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

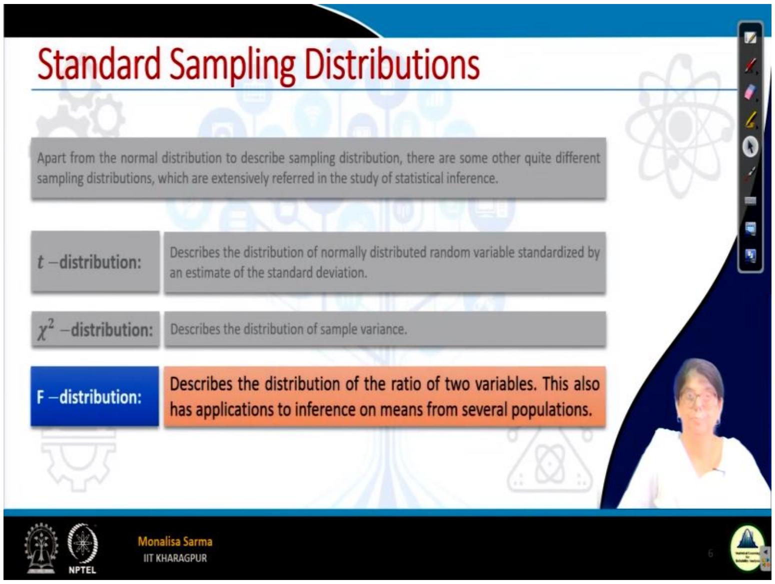
Lecture - 18
Sampling Distribution (Part 3)

(Refer Slide Time: 00:31)

Hello, welcome back. So, in continuation of our discussion on sampling distribution, today we will be seeing few other distribution some other sampling distribution along this. On the last class we have discussed the sampling distribution of mean for which we have used that is the normal distribution that is the Z distribution, sampling distribution of mean it is nothing but it has approximately it is normally distributed. So, basically we have considered the normal distribution that is, I can also say it is not normal basically the standard normal distribution.

So, in this lecture, there are some other distribution for sampling distribution I should not say sampling distribution of mean there are some other distribution which basically, which we use to find out the sampling distributions, that is the t distribution, chi square distribution, F distribution to name a few which we will be discussing and coming in this lectures as a listener coming through lectures.

(Refer Slide Time: 01:25)



So, now firstly, the thing is that what we have seen in the last class, we were interested in finding out the population mean if you are interested in finding out the population mean than the sampling distribution, which will make some we will form the sampling distribution of mean. If we have to infer about the population mean, we for that we will need sampling distribution of mean similarly, if we have to infer about the population variance, we will need sampling distribution of variance.

So, likewise to infer about the different things of the population accordingly, we need the sampling distribution so based on that we will have different distributions now like here as I have mentioned apart from normal distribution to describe sampling distribution that was the sampling distribution of mean. There are some other quite different sampling distributions which are extremely referred in the study of statistical inferences.

There is one says distribution is t distribution, t distribution is also used for the sampling distribution of mean, sampling distribution of mean we use Z distribution we have seen or normal distribution I can say that similarly, t distribution is also used for sampling distribution of mean, but here, sometimes in the population, we cannot say anything about a population standard deviation, like when we talk about the sampling distribution of the mean when we use normal distribution.

That time, you have seen that we have estimated the population mean as well as the population standard deviation, is not it? So, sometimes population standard deviation, it is really not possible to know the possible I mean population standard deviation, when we do not know the population standard deviation, then we cannot possibly use normal distribution, then we will have to use t distribution. Another is the chi square distribution, when we are interested in the distribution of sample variance that is called a sampling distribution of variance.

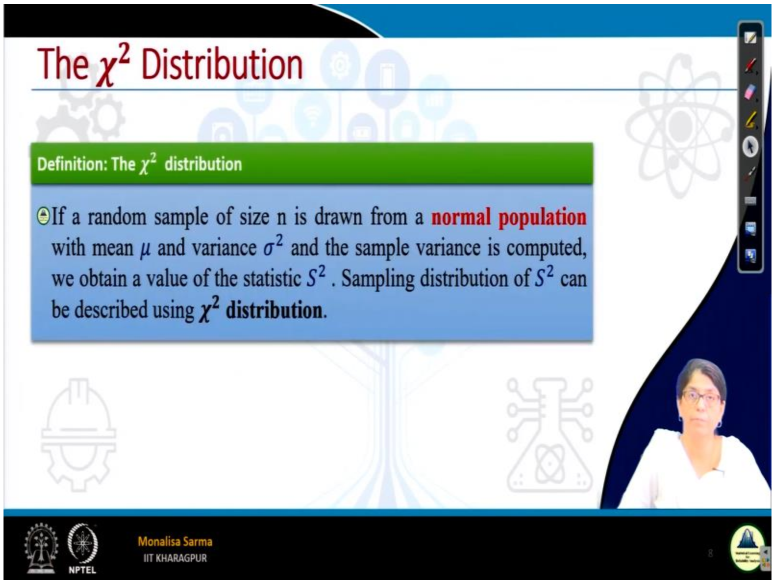
Then the distribution that we will use is chi square distribution. So, then again, there is another one distribution that is F distribution, when we want to compare the variance of 2 different populations, like it is mainly used in what to say this food and beverage industry, where in food and beverage in this industry, like suppose, there are 2 different companies which produces cold drinks now, let us suppose a cold drinks of volume say 1 litre.

