Data Analysis and Decision Making – I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

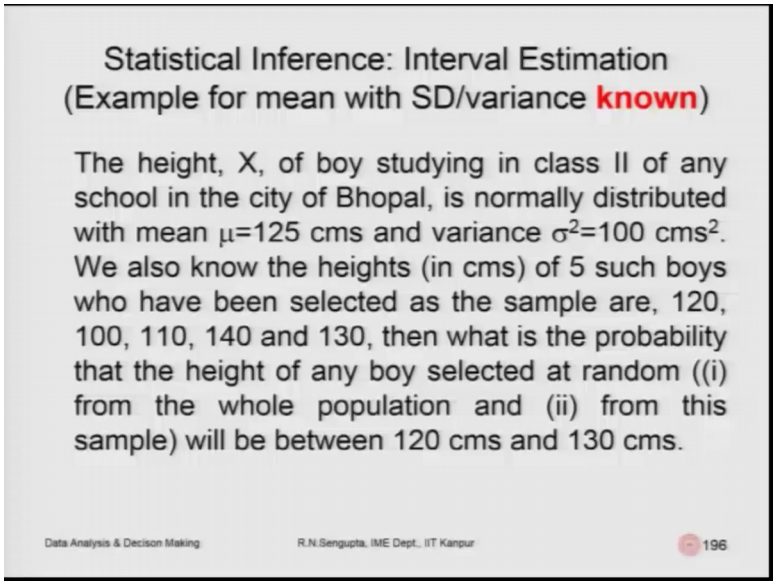
Lecture – 17
Statistical Inference

Very good morning, good afternoon, good evening my dear friends. Welcome, this to this Data Analysis and Decision Making - I course on the NPTEL, MOOC series. And this is a 12 week course of 30 hours, each week as you know we have 5 lectures and each lecture being of half an hour. So, we are in the 17th lecture which is we have just started the 4th week and we are in the second class for the 4th week. And I am Raghu Nandan Sengupta from the IME Department, IIT, Kanpur.

So, if you remember we were just in the last class we were discussing about interval estimation, and I said that to wrap it up I said that for the mean value for the interval estimation given the parameter estimate is known from the sample we will use the z distribution or the t distribution. And if for the standard deviation to be found out we will use the chi-square distribution and the f distribution. And the other distributions considering can also be used using the concept of central limit theorem and all these things.

So, we will try to do a problem and then discuss. So, the problem is basically what we left in the discussion in the last class which is the 16th the last 3- 2 slides. So, you will first cover the concept then do a problem. And in case if it is needed we will just discuss the concept and then go further on continue, but with the promise that we will utilize those concepts where for where we have not utilized the problems to basically cleared them in more details, will clear them in a much more bigger scope where things would be much holistic in their concepts or that I am sure the students would understand it much better. So, this is the example for the mean with standard deviation variance known.

(Refer Slide Time: 02:13)



The height X of boy of boys studying in class II of any school in city of Bhopal is normally distributed with mean 125 centimeters and variance sigma square is 100 centimeter square. So, remember the variance for the whole population for all the students who are studying in the city of Bhopal in class II, the whole population variance is given as sigma square which is there 100. I am only talking about the values. So, I am not going into the units.

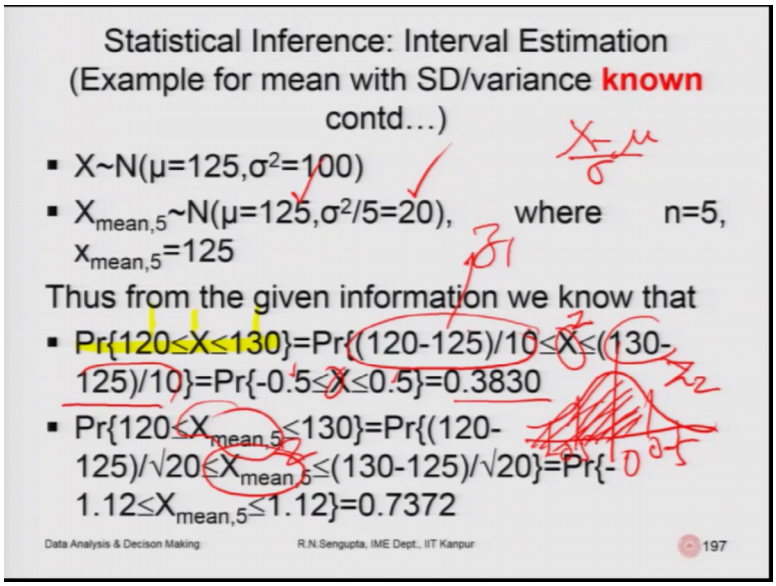
We also know the height in centimeters of 5 such boys who have been selected as the sample are given as 120, 100, 110, 140 and 130. Then we want to find out that what is the probability, there are two questions what is the probability that the height of any boy selected at random number 1, from the whole population and from this been selected will be between 120 and 130. So, basically we have between 120 and 130 we have to find out that the boy being selected from the whole population would be in that group and boy selected from the sample would be in that group. So, what is that group? That group is basically the interval 120 to 130. So, you have to basically solve it.

Before I give you the solution first let us see the problem what are the informations given. So, the informations given are basically the population mean and the population variance. So, population mean and population variance are given and you want to find out something to do with the mean so obviously, we know that as the population variance is given we have to use the z distribution. So obviously, it will be t if the population

variance was not known, not given. So, for population mean we they either we use the z or the t.

