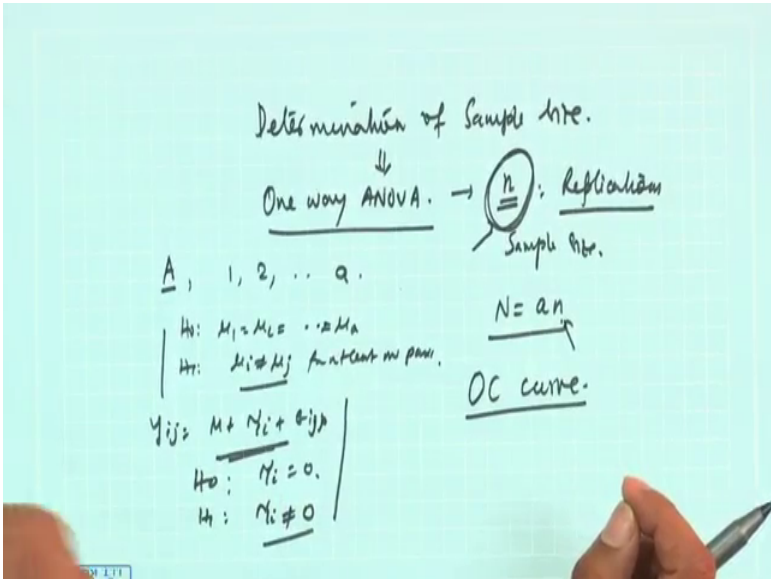
Design and Analysis of Experiments
Prof. Jhareswar Maiti
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur

Lecture – 19
Determination of Sample Size for ANOVA

Welcome we will continue determination of sample size, determination of sample size.

(Refer Slide Time: 00:24)



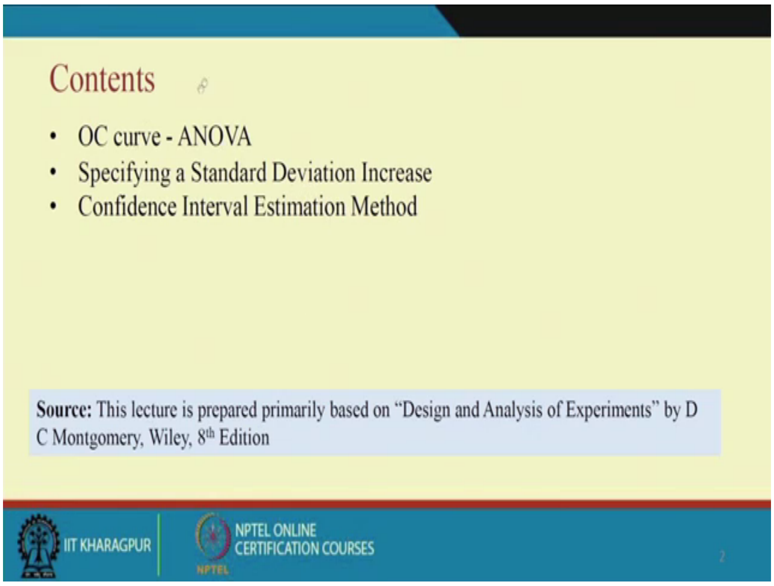
In this lecture we will consider One-way ANOVA, and we will see that how to compute representative n that is the replications. Other way we can say that sample size. So, very quickly I will tell you the one-way ANOVA part. So, you know that there will be only one factor A which will be having a levels and your H0 is mu 1 equal to mu 2 equal to mu A, all means are equal and H1 will be mu i not equal to mu j for at least one pair for at least one pair. And this a with reference to the model y ij equal to mu plus tau i plus epsilon ij with reference to tau i the treatment effect H0 tau i equal to 0 and H1 tau i not equal to 0.

So, we will do we will estimate parameters conduct hypothesis testing test of us many things, but all with reference to basically the number of data collecteds. So, here the total sample size is a in cross n, where n is the sample size across each level of the treatment lecture that in the part one of determination sample size what we have we have learned that given a single population and 2 population case how to compute standard sample size for using 2 approach confidence interval approach and hypothesis testing approach or confidence interval framework and hypothesis testing framework.

And in this case we will see that and in fact, in hypothesis testing framework we have shown you the operating characteristic curve or OC curve, OC curve where beta versus n or del d or 1 minus beta versus n or d that was plotted, and we will see the similar kind of things in ANOVA, here it is not a 2 population case or one population case as there are n number of levels it is a popul A number of population case. So, that is that is what is the big difference coming from the sample size determination of view considering the earlier lecture versus this lecture.