So, we will be definitely when we are buying a cold drinks bottle of 1 litre we will expect that he exactly has 1 litre of cold drinks, but if it has less cold drinks, then the customer will get cheated if it has more than it will be customer will be benefited, but it will be loss to the company. So, basically what is necessary is that it should strict to that 1 litre volume means it has 1 litre with the slight variations. So, like so if we want to find out so, whether this is sticking to the required specification.

Or it is more or less then we will be using chi distribution when we want to compare the variances. Now when we want to compare 2 different populations like similarly for food industry only we are trying to compare for 2 different populations based on their variance. Then we will be using F distribution, of course, F distribution has other application as well, which we will be seeing later in this lecture series.

That is, we have something called ANOVA when we will be discussing a number then we will see F distribution is used however, distribution is used in ANOVA. So, now for your till now, it is sufficient to know that we use F distribution when we try to compare the variance of 2 different populations.

(Refer Slide Time: 05:31)



Now, first we will discuss is the chi square distribution. So, first if a random sample of size n is drawn from a normal population with mean μ and variance σ^2 the sample variance is computed and we obtain the value of the statistics S^2 the sampling distribution of S^2 can be described using the chi square distribution as I told you sampling distribution of variance to x what to say to represent the sampling distribution of variance we use chi square distribution.

Now, what is the sampling distribution of variance? Sampling distribution variances we will call it as S^2 because we are finding the variance of sample so, it is not σ^2 that S^2 and one more thing like when we are doing sampling distribution of mean we were not very much interested in the parent population, I have parent population can be any distribution. It can be normal, it can be non normal, it can be slightly away from normal, it may be very much away from the normal.

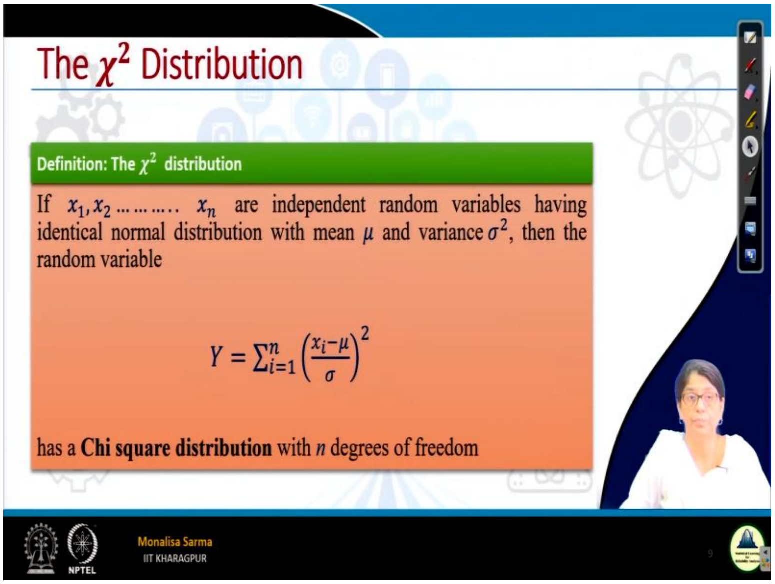
Accordingly this we could solve it by adjusting the sample size into it the population is normal we take a some smaller sample space, They have more near normal we take a bit more sample size, but not very big also. But if the sample size is totally not normal then we take a bigger sample size. But chi square distribution, it is very much sensitivity to normality assumption of the parent population, this is a very important point. Chi square distribution we cannot use with the parent population is not a normal population.

Because in such situation if you use chi square distribution, then our result will not be a reliable result our results will not be precise results. There are in those cases basically, we have other techniques so this techniques of inferring about the populations and is called parametric methods, this methods and there are 2 different methods of is in what is a say one is parametric, what is one is nonparametric. So, what we are discussing this probability distribution sampling distribution, all this comes under parametric method.

So, when the population is not normal, and we want to compare the variation of 2 different populations, then definitely we cannot use chi square distribution, then definitely, we will have to use some other methods, those methods we will be discussing later. Even now, what is it? I will repeat this again if a random sample of size n is drawn from a normal population that is a normal I have marked it red.

If a random sample of size n is drawn from a normal population with mean μ and variance σ^2 and the sample variance is computed, what is sample variance is S^2. The sampling distribution of S^2 is described by chi square distribution.

(Refer Slide Time: 08:21)



Now, question is what is chi square distribution? Earlier was sampling distribution of mean we use Z distribution we knew that this was a normal distribution, but was that we know, but chi square distribution we do not know. So, what is chi square distribution first let us see that, see this is the definition of chi square distribution if x 1, x 2, x n are independent random variables having identical normal distribution, these are different independent random variables having normal distribution with mean μ and variance σ^2 then the random variable Y.

