INDIAN INSTITUTE
OF
TECHNOLOGY
KHARAGPUR

NPTEL
National Programme
on
Technology Enhanced Learning

Applied Multivariate Statistical Modeling

Prof. J. Maiti
Department of Industrial Engineering and Management
IIT Kharagpur

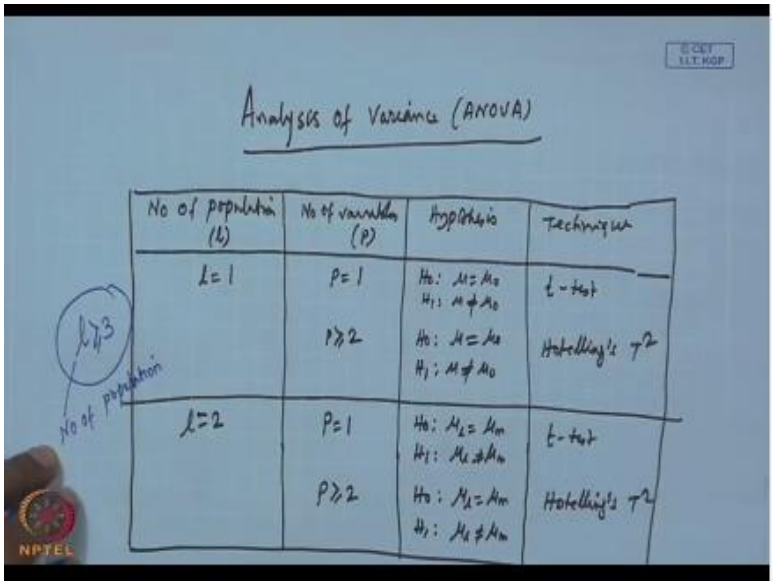
Lecture – 14

Topic

Analysis of Variance(ANOVA)

Good afternoon. Today we will discuss Analysis of Variance.

(Refer Slide Time: 00:24)



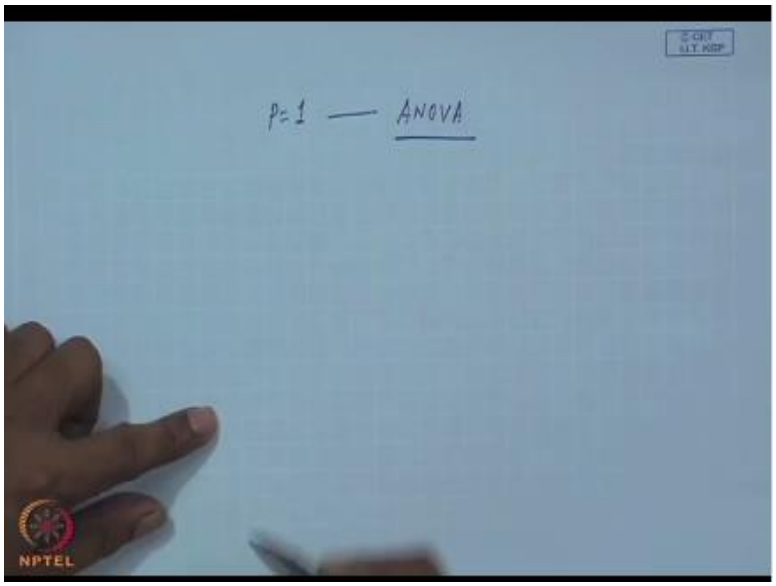
Popularly known as ANOVA. In last class, we have discussed about the difference between two population means and in hypothesis testing we have covered the equality of population means from univariate point of view from multivariate point of view. So if I just recapitulate this what we will find out that, one is number of population, then your number of variables, then hypothesis, then the techniques used, technique what is, what technique is used? If you say number of population l number of variable p and hypothesis as usual.

Then the situation is single population l equal to1, suppose number of variable is also one p=1, what we do in hypothesis testing? We do H0, $\mu = \mu 0$ and H1 $\mu \neq \mu 0$ generally we test this and we use for the small sample case t- test, we assume that population is normal. Now l can be 1, and p can be more than 2 all case also that is single population more than or equal to two variables.

And your hypothesis will be again H0 $\mu = \mu 0$ here μ is a vector quantity and H1 $\mu \neq \mu$ 0 and we have used Hotelling's T^2 for hypothesis testing again for small sample case. Now second issue we have discussed l=2 two population case, under this also we have already discussed p equal to one case and there our hypothesis was $\mu l = \mu m$ and alternative hypothesis we have framed as $\mu l \neq \mu m$ for at least one pair.

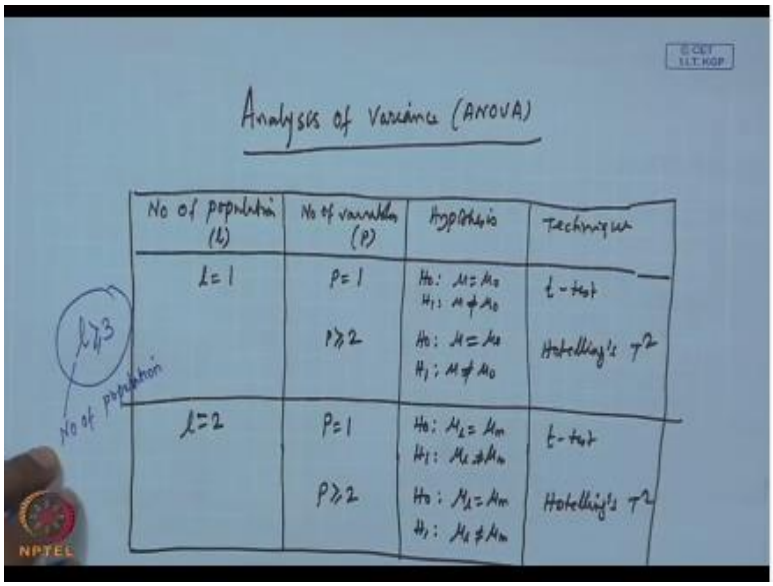
And we have used t- test and for $p \geq 2$ case we have hypothesis like this that $\mu l = \mu m$ here μl and μm are vector quantities and H1 $\mu l \neq \mu m$ for at least one pair. We have also used Hotelling's T^2, okay. So this is what from hypothesis testing point of view we have already completed. What will happen when l is three or more? So our discussion today for $l \geq 3$ case where l stand for population, that is number of population, okay. So, in this case we will first consider the univariate issue.

(Refer Slide Time: 04:39)



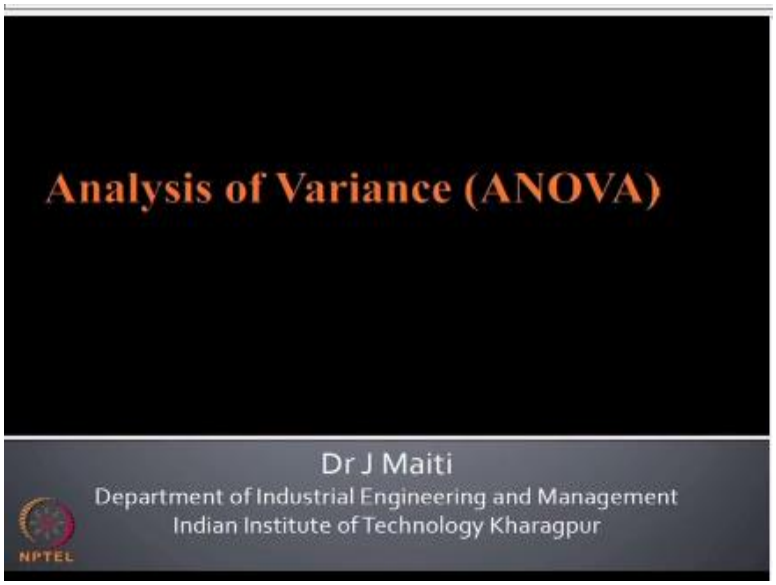
That means we are interested in p equal to 1 case and when we test equality of several population means for one response variable the technique used is ANOVA analysis of variance. That mean what I said that suppose you case is p = 1 that is univariate case there are several groups or population which is definitely three or more, in that case the first t- test is not applicable.

(Refer Slide Time: 05:21)



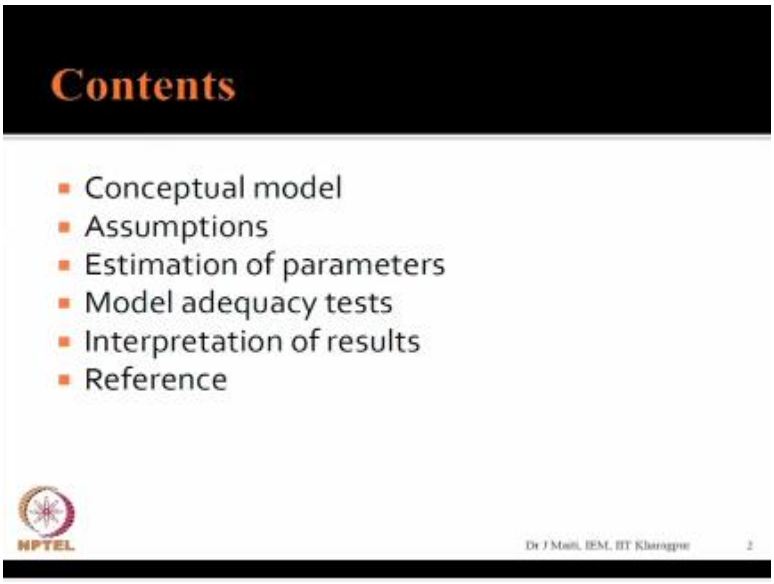
Simultaneously if you want to test the difference.

(Refer Slide Time: 05:33)



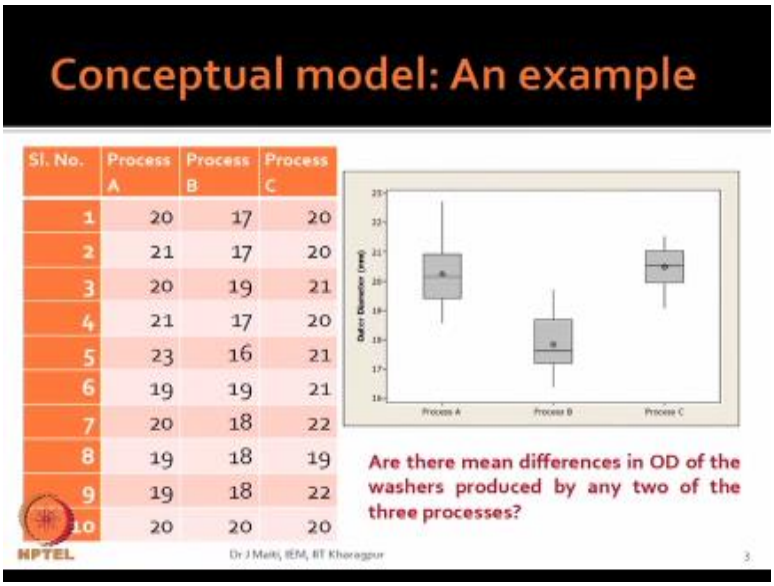
You required to use a special technique known as analysis of variance.

(Refer Slide Time: 05:34)



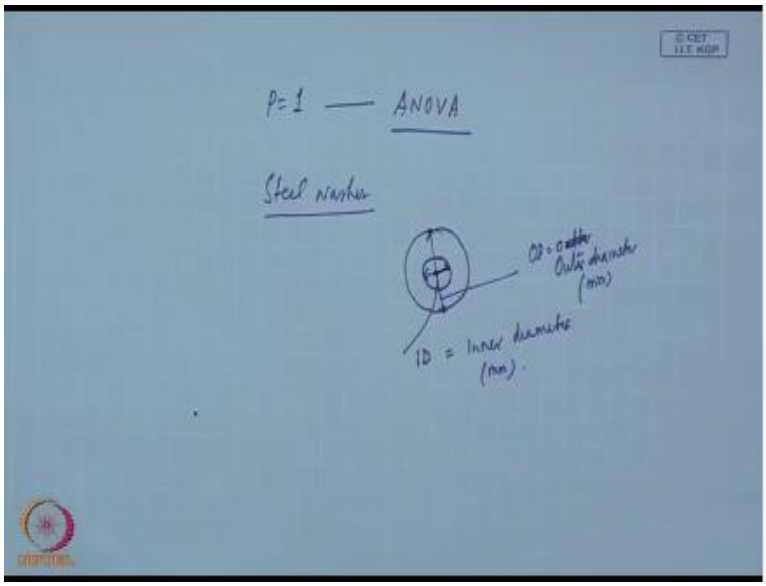
So today we will see in terms of ANOVA for ANOVA these are the contents conceptual model, assumptions, estimation of parameters, model adequacy tests, interpretation of results, followed by references. What is this conceptual model? I will be describing with an example.

(Refer Slide Time: 05:56)



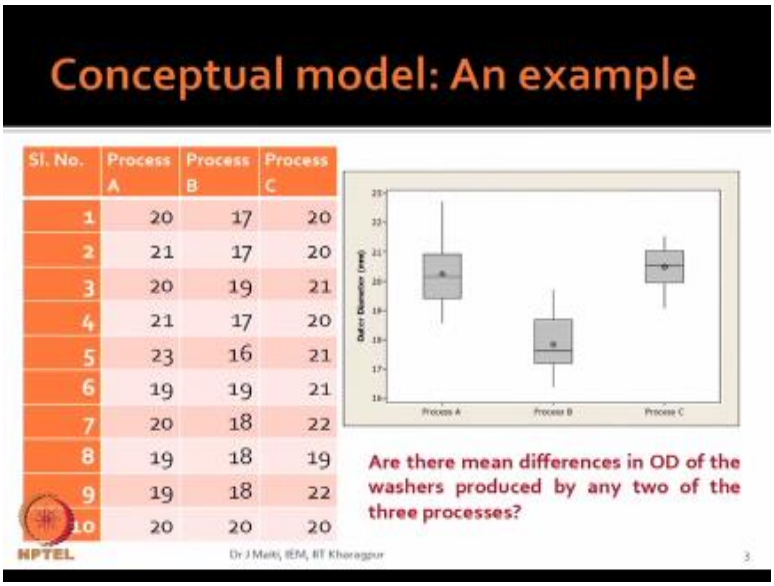
You see this slide here we have three processes A, B and C this three processes are producing let steel washers. So our product is steel washers.