So, the continuation is the first bullet point says then X is basically normally distributed with 125 as the mean and the variance is 100. Now, we need to find out that what is the distribution of 5 observations taken in a group and what is the distribution of their sample mean.

(Refer Slide Time: 04:13)



So, here the suffix X mean means and comma 5 means that we are taking a bunch of 5 finding or they mean and basically trying to find out what is the distribution of that mean.

So, X mean comma 5 means bar 5. So, this is also obviously, we know that will be normally distributed, but 125 mean value, but now in this case the total variance would be if you remember the sigma square by n. So, it will be 100 by 5 it would be 20. So, where n is equal to 5, and the sample mean is given by 125, the mean yeah as we found out.

And now from the information what we need to know is that, what is the probability that the value of X would be lying between 120 and 130, and in the second case that what is the probability that the sample mean would be lying between 120 and 130. So, let us do the problems accordingly in the first case X is the random variable. So, let us ask that

what is the distribution of X; we know the distribution of X is normal. What is the mean? I my answer is 125 and sigma square is 100. So, we you need to convert that into a standard normal and proceed with our calculations.

So, the probability that it X is between 120 and 130 is given by this let me highlight. So, I need to convert X into the standard normal, as I do that this is basically the X 1 and X 2 which I have the corresponding LCL and UCL for the X, they would also be converted into LCL and UCL for the z distribution. So, they would also be transformed accordingly.

So, here how do we convert the X into a standard normal deviate? It will be X minus the expected value of X divided by the standard deviation exactly what we do is basically this X is now converted into Z. So, we have not written it, so my apologies for that. So, this would basically be converted into a Z. So, technically what we are doing is X minus mu by sigma. So, mu is basically given as 125 which is here the sigma is basically 10 we convert that into a Z value the capital Z standard normal and these are basically Z 1 and this is also Z 2. So, these Z 1 and Z 2 are corresponding to 120 and 130.

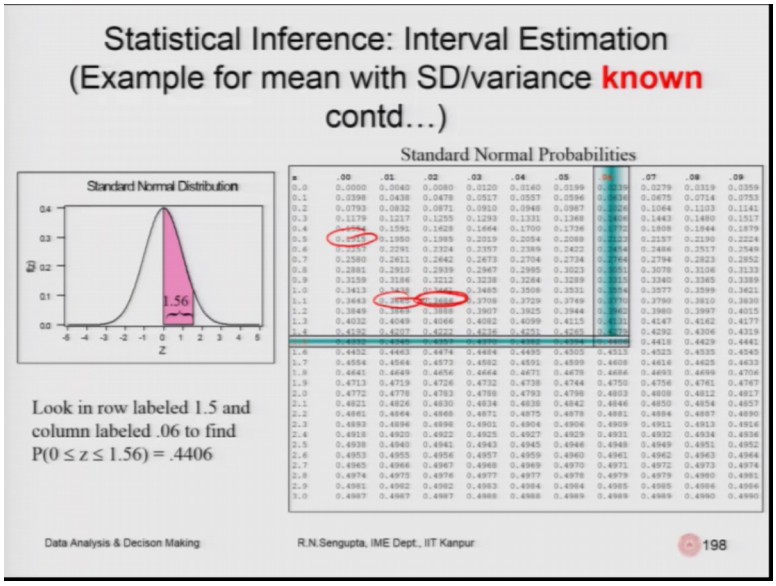
So, hence for the Z it is basically bounded between minus 5 to plus 5 the value basically comes out with the probability. So, here you have the distribution the mean value is 0, this is 0 points minus 0.5, plus 0.5 the whole area bounded between them is about 38 percent at is given and hence the probability is 0.38.

Now, go let us go back to finding out that what is the probability of the mean value being between that. So obviously, again we find out the same stand standard formula probability that X mean of that corresponding to the sample size of 5 you will be between 120 and 130. We know the sample mean is distributed with normal distributed with a mean value as given as 125 and the standard deviation or let us talk about the variance sigma square by the sample numbers which is 5 which comes out to be 20. So, hence it will be the; it will be bounded between 120 and 130.

Again convert this X mean 5 into a standard normal deviate which is basically the Z value and the corresponding lower control and the upper control will be calculated accordingly which will be, in the first case the lower control will be 130 minus 120 minus 125 that is and divided by square root of sigma by root of n which is technically

square root of sigma square by n that which comes out to be square root of 20 put that values. And on the right hand side again you will have that value of 130 minus 125 which is plus 5 divided by square root of sigma square by n. So, put those values and that value comes out to be 0.732 in this case. That means this is deviating more on to the left more on to the right hence the overall coverage would be much more than 0.38. So, it is coming out to be almost double.

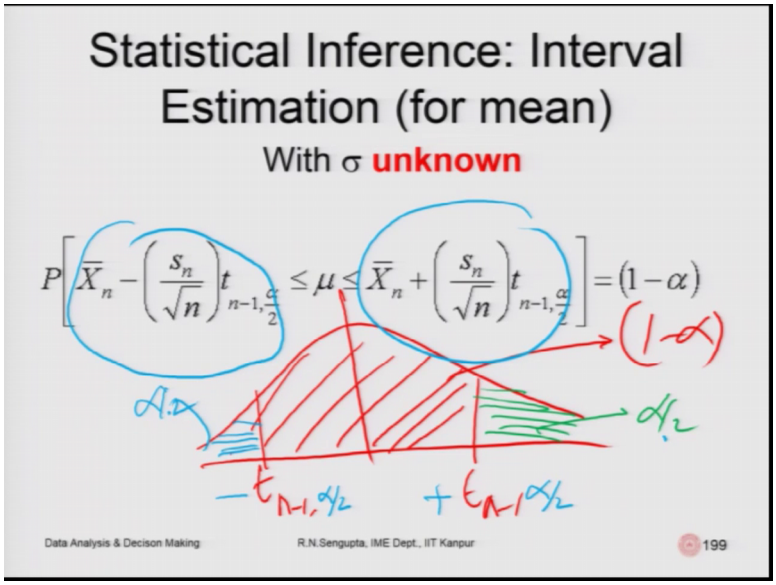
(Refer Slide Time: 09:12)



