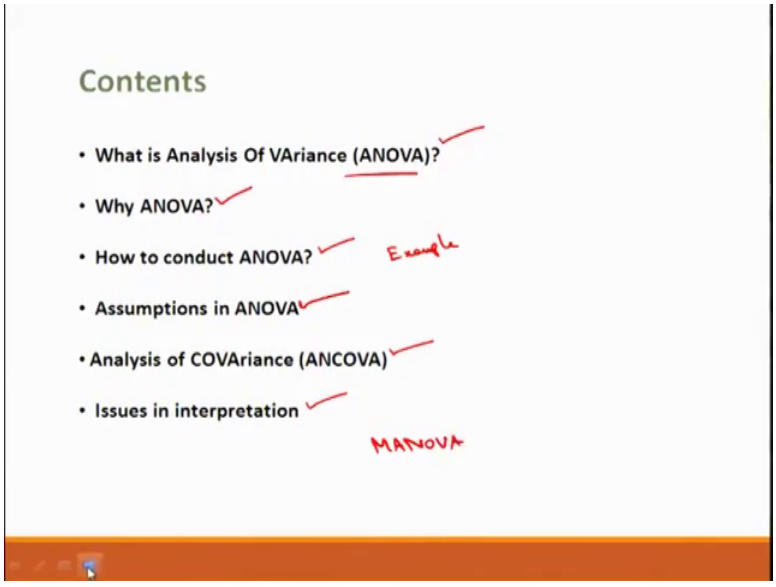
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Mr. Sanjeev Newar
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture - 23
Analysis of Variance (ANOVA) Part 1

Good morning welcome back to the course Practitioners Course in Descriptive, Predictive and Prescriptive Analytics. So, till now you must be very clear. What are distributions? What is normal distribution? Which is the main distribution that can set to be the building block of all other distributions so; you have gone through hypothesis testing t test. Today in this lecture I will discuss ANOVA Analysis of Variance, this can be said to be back bone of analytics, if we are working in predictive analysis here.

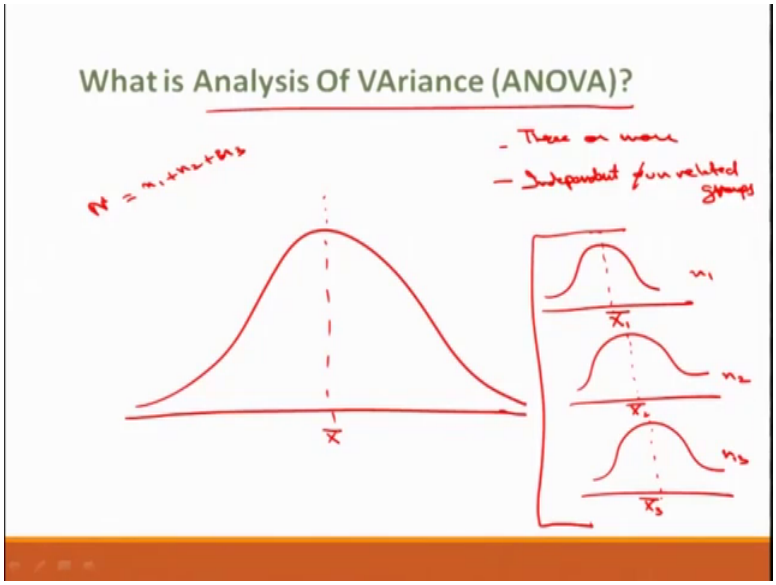
(Refer Slide Time: 00:55)



So, the contents will go like this what is analysis of variance ANOVA, why do we need ANOVA how to conduct ANOVA, I will do this with the help of an example here, then will see certain assumptions which are taken in ANOVA, then we will see analysis of covariance we will see the difference between ANOVA and ANCOVA and certain issues in interpretation also, I will give the term MANOVA. So, what is analysis of variance analysis of variance or ANOVA is a statistical technique that is used to see if there exist

significant differences between means of three, or more independent or unrelated groups, three or more independent or unrelated groups.

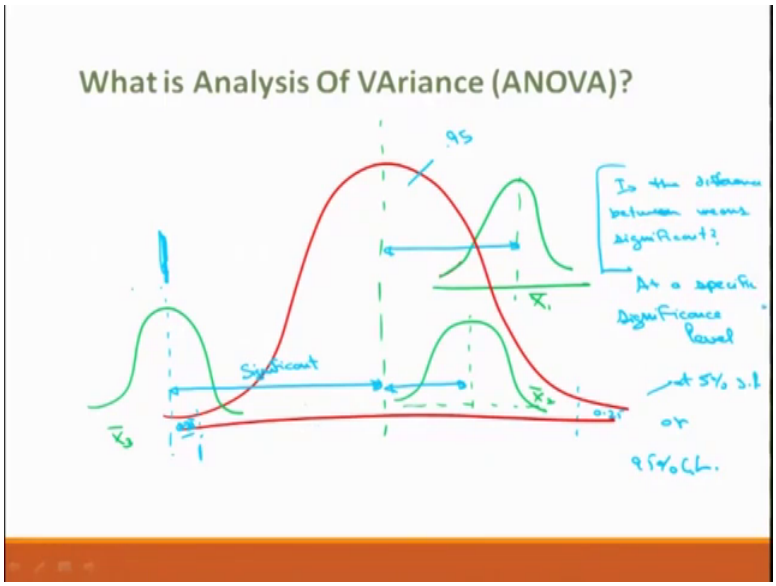
(Refer Slide Time: 01:48)



