INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Marketing Research

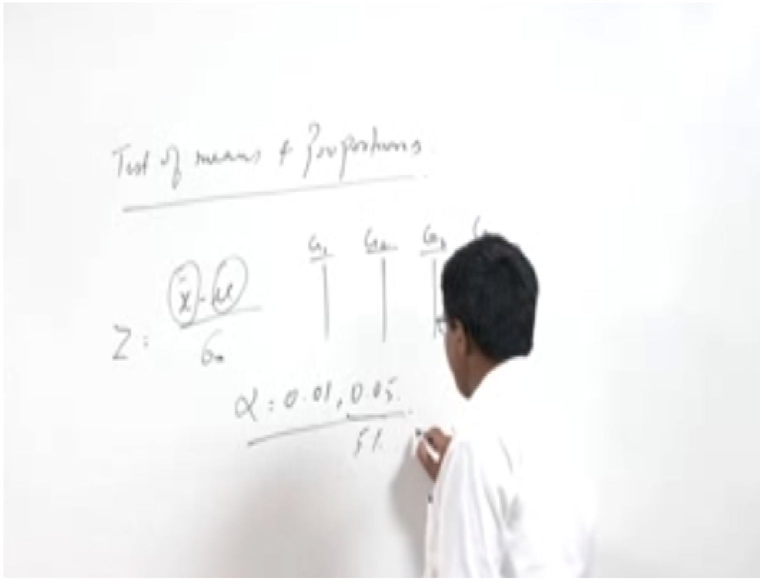
Lec -21
Hypothesis Testing: Anova & Manova

Dr. Jogendra Kumar Nayak
Department of Management Students
Indian Institute of Technology Roorkee

Welcome everyone to the section of marketing research analysis today we will discuss about one of the ways of hypothesis testing which is a very popular method and largely used in all kinds of researchers beat experimental or non experimental or course experimental like serve and all so whenever you do but it is largely applied in all cases right and the application is much more is seen much in larger way in the case of experimental design especially.

So what is this way of testing and let us see and we will discuss about it basically in the last section if you if I if you remember we have discussed about you know the beginning of hypothesis testing the, we talked about the test of means right the test of means.

(Refer Slide Time: 01:24)



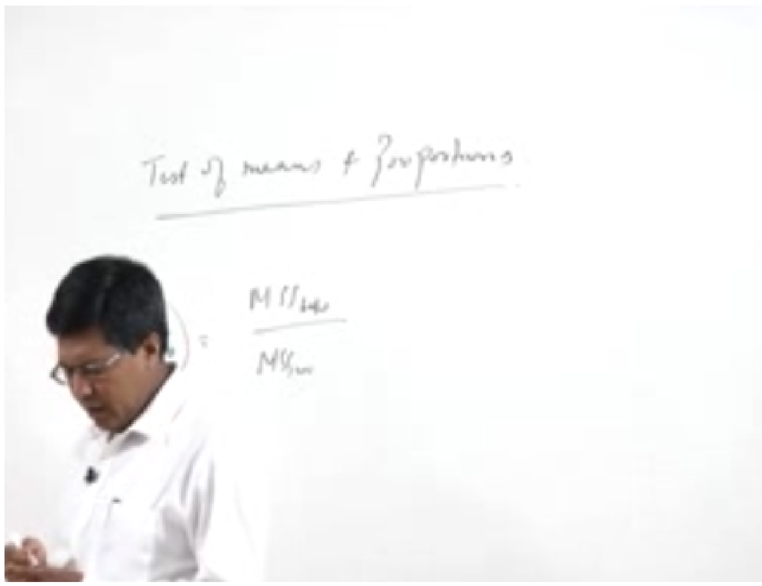
In which we and proportions okay so where we talked about that how do we calculate now we just calculate now we just calculate as Z score which is equal to x bar of the population mean this is the sample mean this I the population mean and over the standard error and we said we would calculate accordingly and we will find out and similarly this also true for a proportional.

So the case proportion right but this case was only possible when they where two levels right that means there where two sample groups now we can say so this was the group one and this was the group two and we could compare it through an independent sample t test or if the sample was only 1 but taken two times then we said it was a dependent sample t test or paired sampled t test but the question raised arises what happens when we have more than 2 groups okay we have more than two groups or more than two levels right.

So in such a condition there is a possibility that there researcher can go for multiple t test right but if you remember I explained you why what is the logic behind not doing a multiple t test and why one should evolved to do that in fact there is something called an you know ball for near in equality test where basically it says that if you do conduct multiple test the problem is that the α which we generally take at as 0.1 or 0.5 or whatever it is this is level of α goes gets inflated right.

So to avoid this problem of inflation so that means if there is a 0.5 that means 5% and you are having 4 times so that it will be around 20% of that means it will get into 20 % right so this is too much of an loss of information or to much of a type one error that is occurring possibility of a type 1 error to avoid the situation fisher the founder of this technique or the one who develop the technique he came up with the technique other than the multiple test and he said we could do it by staring the variants so he said if we are going to use a fishers side if you could use the variance then we can do it better.

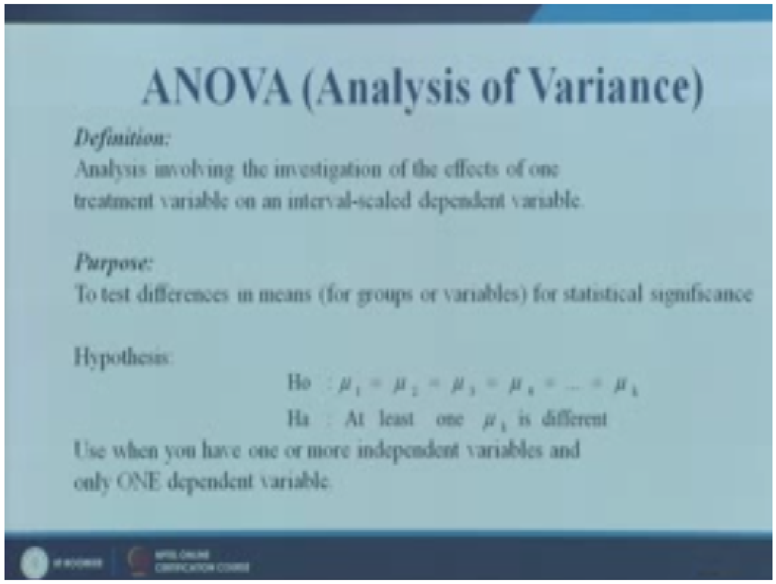
(Refer Slide Time: 04:10)



But to do this he said he calculated something called the he developed the F test or we calculate through that F ratio which I just began in the lat section so the F ration is nothing but I had said if the mean sum of square between right divided by mean sum of square within the groups okay so he said if there are n number of groups right so you need to find out calculate the variance for the entire group right or you can the total variance the between the groups the variance between the groups let us say they are different teams.