(Refer Slide Time: 06:16)



Steel washer is the product which can be in 2 dimensional figures. We can see like this, suppose this is the case and then this part or otherwise this inner circle part if we consider this is ID inner diameter and if you consider the outer circle this will be OD, which is outer diameter outer. Let both are measured in millimeter is this case.

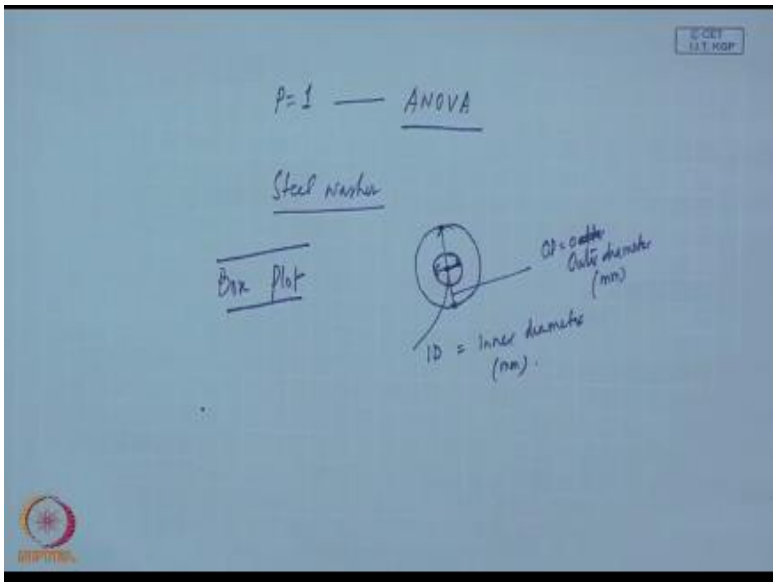
(Refer Slide Time: 07:04)



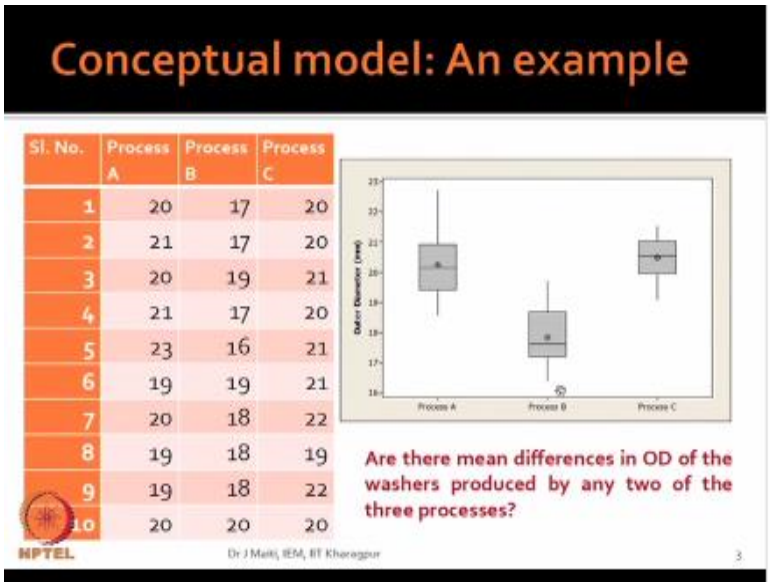
Here these values are in millimeter for outer diameter, the problem is that you are producing through three different processes but you are producing same thing that is steel washer and outer diameter is the quality variable and which is specified by the customers, okay. Now as a manufacturer of this steel washers what you require to know, you require to know that whether all the processes are producing at the same level from quality point of view same level or not.

And the manufacturer is interested to test the differences in means of outer diameter for the three processes. So our objective here is are there mean differences in outer diameter of the washers produced by any two of the three processes, okay. Under such situation the most widely used graphical plot is box plot, okay.

(Refer Slide Time: 08:24)



(Refer Slide Time: 08:34)



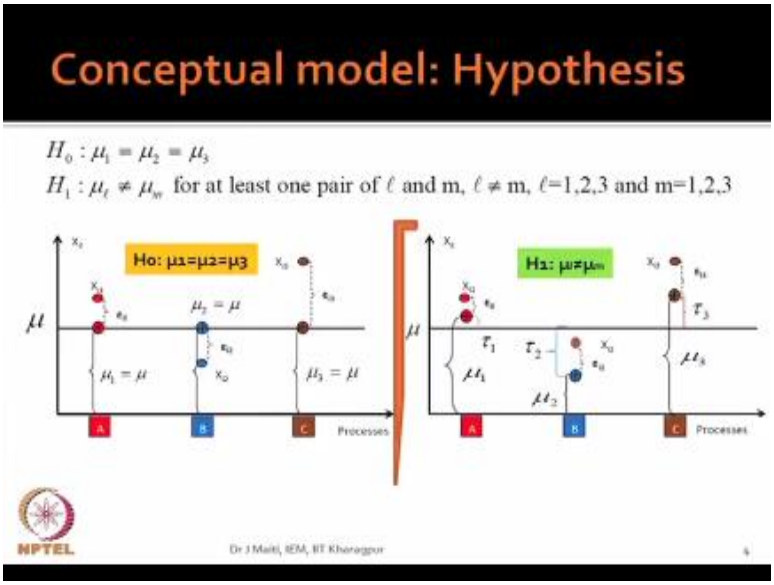
Now see what is box plot this figure shows box plot this box plot is for process A, this is for process B and process C. In the box plot there is one box and two whiskers this one is box the upper portion this straight line is right whisker and this is left whisker, this box is determined by inter quarterly range IQR and the horizontal middle line here this is the median position from the data and the circle plus one that is the mean position, okay.

So if we see the three boxes here for process B and process A as well as process C what we will find out that the box hardly differs in terms of variability except process C but it is to be proven that whether this variability is different from A or B or not apparently it is different, okay. But if you see the mean point the circle plus point this one is quite different apparently it shows that there is mean difference.

You see this is the mean point for process A this is mean point for process B this one is mean point for process C and if I see the value it is 20 point something this one is around 18 and this one is again 20 point something. So even if there is no difference between maybe process A means versus process C means but there is chance that both process A and C means are different from process B.

Which is basically from the graphical plot it resembles like this, but what you want you want a quantitative explanation of this and If we are able to do this then we are able to answer this question are there mean differences in OD of the washers produced by this.

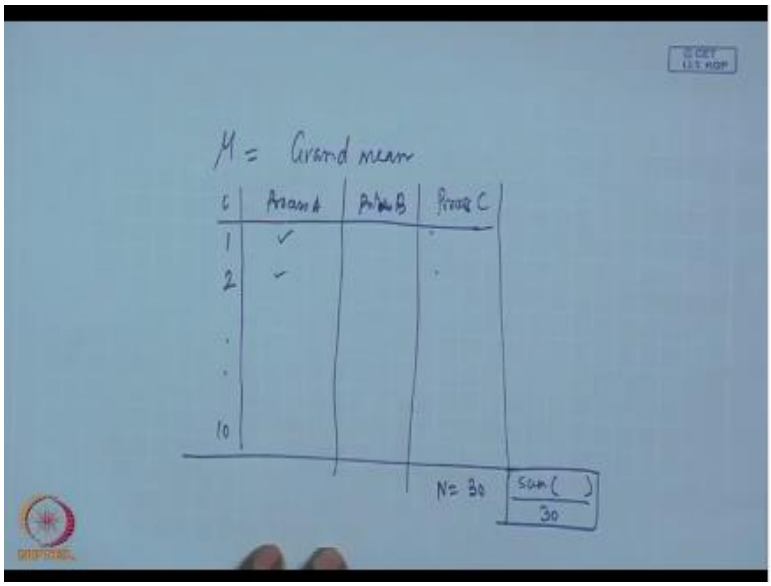
(Refer Slide Time: 10:56)



Then we will use that I told that ANOVA, ANOVA you just two hypothesis, one is null hypothesis and alternative hypothesis what null hypothesis says, null hypothesis says that there is no mean difference as you have taken 3 processes A, B and C. So we are saying that process A mean as μ1 process B as μ2 and process C as μ3 and there is no difference means μ1 = μ 2 = μ3 and our alternative hypothesis is atleast one of the pairs.

That means either μ 1, μ 2 pair or μ1, μ3 pair or μ2, μ3 pair is different that means are different. So pictorially the left hand side figure is your null hypothesis and here the right hand side is your alternative hypothesis, if your situation is as the left hand side figure you say what happened here there is one parameter μ please see the left hand side this is μ here this μ is the grand mean.

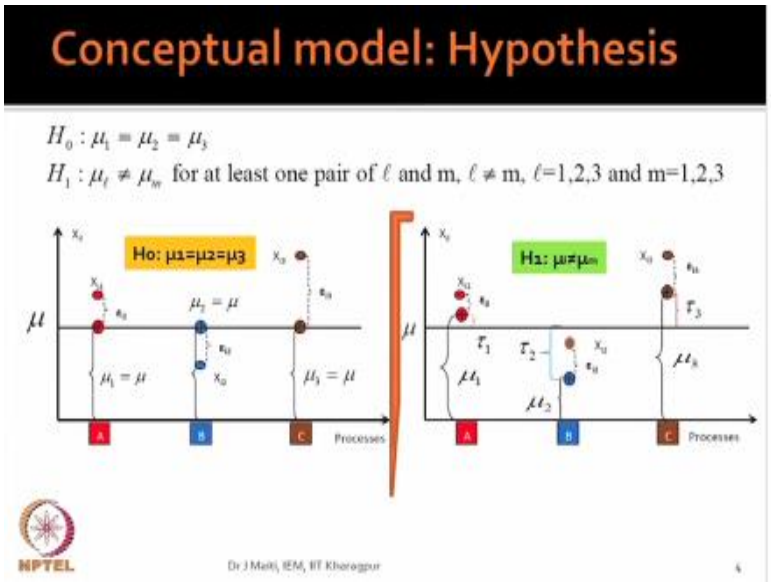
(Refer Slide Time: 12:22)



So there is one parameter called μ, which is known as grand mean by grand mean what do I mean, by grand mean we mean that that you have three processes here in this case. So you have produced several items you measure the outer diameter in this case irrespective of the process you take everything as a whole and compute the means, okay, this is what is grand mean the totality.

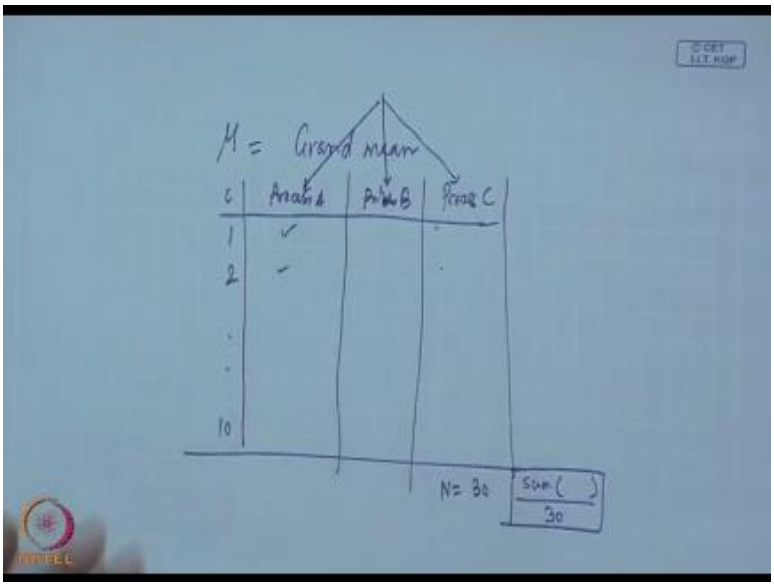
For example, in this case we have seen that you will collect certain amount of data and we have collected ten data points sample size is 10 for process A process B and process C ten each. So what will happen ultimately total data point if I say N that is 30 and you are considering every data points all sum so sum of all the data points divided by 30 that is your grand mean. If you collect 30 data parts 10 from each of the processes that will become your grand mean, okay.

(Refer Slide Time: 13:45)



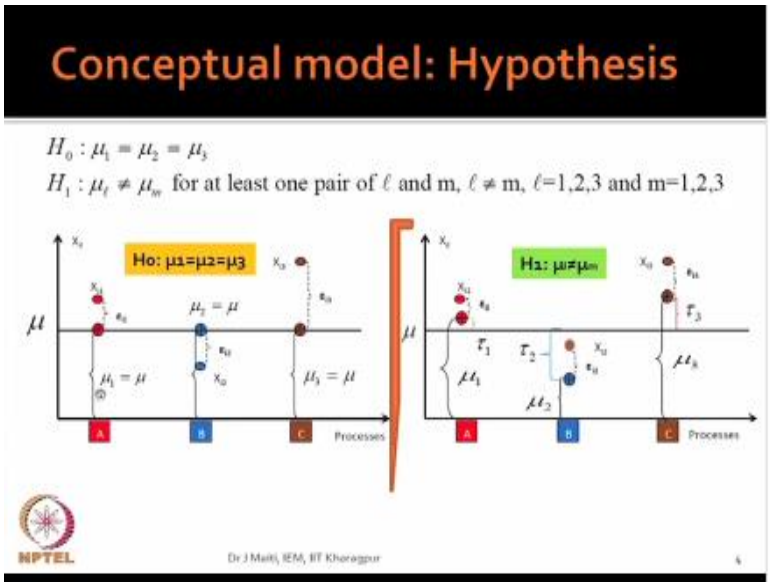
Now there is one another mean which is known as process mean.