So, these values which I have seen I am just giving the standard normal table also in order to understand so minus 5 and plus 5. So, minus 5 value is basically 0.195, so double of that because you are going equally disposed onto the left hand side on the right hand side it comes out to be 0.38 as the value.

For the case when you are using 0.7, so let me go back till the value it is coming out to be 0.732. So, it is plus minus 1.12. So, 1.12 would come out to be this. So, this would be make about this, so 0.3686. So, both 0.3686 on the left hand side and the right hand side add them up or multiply them by 2 comes out to be 0.72 which is 72 percent.

(Refer Slide Time: 10:15)



Now, consider the case of the statistical inference interval estimation for the mean with the standard deviation being unknown. So, none the first question comes is that if you are interested to find out something with this with the mean value for the population. So, what distribution you should be used. To answer as we know is basically the z or the t.

Now, why we are using t because the case is that the standard deviation of the or the variance of the population is not known, if it is not known mean we know we have to use one of the s s. And when we use this one of the s s, I am not going to come into that detail what I am basically highlighting the main points is that we will basically lose one degrees of freedom hence you will basically be utilizing the t distribution. Hence t distribution with me n minus 1 degrees of freedom, but remember one thing here if I draw the t distribution let me draw it. So, this is not a normal, this is t.

So, this is the mean value and we want to find out this limits. So, this is basically so called t n minus 1 and this would be t n minus 1 but the and this whole area as we know is equal to 1 minus alpha by 2; alpha. This area let me change the color, let me change it to blue or green here. This will be basically alpha by 2; this area is also alpha by 2.

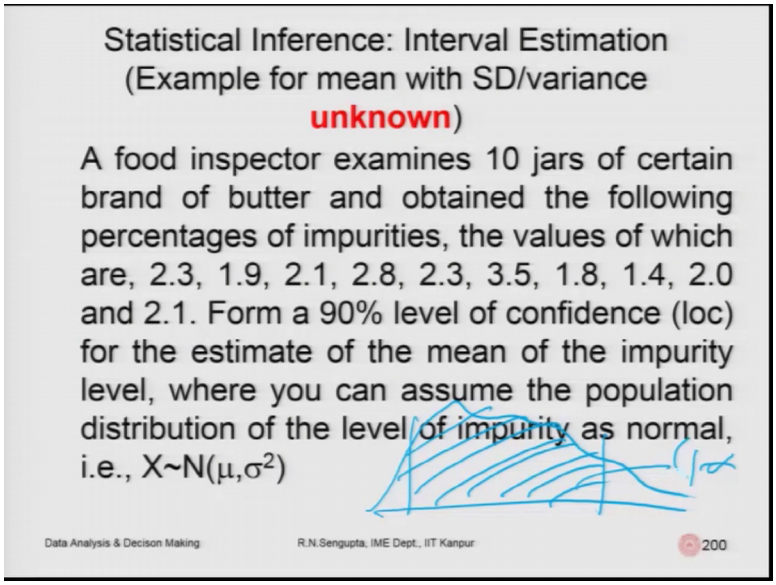
So, hence some of alpha by 2, alpha by 2 and 1 minus alpha would basically be me the total area under that curve is 1. Now, remember t distribution have I mentioned is symmetric. So, if it is symmetric the values on to the left hand side and right hand side

would basically given by alpha by 2 here, alpha by 2 here, but only remember the mean value being 0. So, any value on to the right hand side would be positive anywhere on to the left hand side would should be negative.

So, we use the t distribution, but in this case remember also we would not be using my s dash, because s dash is only used for the case when the sample the population mean is known, but in our case the population mean is not known. So, you will be utilizing s without the dash. So, hence it will be divided by 1 by n minus 1. Why 1 n minus 1? Please bear with me I will be repeating these things time and again because we lose one degree of freedom considering that X 1 to X n have been obtained you will use them for the first time to find out what is the best estimate of the population mean given the sample mean is the best estimate which we know.

So, again we will we will basically formulate the problem in this way where X n bar minus this term and X n bar plus this term would basically with the corresponding lower control and the upper control correspondingly.