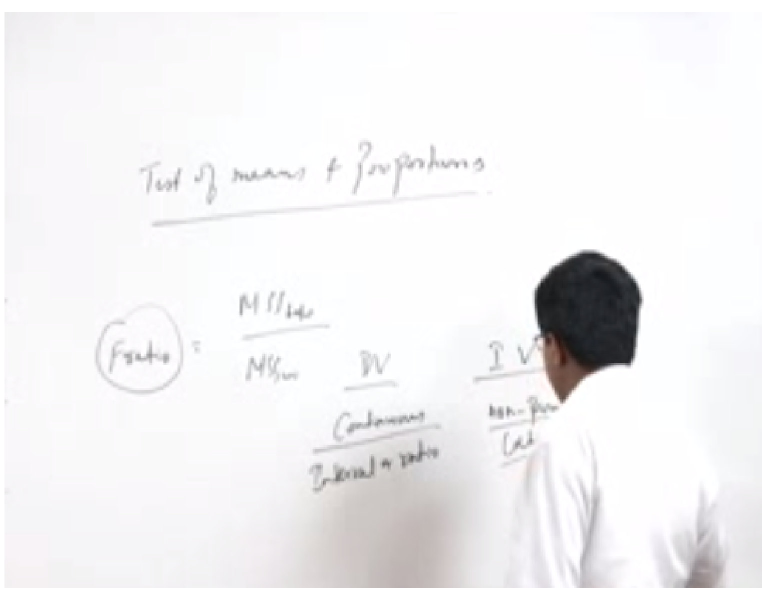
So across the teams what is the variance and within the teams or within the groups what is the kind of variance so suppose 11 players in the cricket team so what is the variance within the team right so when you multiple when you sorry when you find out this variance the you can calculate the F ratio and by then comparing the f ratio the calculated f ratio with the table corresponding table value for the f value then we can say that we want to reject our hypothesis or not, but what is the hypothesis let see let us go by slowly so what is the definitions saying it stays analysis.

(Refer Slide Time: 05:22)



Of variants basically involves investigating the effects of one treatment variable so this is why I had said this is a basically any kind of experimental study it is used so there is a treatment variable now the treatment variable for an yield for example enough agricultural a firm is like suppose you are giving fertilizers so fertilizers could be the type of fertilizer could be the treatment on interval scaled dependent variable, now that is important what is that the dependent variable and the independent variable.

(Refer Slide Time: 06:12)



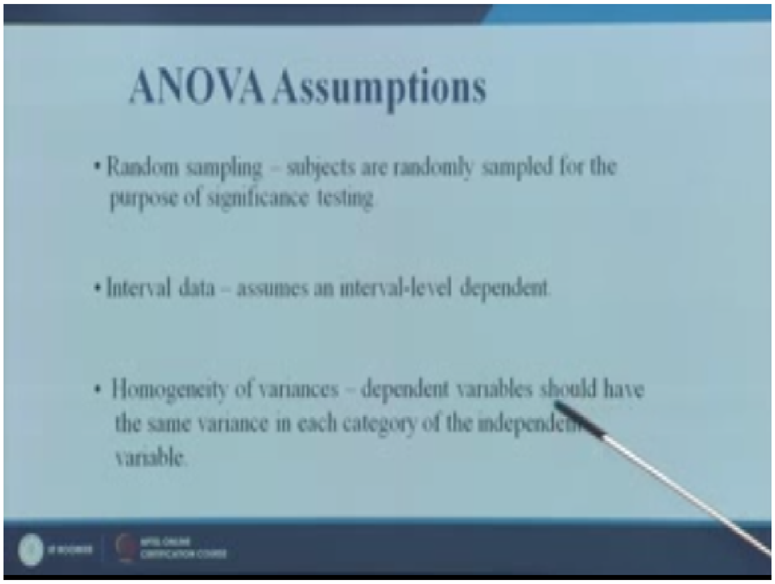
So the dependent variable if I have if I remember I had also said this and the independent variable now the dependent variable in case of analysis of variants is basically measured in a continuous scale, continuous so it maybe and interval or ratio scale basically interval or ratio okay on the other hand the independent variables are basically nothing but they are you know the non parametric right, in nature they might be categorical, categorical in nature right so this is continuous.

So this is continuous this is non continuous or non parametric whatever you say so this is categorically nature right, so this should be in a form of a nominal scale or something okay, so let us go see so what it says to test the differences the purpose is to test the differences in means for statistical significance, now what is the hypothesis the hypothesis is suppose there are four groups or whatever k number of groups so we say there is no difference between the means that means the means of each group are equal in are equal right.

So μ 1 = μ 2 = μ 3 = μ 4 goes on right till μ k now what is my alternative then the alternative says okay so there is at least one μ which is different so whichever it is 1 we do not know but at the moment but at least 1 is different that means I cannot claim that my hypothesis is that there is no difference between the means write is correct right, so it is used when we have 1 or more independent variables and only one depend variable the case ANOVA is basically a one way INNOVA we are saying we are talking about right now.

So we are having one or more independent variables and one dependent variable so let us see right what is happening so if you, you can have multiple independent variables that is one thing so multiple independent variables means suppose you have a one variable one independent variable it is one way INNOVA if it is a which is also called as a factor basically you can understand it has a factor okay factor or whatever one way we say so it is two factor two way INNOVA n factor n way INNOVA okay another assumptions.

(Refer Slide Time: 08:28)

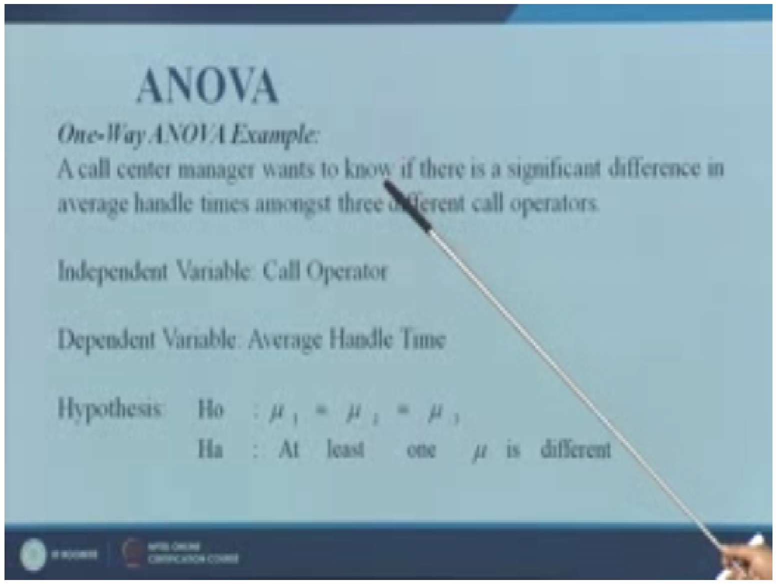


Random sampling subjects are random is sampled for the purpose of significant testing, it is a random selection okay data is interval level dependent so the dependent variable that is in a interval level so which we also said here right, now this is interesting in fact if I if you remember I had explain okay so there is something called a homogeneity and a Hetro scarcity homo means when the data are plotted there around the regression line right, close to the regression line that means the variants.

Within the or the standard deviation or the movement of the data from the regression line is minimal it is quite close right, but if there is a the opposite of that is Hetro scarcity when the data is highly scattered right which is highly unwanted situation which is not reset dependent variable should have the same variants in each category or the independent variable that means this test although it is don if you go to any software they measure it through the this variants is equal the variants is not equal to conditions.