(Refer Slide Time: 13:51)



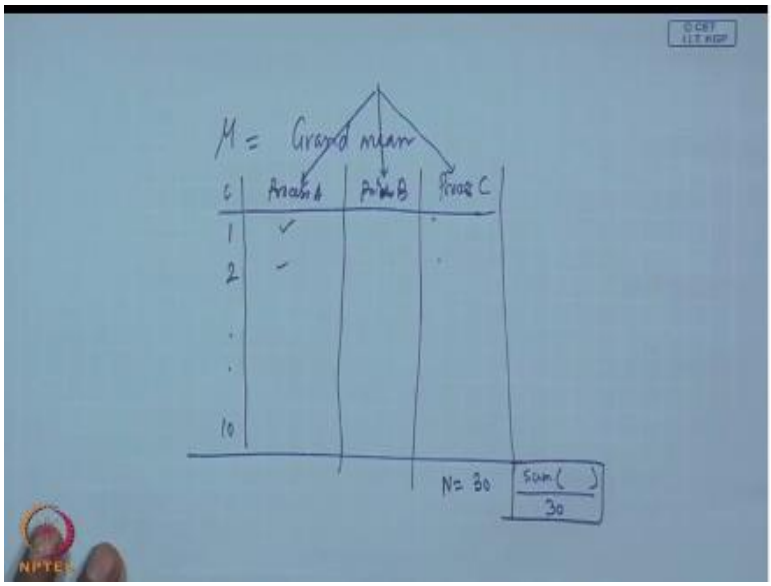
Or in this case we will say the population mean. Because we have assumed that each of the processes are different as if each of the processes are presenting a population, process A is one population, process B is another process B is the other one.

(Refer Slide Time: 14:13)



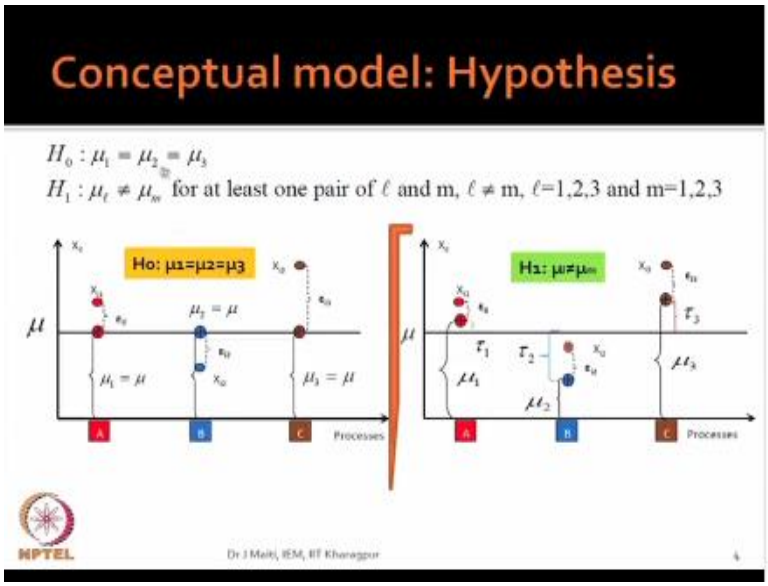
μ1 is the mean of process A, μ2 is the mean of process B and μ3 is the mean of process C. When null hypothesis is 2 I mean μ1 = μ2 = μ3 then it will be equal to μ, okay. Because we are collecting some here ten samples from each of the population and when null hypothesis is true, what will happen?

(Refer Slide Time: 14:47)



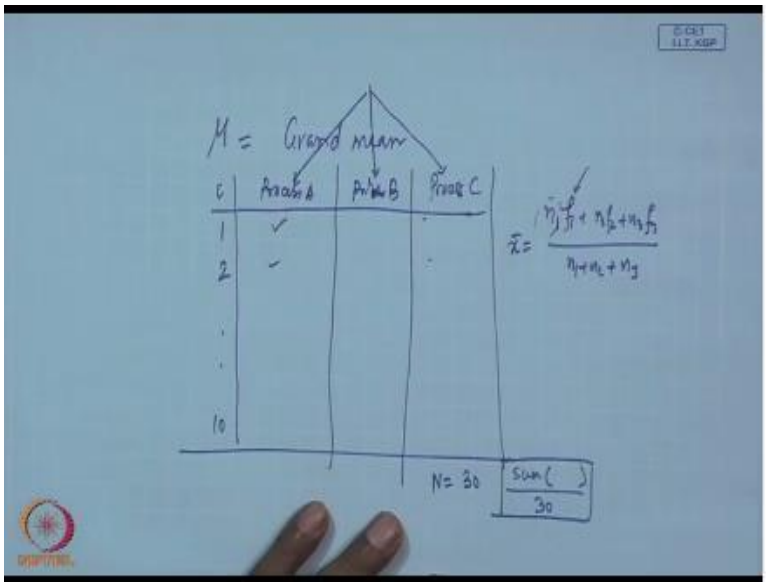
You compute the process A mean process B mean process C mean you will also the grand mean we will be finding that they are equal.

(Refer Slide Time: 14:55)



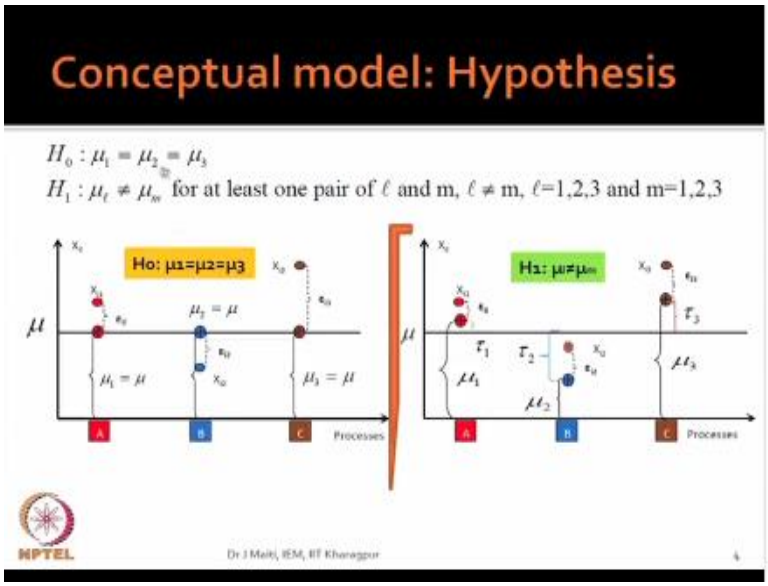
Because it is obvious because what is the group mean calculation do you know that.

(Refer Slide Time: 15:04)



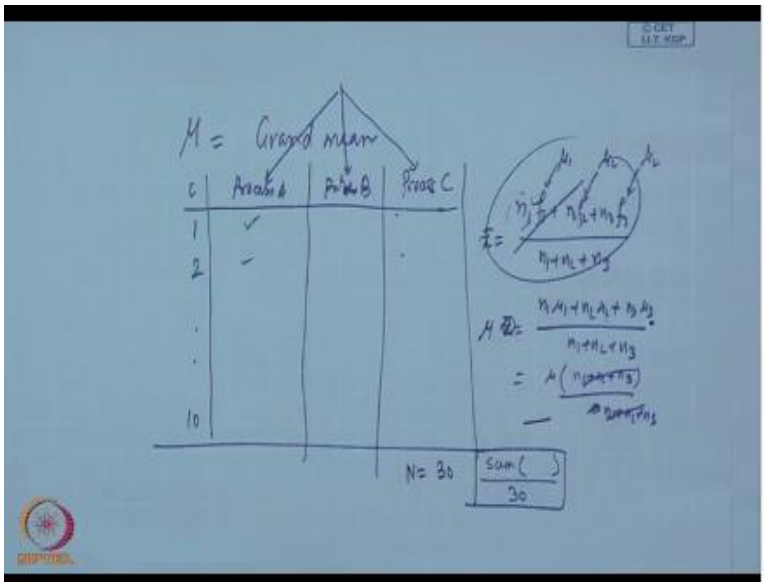
I think all of you know n1f1 + n2f2 + n3fe/n1+n2+n3 that is the group mean if you calculate group mean this is the formula, where for the first case frequency is F1 second case frequency is this is basically f1n1, f1,n2, f2, n3, f3 and n1 is the number of cases.

(Refer Slide Time: 15:35)



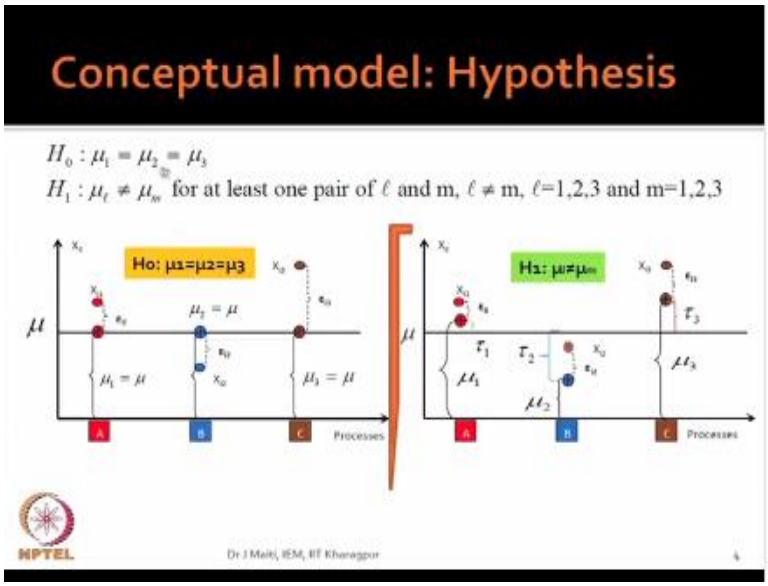
From the that is the I think we have used μ here.

(Refer Slide Time: 15:37)



N1 x μ + n2f1 it is basically μ1, μ2 that will be better. So I will write like this n1μ1 + n2 μ2 + n3 μ3/ n1+n2+n3, okay. So this is the general expression we will not consider this one here. So if μ1, μ 2 and μ 3 are same then what is happening here then it is μn1+ n2 + n3 / n1+n2+n3 this will be cancelled out this is μ = μ, so grand mean will be like this.

(Refer Slide Time: 16:26)

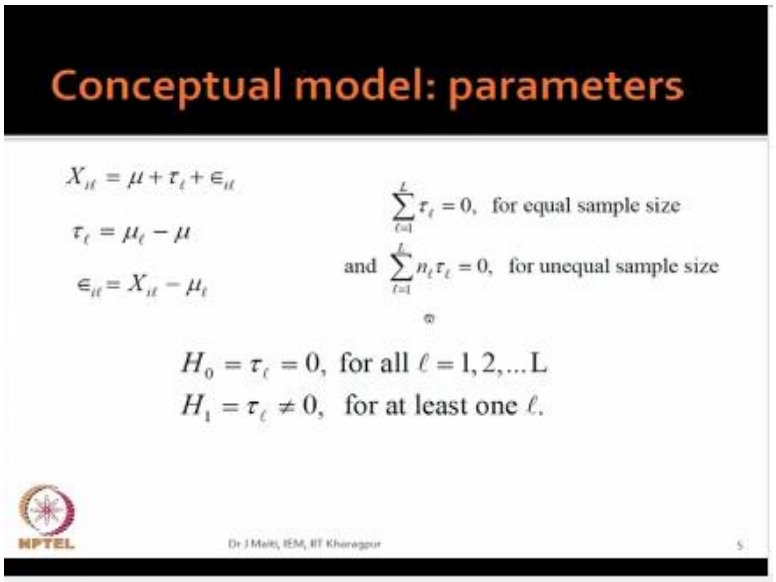


Now let us see that what are the other parameters available here, you see the left hand side again there is another parameter which εil this one, this is the error quantity error in the sense if you take any observation xi1 for the first process from the first process then if we consider that that μ 1 = μ then μ1 will come here the difference between the two that the observed value minus the mean value that is εil.

Later on we will see this is basically error quantity. So similarly for μ2 also εi2 for μ3 that is process C case also εi3 you are getting. Now come to the right hand side in the right hand side what is the figure resembles, that the μ1, μ2 and μ3 are not same. So they are not coinciding with μ and as a result what will happen you will be getting one more parameter here that is known as τ you see this is τ this one is also τ, this is τ.

So τ 1 is this value what is this, this is the process mean minus grand mean and τ 2 is again the process mean minus grand mean for process B τ 3 is process mean minus grand mean for process C. So one more parameter is coming into picture, so what we will do using all those things?

(Refer Slide Time: 18:04)

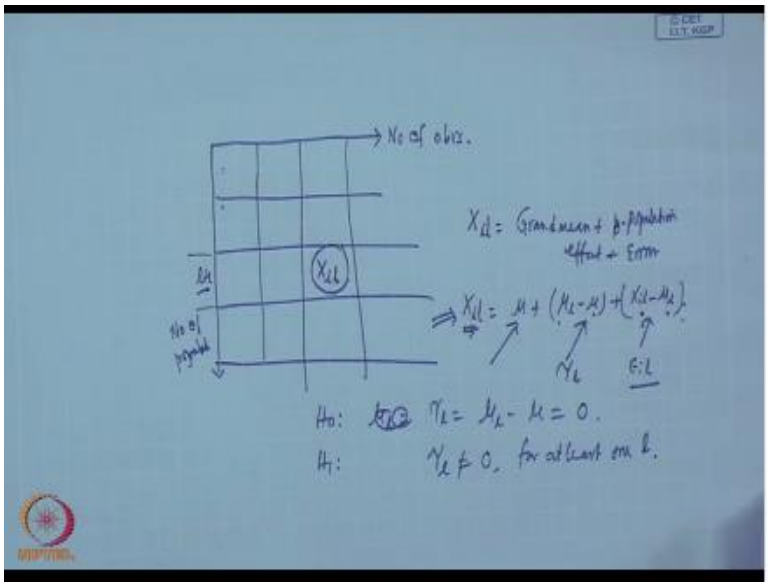


We will redefine the ANOVA model and that is that any observation can be partitions into three parts. That is Xil is the observation that is the i^{th} observation on the l^{th} population to be collected μ is the grand mean τl is the population effect and εil is the error term which is not captured either through grand mean or τl or combining it is not captured basically, okay. So and what is τl ? τl is the population effect.