(Refer Slide Time: 13:10)

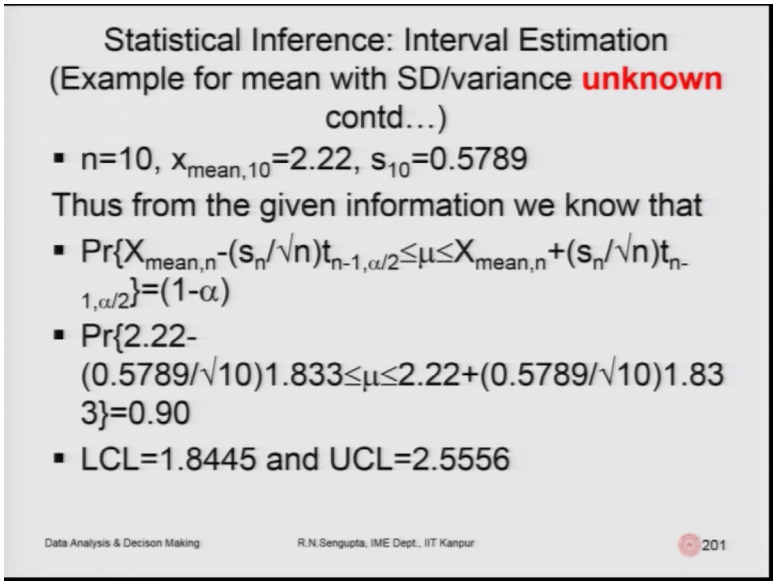


So, we will do a again a problem, small simple problems. A food inspector examines and this is the case for the population variance being unknown. A food inspector examines 10 jars of certain bread of butter, brand of butter and obtain the following percentage of the impurities the values of which are given as 2.3, 1.9, 2.1 so on and so forth. You want to

basically form a 90 percent level of confidence loc, for the estimate of the mean of the impurity levels such that where you can assume the population distribution of the level of impurities basically given in a normal and we want to basal comment intelligently according to that.

So, your values or impurities are given, the level of confidence 90 is given. Level of confidence means the overall area between the lower control limit and the upper control me. So, if I basically have and the distribution; again I am not trying to specify with this a normal distribution, t distribution chi-square distribution or f distribution. So, these are the limits the overall area we in between is basically given by level of confidence which is 1 minus alpha.

(Refer Slide Time: 14:17)



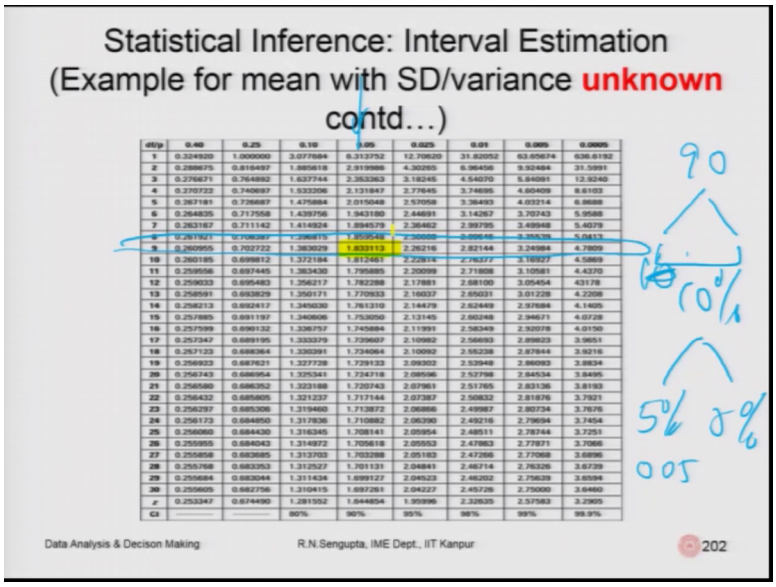
So, from this n value is given as 10, the mean value X X mean of with the sample mean of 10 is given by 2.22, and this s without the dash because we are utilizing the sample mean as the best estimate of the population mean based on which when we calculate we find out the standard error comes out to be 0.5789. Thus, from this given information we know that the lower control limit and the upper control limit are in such a way that X mean of this ends observations.

And on the left hand side if we remember to be s n divided by the square root of n because here again you will basically be dividing by square root of n depending on the

set of observation which is there for the sample. Into t n minus 1 into alpha and t n on the right hand side also it will be t n minus 1 into alpha; alpha by 2. But remember technically they can be written later on we will see for the chi and the z, chi and the f distribution on the right hand side values would be given by 1 minus alpha by 2 depending on whatever area you have to cover on the right hand side I am going to come to that later.

That is for the given information we know that probabilities being such that that the lower control limit which is X mean for this n observations and s n being the square root s n divided by square root of n into t the t distribution the value. This t n minus 1 alpha by 2, alpha by 2 is basically depending on the level of confidence which you have, so that and on the right hand side the upper control limit would be X mean n and again the same thing, but with the same a plus sign here. So, once you put this values the lower control limit and upper control limit you can check that comes out to be 1.8445 and upper control has 2.5556.

(Refer Slide Time: 16:18)

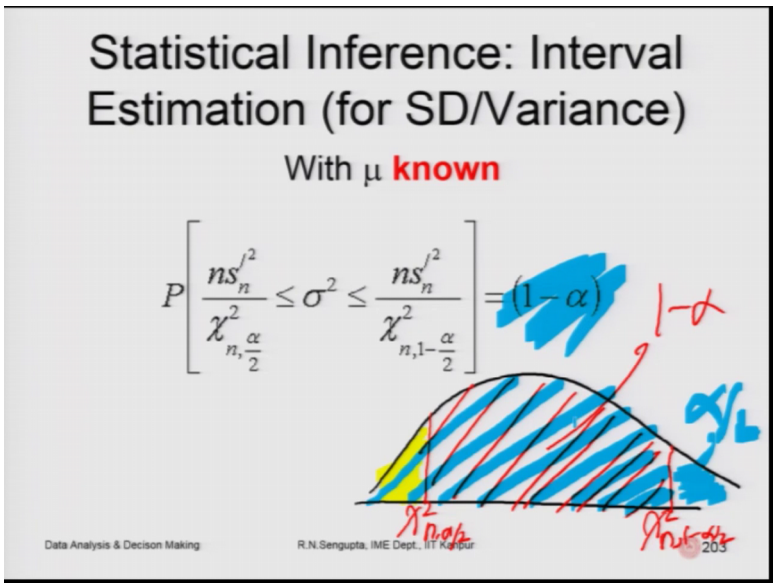


So, again I have added, the this is the first time you is seeing the t distribution table on the t distribution table on the leftmost column you have the degrees of freedom, on the top most value on the row you have the p values and the inside values are basically the t values.