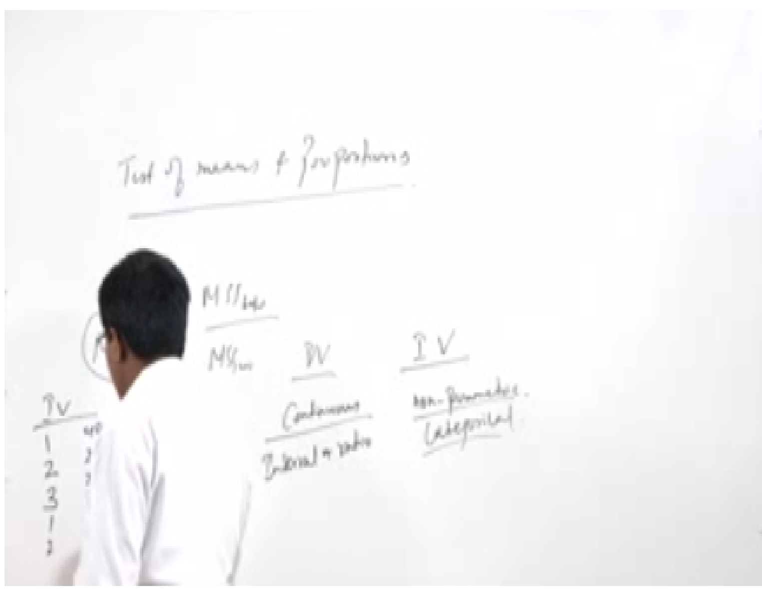
But generally we take the case of where the variants is are equal that means if the variants are equal we would assume then only that this groups can be actually compared okay, the groups are basically those levels right.

(Refer Slide Time: 09:58)



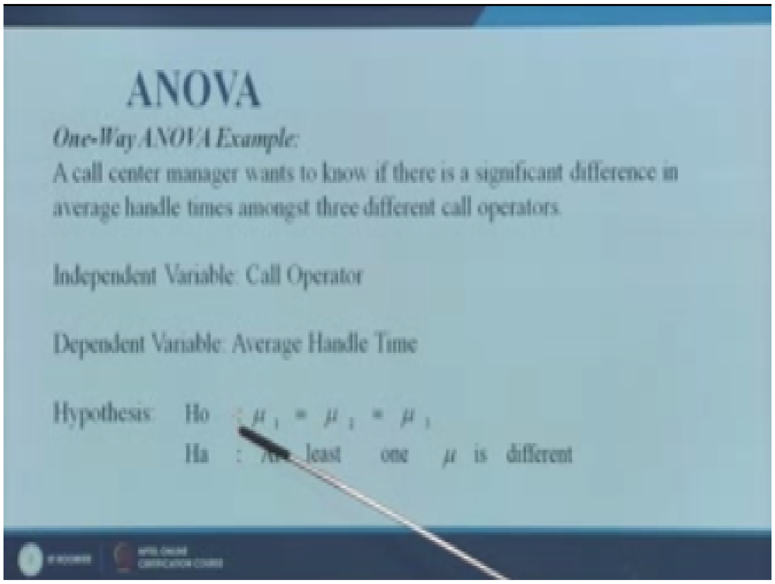
So this is an example we will also solve a problem let see one so what it is saying a call center manager wants to know so if there is a significant difference in the average handle times among three different call operators so there are three different call operators so the independent variable are the call operators, so the call operators could be let us say the independent variable is call variable 1, call variable 2, call variable 3, right.

(Refer Slide Time: 10:26)



The dependent variable is my average handle time so how much of time they are taking to handle the clients, customers is my dependent variable okay, now that means how will it look like now it will look like something let us say, so let us say this is how it is suppose for the moment, so we will say let us say this is 40 minutes, 40 seconds or this is 20 seconds this is 25, 30 seconds, this is again 35 seconds so whatever the time actually they have taken right,42 seconds whatever for time, seconds, minutes that is up to your unit, so that is the different story.

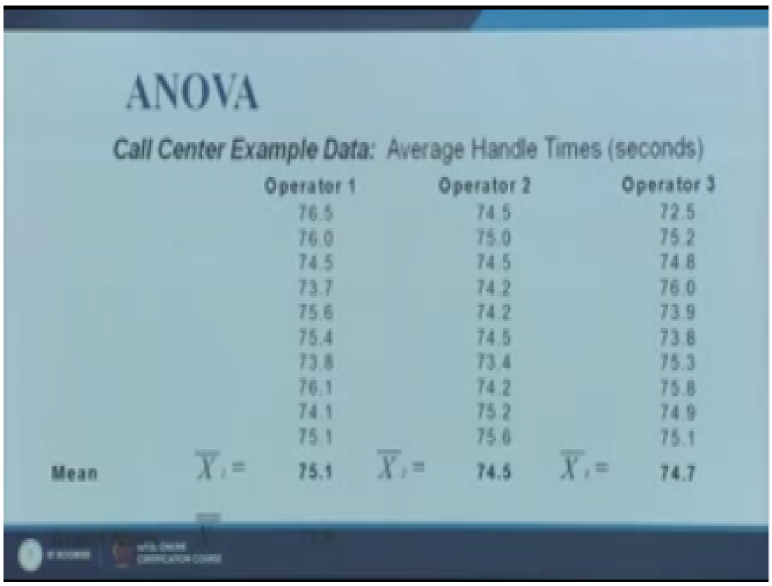
(Refer Slide Time: 11:09)



So now what you are saying the hypothesis is that μ1=μ3 because there are three operators so we are saying the time taken, the average time taken by the first operator is equal to the average time taken by the second operator is equal to the average time taken by the third operator. What is by alternate as a researcher is sometimes I given an example, a researcher is a fault finder is generally his habit an alternative hypothesis through which he finds is early like fault finding, he is trying to find out how come there can be no difference there has to be some difference like he is share look homes, he is like a detective, he is trying to find out.

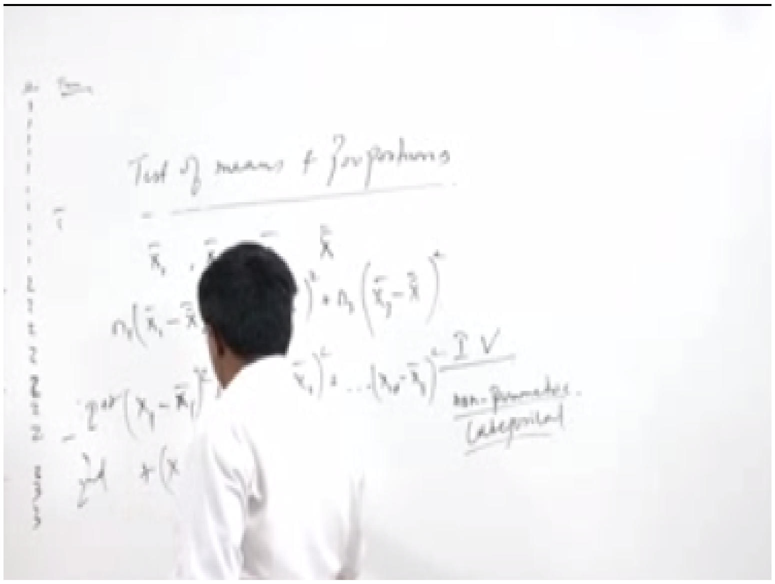
So at least one μ is different he is saying, now let us see this example, so the time id given in seconds.

(Refer Slide Time: 11:54)



So the operator 1, this is the operator 1's data given to you operator 2's data is given to you, operator 3's data is given to you, now you can understand suppose you go to an actual of folder or a file right, how the datas will look like, so you may have, you should also understand this so this is let us say 11111 there are how many 1,2,3,4,5,6,7,8,9,10 so 10 right, 3,4,5,6,7,8,9,10,2,2,2,2,2,2,2,2,2, till 10 again 10 right.