So l population effect is τl which is also known as treatment effect. So, this is μl - μ that means the mean of l^{th} population minus grand mean and the other one εil that is the error part is calculated like this. Here we this right hand side there are two conditions give N1 = one to τil = 0. For equal sample size case means when you collect sample same sample size sample from the three populations or l populations what will happen here.

You will find out that their effect will be total effect will be 0 and one is more than the other some cases some are more than the mean grand mean some are less than the grand mean but the total effect will be 0 and if you go for unequal sample size then your formula will be like this N = 1 to l, Nlτ = 0, okay. I think I will, let me repeat this.

(Refer Slide Time: 20:02)

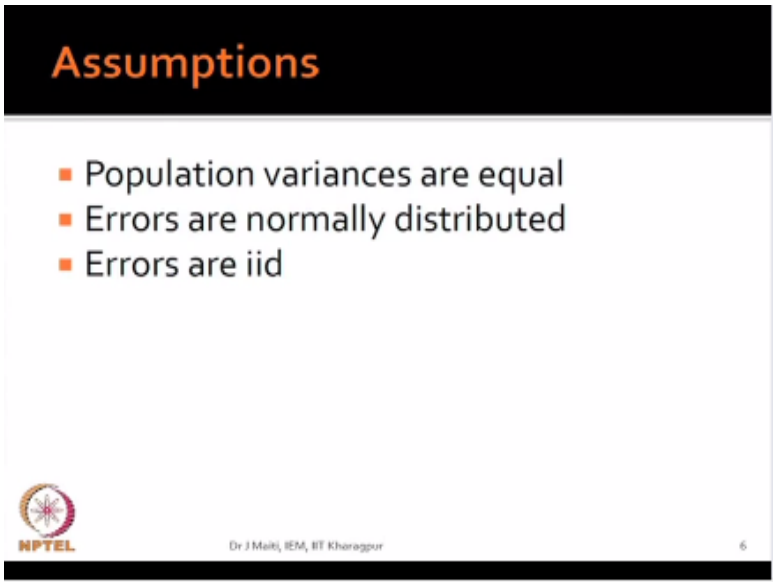


What we are saying in ANOVA case we are saying that you have number of observations in this case, here it is number of population, okay. So you may find out that several populations and also here will be several observations suppose I am saying that one observation is Xil that is the general observation that means i^{th} observation on the l^{th} population this is my l population, what we are saying further that Xil it can decomposed as grand mean plus your population effect plus some error quantity.

How do you get it, we will get it like by this manner Xil equal to, I can write μ plus let write like this μl - μ. So μ, μ cancelled out μl is there but left hand side is Xil, so Xil – μl. See μ and μ will be cancelled out μl, μl will be cancelled out. So Xil = Xil, so this is what is known as partitioning observation into three components here one is grand mean other one is the τl that is the population effect and this one is εil.

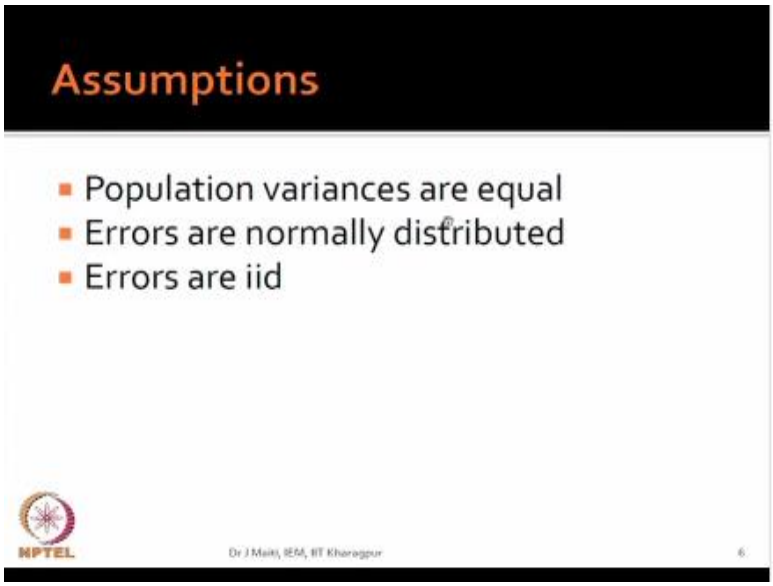
So you are partitioning the observation values into three component here and each component is a parameter of ANOVA. That is from conceptual point of view that there are three parameters one is μ another one is τl and error component settings are there, okay. If this is the case then your hypothesis can be changed to h0 that is μ l τ l = μ l - μ = 0 and h1 τl ≠ 0 for at least 1l.

(Refer Slide Time: 22:48)



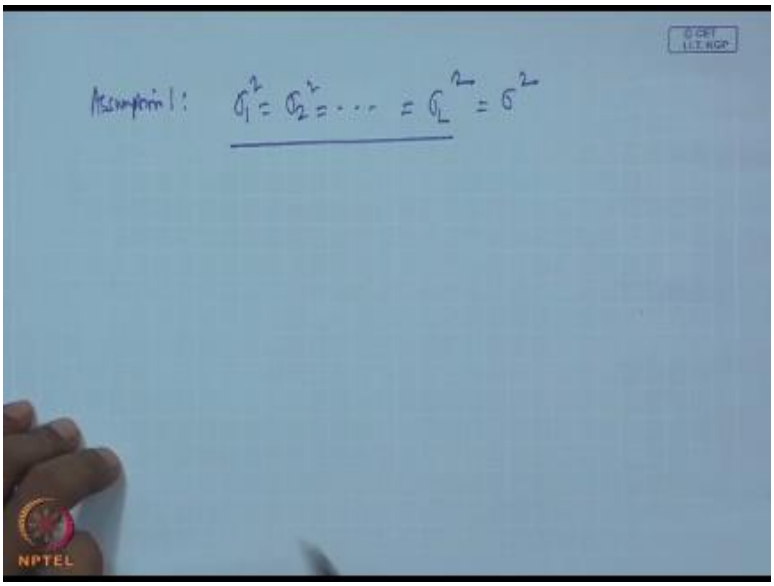
Like other multivariate models, ANOVA comes under univariate model like other multivariate model or like any model any statistical model.

(Refer Slide Time: 23:15)



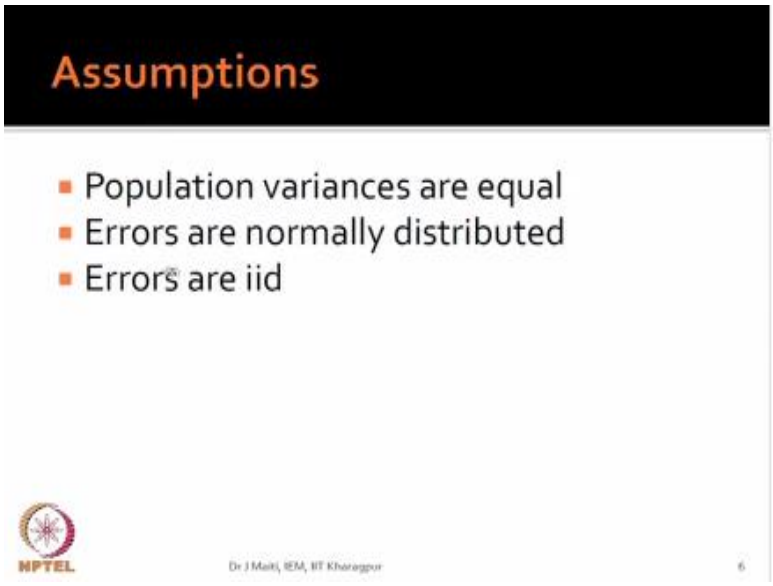
That ANOVA also requires certain assumptions to be satisfied, these assumptions are populations variances are equal. So you are sampling from l populations, the population variances must be equal that means what I mean to say.

(Refer Slide Time: 23:29)



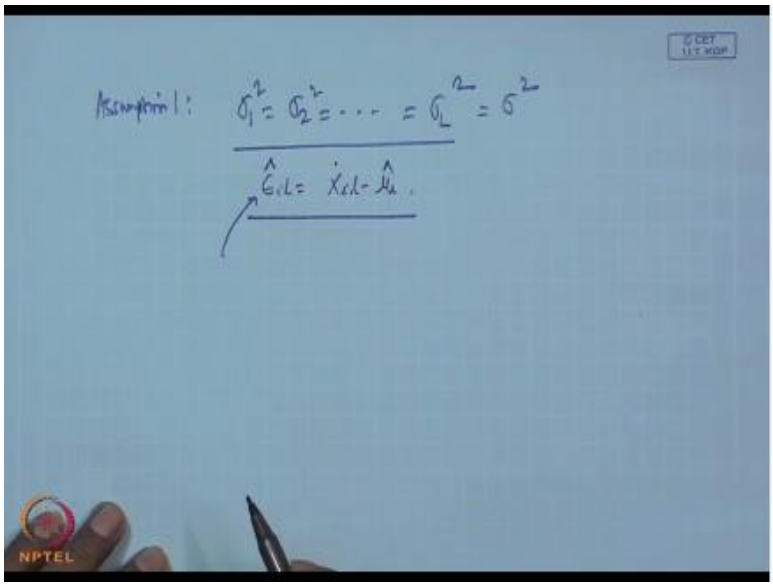
$\sigma 1^2 = \sigma 2^2 = \sigma l^2$, okay. This to be satisfied this is our first assumption, assumption one.

(Refer Slide Time: 23:48)



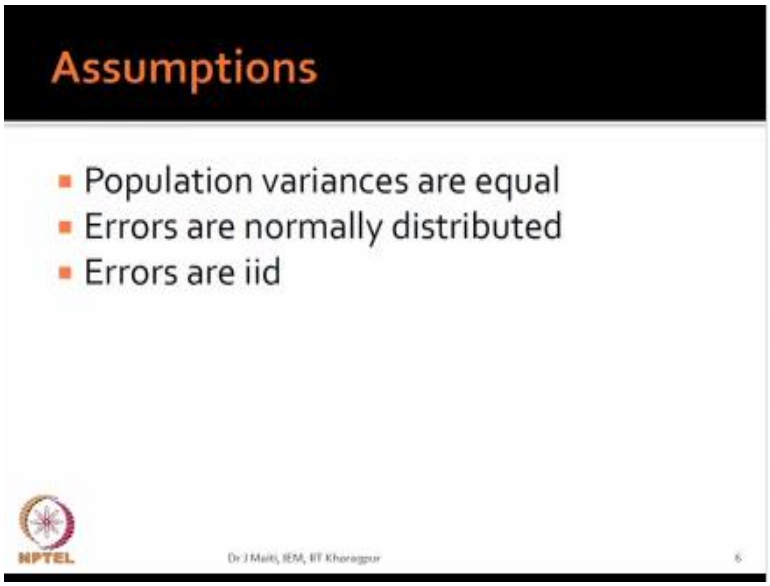
Assumption 2 is errors are normally distributed, what is this error you have seen earlier that we have said that.

(Refer Slide Time: 23:55)



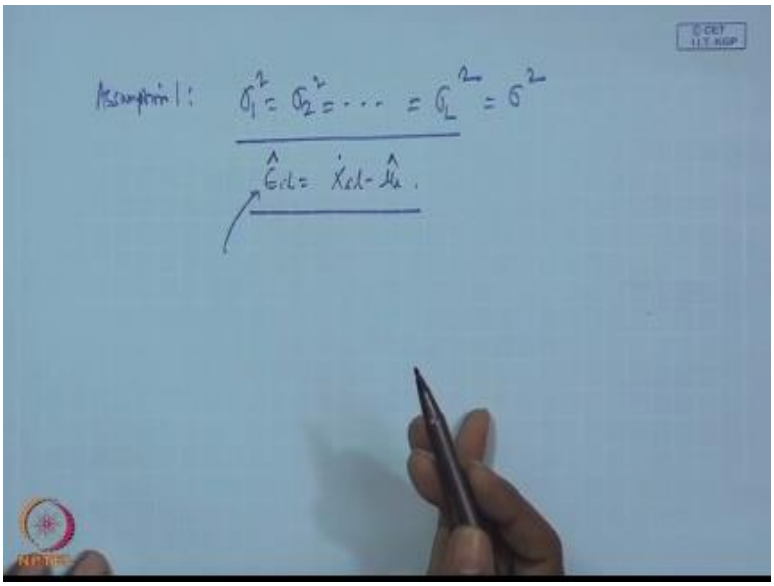
εil = Xil - μl. If you estimate this will be like this, so this error quantity is normally distributed.

(Refer Slide Time: 24:15)



Errors are IID mean independent and identically distributed.

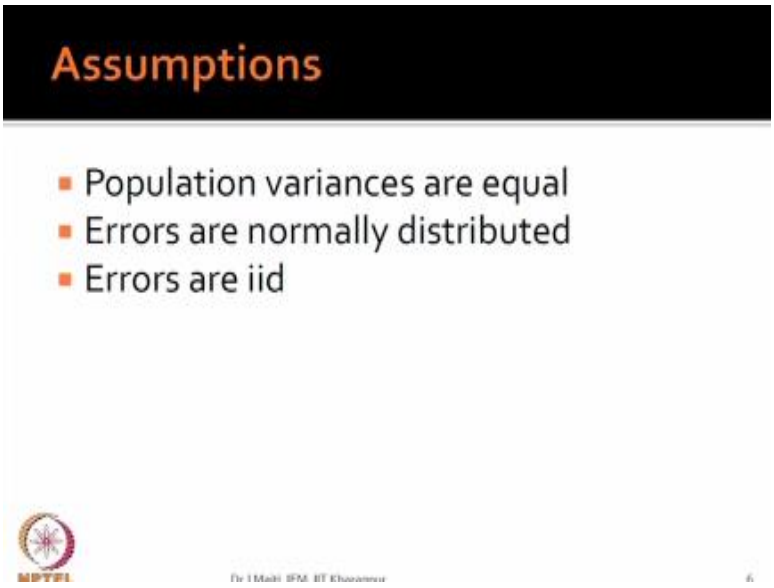
(Refer Slide Time: 24:19)



If I say error one is normally distributed then error N also normally distributed that is identically distributed, independent mean there is no correlation between the errors, okay? So normal distributions as well as IID condition those things you know and how to test the normality we have also discussed earlier in some lectures. Now I will first explain how to test the equality of population variances.