So, you know this is distribution by this shape, the first thing comes in to your mind is normal distribution, this is my normal distribution for the whole population. If I divide a population into groups, I have put here three or more groups, I divided into three groups let me say, they are three groups three small groups. And let me say there are N elements in the population and n 1 here, n 2 here and n 3 here, then N is equal to n 1, plus n 2 plus n 3 that means, these three groups form the whole population.

Now let me say this group has mean X 1 bar group 1 bar group 2 bar and group 3 bar. The mean for whole population is X bar. Now what would ANOVA test ANOVA would test at is the main similar do these groups belong to the same population, or have they come from the different population, these things would be tested in ANOVA.

(Refer Slide Time: 03:37)

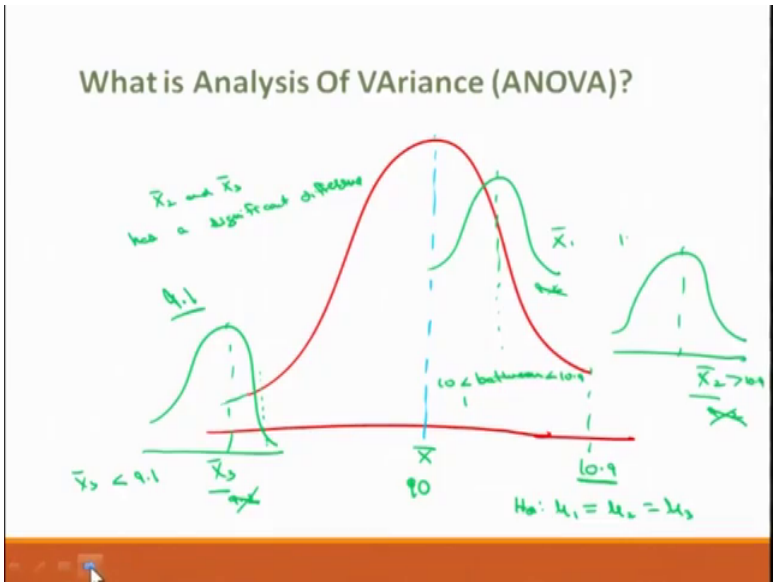


Now, let me try to represent this visually here, if this is my full population and 3 groups are placed let me take a different colour, 3 groups are placed like this, this is the total mean group 1 group 2 let me say group 3 here. Now we can see there is a difference between the means of these groups this is the difference, this is the difference in the means, this is the difference in the means.

Now is this difference significant, this is the question is the difference between means significant, this is the question let me put in a different way here. So, let me say 1 of my group here is lying outside this population, this is the group X 3 though there is some overlapping, but it is lying outside this population; that means, this difference is significant and we put the level of significance as well.

So, is it significant at a specific significance level. So, you know what is significance level, if I put 0.25 and 0.25 here, the area here it is 0.94 other many area this means, it is at 5 percent significant level, or 95 percent confidence level CL, this is 5 percent significant level. So, is this mean this mean away from or falling outside my population mean; that means, this there exist a difference of significant difference of means and do these means lie inside. So, let me take one more example here.

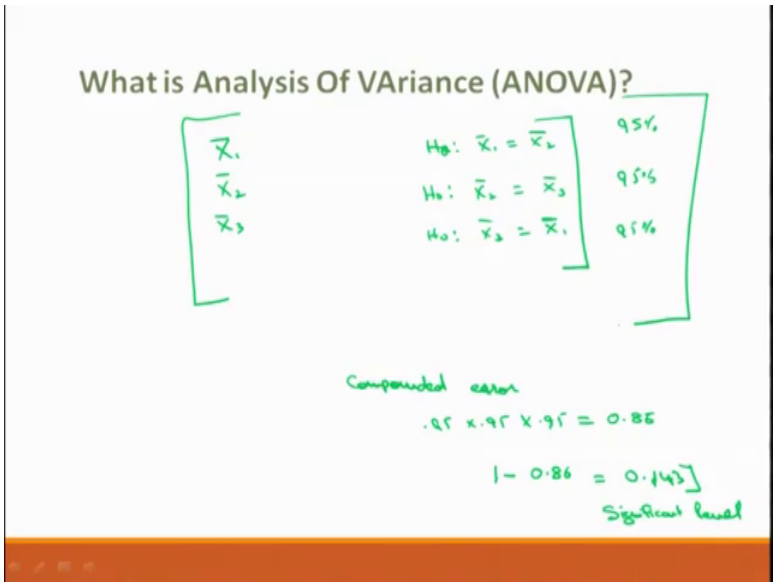
(Refer Slide Time: 06:48)



So, let us say I have X 2 here, and X 3 here, who have a X 1 is lying inside this is X 1 bar here, I have X 1 bar, this is X 2 bar this is X 3 bar in this case X 2 bar and X 3 bar has a significant difference from X bar. So, these are lying outside X 2 and X 3, if I put significant available here. So, means are in different locations here X 3 and X 2, X 1 is quite close to the population mean, but the overall relative mean is away from X 2 and X 3 group means here.