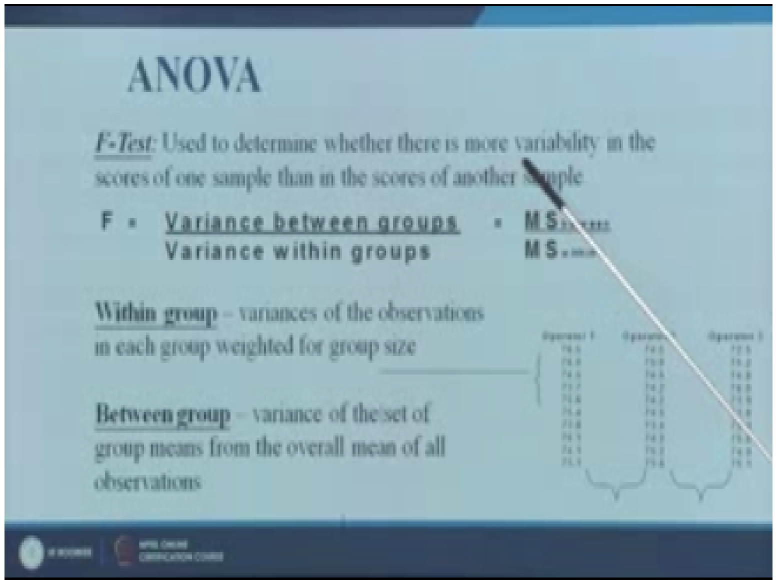
(Refer Slide Time: 12:27)



Then 3 goes on 10 so the values are correspondingly so suppose in any software package you want to use this is how it will look like, so the operator the time okay, this is how you will make because making in the files also in the, you know software files also it is very important how do you put your data that is why I am showing you. Now first is what is saying, let us take the X1 so there are three groups okay, and there are 10 participants in each group okay, so this could be this is the case whether I equal participation it is there could be possibility that there are not equal participation also, okay.

So X1=75.1, 74,2 is the X2 and X3 is the 74.7 so this is the mean, the mean of the first two operator mean of the second operator, the mean of the third operator right, this is something called if you, if it is not visible to you I am drawing it again the $\bar{\bar{X}}$, X double bar is called the grand mean.

So the grand mean is the overall mean right, so either you can add up all this, this, this, this, this till this till all this and then divided by the number that is 30 here in this case so that means or you can simply do it by suppose you have this 75.1x10, so 700.1x10+74.5x10+74.7x10/30 right. So if I do this also I can find it right, so this is my grand mean which is coming 74.8.
(Refer Slide Time: 14:13)



Now F test is used to determine whether there is more variability in the scores of one sample then in the scores of another sample, is more variance there in the score of one operator over the other or something. So let us see now how is using now F, so the F ratio which I have written here right, is nothing but the variance between the groups and variance within the groups so I said mean sum of square is nothing but the variance you are calculating here between the groups and this is the within the groups, okay.

So means it is written here, mean sum square between mean sum square within, so what is the within group let us see, now within group if you can see it is shown here this is the variances of the observations in each group weighted for the group size now this is important many a times you will get equal sizes, group o equal sizes 10,10, 10 in this case there might not be equal sizes so if there are not equal sizes then you have take in to account this factor of group size this has to be waited for the group size if you do not do it then you will make a wrong analysis so whatever the 3 number of groups so size that has to be taken care of okay.

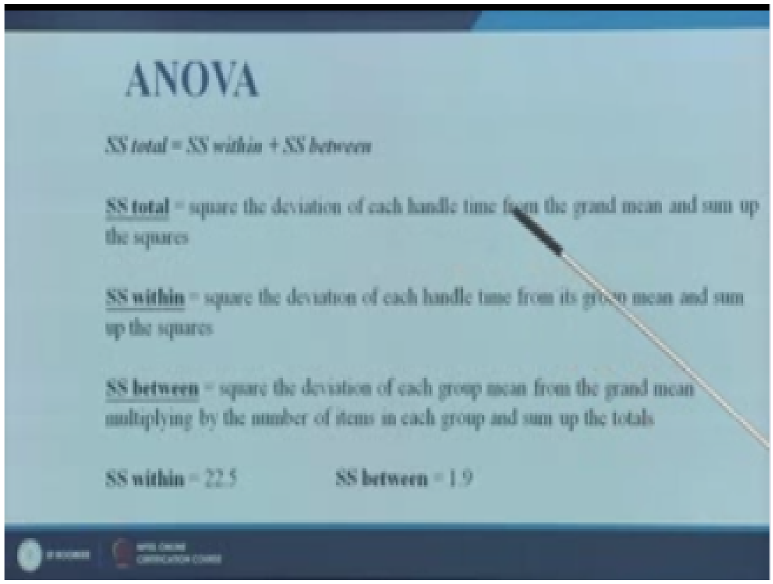
Now this is the between group now between group is this right between this between this and may be this is another one right three possibilities right so there is a variant of set of group means from the overall mean of all observations so what did he saying how much is the variants of all the group means from the overall mean of the observations. Now let me show you here so how it will look like?

How does it look like? So I have three things right I will tell you something the simplest way is to you do not have to remember anything right find out let us say in this case what is set it saying between right son you have the x1 bar you have the x2 bar you have x3 bar right and you have something the grand mean right so you have between groups is nothing but x1 bar minus the grand mean multiplied by the n 1 right.

Similarly xq2 bar so you need a plus you have to add it up all right - $\bar{x}$ so here is n2 is it visible let me do it again so + n2 x this + n3 x$\bar{3}$ - x bar so if you take this so this is what is the between group right now similarly the within group is nothing but x1 – x bar right 2 this is all square okay remember please it is variants this is not a standard deviation this is the variants which is the √ of the standard deviation plus let us say x1 for the first row we are doing only for the first row right.

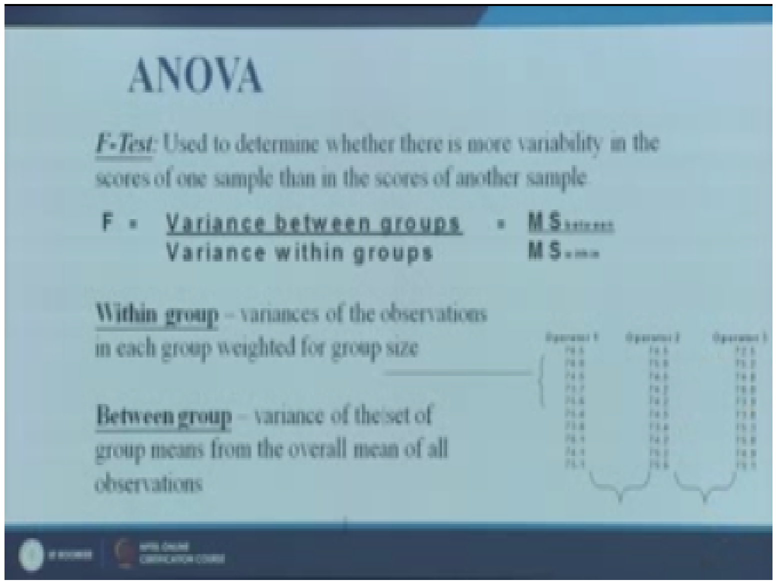
So what will be this? This will be the x 11 x21 x31 x41 x51 x61 x7x8 x9x x10 similarly x first row so the x1 2 x13 x14 so it goes on right sorry this is first second row second column so 12 x12 so this is 22 so x 22 x3 third column third row 3 2 x42 x52 x62 you have to go on right, so this is one is third first row so this is x1 3 first row 3 column right so then is this one let us ay this one second row third column right it goers on till x this is the third row 10th row right x10 and third column right so it goes on you have to add it up right.