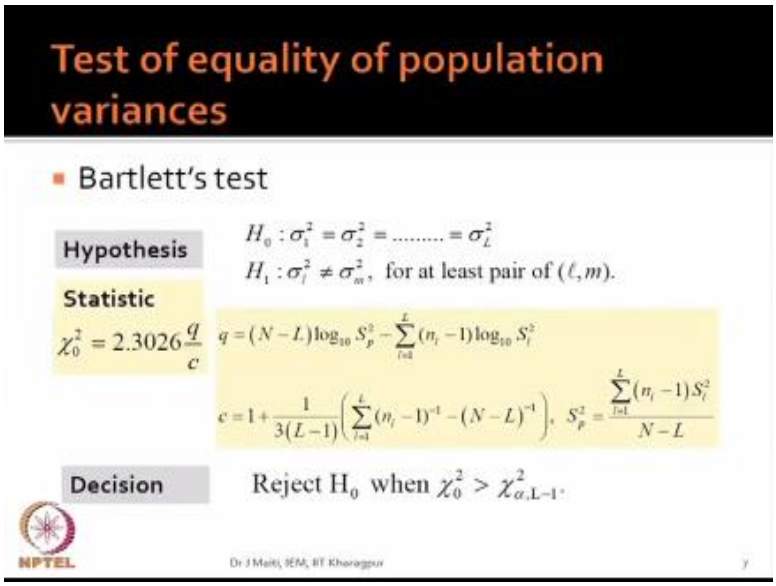
If the one is not satisfied this assumption 1, that is equality of population variances is not satisfied, what will happen? Then you cannot use the traditional ANOVA, okay. It is there some other methods are there but you have to go for different way of doing things.

(Refer Slide Time: 25:18)



But here this assumption is vital for us.

(Refer Slide Time: 25:22)

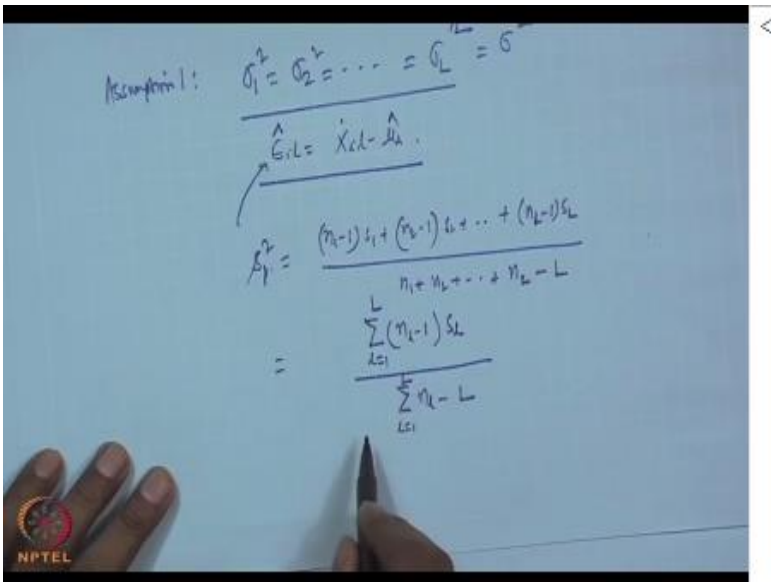


So for equality of population variances you will use Bartlett test, okay. Now see the format what we have written here hypothesis statistic and decision and that we have seen earlier also, okay. Now what is our hypothesis? That there is no differences in the population variances and alternative hypothesis is at least one pair of variances are different, okay. Now here Bartlett proposed one statistic, which is 2.3026 q/c.

This is a statistic where q is N- L log 10 Sp^2 – l = 1 to l nl – 1 log 10 Sl^2. I hope you can recollect what is Sp^2 and what is Sl^2, okay. This is pooled variance and this one is individual population variance. So you require to compute first l S1 $S2^2$ $S13^2$ $S4^2$ like the all individual population variance values, then also you require to compute the pooled one this pooled one is given in this equation.

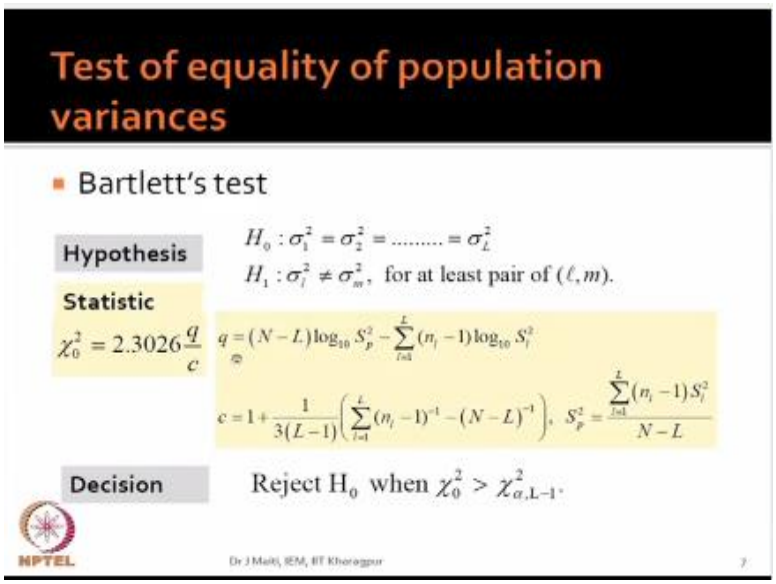
Sp^2 = l = one to capital l N l – 1S l^2 / N - l. I think that we have already seen in last class, what we are saying.

(Refer Slide Time: 27:27)



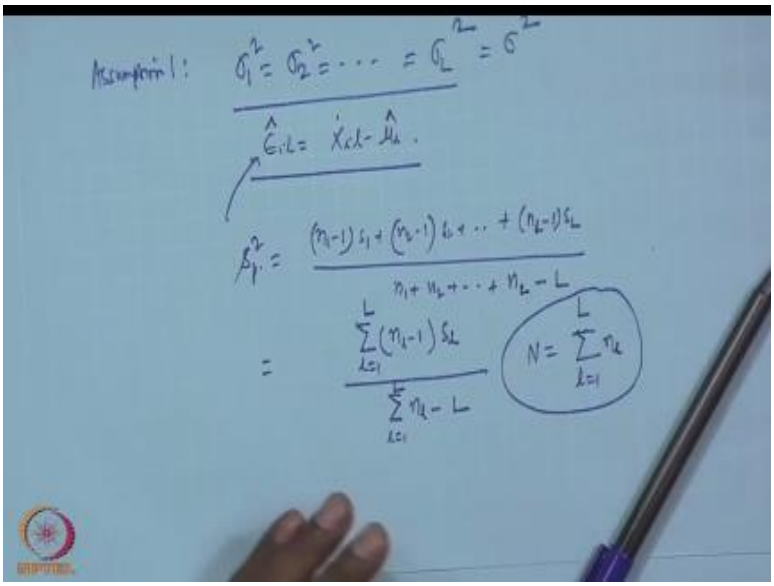
That Sp^2 this is (N1 – 1)S1 (N2 – 1)S2 like this Nn - 1Sl / N1+ N2…..Nl minus the how many populations are there that is l. So, this quantity is nothing but Nl – 1Sl = 1 to l divided by sum total of N l - l where l equal to L. So this is your Sp^2 you have developed earlier also.

(Refer Slide Time: 28:22)



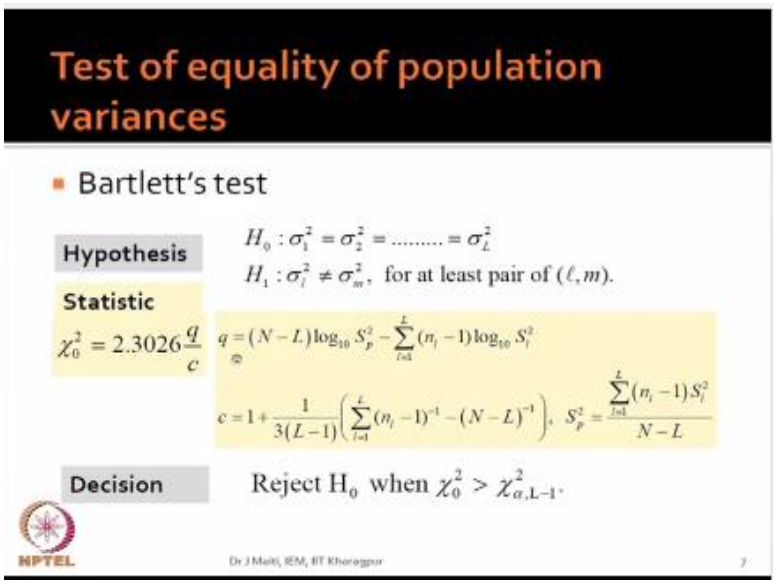
If you know this then you are in a position to calculate q what is capital N here? Which one, N –l - l this capital N is nothing but this one.

(Refer Slide Time: 28:45)



Capital N is l = 1 to Lnl that means the total sample size from all the populations. Although, it is total sample size is not that meaningful mean from every population. You collected certain sample of size some sizes and that totality we are talking about.

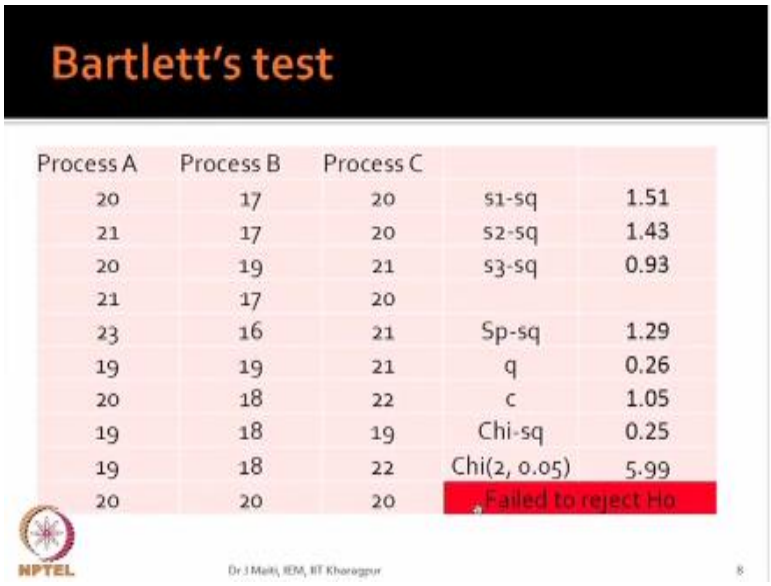
(Refer Slide Time: 29:17)



Then you have to compute C is 1 + 1/3 l-1, l = 1 to l N l – ^{1} – N – l^{-1} this is your C value, okay. So please go through this C there are basically C is another quantity which is basically il and N values are defecting this C and this C will be will be used as the divisor. So then chi square is 2.3026q/c. So given data set you have to calculate q you have calculate C then you find out the chi square value.

Chi square is 2.3026 q/c and what does it mean, when you say chi square because this quantity follows chi square distribution and what will be the degrees of freedom? That you see it is clearly written here in the decision case that reject h0 when the chi square computed is greater than chi square l-1 for a particular alpha value. So what do you mean then that 2.3026 follows chi square distribution with l-1degreesof freedom. I think the similar thing you have seen earlier also.

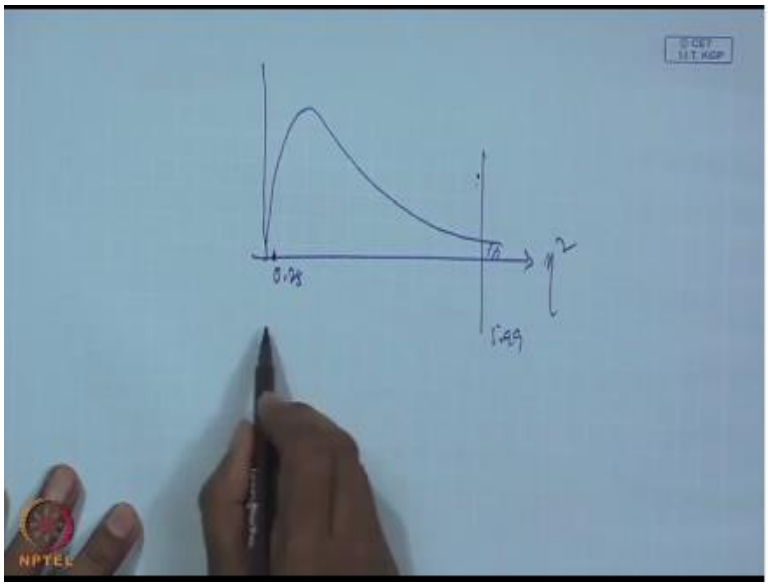
(Refer Slide Time: 30:54)



I have been repeating, see the same problem. So we have 30 observations 10 from each of the processes and what we have done here we have calculated the variance for the first process that is 1.5 one variance for the second process that is 1.43 variance for the third process that is 0.93. and then you have computed chi square by that this one is chi square. So you have computed first pool variance then q then C then chi square, which is 0.25.