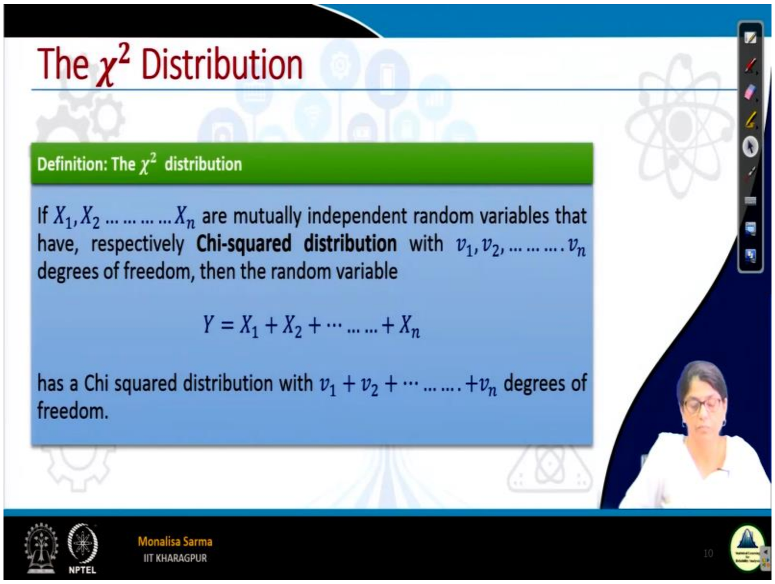
What is Y? Y is nothing but a summation of Z^2 x i - μ / σ what is that? x i - μ / σ is Z is not it? So, Y is nothing but a summation of Z^2 where this x i is a different random variables having normal distribution that means I can say normal random variable x i is a normal random variable. And μ and variance μ is the mean of that random variable and σ is the standard deviation of those random variables.

This here I told it is identical normal distribution why identical? That means, each has mean μ and variance σ^2. That means chi square Y is nothing but the summation of Z^2 Z i^2. So, this is called it has a chi square distribution with n degrees of freedom. So, the n degrees of freedom mean how many what to say logically independent unit there are totally n logically independent unit when you consider a chi square distribution.

So, the degrees of freedom are n so, when we consider a chi square distribution, the only parameter that we have to worry is the degrees of freedom. That is one parameter that is the degrees of freedom. That means for chi square distribution also we will have a table where we can find out the value. In the table, we will have values for different degrees of freedom. Remember, for normal distribution, when we are considering normal distribution, we have the values for different values.

Since it was not possible to have different values of μ and σ that is why we have converted it in to Z value. So, we have different values corresponding to different Z value. Similarly, when chi square distribution will have different values corresponding to different degrees of freedom that is, the only single parameter of chi square distribution that is degrees of freedom, that is the number of logical independent units. So, it is a summation of Z i^2 and i go from 1 to n. So, there are n logical independent unit.

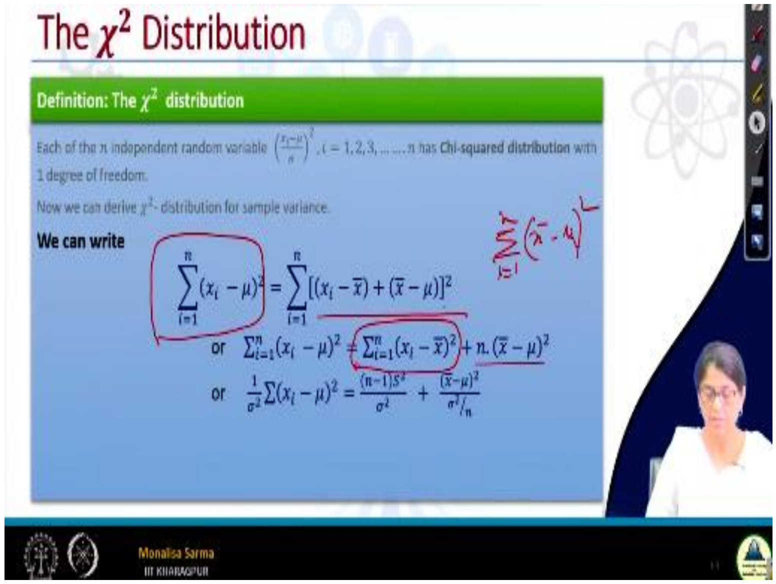
(Refer Slide Time: 10:54)



There is one more theorem here. So, if X 1, X 2, X n are mutually independent random variables that have respectively chi square distribution, if X 1, X 2, X n all have chi square distribution, x 1 as chi square distribution x 2 has chi square distribution with v 1, v 2, v n degrees of freedom, then the random variable Y, Y = X 1 + X 2 up to X n this Y is also a chi square distribution. What will be its degrees of freedom? Degrees of freedom will be the addition of all these degrees of freedom.

X 1 has degrees of freedom v 1, X 2 has degrees of freedom v 2, so, Y will have a degrees of freedom v 1 + v 2 + v n that will be the degrees of freedom of Y. And Y is also a chi square distribution Y will also have a chi square distribution, this is the important theorem I should say corollary of this previous definition of the chi square distribution.

(Refer Slide Time: 11:54)



Now, here when we consider the chi square distribution and there are as I told you in a chi square distribution, there are totally n independent units. So, each independent unit what are the each independent unit is Z i is one independent unit Z i is nothing but $(\sigma\ i - \mu)^2$ $Z\ i^2$ this one independent unit but in $Z\ 1^2$ is one unit and $Z\ 2^2$ is one unit. So, this is one independent unit $(x\ i - \mu / \sigma)^2$ this each unit has a degrees of freedom of 1.