So, remember one very very interesting thing when we using the z distribution it gives you the actually the overall coverage of the probabilities. Hence, if you add up all the values starting from minus infinity to plus infinity the probability values for the z distribution there would be 1. But the values which are inside here are the corresponding t values based on which you are going to do with the calculation it does not give you the cdf value. So, be careful about that that would also be true for the f distribution that should also be true for the chi-square distribution.

So, for a value of 29 and if you see the value being 29 because we have lost 1 degree no this would be basically sorry my apologies it would be 9. So, basically we go to 9. So, the first we basically mark this, then let us see what is the degrees of freedom, degrees of freedom is basically given by 9 which we know 1 minus alpha is known. So, if we have, so basically it was 90 percent so the overall area left and right is 10 percent. So, this is would be 10 percent is the total coverage. So, it will be divided into 2 equal halves 5 percent, 5 percent. So, if you see 5 percent, 5 percent technically 0.05. So, for the 0.05 value it comes out to be 1.83313 utilize that value to do the calculation.

(Refer Slide Time: 18:19)



Consider the case when you want to find or do some statistical inference, but with respect to finding or something further for the variance of the standard deviation given the mean value is known. So, this is the first instance where we want to find out the interval estimation or the population variance.

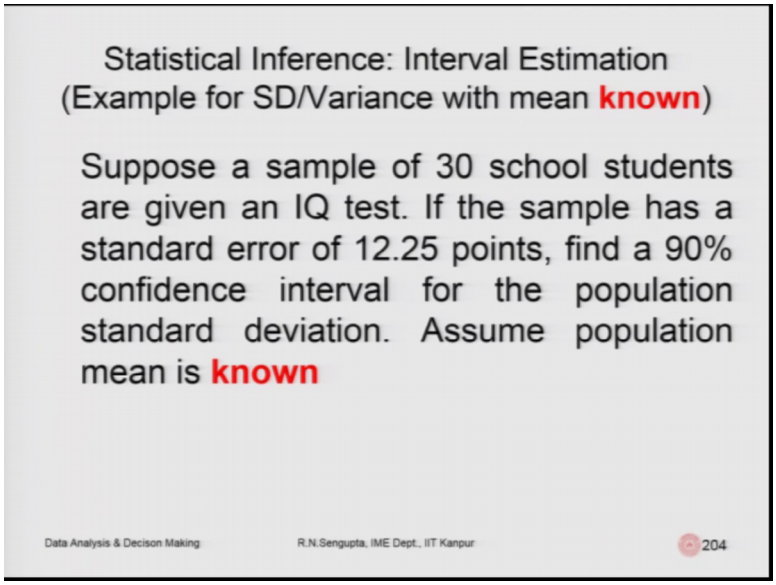
Now, if the mean value is known immediately it will occur to your mind that we should be utilizing s dash and not s, because s dash basically means that the population mean is known, so we do not use the sample mean and the degrees of freedom remains as n, point one noted. Now, when we utilize that remember that the chi-square distribution which is being formed would also would not lose 1 degrees of freedom. So, it will basically be chi-square n comma alpha by 2 and the right hand side it will be 1 minus alpha by 2 and this is the first time I am trying to basically explain why it is like this, and see that this is the chi-square.

So, you will have chi-square on the left hand side this is n this is alpha by 2, this is the second one chi-square n 1 minus alpha by 2 this whole area inside is 1 minus alpha. Now, see how the nomenclature is this, if you see the suffix it gives me gives me alpha by 2 that means, alpha by 2, alpha by 2; yes, alpha by 2 is the area on to the left of the chi-square first value and 1 minus alpha by 2 is the overall area which is on to the left.

So, we are basically following the same nomenclature as considering what is the overall area covered on to the left. See if I see this, so 1 minus alpha by 2 technically means the whole area we have covered from the minimum value chi-square to that value where it is chi-square this whole area is 1 minus alpha by 2. And the right hand side which 1 minus alpha by 2 would be for the case wait, alpha by 2 should be 1 minus alpha by 2. This would be alpha by 2, yes, this area should be also alpha by 2 because the total sum of the areas on the left hand side the middle portion and the right hand side should be alpha 1. So, I will basically highlight this in more details because let me do the problem I will come to this conceptual framework in more details later on.