So, the hypothesis in ANOVA is put like H not null hypothesis here is mean 1 is equal to mean 2 is equal to mean 3, please remember we are not saying that they are exactly equal; that means, the means are similar they belong to same population, exactly equal means for example, this mean is 9.2, this would also be 9.2, this means also be 9.2 this is not the case, if this mean is suppose 10, this 95 percent limit is let me say 9.1, and here I have 10.9 as my 95 percent interval limits here so; that means, this mean is between 10 and 10.9 between 10 and 10.9 and this mean X 3 bar is less than 93.1, in this case X 2 bar is greater than 10.9, which is my limits here right hand side and left hand side limits.

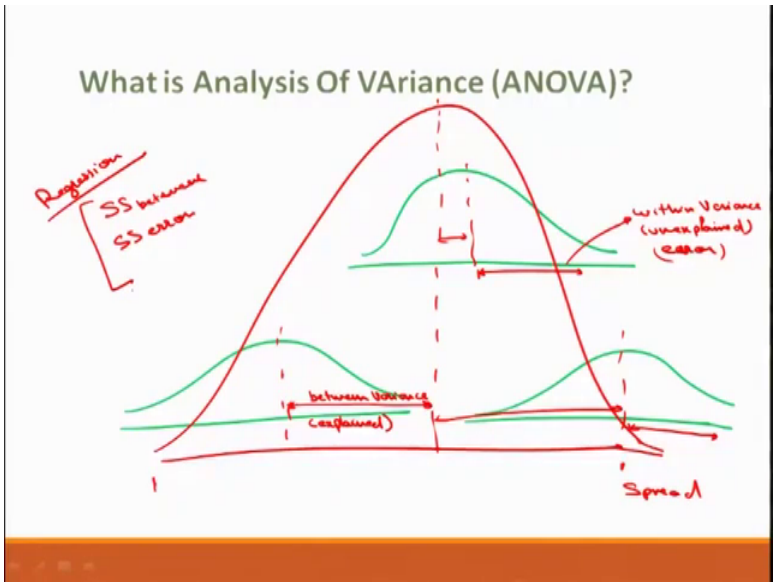
(Refer Slide Time: 09:21)



So, we can see here is X 1 and X 2 and X 3 group means, we can compare them individually using t test t test hypothesis may be states that for comparing X 1 and X 2 we can say X 1 bar is equal to X 2 bar, or second hypothesis may be X 2 bar is equal to X 3 bar, third hypothesis may be X 3 bar is equal to X 1 bar. So, all these three means are compared individually.

And if I say I have 95 percent confidence level here, 95 percent confidence level here 95 percent confidence level here overall population confidence level would be compounded compounded error I could put here, that would be 0.95 into 0.95 into 0.95 which comes down to 0.86 something. So, 1 minus 0.86 which is equal to 0.14 becomes my significant level here. So, we can use ANOVA to test all these 3 groups together 3 or more or multiple groups can be tested together. So, this is what ANOVA would do.

(Refer Slide Time: 11:16)



So, you know what is variance; variance, is the spread this spread, this spread from this point to this point is variance or. So, let me say my three small groups which am taking here have larger variance, let me say this variance is large and, this variance is again large and this variance is again.

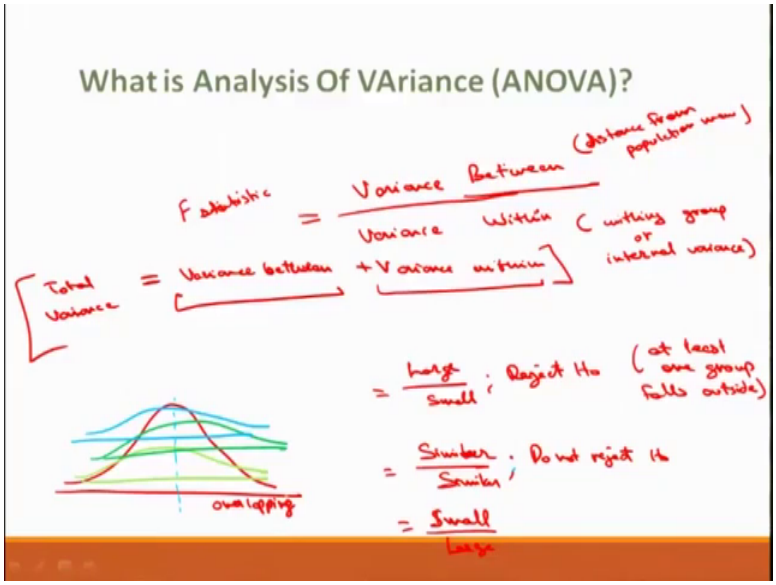
So, where does this bring us to, if my small groups the groups are small, if this spread is larger that is the variance is larger it is more likely to be within the group, it is that is the overlapped be higher, if this variance of small group I am talking about the small group 3 small groups that I have if the variance of small group is smaller, and they are away from my population mean; that means, that is significant difference.