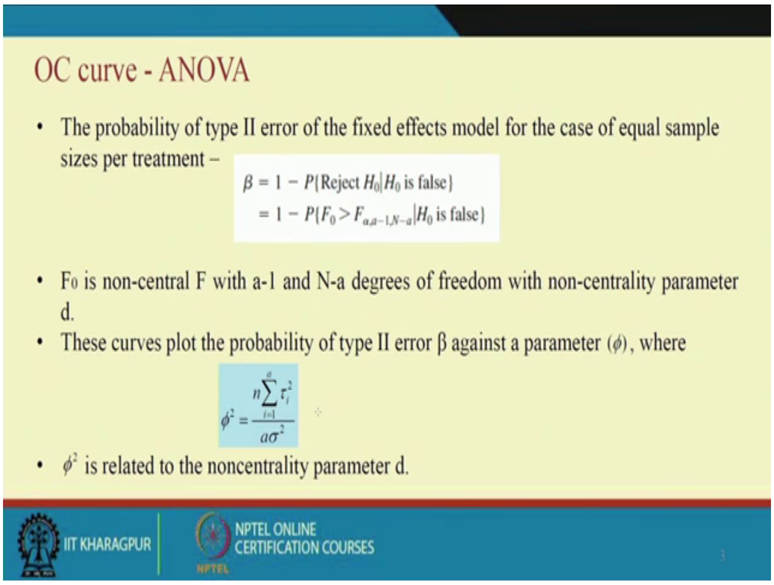
So, we will see that how OC curve can be will be used in calculating

(Refer Slide Time: 03:38)



The sample size in ANOVA, and apart from OC curve there are some other way of finding out the sample size like specifying a particular standard deviation increase and also, we will use the confidence interval method approach, what we have seen earlier. So, I as I told you earlier also that most of the lectures what I am giving to you we are heavily dependent on the book by DC Montgomery Design Analysis of Experiment this lecture also has been taken from that book.

(Refer Slide Time: 04:21)

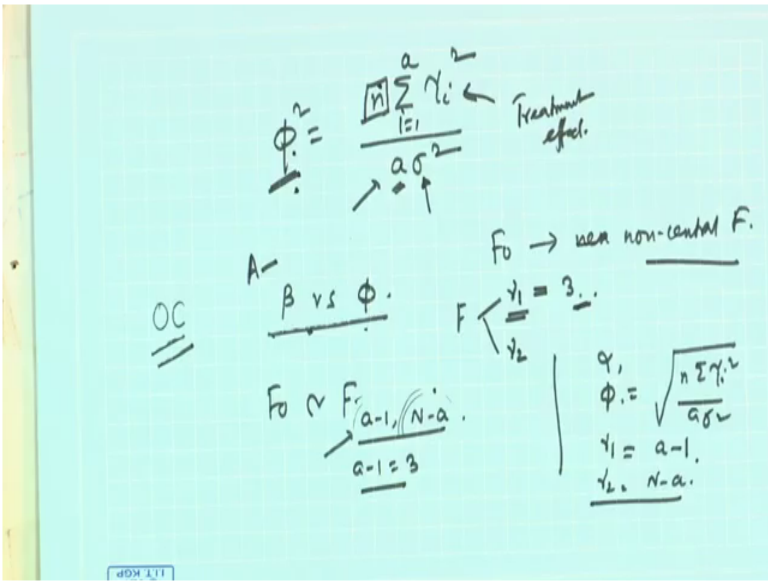


So, I have you know that what is beta. So, beta is 1 minus probability reject H0 given H0 is false, or 1 minus probability that F0 greater than F alpha N minus N F alpha a minus 1 N minus a given a 0 is false. So, you know how to come do this hypothesis testing part. Interestingly what happened when H0 is true this F0 follows F distribution with a minus 1 n minus 1 degrees of freedom, but that will be is that is that is with central non-centrality parameter d equal to 0.

But when H0 is false means; that the means are not equal there is some difference between the mean some means shift is there if I consider the difference is a shift. Or other way I can say that from one mean to another mean there is a there is there is there is difference, and then what will happen this quantity when H0 is false F0 will beca will become non-central f or it basically follows non-central F distribution with a minus 1 and N minus a in numerator and denominator degrees of freedom.

And here the non-centrality parameter is d. So, this situation can be plotted. This situation can be captured using a curve called type 2 error beta versus a parameter phi where the phi is defined like this phi square equal to N sum total of I equal to one to a tau i square by a sigma square. So, what is phi square? Phi square is

(Refer Slide Time: 06:29)

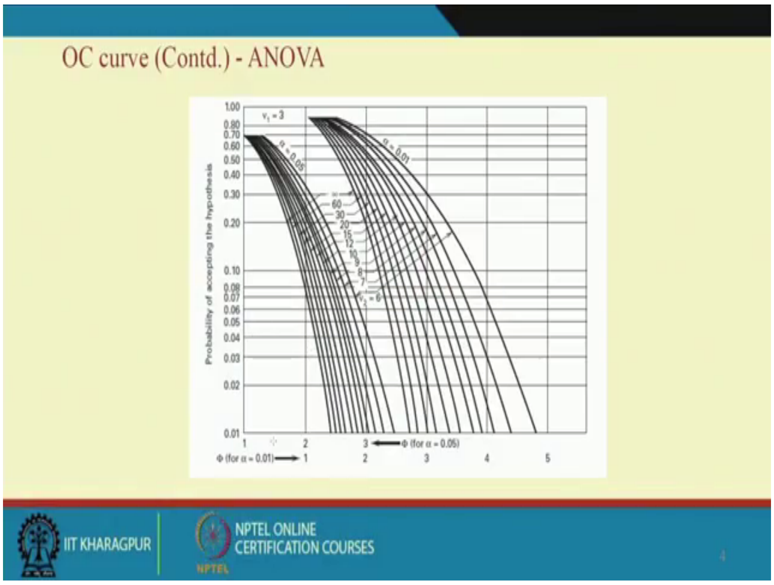


N sum total of tau i square i equal to 1 to a divided by a sigma square.