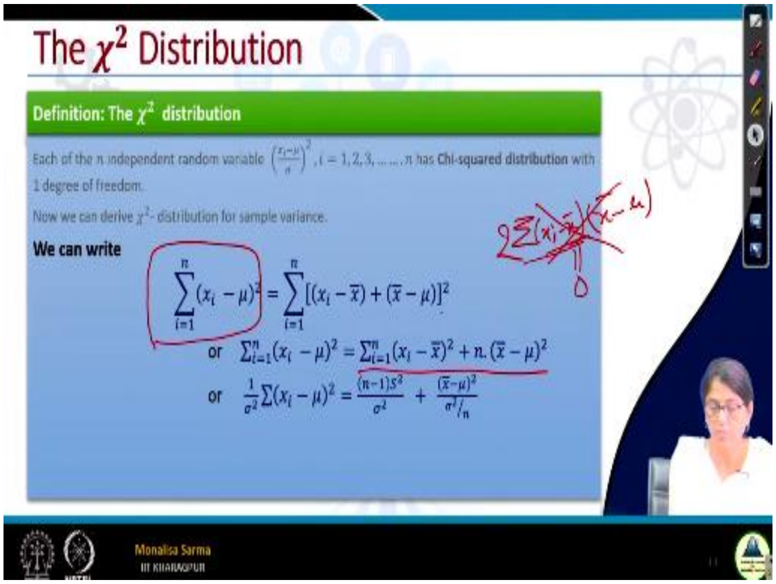
So, each unit has a degree of freedom 1 that is why chi square distribution as a degrees of freedom of n. Now, we got 2 different definitions one is this definition that is this definition and another we got this definition we also got this definition with these definitions with these 3 definitions. Now, we can derive this chi square distribution for sample variance here see, I already mentioned we use chi square distribution to find out the sampling distribution of variance means sample variance.

See in this chi square definition what we have seen till now, there is no mention of S^2 anywhere somehow we will have to bring the quantity S^2 to it then only then only that will be a distribution of because if that term is only missing, how can it be a random variable of that term. So, somehow we will have to bring S^2 to it, how do we get that. So, basically, first just simple consider this term $(xi - \mu)^2$ that I have done is that.

Here I have just done some sort of manipulations here this can I write it in this way x i - x bar + x bar means x bar I have substituted and added it. So, I got this now, this is $(a + b)^2$ what is $(a + b)^2$? $(a + b)^2$ is $a^2 + b^2 + 2ab$. So, I am writing it here only it will be easier. So, what do I get from this summation of first is a square. So, this is my first term. This is a square. Similarly, b^2 what I will get b^2 I will get is i = 1 to $(n \bar{x} - \mu)^2$ but now way i is there.

So, what it will be a? It will be just n into x bar - μ. So, this is the term I got this is my b^2. Next what is the remaining is 2ab, summation of 2ab so 2ab is what is this?

(Refer Slide Time: 15:00)



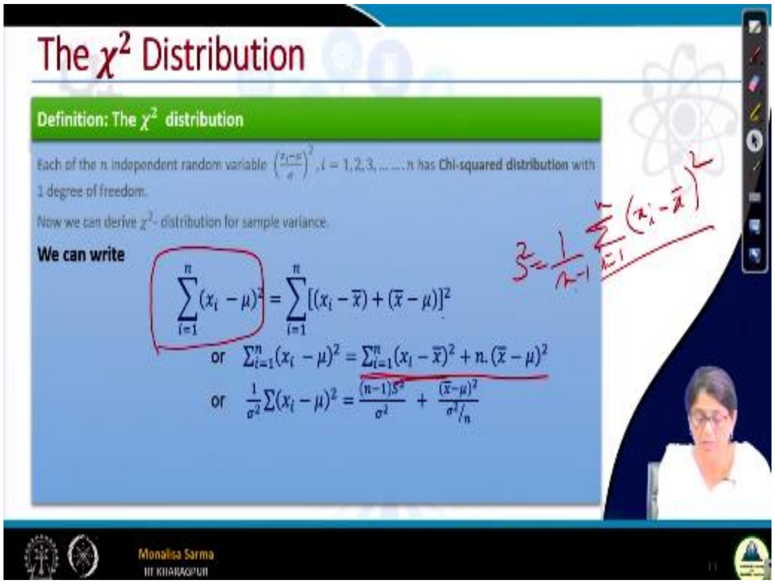
2 summation I am putting it inside say a summation of x i - x bar into x bar - μ is not it? This is my 2ab now, see this term summation of xi - x bar. While calculating variance I think you have already learned in class 9 and 10 while calculating variance why we use square $(x_i - \bar{x})^2$ you know that can you remember why we? Because if we do not use square because this minus plus minus plus updating will be done and we know what the end result will be we will get 0.

When we because some value will be greater than mean some value will be less than mean what are the different x i, x i are the different values x bar is the mean when you try to calculate the variance when we subtract each value from the mean some value will get greater than mean some value will get less than mean and everything sum together when you say we will get 0.