And if this variance is large that is trying to bring it closer, trying to bring it a little overlapped to the, my population mean here. So, if this variance is large. So, this variance that is in my groups. So, better to say within my groups is my internal spread that is known as within variance, and for the whole population this difference, the variance of the groups from the population this difference is known as between variance between the population mean and my group.

So, this is also between, this is also between this is within. So, overall mean has distances from my groups and my groups individually also has spread. So, this variability between variance is also known as explained variance, this within variance is unexplained. So, the whole population is not able to explain this.

So, we call this as error. So, if you recall we had SS between and SS error in our regression. So, these terms will use here to see how does ANOVA work.

(Refer Slide Time: 14:44)



So, this between and within variance tell us about the significance of difference. So, this is distance from population mean to make it clear again this is within group, or internal variance. So, they can be certain cases here. So, we know that variance between and variance within variance between plus variance within brings our total variance in the data. So, this is explained by the data this is within group this is not explained by the data.

So, if the variability between the means that a distance between the oral mean that have put in a numerator here, I will put this as a term and put F statistic here. So, I am not putting equal to sign here, am saying statistic is proportional to this relation what is it exactly equal to we will see. So, they can be certain cases here, when variance between that is numerator is large and denominator is small.

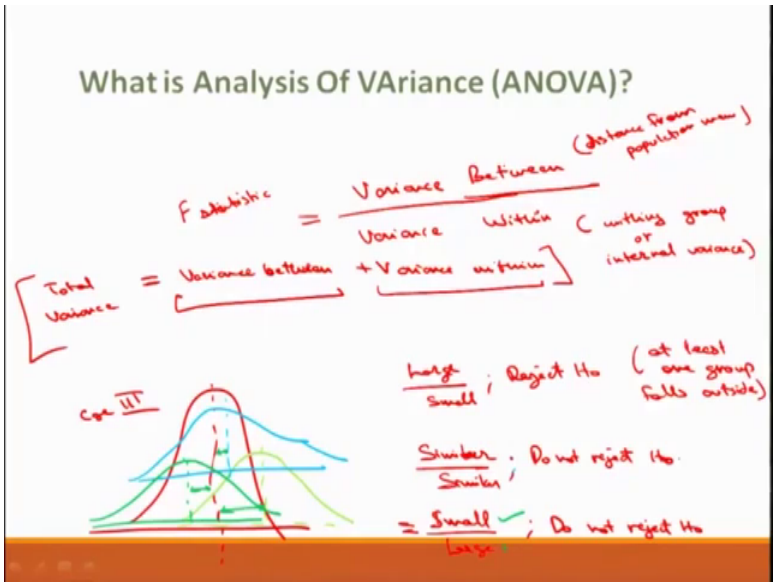
So, there can be three cases 1 is the numerator is large denominator small other thing I have considered say these are similar, third case is this is small and this is large and remember what was our hypothesis here that these means are equal. So, this hypothesis will be rejected if the means are not similar. So, if the means are not similar what is a this case this case is variance between that is that is this distance the distance between the

group means and the population means is large is very large; that means, these groups are located far away from my population mean.

So, what does that do that makes it not to fall in the population? So, in this case the null hypothesis of similar means is rejected. So, what we will see here we reject the null hypothesis. So, I did at least 1 group falls outside and, we have small or narrow within group means here. So, if they are similar means, if the cases like this. So, what is here in this case this is my population mean, if I say and I have three groups here, this is let me say group 1 group 2 and group 3.

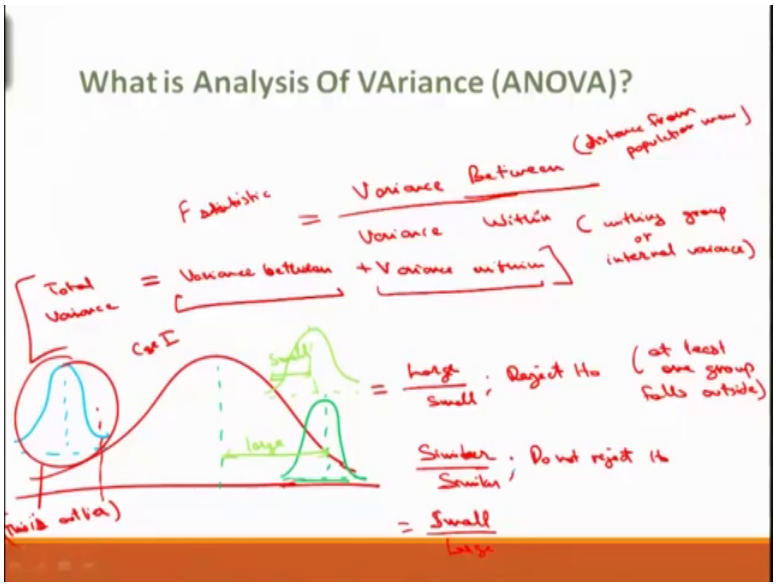
So, what we can see here is the means are very much close; that means, we fail to reject the null hypothesis do not reject H not and, we can see overlapping here and, if the population mean is close to the group means.