So, you know a this is the number of levels the factor a has you know sigma square that is the variability or variance of the response variable, and you know tau i this is the treatment effect, treatment effect. So, what we require we require to know a representative n. So, fine if I consider a particular n you will be able to compute phi square once you able to compute phi square. So, there is OC curve which basically with reference to beta versus phi.

So, beta versus phi and here you see that phi is dependent on n. So, there also beta versus phi curve will tell you for a particular threshold or if you consider a beta threshold beta how what will be the sample size and how it is to be computed. So, that is what we want to know here

(Refer Slide Time: 07:55)



So, one such operating characteristic curve is shown here you see carefully the curve what is x axis? In x axis there are 2 things phi for alpha equal to 0.01 and phi for alpha equal to 0.05.

So, what is alpha? Alpha is type one error; that means, probability of rejecting a true null hypothesis. Now what is y axis? Probability of accepting the hypothesis when the hypothesis null hypothesis is wrong that mean that is beta. So, y axis is giving you beta, y axis is giving you beta and x axis is giving you phi where phi is phi square equal to n sum total of tau i square by a sigma square for different values of alpha. So, essentially alpha is 0.01 or 0.05 considered because these are the 2 level of significance.

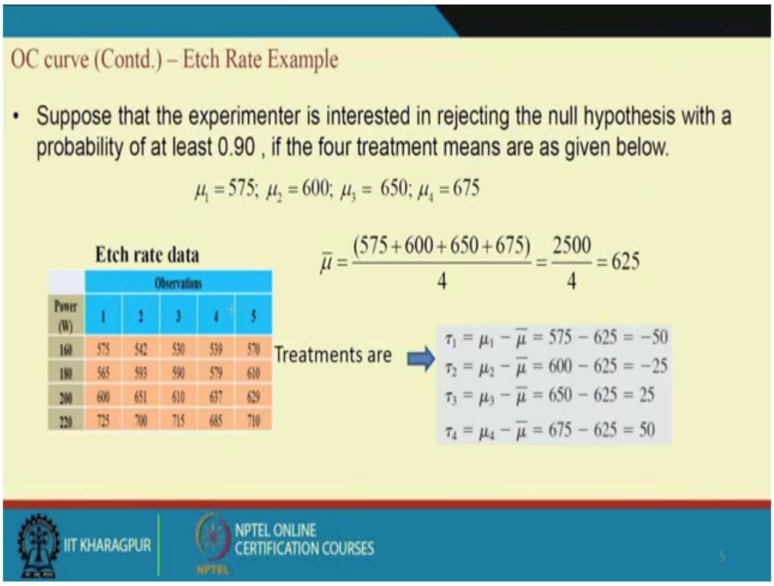
We consider in most of the experimental studies. Now so, x axis and y axis is known to you now, what is this? This nu 1 equal to 3. So, all of you know this is related to because we say that if H0 is false then the F0 the test statistics that is non-central non-central, F non-central F. Now F has numerator degree of freedom and denominator degree of freedom numerator nu 1 denominator nu 2. Now this nu 1 is numerator degrees of freedom. So, it is 3 means an numerator degree of freedom usually here in case of a this F0 this follows F a minus 1 and n minus a degrees of freedom where a is the number of levels the factor assumes and N minus. So, this is basically with reference to the treatment and this is with reference to the error.

So, a equal to nu equal to 3 means a minus 1 equal to 3. Then you see there are there are for alpha equal to 0.05 there a large number of curves are there. These so many curves and here it is written nu 2 equal to 6 7 8 9 10 up to infinite irritating. So, that mean each of the curve represent that the relation between beta and phi for given alpha given nu 1 and given nu 2. So, you must know what is your alpha value? What is your phi value? Phi will be computed by using this equation and then you must know what is nu 1. Nu 1 is in case of ANOVA 1-way ANOVA it is a minus 1 what is nu 2 n minus a.

And then using this information you will be able to find out the beta value. Just for a for sake of calculation suppose if I say my alpha is 0.05 then this is the curve these are the curves what you will consider and suppose phi value is coming out to be 3. And let your nu one is 3 and let nu 2 is, let nu 2 is your this let it be nu 2 is 10, 10 means which one 10 is this last this one.

So, if 10 is this one mean this one this one. So, under this situation if phi is 3 you go up it is meeting the curve here then project towards left and you will be getting a value of this which is maybe 0.013. So, this is your beta value. So now, this is the way you have to read the OC curve.