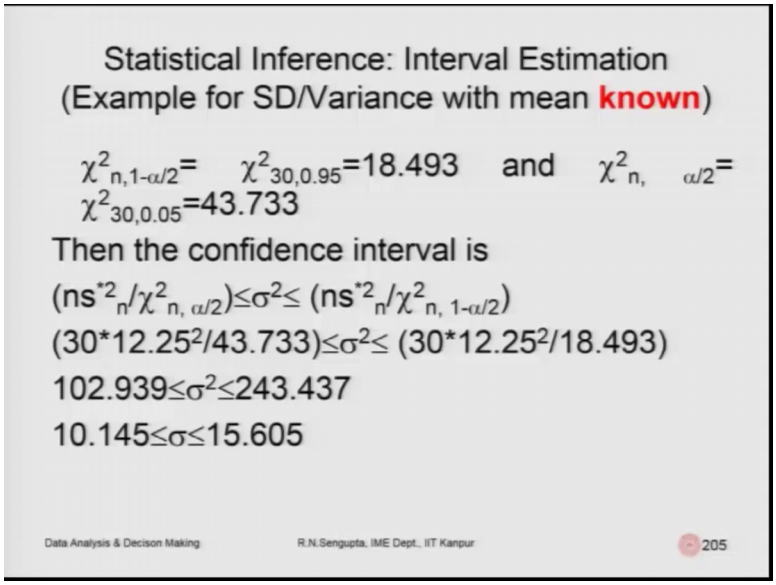
So, in this case we use the chi-square with without losing one degrees of freedom to recap important things. We will use s with the dash that means, we are not losing any degrees of freedom, so which is basically mu is mu not X n bar because the sample mean is not to be utilized. And once we have that the overall rule is probability of n s dash square by n divided by, n means the sample size divided by chi-square with degrees of freedom n and alpha by 2 b where the position and is depending on the level of confidence which we have. And the right and the right limited the UCL would be again same thing, but the chi-square value would be n into 1 minus alpha by 2 depending on the nomenclature is and that overall area would be in between would be level of confidence which was 1 minus alpha which is this.

(Refer Slide Time: 22:25)



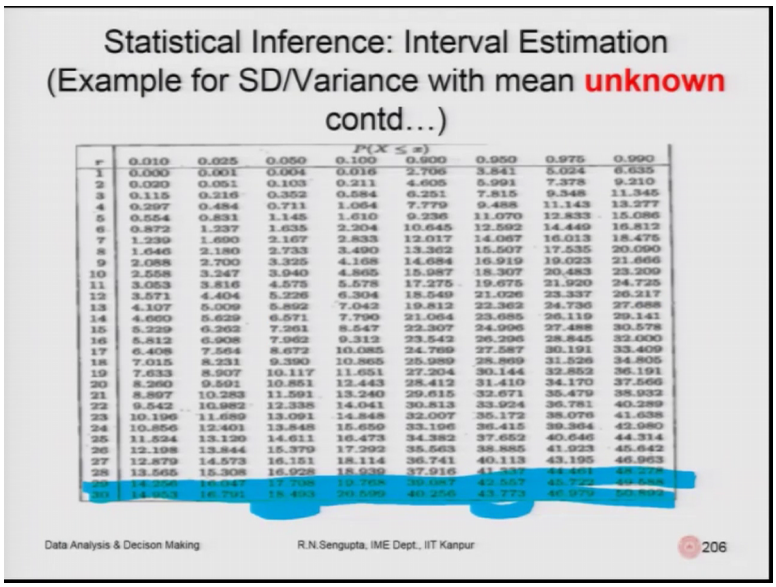
So, let us do a problem and I will highlight it and come to the basic concepts more in more details. Suppose a sample of 30 students, school students are given an IQ test. If the sample has a standard error of 12.25, so we are basically considered the stand the sample standard error is 12.25. Find a 90 percent confidence interval for the population standard deviation assume that the population mean is known. So, if the population mean is known then this 12.25 is basically s dash suffix n and being 30. That means, given the standard error for the sample and provided the population mean is known we know that is the standard error such that we have not lost any degrees of freedom.

(Refer Slide Time: 23:24)



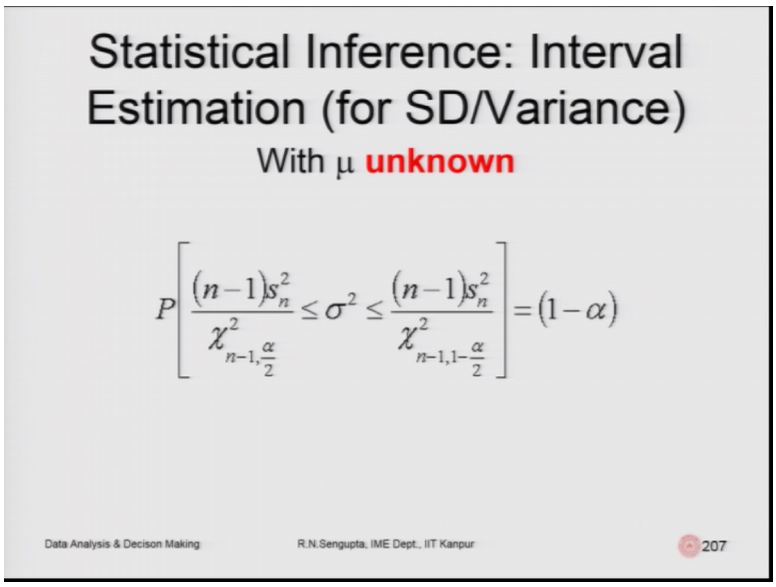
So, in this case continuing from the table we will come to the table, chi-square n and 1 minus alpha by 2 this is beings on the right hand side would basically be chi-square, what is n? N is 30 because you are not losing any degree freedom and 1 minus alpha by 2 basically be 0.95 that comes out to 18.493. And chi-square n alpha by 2 on the left hand side basically comes out to be 43.75. Then the confidence interval is found out in the same way as we have just noted down, and the confidence interval comes out to be for sigma comes out to be 10.145 and 15.605.