(Refer Slide Time: 19:48)



So, this is case three where the denominator is small; that means, the population mean is not very much distant apart from the group means I will put group 1 group 2 and group 3 here. In this case you can see the dark green, blue and light green, they are three groups with means X 1 bar X 2 bar and X 3 bar the small distance is from the population mean; however, their within group mean is large, in the previous case I had the similar means this is both the similar the spread and the spread for both the population and the distances are all similar. So, in the third case also we do not reject H not.

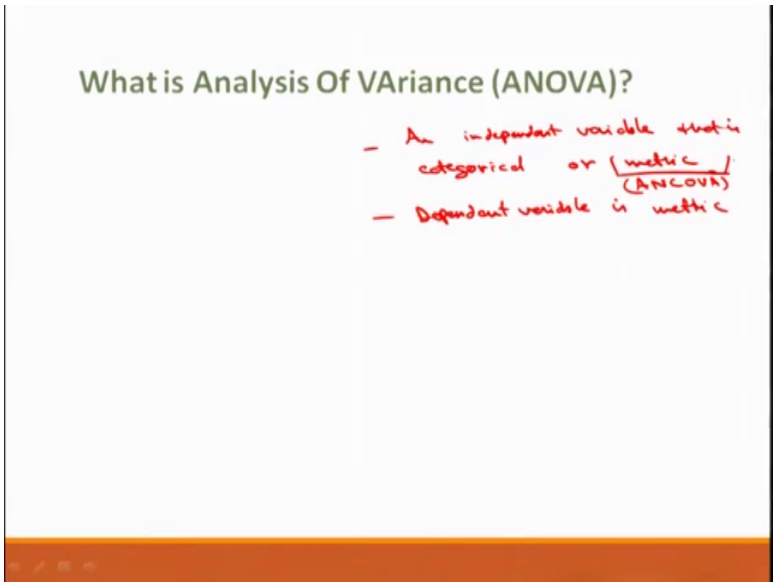
(Refer Slide Time: 21:26)



So, what would be the case 1 case 1 is population am having the narrow, or small within group variances, in this case within group variance is small this is small; however, this is large. So, this is my case 1 case 1 provided this one of the group is outlier this is outlier. So, these are my 3 cases.

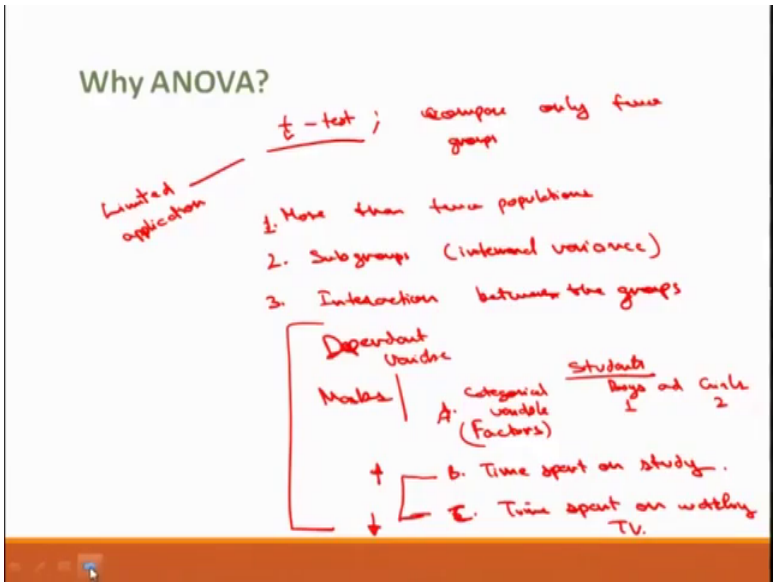
So, my H not was all the variances are equal and attached ANOVA would be conducted for more than 2 populations.

(Refer Slide Time: 23:00)



So, ANOVA should have an independent variable that is categorical, or metric however, the dependent variable is metric in business terms if this variable is metric; we call the term analysis of co variance.