(Refer Slide Time: 12:45)

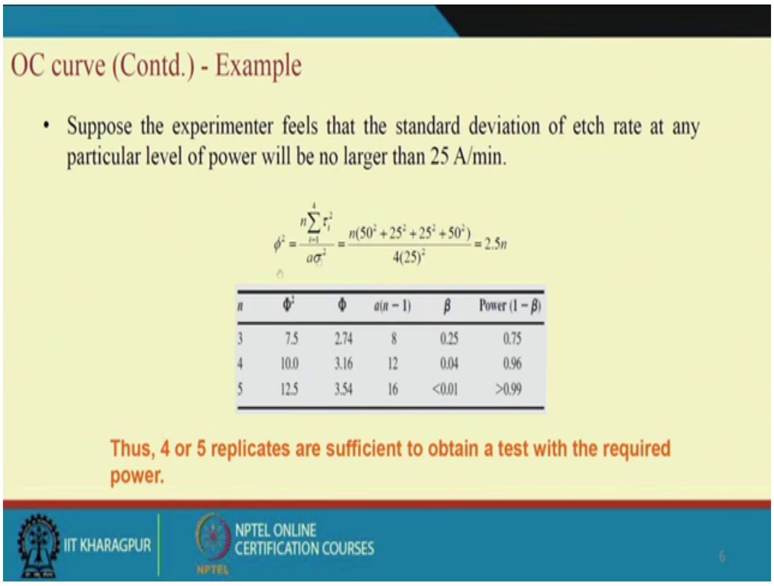


Now let us see the example. This example we have solved earlier this is the etch rate example where different the power levels are considered one 60 to 220 with a gap of 20 each, and then 5 number of observations were collected in the same experiment we are done and this was the data available. So, that you have already seen.

Now, let us hope that for these particular case the 4 mean values population mean when a when power is at 160 180 200 and 220 the population means are the treatment means are known that is mu 1 575, 6 mu 2 600, mu 3 650, mu 4 675 then if this is known and then mu bar will be this 625, and given the data you are in a position to calculate tau one tau 2 tau 4 and these are all population tau one tau 2 tau 4 because you have mu 1 mu 2 mu 3 all are known.

So, mu 1 minus mu bar is 50 mu 2 minus mu bar is 25 minus 25, 25 50 like this. So, given this tau 1 tau 2 tau 3 and tau 4. So, you are in a position to compute

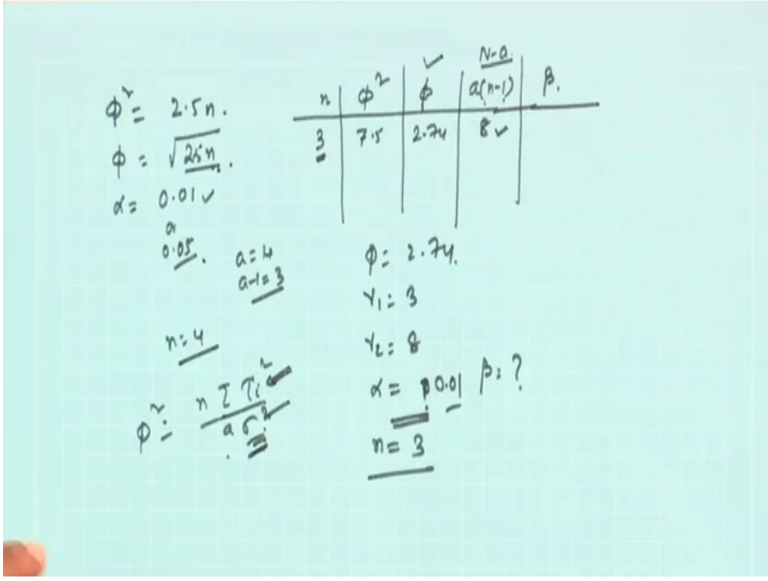
(Refer Slide Time: 14:24)



Tau i square and a is known a is 4 sigma square is given which is no longer than 25 angstrom per a per minute angstrom per minute this is angstrom. So, then phi square is coming as phi square equal to 2.5 n.

So, what do you require you require to know what should be the sample size. So, that your power test is at least 0.9 means beta is one or less. So, under such condition what you how do compute how do you go about your case is you first know 2 things first you calculate phi square which is 2.5 n.

(Refer Slide Time: 05:21)



Then you call you know phi equal to root over 2.5 n, this is phi once you know the phi you say what is your alpha value if you consider alpha may be 0.01 or 0.05 or anything else provided that OC curve is available.

So, then what do you how do you calculate you first choose n. So, n if you consider 3 then what is phi square value 2.5 into 37.5. So, this is 7.5 what is your phi value square root of 7.5 is 2.74, then what is you are a into n minus 1 in n minus a value. So, what is a? A is 4 this is 3. So, n minus 1 is 2 into 4 this will be 8. So now, you are in a position to use the OC curve because you can fix alpha some value you phi is known also and n minus a is 8 and you're a minus 1 is 3.

Because here a equal to 4. So, a minus 1 is 3. So, all relevant information to use the OC curve is available now use this one. So, what will be then the value of beta. So, let us see the OC curve no not this one let us see the OC curve here. So, suppose we consider now let me also check that what will be is our alpha value you have consider we have considered 1 minus beta is 0.90 then you will be accepting it and an alpha value we have not written here. So, let alpha equal to 0.05 what will happen let us see.