(Refer Slide Time: 24:02)



So, this is the value of 30 and this is the value of 29 we come to that later on. So, here if it is basically 30 let me go back what was we wanted we wanted a value at n is equal to 30 and 0.05; and 0.95; 0.95 is this 43.75 and 0.05 is basically 18.493. Check these values and this is basically 18.493, and 43.733 you utilize these values to solve your problems accordingly.

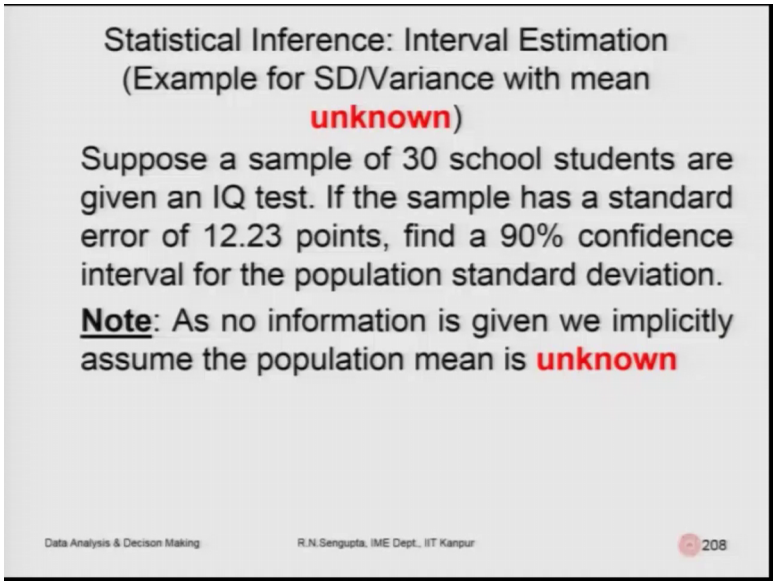
(Refer Slide Time: 24:50)



Now, let us consider the problem. The same time, but with the population mean being not given. So, consider that statically inference problem is their interval estimation is their for the standard deviation of the variance, but with the population mean being unknown. So, here again there we loose 1 degrees of freedom because the population mean is not known we use the sample mean to find out the population mean as we do that we have a chi-square which has 1 degrees of freedom less hence it is n minus 1.

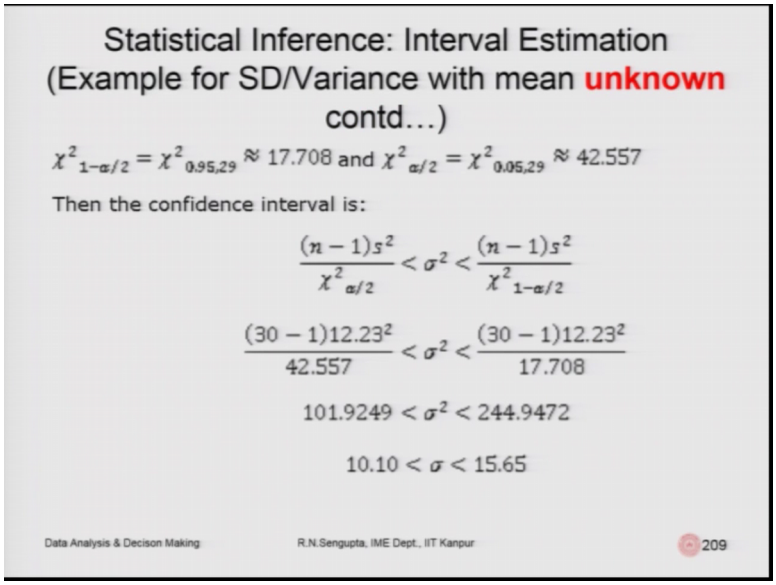
In the numerator in place of s dash it now would be s without the dash because you are utilizing the different standard error where mean values of the population are not known and repeating things time and again please listen to me carefully. And obviously, in place of n which was pre multiplied before s it will be n minus 1 depending on the formula. So, you one can check any good book for the formulas. So, the formulas are basically to be utilized not to be derived.

(Refer Slide Time: 25:53)



Here again the same problem suppose a sample of 30 students exactly if the sample has a standard deviation of 12.23, now it is not 12.25. Find a 90 percent confidence interval for the population standard deviation again and now nothing is mentioned about the population mean. So, as no information is given we implicitly assume that the population mean is unknown.

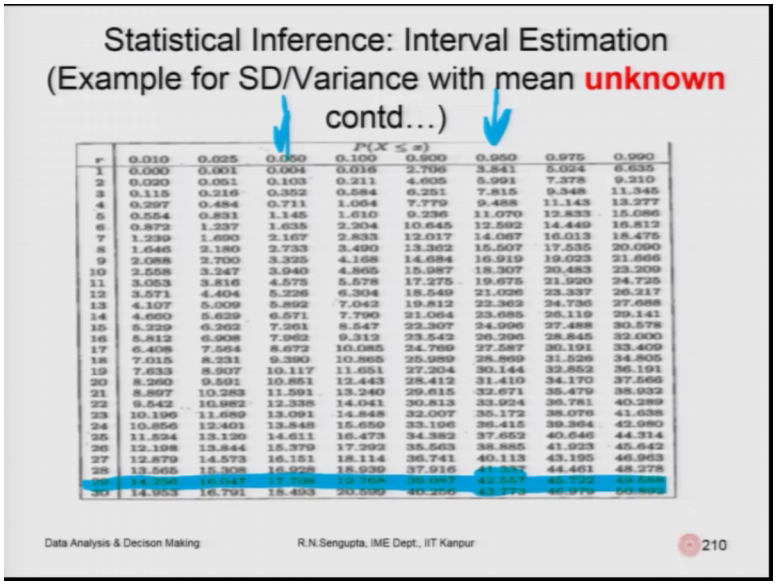
(Refer Slide Time: 26:14)