That is why to nullify the effect of minus sign we have used square that is why in variance we use square there that is the reason.

So, now, this term that means, this term will be equal to 0. So, this term is gone. So, what remaining is this 2 term only. So, what remaining is this 2 term. Now, what I have done now, this left hand side and right hand side I have divided by σ^2. So, this term divided by σ^2 this term divided by σ^2 now, if this term I divided by σ^2 what do I get. What is my S^2? S^2 is nothing but variance only is not it? Variance of the sample so, what is my S^2 I am not using the board I am writing it here because then it will be it will be easy for you to relate.

(Refer Slide Time: 16:55)

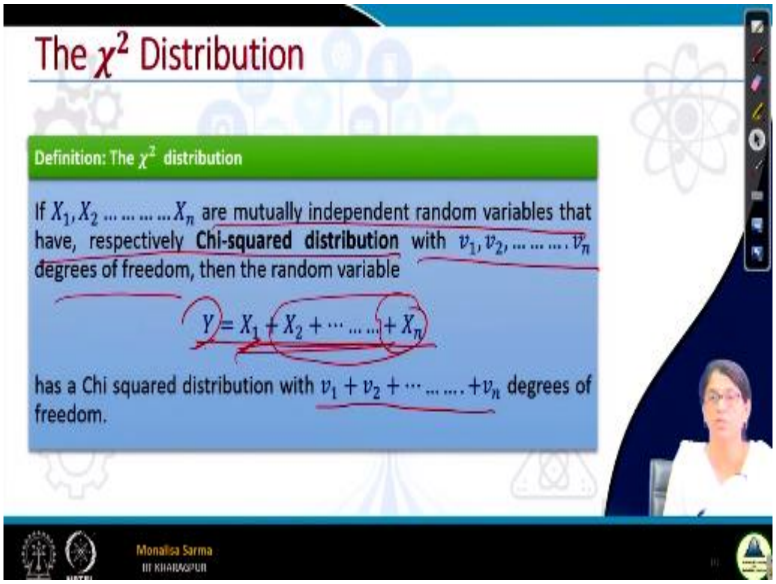


So, what is my S^2? $S^2 = 1 / n - 1$ it my sample size is n summation of this is irritant actually i is equals to 1 to (n xi - x bar $)^2$ this is my s^2 is not it? Somehow why, we are doing this manipulation. So, as we have somehow we need the S squared term to what to say, so, that we can find out the sampling distribution of variance. So, this is my S^2 now, this term is nothing but this term. So, what is this term this term is equals to S^2 into n – 1 will give me this term.

So, that is why I said instead of this term n - 1 S^2 / S^2 because σ^2 we have divided both left hand side and right hand side. So, this is what I got. So, this is clear. So, from this that is why I got this. Now, what is this? This is a chi square distribution. What is chi square distribution? Chi square distribution is this summation Z i ^{2} is chi square distribution this is summation i = 1 to n it is implicit not written here.

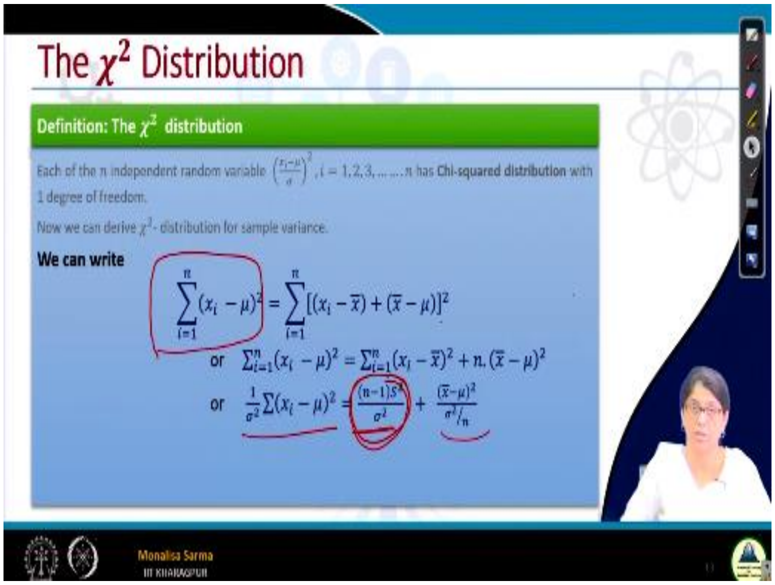
So, this is a this term is a chi square distribution with what degrees of freedom n degrees of freedom i = 1 to n it is n degrees of freedom and what is this? This is also a chi square distribution, but just 1 degree of freedom just 1 term. So, this is a chi square distribution with n degrees of freedom this is a chi square distribution with one; degrees of freedom. So, that means, this has to be a chi square distribution with n - 1 degrees of freedom.

(Refer Slide Time: 18:43)



This theorem if this all is a chi square distribution, this will be a chi square distribution so, now, I have this as chi square distribution this may be this as chi square distribution definitely this will be also a chi square distribution.