But I have computed this values. So, beta is 0.25 anyhow, then I know that what alpha I have consider. So, phi is 2.74 a is a nu 1 is nu one is. So, our case is phi is 2.74 nu 1 is 3, nu 2 is 8 and we require to know alpha what because alpha I have not consider, but we actually we want to know beta that is beta is what? And n is 3 you see nu 1 is 3. So, this is the curve then our nu 2 is 8 later let me check this one alpha equal to 0.05.

So, nu 2 equal to 8, 8 mean this one, this one is nu 2 and then our let alpha equal to 0.05 fine and then what is the phi value? 2.74. So, 2.74 somewhere here. So, whatever is the 8 this is 8 suppose 2.74 somewhere here somewhere here what is the beta value beta value is coming in between this.

Suppose if I consider alpha equal to 0.01 then what is the situation one to 2.74 means somewhere here and your 8 nu to 8 means this one and your alpha given nu 1 nu 2 given phi is 2.74. So, phi will be 2.74 somewhere here if we see that 8 means this is the curve. So, somewhere here we are we are coming like this so; that means, we are our beta is 0.25. So, we are considering alpha equal to 0.01 in this case now it is clear.

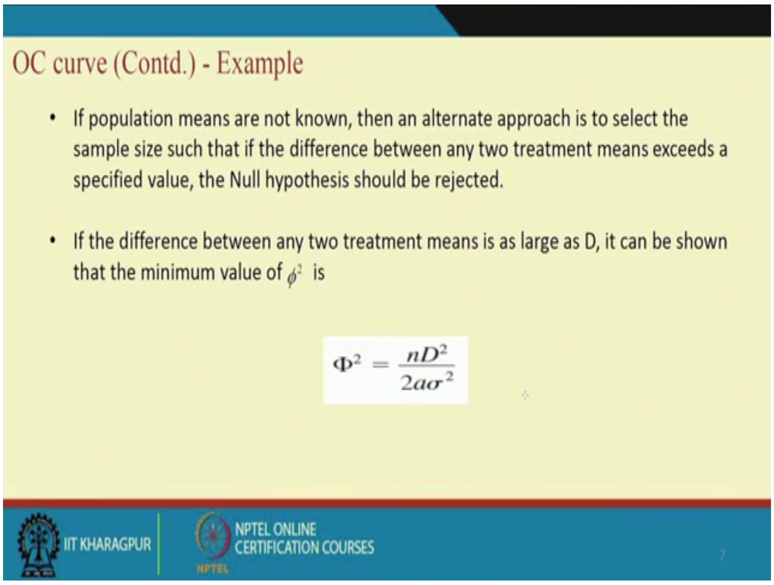
So, 0.01. So, given alpha 0.01 phi equal to 2.74 a equal to 3 a nu 1 equal to 3 and nu 2 equal to 8 you are getting beta equal to 0.25 somewhere here it will go to here then this is 8, 8 will be this one and then when you when you come down you see that it is matching around 2.74. So, it is a there is some approximation obvious visually we are seeing. So, 2 what happened this is 0.25.

Then what is one a power of 0.75 what is our condition? We want that the reject null hypothesis with probability at least 0.90. So, if we consider n equal to 3 we our power is 0.75 which is below 0.90. So, 3 is not sufficient now you consider 4 if you consider 4 phi square become 10 and phi become 3.16 a into n minus 1 that is nu 2 become 12 nu 1 is always 3.

So, that alpha is 0.01. So, using all those things you will find out that beta value will be 0.04. So, how nu 2 equal to 12. So, nu 2 12 means this one this is basically nu 2 equal to this, this one is 12 and then what is the phi value here? Phi value is 3.16. So, nu 2 12 this one phi value 3.16 somewhere here then maybe you will when you go up and project this side this side you will be getting the value 0.04. So, this is one then 1 minus beta is.9 6 it is it is more than.9.0 So, this is sufficient.

Now, you check one more if you take 5 it is greater than 0.99. So, 44 replications. So, n equal to 4 is sufficient how many we have considered 55 replications. So, this is the way you.

(Refer Slide Time: 22:22)