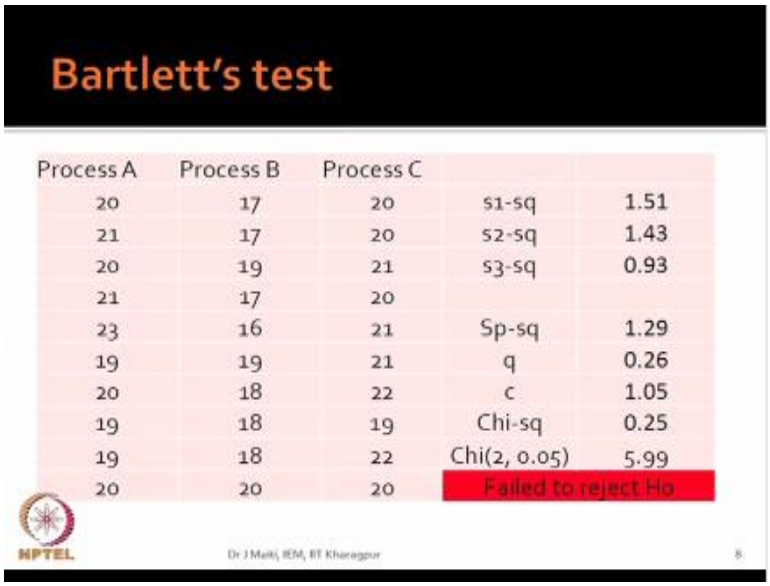
Now we all know that chi square two with two degrees of freedom and alpha equal to 0.05. This value is 5.99 we have seen earlier we have seen this one earlier also. Now the computed value is less than the tabulated value. So what will be your decision you will accept null hypothesis or reject null hypothesis accept null hypothesis you fail to reject null hypothesis.

(Refer Slide Time: 32:09)



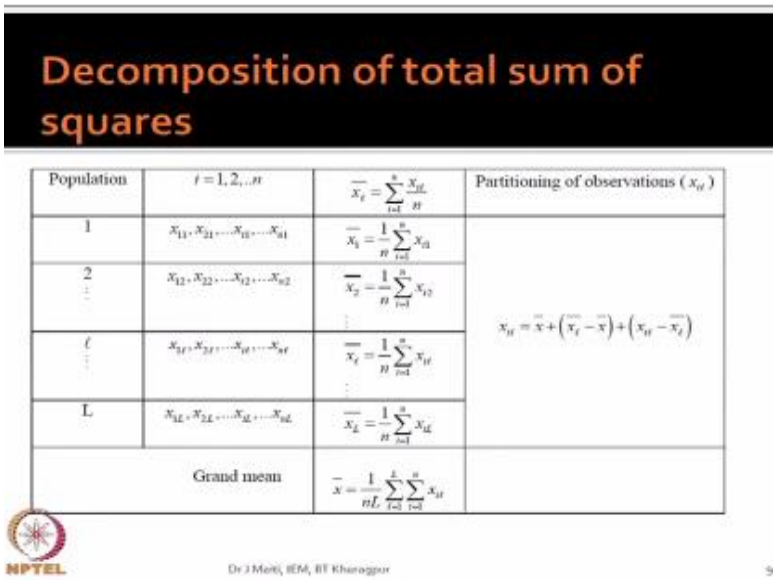
So the same principle this will be my chi square suppose this is your chi square distribution. Your alpha tabulated value here this is 5.99 but your value is how much? 0.25, this is very close to 0. So what does it signify?

(Refer Slide Time: 32:33)



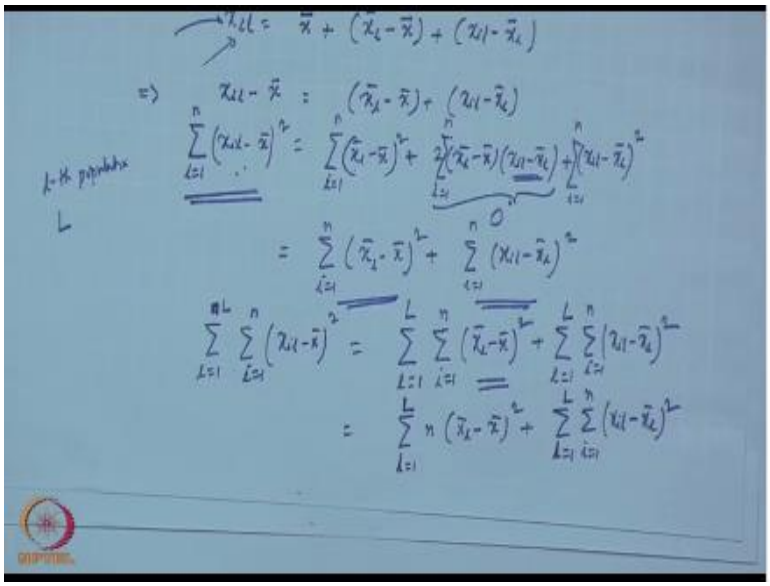
It signifies that there is no population variance differences all the population variances are equal same. So we are fit for ANOVA case, okay.

(Refer Slide Time: 32:50)



In ANOVA it will not go in the same manner like Hotelling t- square or your other one that t-test the way we have developed in ANOVA it is a different ball game altogether, because here the primary concern is partitioning the observation.

(Refer Slide Time: 33:11)



We have seen earlier, I told you that from the population point of view when I am planning to collect some data let Xil this is the observation to be collected from a population where the grand mean is μ and population mean is μl then we have partition the observation like this, you have seen this one see the left hand side and right hand side both are same. Now you have collected data.

That means you have a fixed value Xil .So you have to go by sample grand mean plus your population mean minus again grand mean plus Xil that value minus population sample means not population sample mean. So it is now you do little manipulation Xil - $\bar{X}$ =$\bar{X}$L - $\bar{X}$ + Xi - $\bar{X}$L square the term, If you square it Xil - $\bar{X}^2$ = $\bar{X}$L - $\bar{X}^2$ + 2 x $\bar{X}$L - $\bar{X}$ x Xil - $\bar{X}$L + Xil - $\bar{X}$L ^{2} fine, you take summation over I, okay.

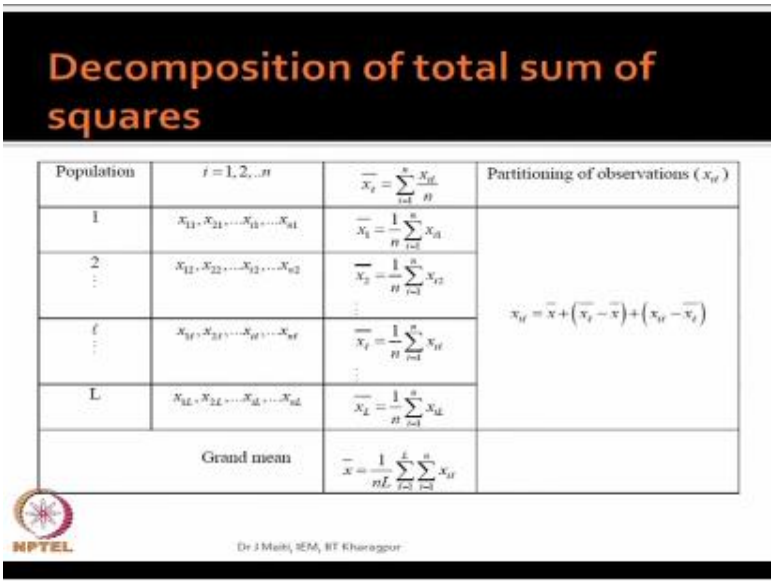
So i = 1 to N we are assuming that equal size samples are collected from the l populations. So again i = 1 to N then this side will be i = 1 to N this side will be i -= 1 to N. Now what will happen to this middle one? The middle one will become zero getting me because of this quantity. So when you write down sum total i 1 to N some Xil this will be N $\bar{X}$L - $\bar{X}$L for N times that will be also N X l bar.

So this value will become zero so your resultant quantity will be i= 1 to N $\bar{X}L - \bar{X}^2$ + i = 1 to N Xil - $\bar{X}L^2$, okay . Now see that this one this is in the left hand side some quantity right hand side also, me quantity two squares is given here, okay. This is for a particular population l population yes or no? Because we have taken Xil every where l is there but how many l populations are there?

Capital l populations are there. So that means your observation is spreaded overall l populations. So if take one more summation what will happen? What will you do? Then you will write down l equal to 1 to capital l sum total of i equal to 1 to NXil - $\bar{X}^2$ = l = 1 to L, I = 1 to N $\bar{X}L - \bar{X}$ plus sum total of l = 1 to L, I = 1 to N Xil - $\bar{X}L^2$ this is also square. Now in this case is there any i $\bar{X}L$ and $\bar{X}$ there is no i component.

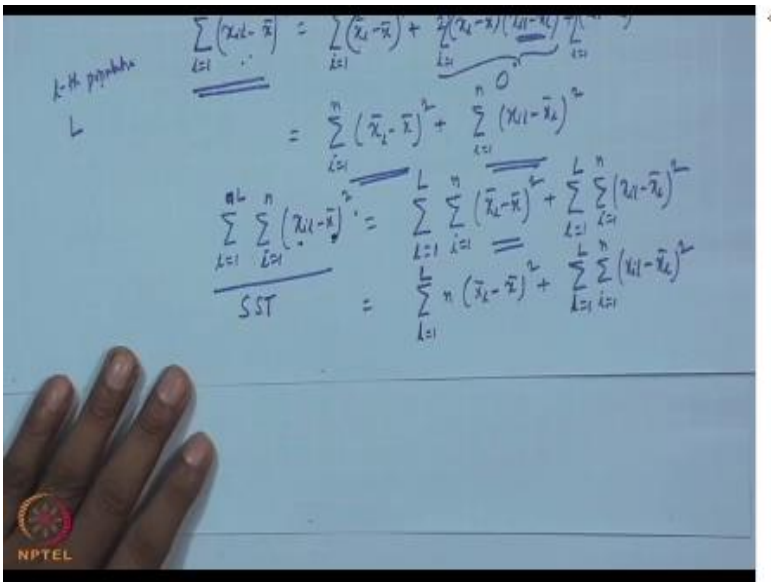
So that mean the same component for N times, so what you can write like this l = 1 To L, N will be there $\bar{X}L - \bar{X}^2$ plus sum total l= 1 to L, I = 1 to N Xil - $\bar{X}^2$, okay.

(Refer Slide Time: 38:57)



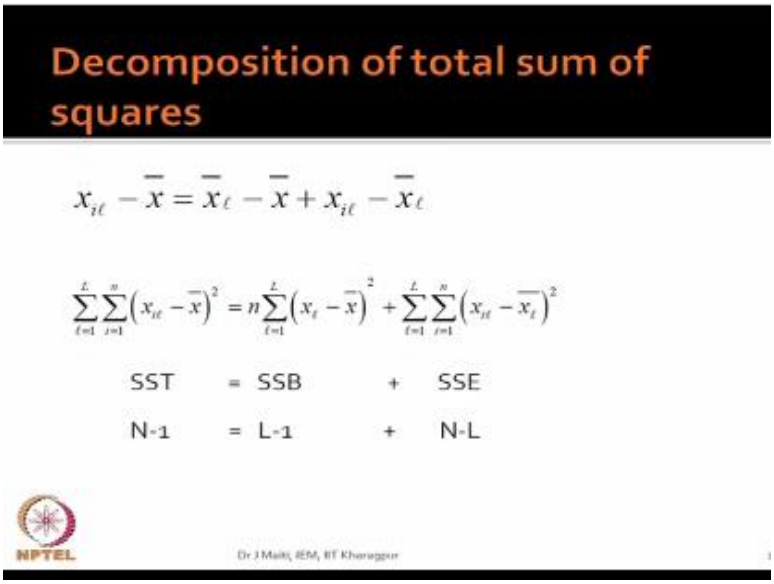
So what will be the left hand side and the right hand side?

(Refer Slide Time: 39:04)



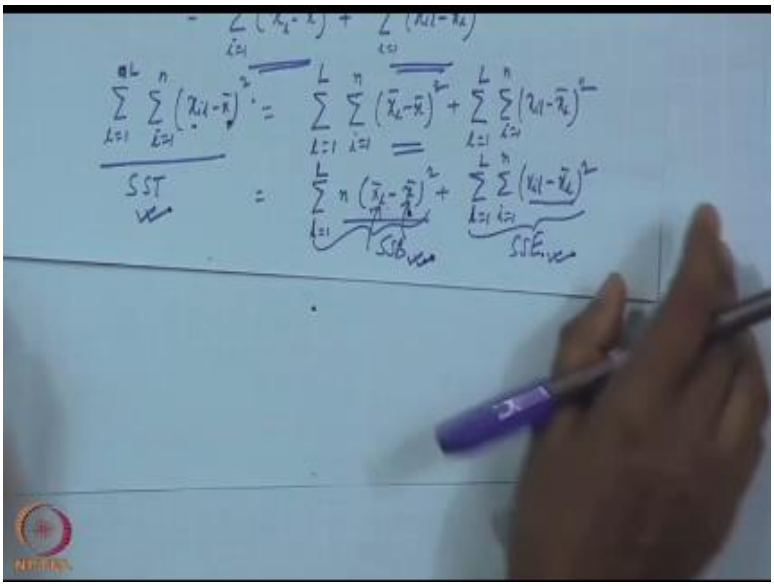
If you see when we calculate the variance what we will do? We will subtract he mean value from all the observed values and then square it and then we divide it by the degrees of freedom here we have not divided it into degrees of freedom but actually what is happening here, this Xil - $\bar{X}$ this was the variability this square. So for all the observation case you have subtracted by the grand central average. So this quantity is known as sum square total, okay.

(Refer Slide Time: 39:55)



You see here that l equal to one to l this i = 1 to lXil - $\bar{\bar{X}}^2$ this is the sum square total and we have already seen that other one N into l equal one to lXl - $\bar{\bar{X}}^2$ this is what is the difference the variability part from grand mean to the individual population means that sum we have taken you see this one what you have done here?