(Refer Slide Time: 19:00)

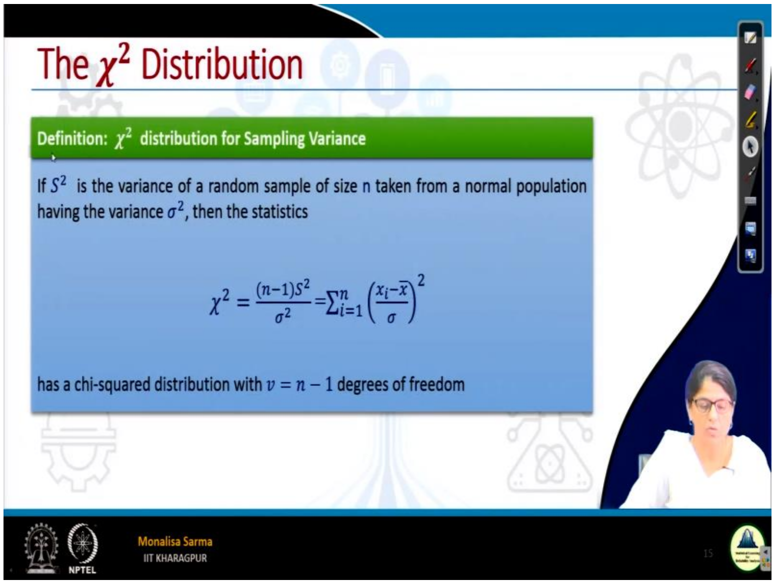


So, this is a chi square distribution with n degrees of freedom this is a chi square distribution with 1, degrees of freedom. So, definitely this is a chi square distribution with n – 1, degrees of freedom. So, now I got it. So, this means my this value will have a chi square distribution and sampling distribution of mean my x bar has normal distribution is not it? x bar is approximately normally distributed, What is x bar? x bar is sample mean now, here I have this value.

This value is chi square distributed this is not only S^2 but something added some linear combination n - 1 S^2 / σ^2. So, now this is my sample statistics, when while finding out a sampling distribution of mean my sample statistic was the sample mean, here, my sample statistic is not only sample variance, but as well as I am multiplying sample variance with the size of the sample -1 n -1 into S^2 and divided by the population variance.

Now, this population variance, and the sample size these are constant values, if we take different values different sample also, if we take different samples, this n - 1 and σ^2d will remain constant. So, this is the constant factor. So, what will vary S^2 will vary, so that means S^2 is chi square distributed. So basically, I can say this whole thing is chi squared distributed.

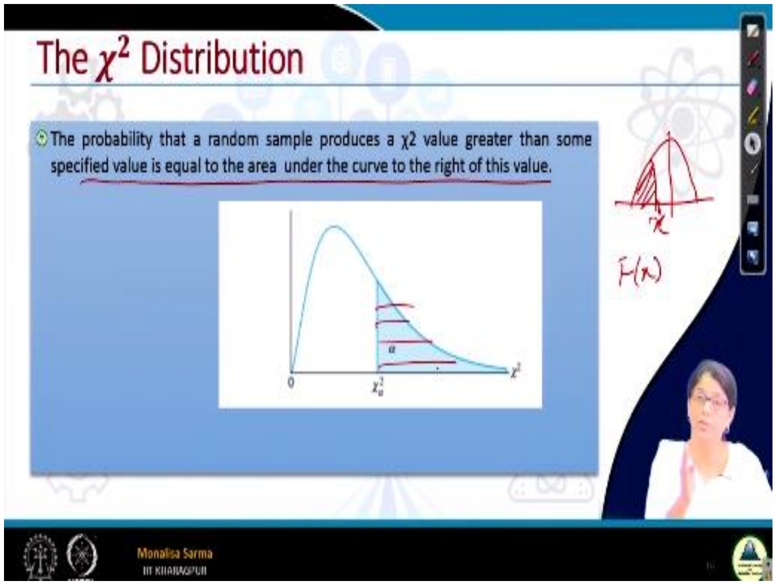
(Refer Slide Time: 20:49)



So, if S ^{2} is the variance of a random sample of size n taken from a normal population having the variance σ ^{2} then the statistics this statistics, these statistics I am calling it chi square, this whole statistics n - 1 S ^{2} / σ ^{2}. This statistics has a chi square distribution with what is the degrees of freedom? Degrees of freedom n - 1 degrees of freedom. Why I am telling this statistics, because I am using the sample data to calculate this. What this sample data x squared calculated in from the sample is not it?

So, I am taking the sample data to calculate it. So, I am calling it sample statistics. So, this is my sample statistics, this have a chi square distribution with degrees of freedom n - 1. So, now this chi square distribution, I will be using it to predict about the population variance. Sampling distribution of mean I have used to predict about the population mean similarly sampling distribution of variance I will be using to predict about the population variance.