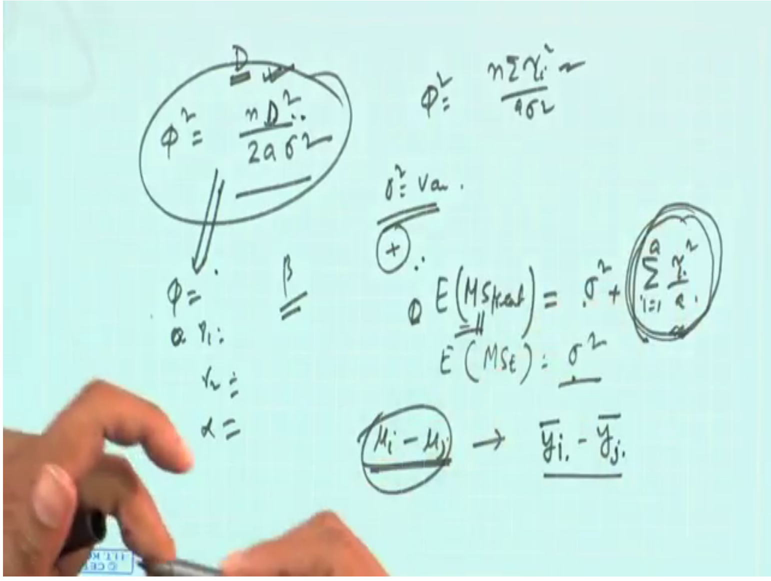


Will be using OC curve in ANOVA, but up to this discussion for in using OC curve what you have use why we are computing phi phi square equal to n sum total of tau i square by a into sigma this is what you are you are using let me let me go back a sigma square, a sigma square that is what we are using.

Now, that mean you require to know the tau i. So, tau i is population parameter. So, if you estimate and then you can find if the estimation is significant representative one, but we only know when it will with our estimate will be representative provided the sample size is representative we are computing sample size. So, that we cannot say that is the representative one. So, that mean actually what you require you require to know the population tau it is not possible.

Under such situation what is done the a the difference between 2 population means for difference between 2 public a threshold is considered that what should be the that the may be allowable differences between any 2 population means if we consider that one is d, then the theory says that suppose difference

(Refer Slide Time: 23:53)

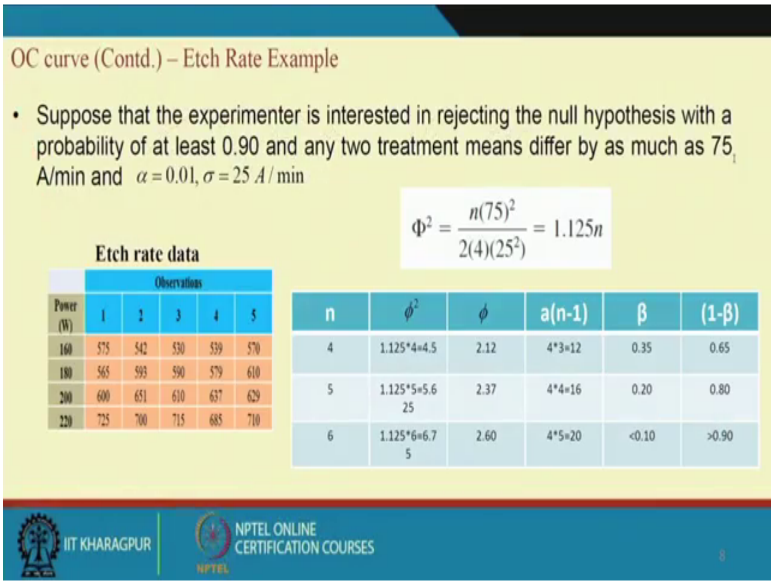


Between any 2 population means the threshold value should be D, that mean should be D or less or the difference should be as large as d, under such situation this pi square calculation become n D square by 2 a sigma square.

So, instead of using phi square equal to n some tau i square by a sigma square you can use this equation calculate and then approach of finding out what should be the value of n will remain same the way we are we have discussed the only difference here is that instead of tau I square you are you are using a pre pacified difference D and then find out what is the value of phi once you know phi you know you know nu 1 you know nu 2 and your alpha is given. So, you will be in a position to find out beta and then you will be able to find out the you can see the observe that whether it is satisfactory or not.

So, this example suppose

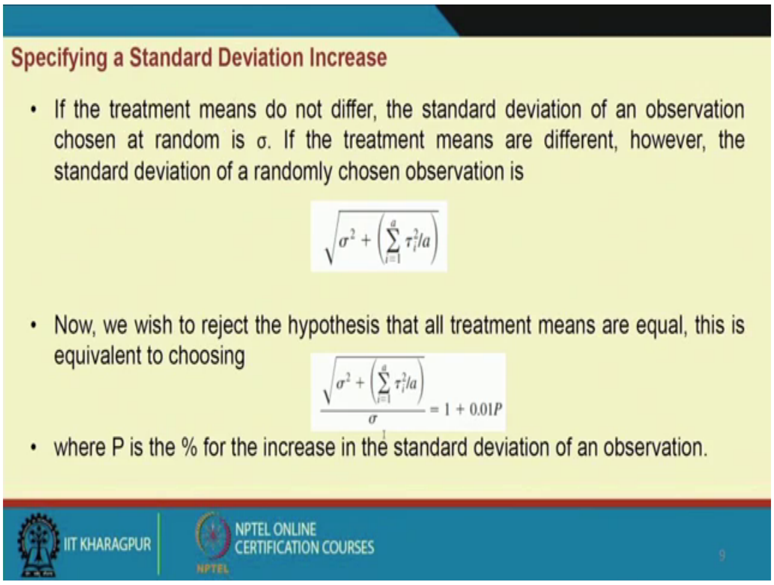
(Refer Slide Time: 25:10)



If you say that the difference maximum difference 75 is acceptable then if you use this formula you are getting phi square equal to 1.125 n it is not matching with the previous 1 the reason is we have considered 75 is the maximum difference that is the threshold value. So now, if you follow n equal to 456 and then phi square phi and this is nu this is nu 2 and then beta value from the OC curve and then 1 minus beta and you see that you require 6 at least 6 number of replications or the sample size should be 6.