So once you do this once you are making it you calculate so you have to find it out from the mean of the group in between the group so you have done this, this is the for the first one now you are doing it for that individually so x1 – x bar or just do it where the group it is simple x1 – x bar 2 + let us say this is only for the first right first set this x2 – x bar right square goes on till x10 – x1 bar 2 + for the second so x let us say again the x1 the first means this one I am saying whether you it independently or you write the way I was writing –x2bar 2 it goes on right. So you have to find out the within group for the entire three okay.
(Refer Slide Time: 19:38)



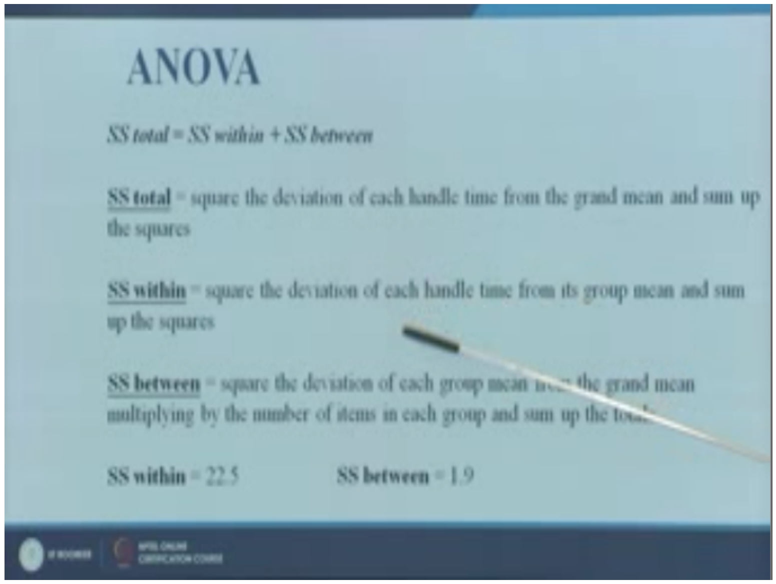
Now this is very simple right, so you are finding the total there three things so total now what is total? Now total is if I am taking the every value each value minus separating it from the grand mean so 76.5 – 74.8 I think it was there.

(Refer Slide Time: 19:58)



Yeah 74.8, so 76.5 – 74.82 76 – 74.8 and 75.1 again till this one then start this one entire group 20 to 25 has to be detected individually from the grand mean okay.

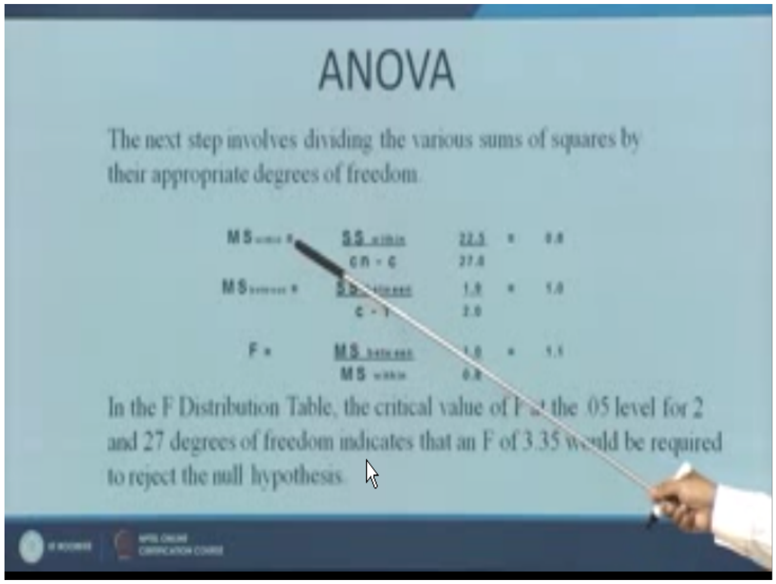
(Refer Slide Time: 20:18)



So I have calculate this is SS total this is SS within this is SS between so if I have the total and if I have within I am not find between also or if I have the total and if I have the within I am not find the between also or if I have the total and I have the between I did not find the within also because this two will sum up to become automatically this that means what, what I am saying is sum of squares total is equal to sum of square within+ sum of square between.

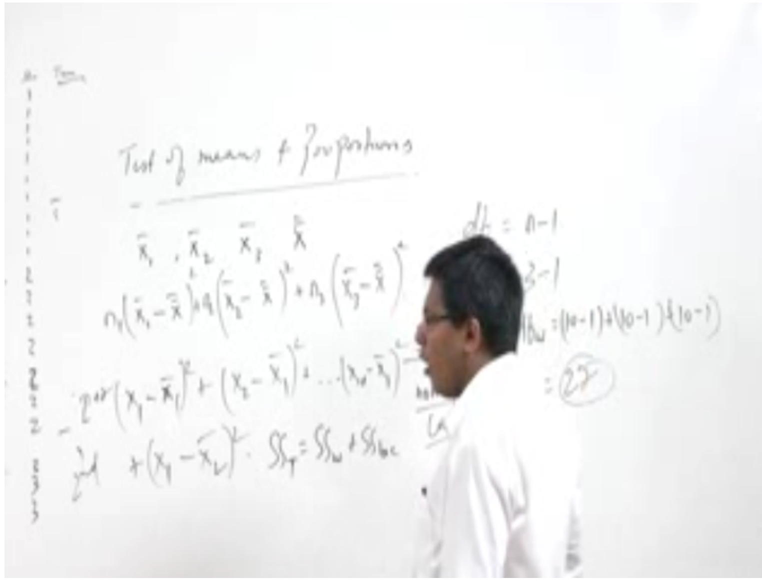
So in case you have the total then and you found one of these the third one you might not also calculate it is automatically you can deduct it and find out what so this is suppose this is 22.5 within it is calculated and 1.9 is this one so what is the total now sum of the total will be so you have to multiply and find out okay so you can say over all this some were in between you cannot do this.

(Refer Slide Time: 21:27)



You can just add up right so 22.5 23.5 24.4 so in some of squares total is equal to 24.4 out of which 22.5 is for the within the groups and between this 1.9 okay so now let us see what is the mean sum of squares now mean sum of squares is the sum of squares within divided by the degree of freedom.

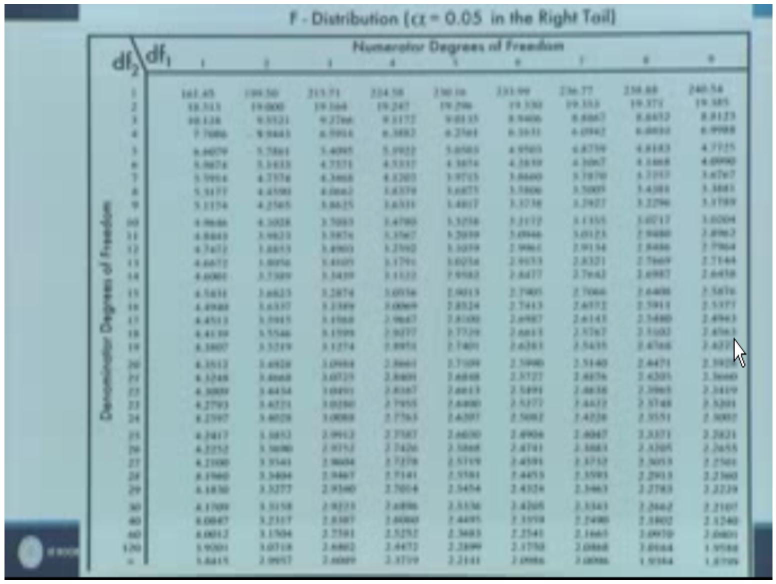
(Refer Slide Time: 21:43)



So I have said now the degree of freedom to degree of freedom is equal to the number of elements -1 right so number of elements -1 so for the degree of freedom between the groups you have let say this case three groups so 3-1 right but when you all doing the within the group let say degree of freedom within it has to be0 for each column you have to deduct 1. So 10-1+10-1+10-1 okay so this is equal to nothing but 27 right or you can send n-k in simple terms right.