(Refer Slide Time: 23:51)

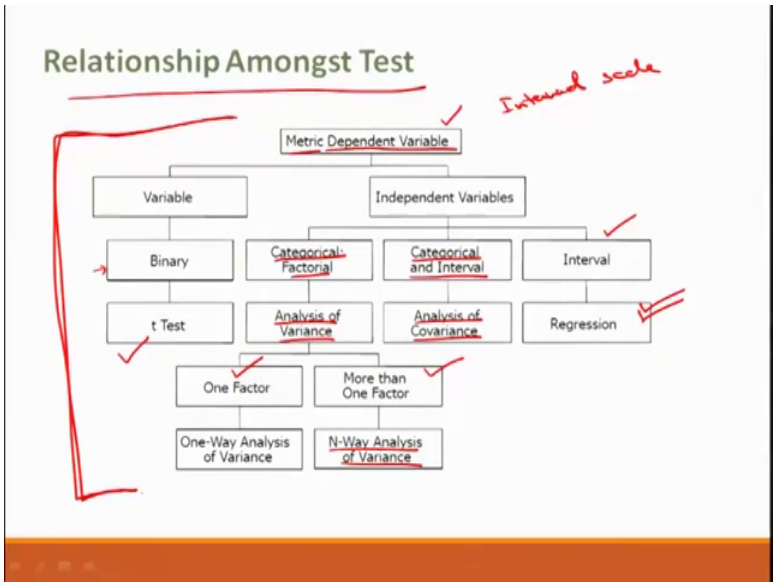


So, next terms, why do we need ANOVA earlier what we did we had t test t test could compare only two groups, if groups are more than 2 number 1, we have sub groups that is internal variance also exists and, we will see that interaction between the groups that is 1 group is interacting with other group as well for instance. If I am saying my metric variable is the final marks of the student and variables are let me say boys and girls is a categorical variable and metric variable here may be the time that spent on study and of one of the variable may be the how much time they watch TV ok. If I say marks ok, this is my dependent variable ok, then I say my students that divides into boys and girls. So, this is a categorical variable that has 2 levels 1 is boys second is girls.

So, next is let me say there are variables, this is variable A variable B these variable also known as factors time spent on study time spent on watching TV. So, there might agree some interaction between the time spent on study and time spent on watching TV. So, ANOVA without interaction consider that there is no relation between these two however, if he spends more time on study he would spend less time on watching TV. So, there might exist an interaction here.

So, ANOVA is used to compare more than 2 populations and sub groups internal variances also consider there and we can also the interactions. So, all these things t test cannot t test has a limited application.

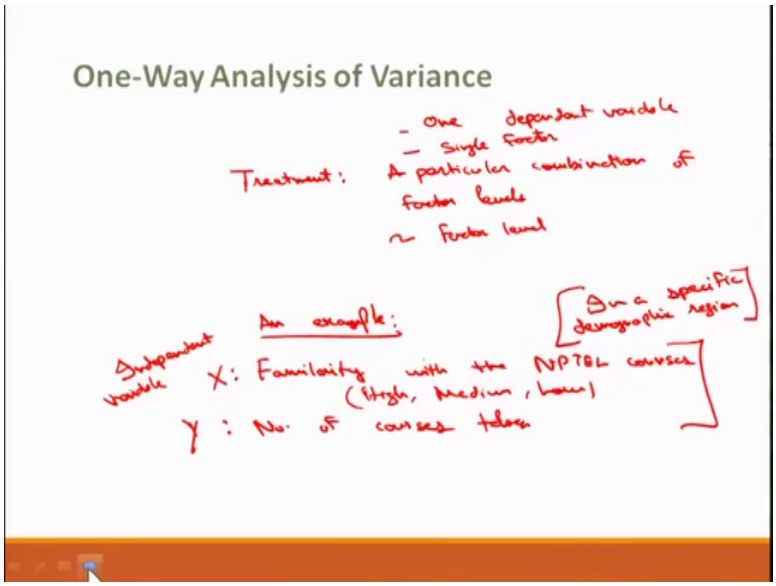
(Refer Slide Time: 27:21)



So, next comes the relationship amongst test. So, we have seen t test we have seen regression, where did we do t test when the depended variable was a metric that is it was in interval scale and, my independent variable was just variable it may be metric non metric that binary test was used that was a t test. And if my independent variable is categorical that is factorial, we use ANOVA technique and if we have categorical as well as interval variable that is metric l as well as non metric variable here and, it is of co variance is used for instance, in this case analysis of co variances would be used, we have categorical variable as boys and girls 1 categorical variable that is the gender, 2 non categorical variables which are the metric variables time spent on study time spent on watching TV.

So, if both the variables dependent variable and independent variables both are metric or both are in intervals scale, then we use regression ok. So, if we have only 1 factor we call it a one way ANOVA, if we have more than 1 factors we call in n way ANOVA. So, please keep this in mind which test or which technique is applied and in which situation. So, next is one way analysis of variance. So, one way analysis of variance is a technique to test 2 or more variables and it is influence on a single independent variable.