(Refer Slide Time: 22:00)



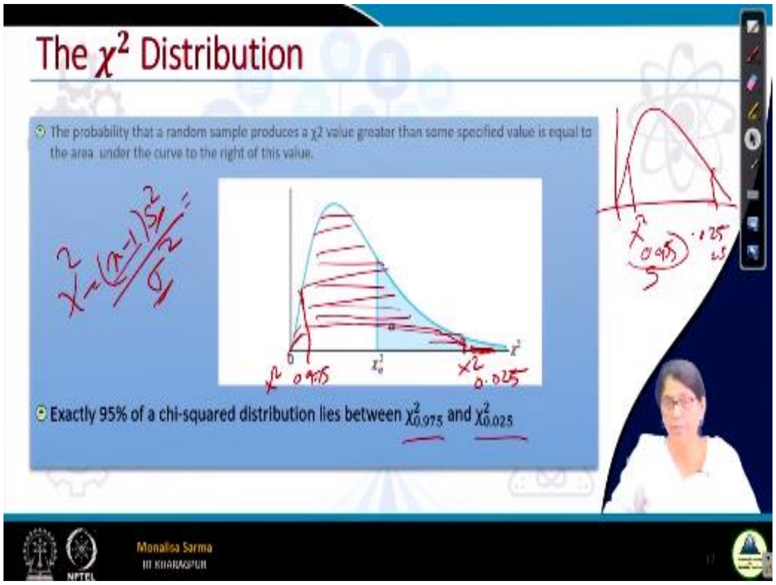
Now, this chi square also like normal distribution how we can refer the table to get the value. Similarly, for chi square distribution, also, we do not have to calculate it ourselves, we can very well have the standard all the standard books have this statistical table available chi square table, from the statistical table, we will be able to calculate the chi square value, but there is a differences in the normal table, we had that cumulative distribution value.

Remember, if it is a normal table, if this is as the normal Z table Z value, so when we if we are interested in finding out of f of x, so that means suppose this is x value, we got this value ∞ to that value. That was in case of normal distribution, we got the in binomial distribution also, we got the cumulative distribution value from -∞ to that particular value x value, whatever that. But in chi square distribution, we just get the opposite table just give the opposite value.

So, if you have a chi square value, so see the probability that the random sample produces a chi square value greater than some specified value is equal to the area under the curve right of this value. So, in normal distribution, when we are interested in finding out a greater than value, then we have done 1 minus of less than, now, because in the table, we have the less than value, but in the chi square the table gives us the greater than value.

So, if you are interested in finding out the probability that a particular sample will have a particular value greater than a particular chi square value, so it will be this area. From the table, we will be getting this greater than value, if you are interested in finding out the less than so we will have to do 1 minus of that.

(Refer Slide Time: 23:52)



And one more important point to note here in a chi square distribution is a very much you see it is a skewed distribution that is not a symmetric distribution it is very much a skewed distribution. In this distribution, what we have seen is that 95% of the value falls within this range. So, this is chi square 0.975 this is chi square 0.025, 95% of the value lies in this range. So, well calculating the probability if we find our probability values lies in this range. That is not what to say not more than 0.975.

If we can decide it will be more than 0.975 is not it because we consider the area to the right, it should not be more than 0.975. And moreover, it should not be less than 0.025 if a chi square value falls within this range then it is an acceptable way. See how chi square we write here it is for α we write is chi square α we get it a subscript. So, similarly, so, this is how we are writing in this. So, if 95% of the value lies in this.

Now, while calculating the probability if we find our value lies in this region, then we can say this is our whatever we have our estimate is correct, if under what case you see we will see our value will lies in this range or in this range under what situation. So, if you see the value will lie in this recent situation means, this is the chi square, so, this is chi square 0.975 corresponding to this say particular value chi square 0.975 a corresponding to this suppose, if it is starting from 0 suppose this value is say 5 let us say any value any unit say 5.

And suppose this value that is equals to 0.025 let us say does this value is it 25. If we get a area greater than 0. 975 when we will get a getting? Get area greater than 0.975 when my chi square value will be less than 5 suppose, if this is equals to 5 when my value will be less than 5 then I will be getting area greater than 0.975. Similarly, 0.025 similarly, I will get a suppose this is area corresponds to 0.025 is suppose as I told you it is 25 when I will get a value area less than this when my value will be more than 25.

Under what condition see what is my chi square? chi square is equals to n - 1 S^2 / σ^2 is not it now, my S^2 is something that I have calculated from the sample I have already taken a sample from the sample I have calculated the S^2 value. What is n - 1 the sample what I have taken that is the sample size - 1 there cannot be any error in that. So, I am getting a value which is greater than 0.75 under what situation I will get a value greater than 0.75?

When this σ^2 the σ^2 is what? Does a population σ^2 is we have just guessed, it is an estimation, it may be an educated guess is we have just predicted when I will get a value greater than 0. 975 When this σ^2 value is very less if denominator is very less than what happen, I am sorry, if this denominator value is very less than I will get a value chi square value more bigger value that means maybe more than 25.