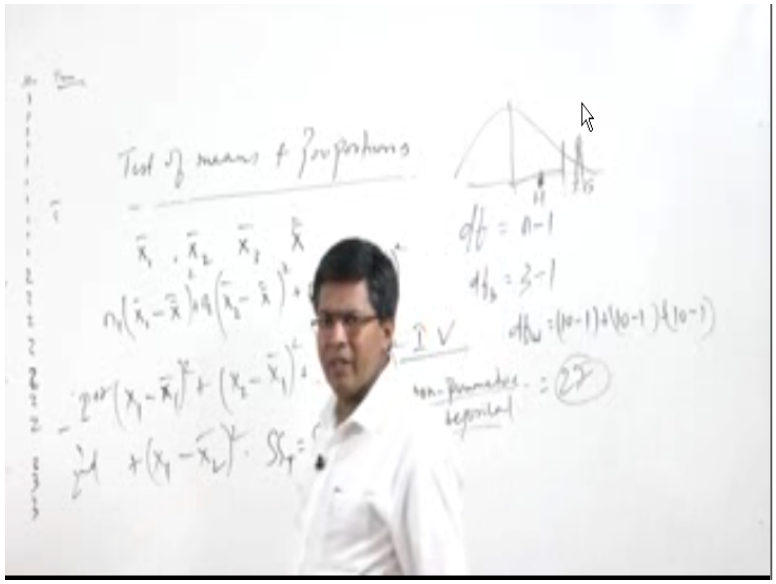
So now the f is coming to we have calculated so this is 0.28 this is 1 so 1/0.8 is 1.1 so if you take the f value at this is the f value let check this how to check I will show you .05 level for 2 and 27 degrees of freedom right 2 and 27 right so did you understand 2 and 27 this is 2 and this is the 27 so between is 2 degree of freedom 27 is the within the group right.

(Refer Slide Time: 22:48)



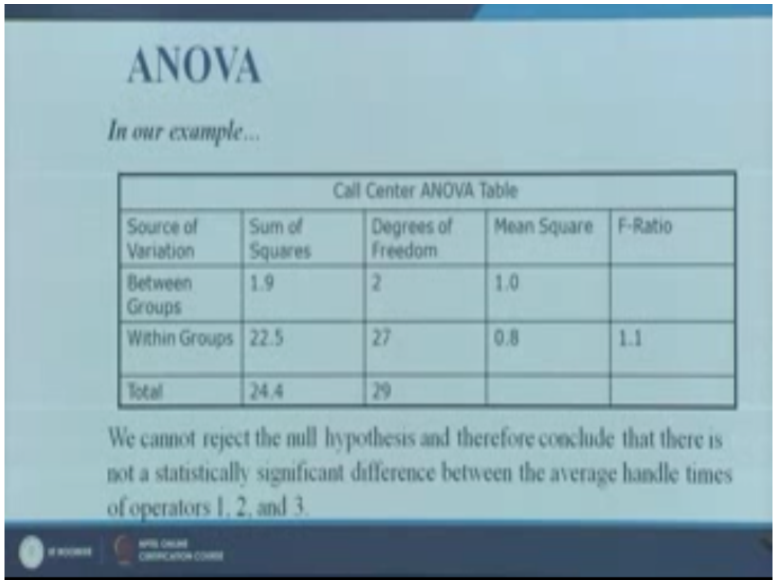
So now let us go for 2 and 27, 2 and 27 so this is something here right sorry 27 is here 3.3541 right so the value I think it is visible 3.3541 yes so 3.35 would be require to reject the null hypothesis but what have we got 1.1 so if you have got 1.1 can be reject the null hypothesis base case.

(Refer Slide Time: 23:14)



The null hypothesis it is coming 3.35 right and our 1.1 so this is anyway you have to understand this is some were here and your value is here so it is well within the home it is well within the boundary so you cannot reject the null hypothesis in this case right there could be some case in which it can cross step boundary okay.

(Refer Slide Time: 23:42)



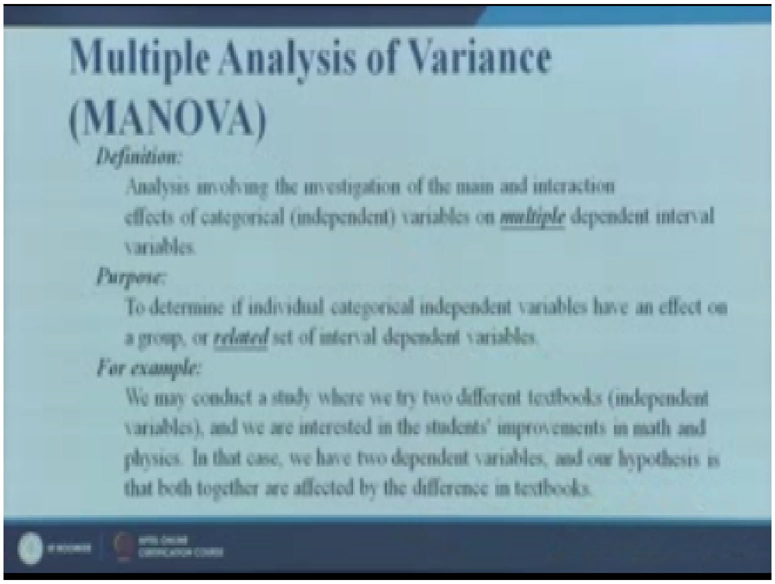
So this is how the if you go for the Anova table it something it sometimes look like this sum of squares if you are using excel, SPSS something sum foe squares between the group is this much within the group this much so total was what I was adding up that time 24.4 degree of freedom is 2 and 27 so total is 29 mean sum of square is 1 0.8 so if I show this one.

So we cannot reject the null hypothesis therefore conclude there is not a statistical significant difference between the average and time of an operator 1 2 and 3 at this case we cannot say but suppose there would have be an difference suppose let us say there would be an difference then you have said at least there is a difference between the mean of first or the second, second or the third whatever it is.

To test this suppose how do you find now to test that we use something called a although manually we are not doing it we do something called a post of test so please remember this so if you go to any you were suing nay software packages like SPSS or something then you are using this post of test which basically does nothing but if it I uses it calculates the mean and it uses the mean to find out what is the which of them is the most let say as the highest value and which one has the lowest value and that it can tell out of this in which significant manner right.

Which one is actually strongest or the highest and which one is lowest as good as that. But now let us say, you have a case in another one more important thing we measure but I will go to it later on it is called interaction effects okay. but before that let me also come with.