If we consider this then the 5 observations are not sufficient you have to go for 6 observations.

(Refer Slide Time: 26:02)



So, this is what using OC curve the second one is there suppose no we will use another criteria is that you are pacifying that what is if the sigma square is the is the variance of y, suppose you will you will accept certain increase of variation and that one you specify then what should be the sample size for that particular that percentage increase in variation that is known as specifying standard deviation increase.

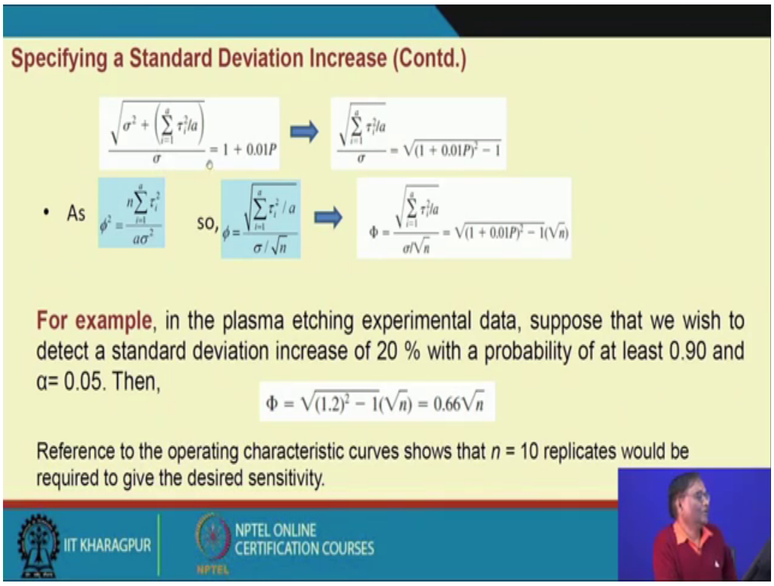
Now, you were seen that if there is a different there is treatment difference treatment difference is there then what will happen suppose that if I want to know MS treatment what is the expected value of MS treatment this will be sigma square plus sum of I equal to 1 to a tau i square by a. Then if an expected value of ms error is sigma square that is we have seen earlier.

So, if there is no treatment difference then ms treatment expected will these 2 expected value become equal. So, if there is really difference is there then this become significant. So, if this increase what happened from sigma square to sigma square plus this, this is the increase. If this increase is significant then there is significant difference between the treatment means. So, what is that for we want to we want to get this a pre specified increase and what should be the value of n.

So, that that increase can be can be understood or it can be when it this entities can be computed or explained. So, then what is the increase is that if this is the this is the with increased variability this is the case if you divide by sigma then this is 1 plus 0 1 p p is the what is the percentage increase. So, this will be pre specified this will be specified.

Now, from here you will know what will be the value of n how,

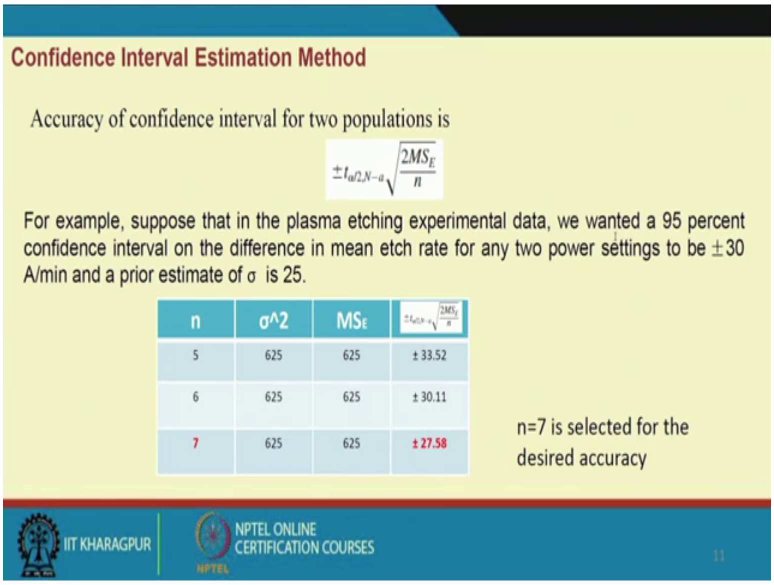
(Refer Slide Time: 28:41)



Because you know phi square equal to this and from the specification of standard deviation increase using this formula you will be able in a position to calculate i equal to 1 to a square tau square by a. So, this quantity is this now phi square equal to n into this means phi equal to this then phi equal to this if I put this quantity as this then this is nothing but this value. What do you? Do you find you choose a one n find out phi and then see the table OC see the OC curve and you see that whether the required beta value you are getting or not whether you are achieving the power 1 minus beta equal to 0.9 or not.