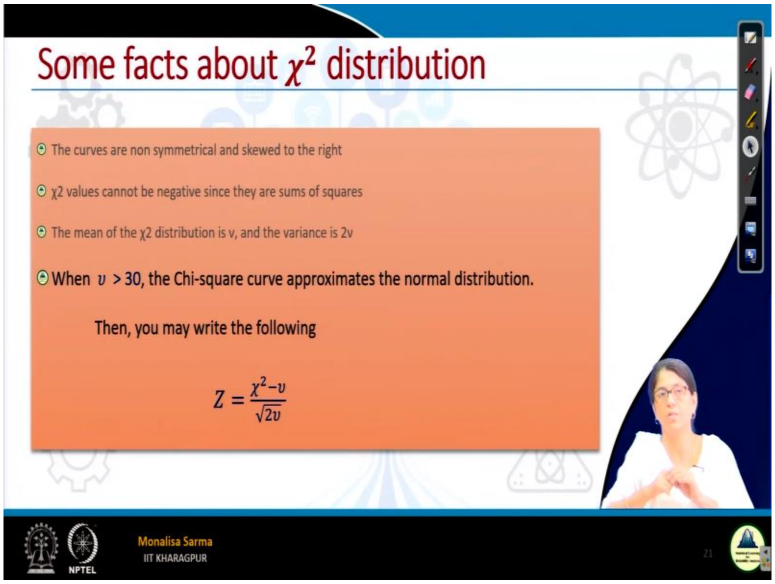
When my chi square value is very big, then I will get a chi square very small value, if my chi square σ^2 value is very small, I will get a bigger chi square value. In both the case there is this is because maybe my I have estimated my σ^2 value wrong why? If the σ^2 value is very big, getting a very big σ^2 means I am getting area greater than 0.975. Getting a σ^2d value very big that means my population is very much variant population.

When the population is very much variant, usually it is like this sort of statistical inference is usually not done in most of the case when the population is very much variant population that means that is not a stable population. So, before coming to this sort of study only steps are already taken there it may be correct, there may be case when my population variance is very big that may be correct for some populations, but usually in general and the population variance is very big this sort of study is not done at all.

Because in that case, the population is very unstable, some steps has to be taken first to stabilize the population, then talking about statistical inferencing and all those stuff. So, that is one thing, second case, when my population variance is very small then what will happen my this value will be very big that means my area will be very less it will be in this area. So, that is also very unlikely usually in a population the population cannot be so stable that variance is so less very unlikely it may happen it is not that it will not happen.

It may happen. But that probability of that is very, very unlikely. So that is why it is considered that if my chi square value falls in this range, then that is that means that estimate what we have estimate what we have estimated? We have estimated the population parameter rest are things we have not estimated we have got it from the sample that means, if we got it if our probability we got in this range that means, our estimation was correct, if we get it in this range the side drains then maybe our estimation is wrong.

(Refer Slide Time: 30:24)

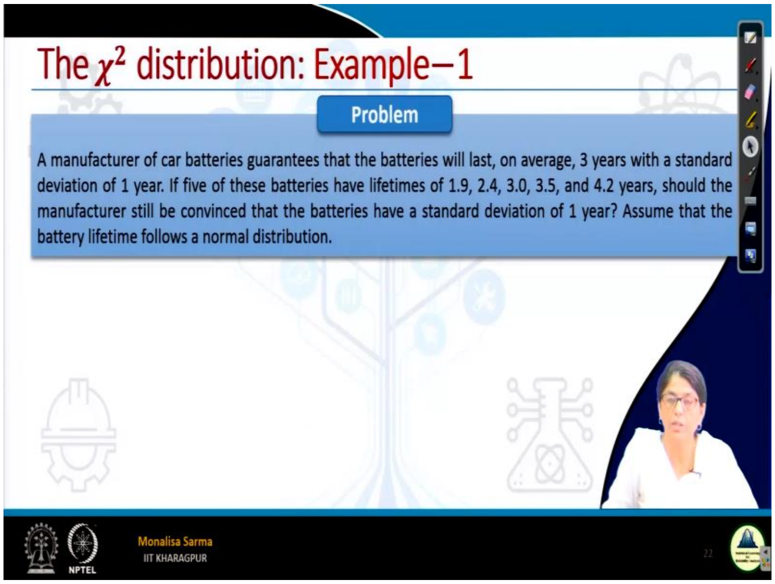


So, some facts about chi square distribution the curves are non symmetrical and skewed to the right we have seen that chi square value cannot be negative since their sum of squares obvious. The mean of a square distribution is v and a variance is 2v, v is the degrees of freedom why does mean and variance why it is necessary? If when v is greater than 30, the chi square curve approximates the normal distribution. When our sample size is more than 30 then what happens instead of using the chi square curve, because chi square curve has very limited table.

In the table, we get very limited value for the degrees of freedom. So, when the sample size is big, instead of using chi square distribution, we can also use normal distribution to approximate the probability we are using all this probability distribution to find out the probability that is the only objective. So, when the sample size is greater than 30 instead of using chi square, we can very well use the normal distribution.

So, in that case, so, normal distribution means we need the value of Z. So, Z means what is Z? Z means x bar - μ / σ, so, what is x bar here x bar is my chi square value this is not x squared this is chi square, this is x bar is my chi square value x bar - mean, this is the mean of the chi square distribution is v and the variance is 2v, v is the degrees of freedom and 2v is the variance, then √2v is the standard division. That is how I find what to say that is the Z value once we know the Z value we can find out the probability distribution.

(Refer Slide Time: 32:14)



So, now one minute quickly we will solve 1 problem, or in fact, instead of doing it quickly let us stop this class here. I will start it from next class. I will start it from this point. Let me stop it here. And then in the next class first we will definitely solve this problem and then we will go to the next distribution. With this I end this lecture. Thank you.