(Refer Slide Time: 25:28)



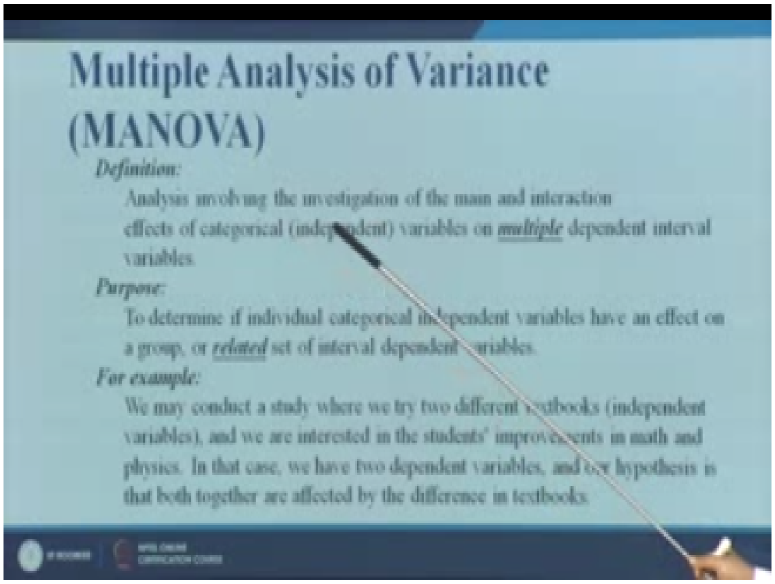
Okay let explain the interaction also many times what happens is there are suppose two levels or two groups okay, or 3 groups right, so in such a condition what happens there might be an, there are two kinds of effects, one is the main effect and there is something interaction effect, now which is important to study. Suppose there are two things right, 2 things individually had an effect on the depended variable, individually they do have an effect.

But what if when these two things come we say in dictionary or energy sometimes the relationship becomes week also, so if two things are coming together they give a third kind of effect, what is the 3rd effect? The 3rd effect which we say that which happens to be the present of different of matters may be in a chemical lab or in compound something. So we say when two things come together automatically.

Let say take the example somebody is enjoying the party, so when his friends are there he is enjoying the lot, if he individually goes with the family also he is also enjoying right but what if when his family and friends come together? Suppose in the same party, will it be the similar effect. So in such a condition the interaction comes into the play, so that is where one need to study that interaction effect can have vary on any study.

So if the researcher is doing any kind of study in the experimental designer or anything, they need to conduct and show it has a result in the many research outcomes or the research paper or in the cases any where okay. Because the interaction has a larger effect in the real life then in sometimes the main effects, it is possible right. Now we come to a situation where we say.

(Refer Slide Time: 27:40)



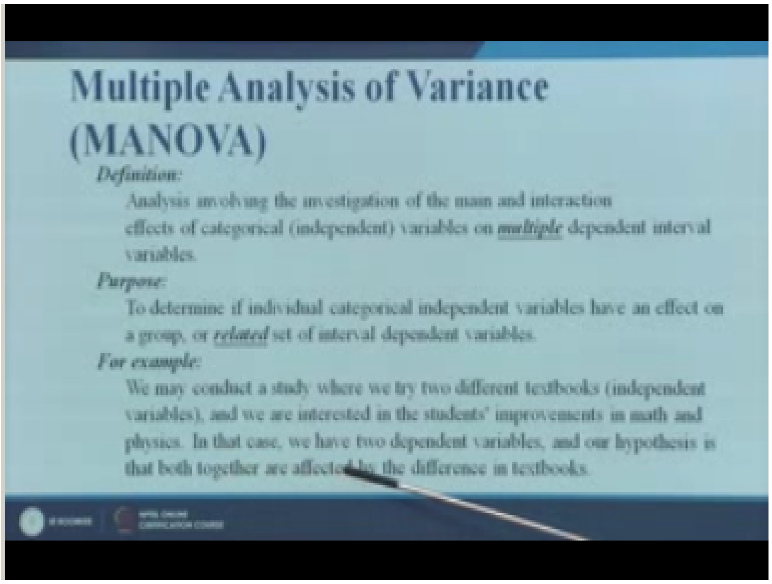
Multiple analyses of variance, now earlier you are talking about one dependent variable and multiple independent variables 2, 3, 4 whatever. Now what if I have more than one dependent variable let see this case, analysis involving the investigation of the main and the interaction effects of categorical independent variables, on multiple dependent interval variables. So if multiple dependent interval variables how would you make the study, so this is the case where we are talki8ng about basically we say is called the MANOVA.

So there are many suggestion, there are many studies, in fact most of the people generally do not do this test because they are not aware but they are not difficult at least if you are using any software everything is there, you will go to general linear body and you can do a SPS of MANOVA which can easily tell you when two dependent variables are brought into together in the same time, how will the independent variables will affect them?

So to determine individual categorical independent variables had an effect on the group or related set of interval dependent variables are not, so this is the purpose take an example. We may conduct a study, where we try to use two different text books, so we are using different text books, so which are the independent variables because the change in text book will affect the change in the dependent variables.

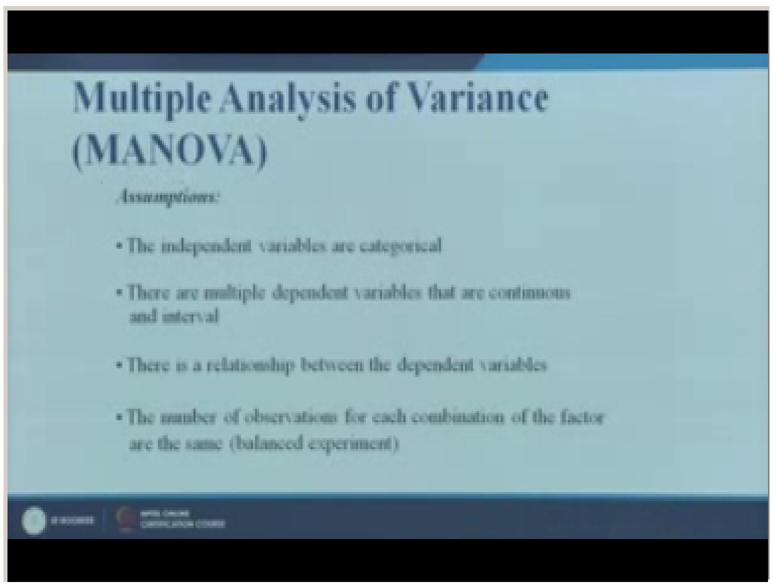
And we are interested in the outcome in the students improvement in math and physics score okay, so in this case that means the math and physics becomes my two dependent variables right, we have two dependent variables and score, score is obviously the continuous variable so we are measuring in 50, 60, 65, 70 whatever the scores are and the hypothesis is that both are affected by the difference in text books, so we are saying that in such conditions.

(Refer Slide Time: 29:58)



The effect of you know the interactional effect comes into a larger play right so we are saying that let this two text books are having an impact on the dependent variables which is the math and physics score.

(Refer Slide Time: 30:21)



Now what are the assumptions the assumptions are the independent variables are categorical the multiple independent variables are continuous and interval now continuous and interval okay I would have gone to the 3rd case there is a relationship between the dependent variables so this is the assumption so you just cannot.