So, here again we find out chi-square 1 minus alpha for n, and now being n minus 1 because it will be your last 1 degrees of freedom. So, it will be n minus 1 will be 29. So,

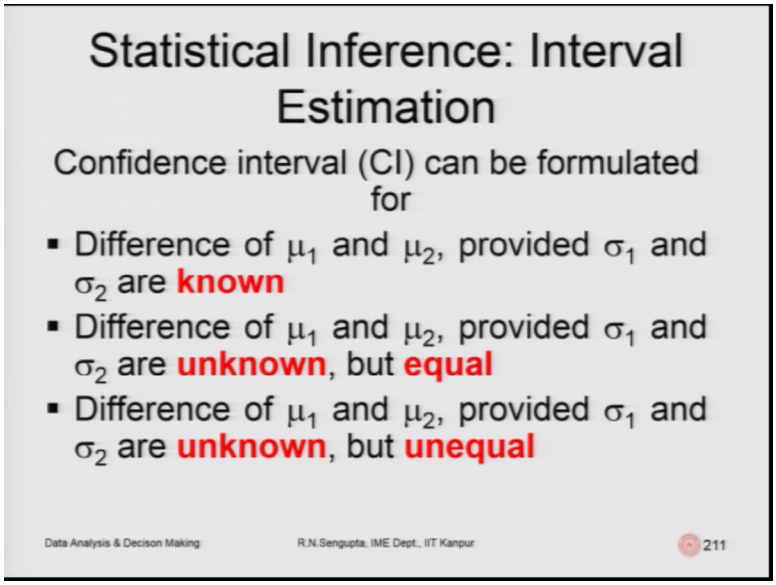
that value comes out to be for a chi-square 29 and 0.95 comes out to be 17.0708 and chi-square with degrees of freedom of 29, but alpha by 2 now being 0.05. If you look into the table I am going to come to that again it comes out to be 42.557 you utilize that in the formula and solve your problems accordingly. So, let me check the table.

(Refer Slide Time: 26:54)



So, here 29 this is 42.557 for the value of 0.9, 0.95 and the value is basically for 0.05 it is coming out to be 17.708. So, let me check, 42.557, for 0.9 for 0.95 depending on the total level of confidence being 90 because you have divided it accordingly. So, you can use these values to solve it.

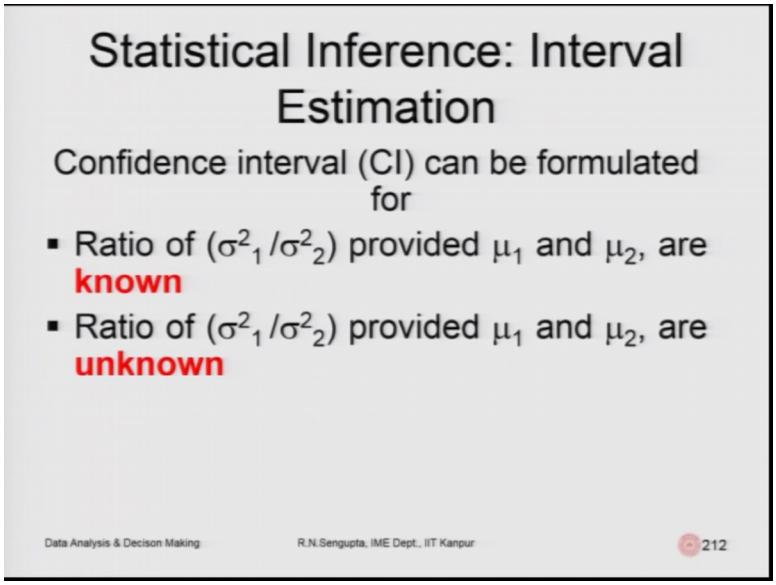
(Refer Slide Time: 27:35)



Statistical inference estimation can be there, confidence interval it can be there when you are trying to find out the difference of the means provided sigma 1 and sigma 2 are known. So, if it is difference on the mean and standard deviation obviously, we will use the z distribution. I am going to be just repeat it bullet points come back to the problem solving later on.

When you are trying to find out the difference of the means provide a sigma 1 and sigma 2 are unknown both being equal case an unequal case we will use the t distribution with degrees of freedom being lost, one from the first and one from the second. So, be careful it would not be now no more n minus 1; it will be n minus 2 depending on a number of odd degrees you have basically lost from the first population and the second population.

(Refer Slide Time: 28:20)



Similarly, you can use the to find of the ratios of the standard deviations or square which is the variances provided mu 1 and mu 2 is known, we will use the f distribution without losing any degrees of freedom it will be m and n. And in the case when you have basically the finding of the ratios of the variances provided mu 1 and mu 2 are unknown you will use 1 degrees of freedom for the first population and 1 degrees of freedom from the second population. So, it will be f distribution with m minus 1 and n minus 1. I will come to that in more details later on.

So, with this I will end this 17th lecture and continue more discussion about the hypothesis testing and go in to the multiple linear regression. And then definitely go into the multivariate statistics which would be to one of the main part on the discussion how they can utilize for different type of problem solving.

Have a nice day and thank you very much.