(Refer Slide Time: 29:23)



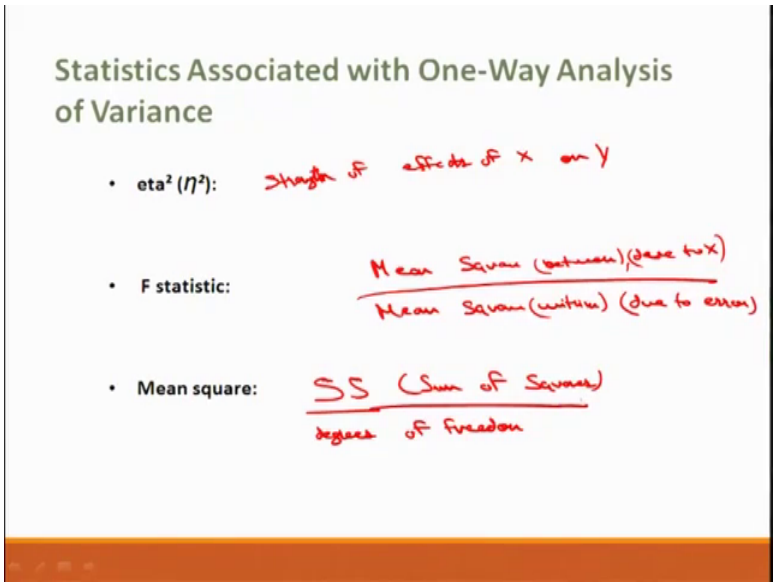
We have one dependent variable and two or more independent variables ok. So, there are certain terms that we will keep using one is treatment, treatment is a particular combination of factor levels or categories for example, if I say if I made the categories boys and girls and I have distribute the time into in to a ratio scale. If I say time from 0 to 5, 5 to 10 and so on up to 45 to 50 10 different classes are there. So, one category or one treatment would be boy whose spends from 25 to 30 minutes this is 1 treatment another treatment may be boy who spends 15 to 30 minutes 15 to 20 minutes that class that class size is 5 minutes here and, it may be a girl who spends 0 to 5 minutes that one set is my treatment here, it is a particular combination of factor levels.

So, one way analysis of variance involves only one categorical variable that is the single factor and, in this ANOVA a treatment is same as factor level a treatment is same as factor level because, we have only one factor. So, they have certain examples of one way ANOVA, for instance do the various product segments differ in terms of the volume of product consumption, or what is the familiarity of the engineering students to NPTEL courses is it medium low high.

So, in this case their familiarity to NPTEL could be might variable and, it is category would be medium high or low this is my one categorical variable. So, this is one way ANOVA. So, I can put it here an example familiarity with the NPTEL courses ok. So, this is my independent variable that is familiarity may be high medium, or low and my

dependent variable here might be the number of courses taken. If I am testing it in a in a specific demographic region, let me say in a specific state in Uttar Pradesh or may be you can say in a specific part of the country.

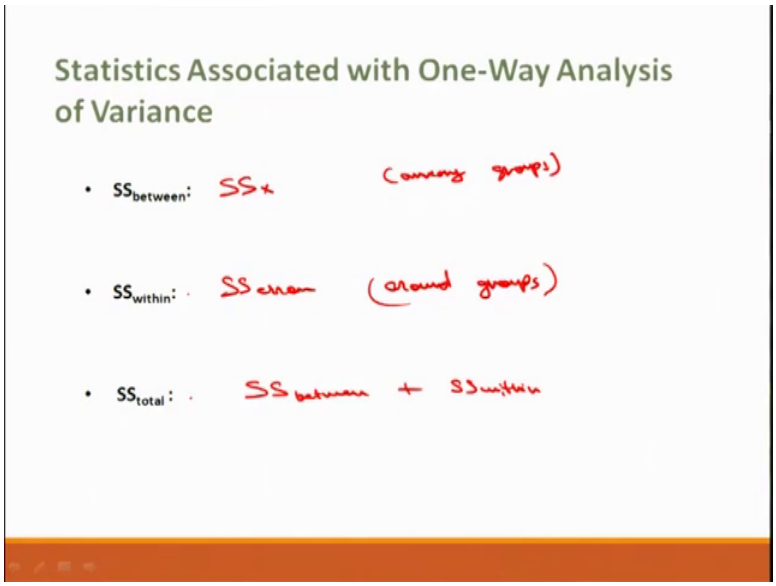
(Refer Slide Time: 33:08)



So, there are certain terms that will use here in ANOVA number 1 is eta square eta square is a strength of effects of X that is the independent variable here, strength of the effects of X independent variable on Y that is made by eta square. And it is value varies between 0 and 1 strength of effects of X on Y. Now F statistic F statistic is nothing as we just discussed it is the ratio between the variance between and variance within.

So, in ANOVA (Refer Time: 34:09) ANOVA table that we will just see we have mean square, mean square between that is mean square due to X and, mean square within that is within group that is due to error. So, what is mean square, it is the sum of squares sum of squares that we did in the regression sum of squares divided by their corresponding degrees of freedom.