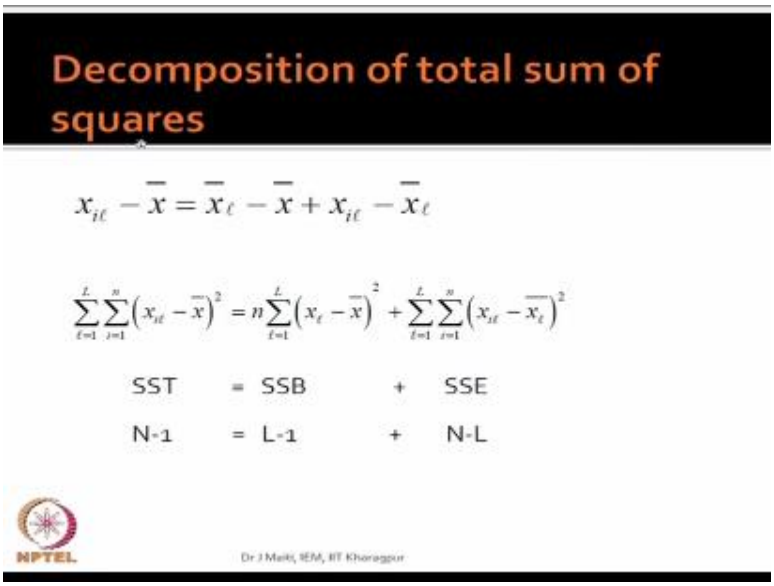
(Refer Slide Time: 40:29)



$\bar{X}l$ this is the population for l^{th} population sample average and $\bar{X}$ is the grand sample average the difference between the two and then this observations you have squared. So if there is no difference between the population means we assume that the sample average also will become same and it will be equal to the grand sample average, so then this quantity will become 0.So if there is any variability this is because of the variability between the population.

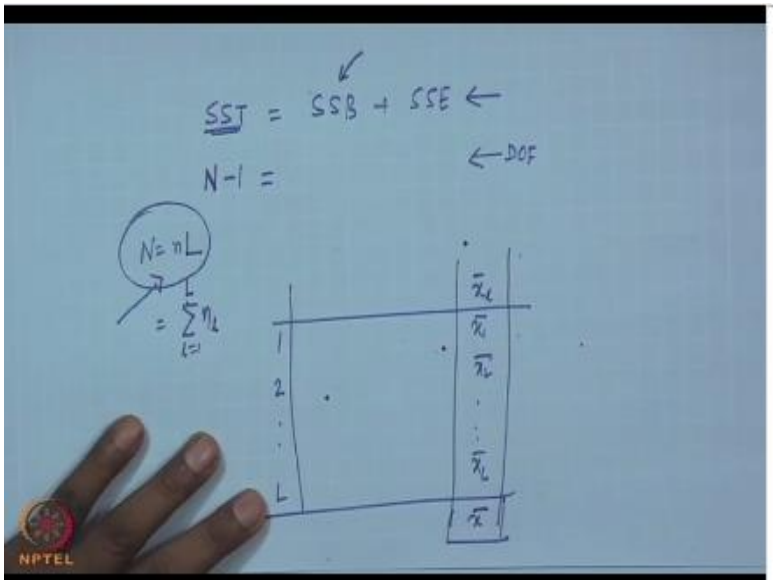
So that is why this quantity is known as SSB and we have seen earlier also Xil - $\bar{X}l$ this is nothing but the error part which is not explained by the which is not explained by SSB. So this is SSE sum square error, now sum square total is equal to sum square between populations plus sum square error. So, what is happening here?

(Refer Slide Time: 41:48)



Then we are basically.

(Refer Slide Time: 41:50)

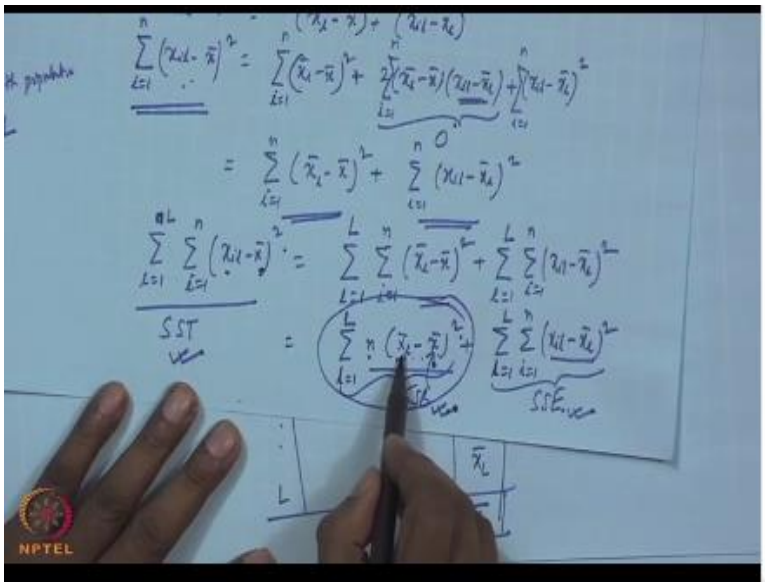


Dividing the total variability in terms of sum square total equal to variability explained by the population plus variability not explained by the population correct. So this is the known as decomposition of decomposing the total variability into between population variability and error variability, okay. So now we have to see the degree of freedom part degrees of freedom. So when we compute SST what we require basically we require $\bar{X}$.

So there are total N observations and we have sacrificed one to calculate this one $\bar{X}$. So degree of freedom for SST is N minus capital N minus one where capital N is N into l when you sample from l population and sample size is equal. If sample size is unequal then this will be your l = 1 to N nl getting me. So this is for equation for the equal sample size case this is the equation for computing capital N for the unequal sample size case.

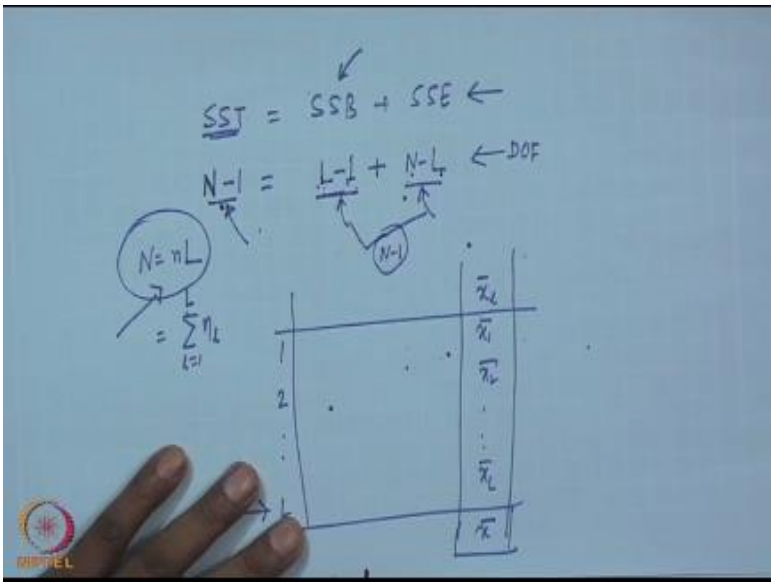
Now come to the between sum square now in between sum square how many levels you have if you see the computation what we will be finding out? We will be finding out the population1 to population l and you will be computing here that bar $\bar{X}$. So here $\bar{X}1$, $\bar{X}2$ like $\bar{X}l$ and then you find out the $\bar{X}$.

(Refer Slide Time: 44:07)



And if you see the computation of SSB you are finding out that this is nothing but $\bar{X}l - \bar{X}^2$. Then multiplied by N and that summation, essentially you are using $\bar{X}l$.

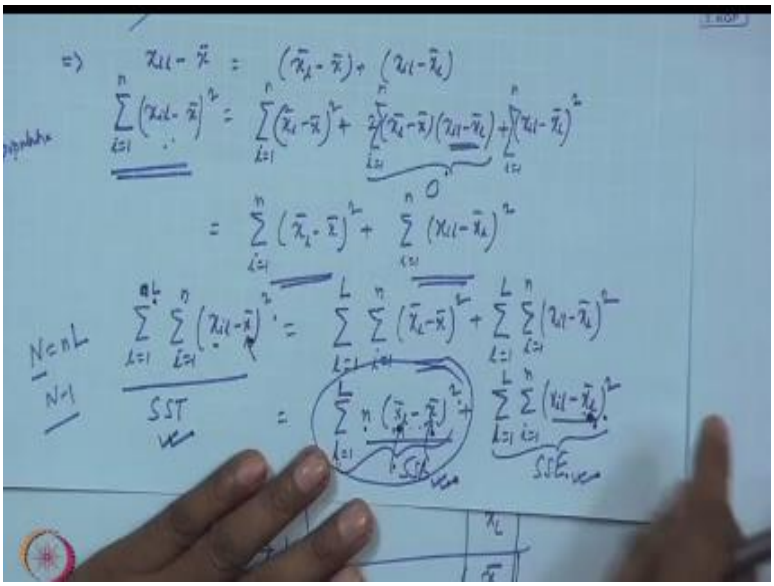
(Refer Slide Time: 44:25)



So as there are there are l populations and there is relationship also with the individual population mean this sum is grand mean. So one degree you are losing here and as a result you have total l data points and one degree is lost by computing X bar grand mean. So plus N - l because this is the rule that the total variability will also be decomposed into the component degrees of freedom.

So total degrees of freedom is N - 1 and your SSB degrees of freedom l-1 SSL degrees of freedom N - l. So if you make a sum l-1 + N-l it is nothing but N - 1right hand side N – l + l – 1 sum total will be N-1 and which is equal to this keep in mind, this one.

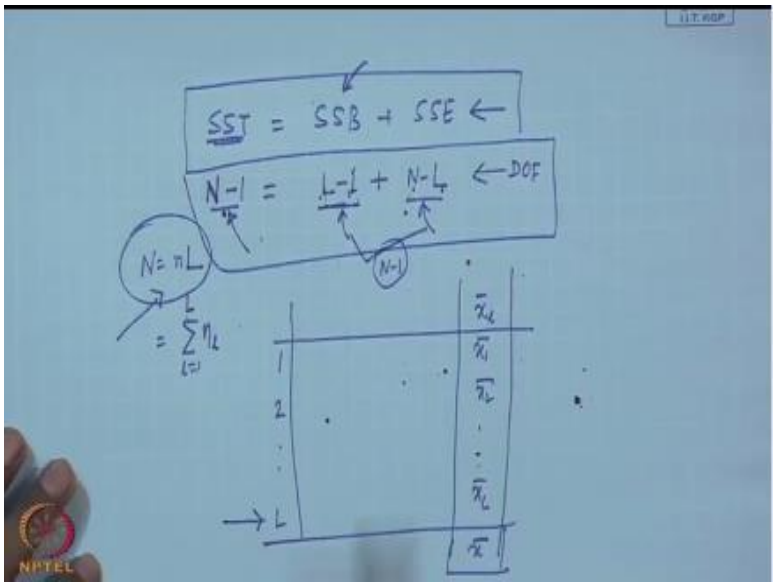
(Refer Slide Time: 45:46)



Yes you are finding out N a, what you are you are using you are using l populations. So, you have l mean values. Second one what is happening N into l this is the total observations. So we are talking about N = n x 1 this is the total observation this side is the total observation is this, but you have already computed X- bar. So this one this portion see how many Xl you have computed?

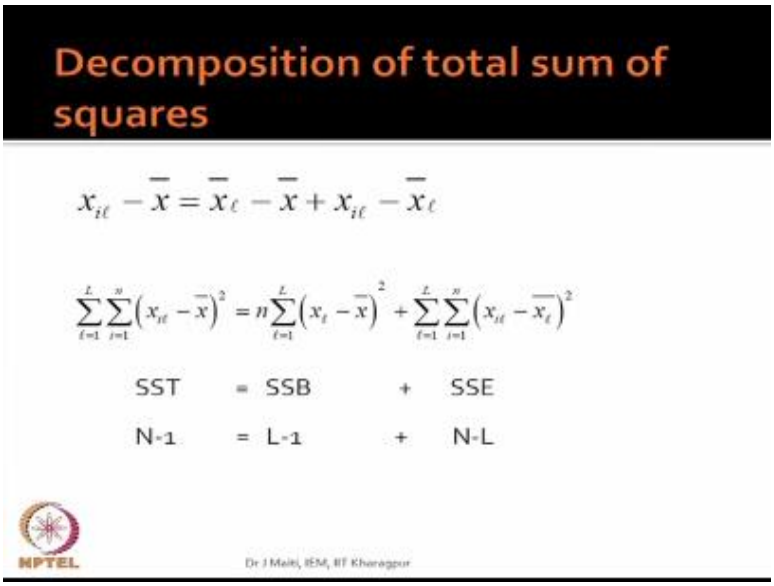
Bar L populations. So total is N - l, so there is l $\bar{X}$l to compute this one you require to know all those $\bar{X}$l. So already l degrees of freedom you have lost you have capital N degrees of observations l is lost.

(Refer Slide Time: 47:06)



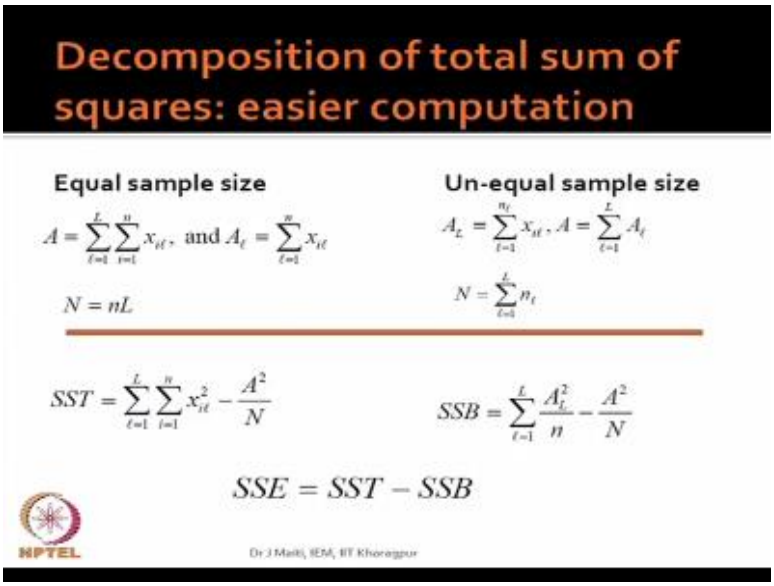
So N - l so this is what is the decomposing the degree of freedom, right.