Obviously you specify a particular alpha you let it be 0.05 then. So, long you are not getting the required power you increase and from for this example earlier example flood matching example we found that if n equal to 10 then it is giving the required percentage increase in standard deviation which we pre-specify as 20 percent.

(Refer Slide Time: 30:07)



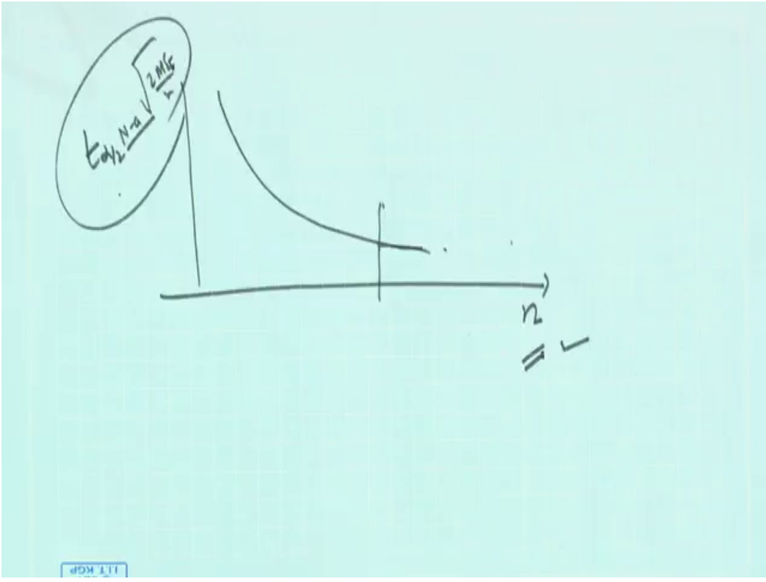
And then the last one is the confidence interval approach. So, what you are doing here I put that mean your mu I minus mu j that difference you are calculating and the random variable is y one dot y I dot bar minus y j dot bar this is the random variable.

So, if the expected value of is this variance is nothing but 2 MEe by n and that difference will for actually the margin of error will be plus minus t n minus a alpha by 2 root over 2 mse by n this one. So, what do you do you first consider one n find out this value and then find out the second n find out this value and like this. So, suppose the margin of error is given a particular value then you see. So, long the this this margin of error is less than that particular less than that particular value, unless than equal to that particular value you go on increasing the number n values and that is what is done here.

For this example we wanted to be 95 percent confident that that the each rate or any 2 power setting be plus minus 30 angstrom per minute, and a priority estimate of sigma is already known which is 25 then using the formula earlier given that is the this margin of error given sigma square mse then this value you calculate if your n is 5 it is 0.33 which is more than 30 n 630.1 one more than 37 it is less 27.58 which is less than 30 well.

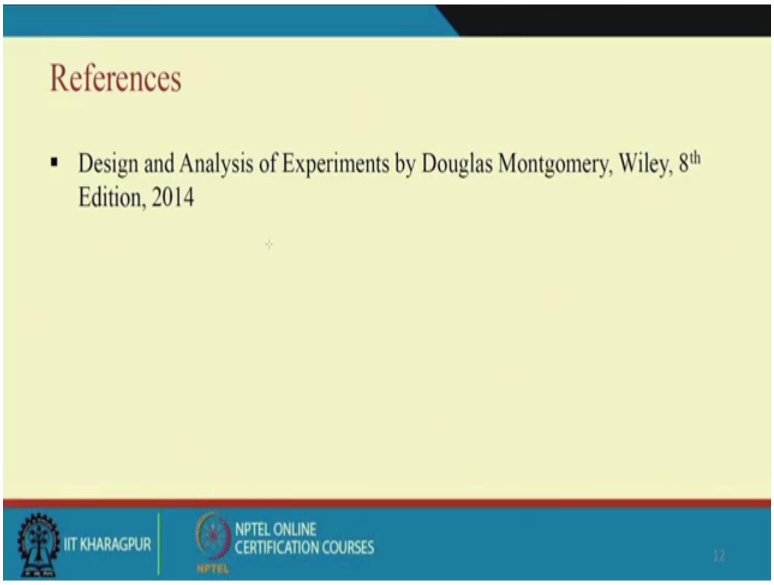
So, say see n equal to 7 is selected for the desired accuracy. So, you can other way you can plot you can have a plot if you if you if you are you will have a plot

(Refer Slide Time: 32:19)



This side n and this side t alpha by 2 n minus a and root over MS E by n if we write like this what will happen. You see that that this value is gradually decreasing as you are increasing value of n, but here we have pre-specified the value 30. So, that is why where it is becoming lower than 30 you are choosing that n if nothing is specified then you see that when it is becoming stagnant, when even if you increase the n, this quantity is not increasing. So, this way you can find out the sample size sample size calculation is very, very important.

(Refer Slide Time: 33:19)



And in all other say in the subsequent lecture. So, whenever when we will be discussing with the CRD we have discuss suppose we will discuss RCBD and other thing then factorial design all the end there the sample size calculation is important and we assume that you know the how the how OC curve will be used or some other criteria measures will be used to find out the sample size.

Thank you.