n value.

(Refer Slide Time: 36:39)



So, this all this I am not repeating it because we have done this sort of questions.

(Refer Slide Time: 36:46)

So, these are the references and thank you guys.