(Refer Slide Time: 47:17)



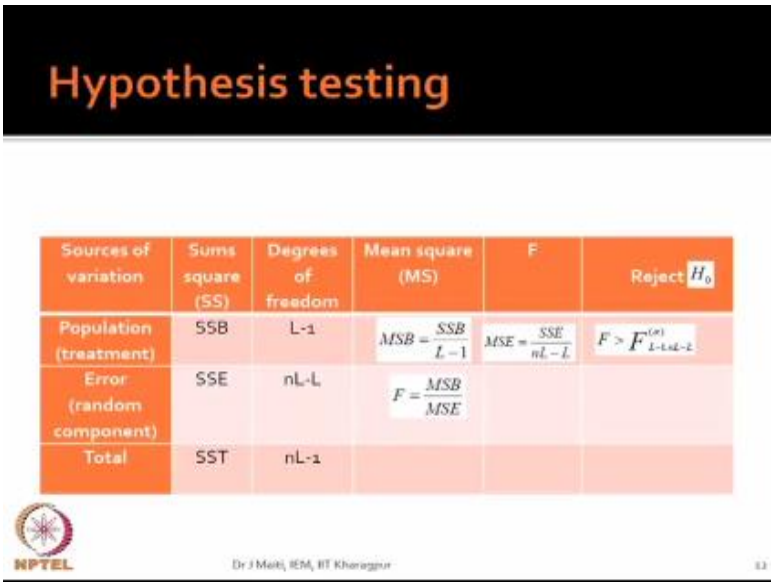
This total concept whatever we have discussed so far.

(Refer Slide Time: 47:24)



Can be seen in a table.

(Refer Slide Time: 47:27)



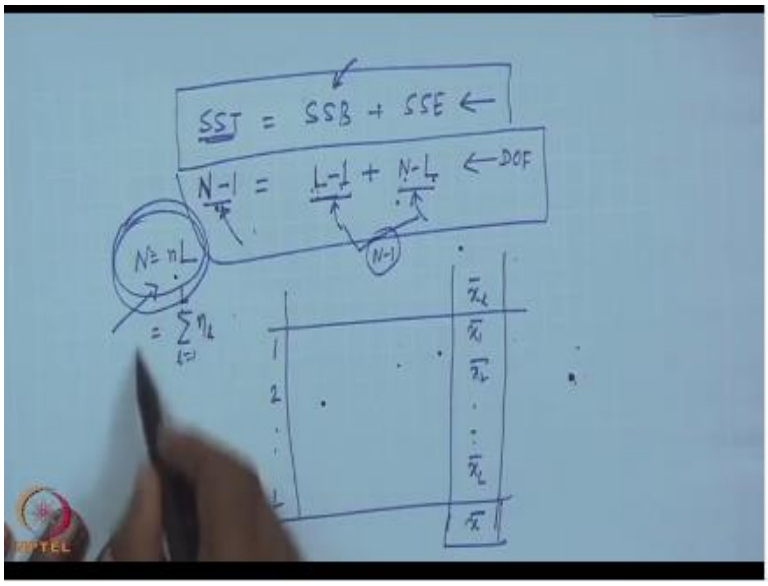
Which is known as ANOVA table, whenever you go through any textbook on ANOVA or any software you use for ANOVA what will happen you will be finding out that one table will be formulated and this table is known as ANOVA table which is very popular table, later on even in regression in other cases also we will be using ANOVA table. This ANOVA table for the problem we have considered now.

They there are few items one is sources of variation then the sum square what we have calculated SST and SSB all those things degrees of freedom also. We have seen then you have to compute mean square you have to compute f value then based on f value we will be accepting or rejecting the null hypothesis, okay. So we have considered l population no other factors we have considered in this present case.

So sources of variation is the different populations and apart from this you cannot nullify the random effects. So random effect is coming under error, so another source of variability is error and this population and error this two sources are making the total variability. So as a result your table looks like this population error and total then SSB SSE and SST you have already seen that,

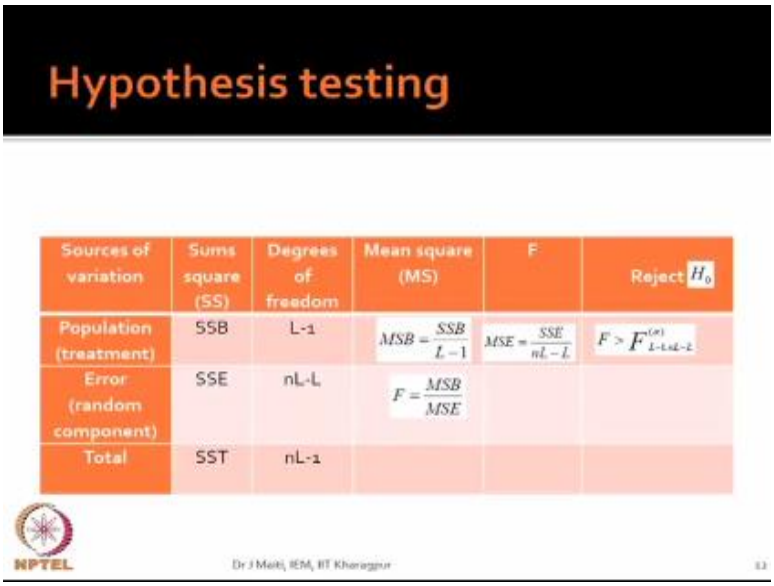
what is the degrees of freedom case l- 1for SSB for SSE N l - l and for SST N l – 1.This N multiplied by capital l this is the capital N, what I have discussed earlier that is capital N.

(Refer Slide Time: 49:39)



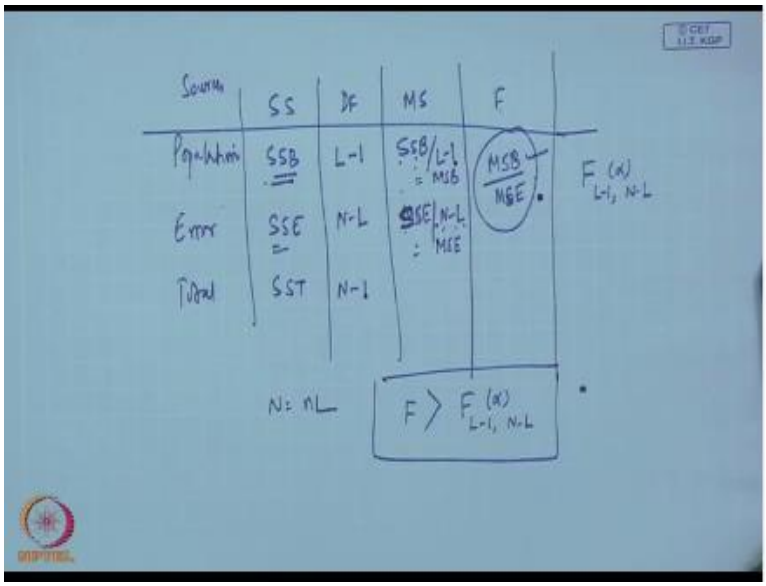
This capital N this is N into l, but if you take unequal sample size this is the different one. Now what is MSB?

(Refer Slide Time: 49:54)



MSB is mean square between populations this is nothing but sum square between population divided by its degrees of freedom, okay. Then what will be your MSE? MSE is SSE by its degrees of freedom. So this MSE will come here you write like this MSB is this MSE here under error you write under error MSE. So, I am writing again what will happen here?

(Refer Slide Time: 50:35)

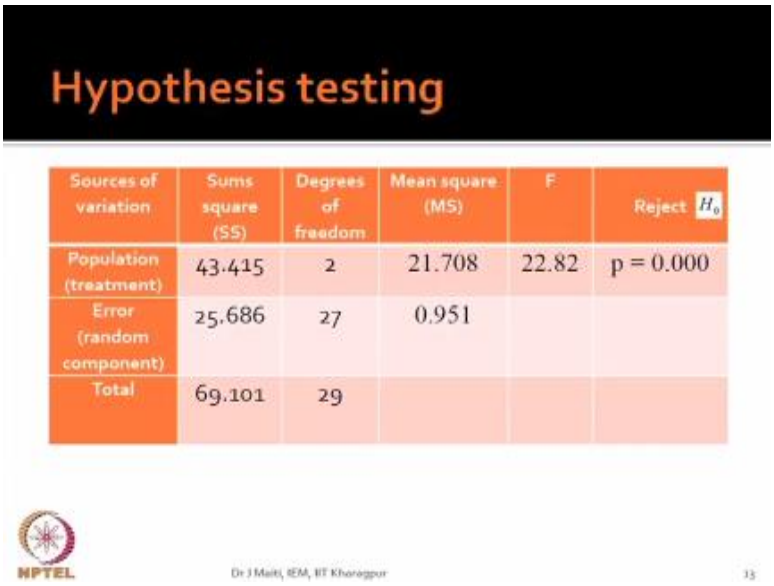


Population then error then total these are the sources of variation sources we are computing sum square that is sum square between populations sum square error and sum square total what is our degree of freedom? Degree of freedom it will be l- 1 and this will be your N- l and this will be N- 1 where N is nothing but capital l x n and then we are talking about MS mean square which is SSB/ N – l – 1 and this be MSE, SSE / N – l.

Then what is this, SS B/ l - 1 is MSB, SSE/ lN - l is MSE, okay. Now you are finding out a statistics called f which is MSB / MSE, why it is f distributed? Ratio of 2 chi square variable because SS B is the square this is sum square and sum square this is f distributed, okay and what will be the degrees of what will be the degrees of numerator and denominator degrees of freedom for f here?

Definitely what will be there in numerator SSB by degrees of freedom l - 1. So, we will be comparing this with l - 1 and what is the denominator degrees of freedom N – 1 and you will find out certain value of alpha and if your computed f that f computed here is greater than f l – 1, N-l α you will reject the null hypothesis, okay. So once you reject null hypothesis you will accept the alternative hypothesis that is the issue.

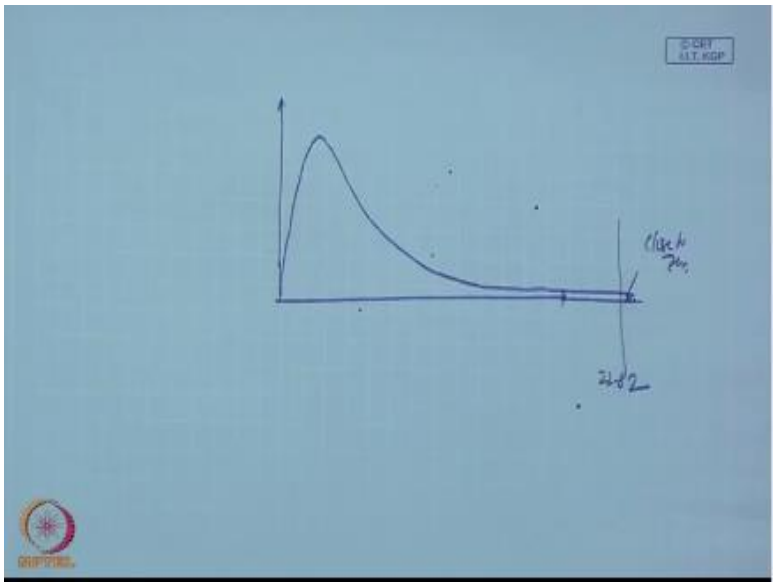
(Refer Slide Time: 53:07)



The problem in the same problem we have done the hypothesis testing and population that variability sum square between population is 43.415. Error SSE is 25.686 and total is 69.101 degrees of freedom because there are three processes. So 3-1 that is 2, 3-1 is 2 and there are thirty total observations that minus one is 29 and you the difference between 29 and 2 is 27 that will be definitely the error's degree of freedom.

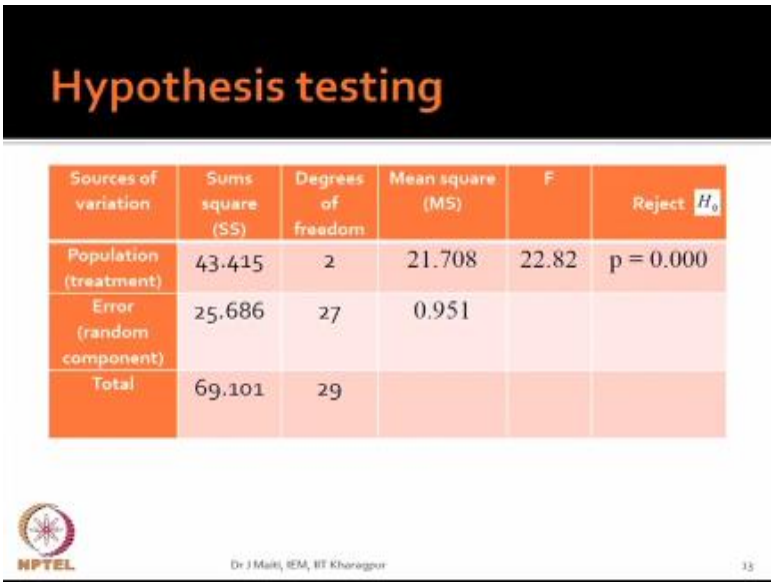
Then mean square is 43.415/ 2 that is 21.708 and for error that is again 25.686/27 which is 0.951 you see the mean square value is very less it is, so less because there is effect. So then f is MSE/ SS, MSB/ MSE. This is 21.708 is MSB and 0.951 is MSE and that ratio is 22.82 and which is far away from 0.

(Refer Slide Time: 54:46)



If you see the f table you will find the probability is almost close to 0. That mean if I say like this is my f distribution that mean you are somewhere here this one which is your 22.82. This one is close to 0.

(Refer Slide Time: 55:21)



So there is difference this is what is tested through tested through ANOVA first level test is this there is difference in population our this one the means population means that is what you are testing, okay. Now here when you test this f test when you are doing here your what is your null hypothesis? Null hypothesis no population no differences in population means correct. What is alternative hypothesis? At least one pair is different.

So when you complete this one and if you find that you are rejecting h 0. It simply says that there is population difference interms of population means, but you do not know which they are, so we have to know which pair are different, okay. So next class I will explain this which pair is different and other things.

NPTEL Video Recording Team

NPTEL Web Editing Team

Technical Superintendents

Computer Technicians

A IIT Kharagpur Production

www.nptel.iitm.ac.in

Copyrights Reserved