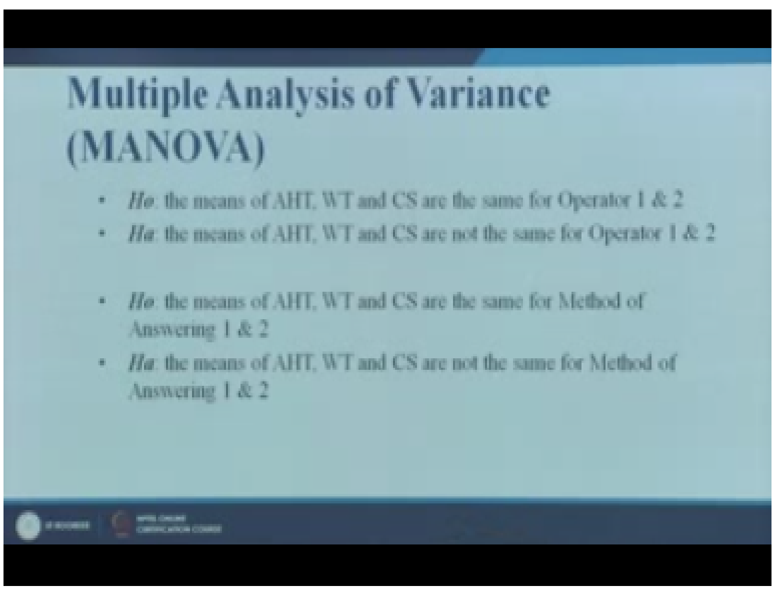
But in any dependent variable that you like know that has to be a theoretical justification why you are using it as a dependent variable and why you are using a manova if you are feeling that there is a relationship between the two dependent variables right a and b here dv1 and dv2 then such a condition manova fits into the situation number of observation for each combination of the factor are the same it is the balanced experiment right.

Now same example I would just show you how it will look like the call center manager wants to know if the operator or method of answering calls makes a difference on average handle time wait time and the customer satisfaction earlier I think we were talking about only the average handle time right so now we have brought in two different things now the wait time and the customer satisfaction so there are three basically dependent variables now earlier we had only one right so one this one is this one are they not related yes they have a relationship the average handle time how much of the wait time and finally what is the customer satisfaction they are the dependent variables and the independent variables are now only two things call operator.

So now who is the call operator let us say who is the call operator let us say when we give promotions we find out the persons you know how effectively he works or how nicely he performs his job so how is the call operator performing let say in that case right and is the method of answering so sometimes the call operator might not be the only factor that can affect the dependent variable the satisfaction and all.

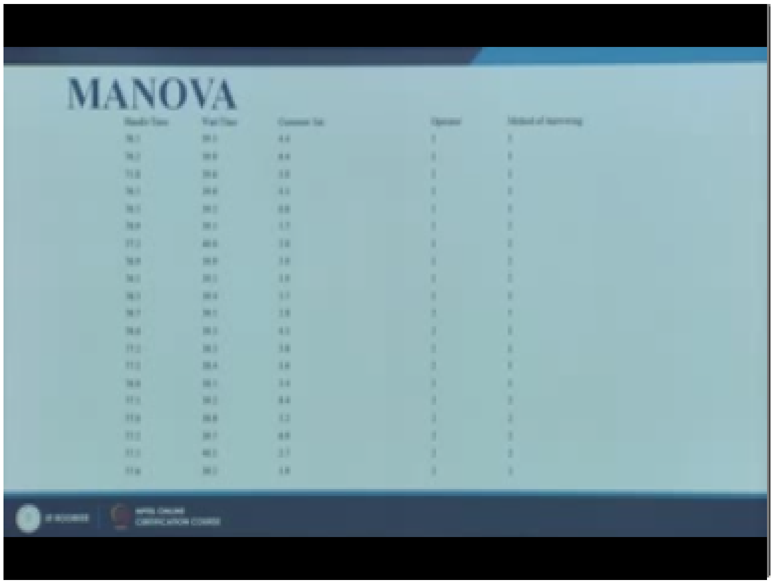
So it could be the method of answering and so how is he answering is he answering through some other some device which is not very clear sound is not going well or some other device or some method which is using which is more clearer and you know clearer to the customer.

(Refer Slide Time: 32:26)



So my hypothesis now is that average handle time wait time and customer satisfaction are the same for both the operators 1&2 what is the next hypothesis null hypothesis the average AHT, WT and CS are the same whether you have a used method 1 or method 2 there are two methods right similarly the alternative is not the same for operator 1&2 the alternate is not the same for method 1&2 right.

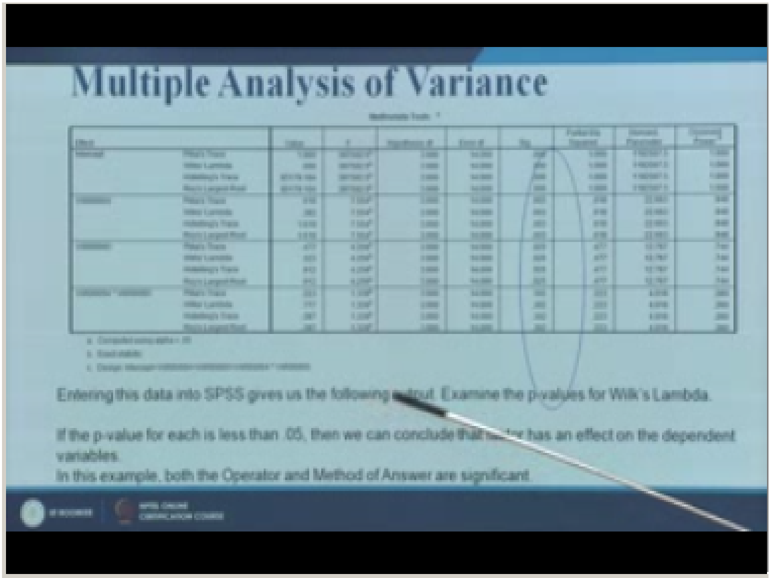
(Refer Slide Time: 32:53)



So this is how it looks like so the total time waiting time handle time customer satisfaction operator 1&2 method of answering 1&2 right so if I'm using this method so here an annova will not fix okay the question is then why could not you do two ANOVAs right you might be asking yourself in your mind may be possibly that why did I do two an nova individually one taking this group individually one depended variable then one of them and so if I do it this also if you see how many times how many combinations are coming each time we are argue we say handle time with this two right only operator let us say or method of answering waiting time again.

So again we are doing the same thing that we were doing in the case of tea test and something right so in such conditions the combinations will increase and more the combinations or more the number of sorry you know ways of doing more that number of reputation you are doing again and again individually when you are doing so the errors will go gone increasing so manova becomes a very good techniques so this is something am just showed you this is something when you do in a software I have brought it I manually cannot do it now.

(Refer Slide Time: 34:10)



So if you do it by this is called something λ if you look at this table and these are the variable four and five right so if you look at this now you have to see this significant values now the significant values I think this is also important for you to know right this significant value is the value which helps you to reject or except a hypothesis in this condition 0.003 for the variable 4 suggests that the null hypothesis is to be rejected because it is less than 0.05 right if we have taken 95% confidence and similarly this also right.

But look at the 3rd one so what we have taken we have taken a interaction between the independent variables 4&5 now what if I take an interaction effect of operator and the method of answering right now when I'm taking the interaction effect if you look now the result is no more significant so that means what we are saying there is a main effect.

But there is no interaction effect in this case we cannot say that there is an interaction effect it is good if there is no interaction effect it is good but what if interaction effect would have been very significant one then that means you would have said two things when combined together only do a better job of sometimes or may be do a job sometimes whatever it is right so this is all for analysis of variances and multiple variance thank you.

For Further Details Contact

Coordinator. Educational Technology Cell
Indian Institute of Technology Roorkee
Roorkee 247 667
E-Mail Etcellitrke@gmail.com etcell@itr.ernet.in

Website: www.itr.ac.in/centers/ETC. www.nptel.ac.in

Production Team
Sarath Koovery
Mohan Raj. S
Jithin. K
Pankaj saini
Graphics
Binoy. V.P
Camera
Arun. S

Online Editing
Arun.S
Video Editing
Arun.S

NPTEL Coordinator
Prof B.K Gandhi

An Educational Technology Cell
IIT Roorkee Production