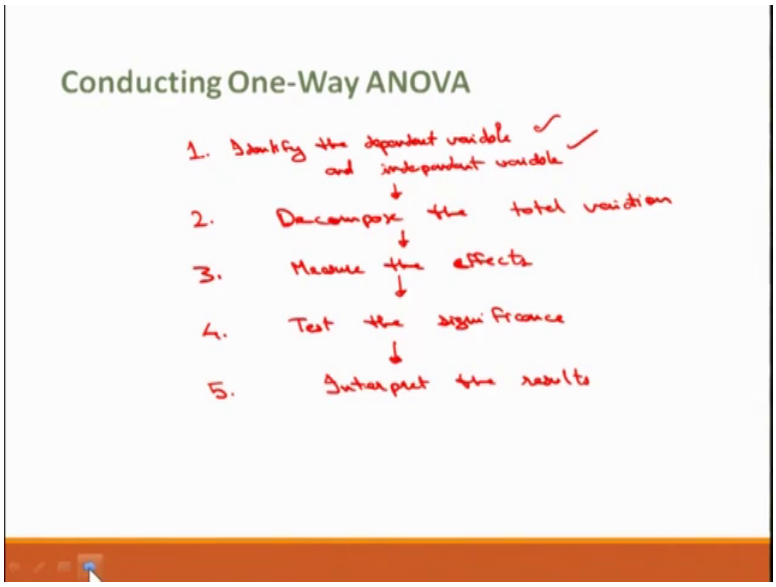
(Refer Slide Time: 35:04)



Next three terms are again SS between SS within and SS total SS total is equal to the total sum of squares, it is SS between plus SS within just recalling and this is nothing, but the variation between groups that be also denoted by SS X this is also to denoted by SS error.

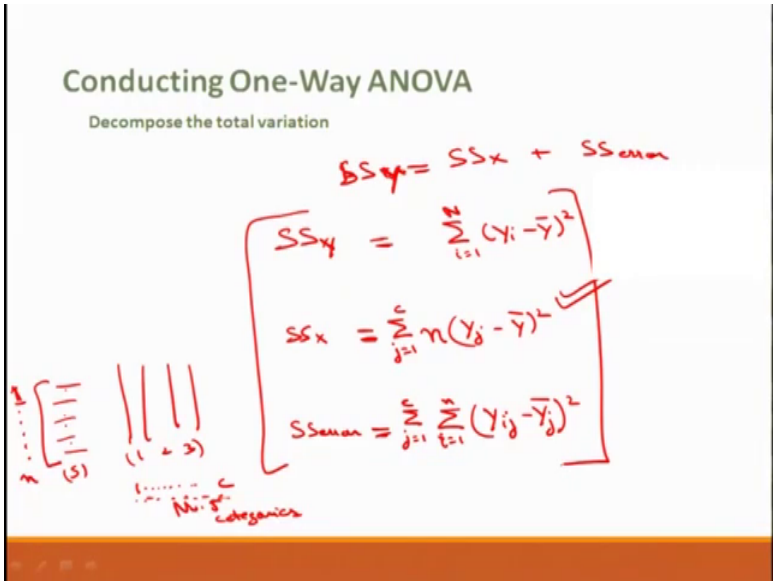
So, we call them around groups and, this is among groups. So, next is how to conduct one way analysis of variance. So, I will put a flow chart here.

(Refer Slide Time: 36:02)



Or the steps here first, first step is we identify the dependent variable and independent variable. So, please note independent variable comes dependent variable is the result is the response here. So, we need to identify which is the dependent variable that is coming because, of the independent variable. So, second step is we need to decompose the total variation, decompose is we decomposintive 2 components between and within third step is we measure the effects, effects of the independent variable on dependent variable here, then we test the significance and finally, we interpret the results all these steps go in a progression 1 after the other. So, I will take an example and to use all this steps to explain how ANOVA works. So, first thing is identifying dependent and independent variable.

(Refer Slide Time: 38:10)



We will see in the given table the example which will take here, what are independent and dependent variables decomposing total variation is nothing, we say that SS total is equal to SS due to X they is explained plus SS due to error, this is decomposition here.

So, SS between is the variation in the dependent variable, that is a related to the variation in the means of the categories of S, that is with the within this categories here. So, for this reason SS between is also denoted by SS X, SS error is the variation in Y related to the variation within each category of X. So, SS within is not a counted by X so, but it is within the category. So, there therefore, it is known as SS error here. So, as we know that SS X SS I will put Y here total is equal to SS X, SS error we saw these relations in

regression this recalling those this is Y i minus Y bar whole square i varies from 1 to capital N here ok. So, SS X was Y j minus Y bar whole square and, we multiply this by N that is the group size, if we have number of groups and these are inputs.

So, what is the group size this is my group size, in this case 1 2 3 4 and 5 group size is 5. If I say I have these groups 1 2 3, there are three categories and this group size is n and in this case this is for each group j varies from 1 to c here to c is my 1 to c is my number of categories ok, then SS error is within group for all the groups. So, within group could be Y i j minus Y in that group the mean of that group the sum of squares for I varying from 1 to the group size, this is 1 to n 1 to n and j varies from 1 to number of categories here.

So, we will use these relations, we will see how the calculations or the mechanism of ANOVA work in the following example. So, with this I will finish the part 1 of my analysis of variance session. So, I will come up with the second part, where I will take an example and do the calculations to conduct and analysis of variance.