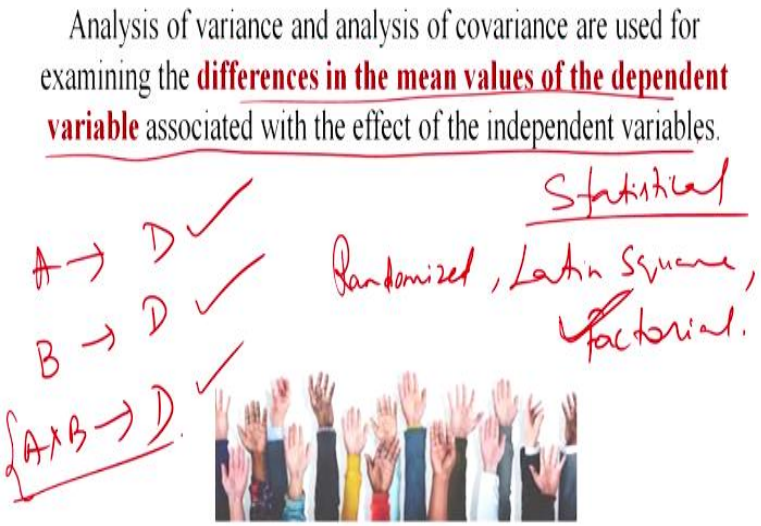
Marketing Research and Analysis-II
(Application Oriented)
Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 33
Introduction to ANOVA and ANCOVA

Welcome friends to the lecture series of the course of Marketing Research and Analysis. Today we will be discussing about the hypothesis testing on a different setup when the researcher is facing a condition of more than 2 groups. So first of all let us understand what do we mean by 2 groups and how do we test a hypothesis in such a condition? So in one of the lectures when we were talking about experimental design, if you remember we were talking about that in experimental design there are basically various kinds of designs among which you studied something called statistical design.

(Refer Slide Time: 01:09)



So what is this statistical design? So we talked about statistical design, under statistical design we talked about the randomized or randomized block design and then we talked about the Latin square design and then we talked about something called a factorial design. So these experimental designs are helpful when the researcher wants to know the effect of an independent variable on some of the dependent variable.

So statistical designs are very helpful especially in the case of factorial design. It helps you to understand the effects of 2 or more independent variables on the dependent variable along

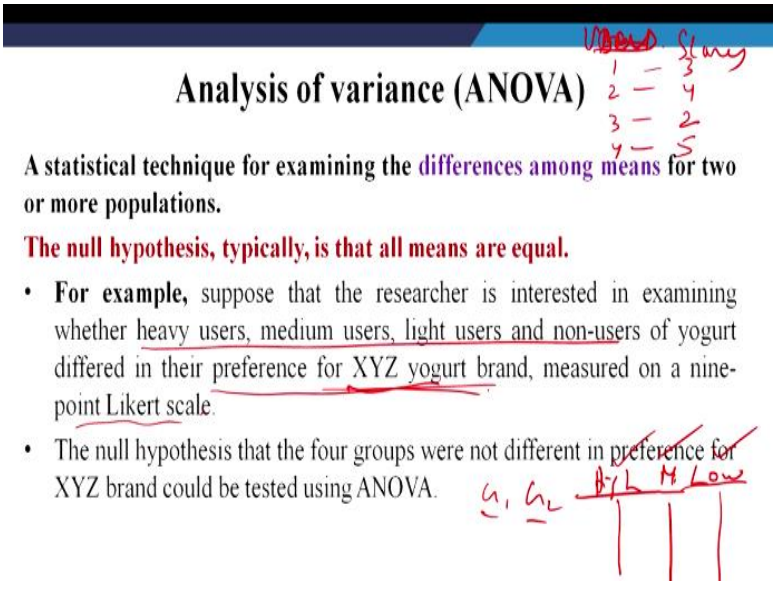
with the interaction effects. That means what? A has an effect on let us say D, B also has an effect on D, and A and B together also have an effect on D. So these are called the main effects and this is called my interaction effect. So let us understand the situation.

For example let us say when you produce a particular kind of chemical, it has to be produced at a particular temperature and pressure. So temperature has an effect on the final production of the material and pressure also has a direct effect on the material, and temperature and pressure together have an effect on the outcome of the final material. So these effects you cannot find out in generally.

So in such a study we use this kind of experimental design we talk about, it is called a factorial design which is a very easy and a very better method so that you can compare 2 groups or more and you can see the interaction effects in that. So let us understand. So how do we test this hypothesis and what are the basic requirements to go for it? So the first is called analysis of variance and analysis of covariance.

So what is this analysis of variance and covariance mean? These are the techniques which are used for examining the differences in the mean values of the dependent variable associated with the effect of the independent variables. That means the change in the dependent variable with the change in the independent variables is what we are trying to find out.

(Refer Slide Time: 03:52)



So if there are 2 groups for example let us say group 1 and group 2, then generally we test a hypothesis, 2 levels or 2 groups whatever you understand, 2 levels also you can understand.

So with 2 levels, we can say let us say high income and low income. So the effect of income, high income and low income, on let us say the tax payer's honesty, let us say how honest is the tax payer and how income effects it.

So we have taken 2 groups of people, let us say high and low, and we are trying to check whether this has got an effect. So when we are having 2 groups only, in such a condition we can test it to a simple t test, but then if I introduce a third group called the middle group or the middle income group, then I have 3 groups now. So when I have 3 groups, I have only a choice that I can do multiple t tests between high and medium, high and low, and medium and low, so 3 tests I can do.

So there is a problem. So this problem is actually because if you do multiple t tests, then what happens is the type I error gets inflated because every time you take a 5% significance level, so each time with each test so 5%, 5%, 5%, the test becomes more inflated the significance level and the type I error gets increased. This reduces the power of the test. So in such conditions, analysis of variance is the technique which is to be used.

So today we will be dealing this analysis of variance and I will try to explain how it is used for the test of hypothesis. So what it says here? A statistical technique for examining the differences among, means what 2 or more populations. So here we have taken the high income, low income, and the middle income. The null hypothesis is that all means are equal. For example suppose that the researcher is interested in examining whether heavy users, medium users, light users, and nonusers of yogurt.

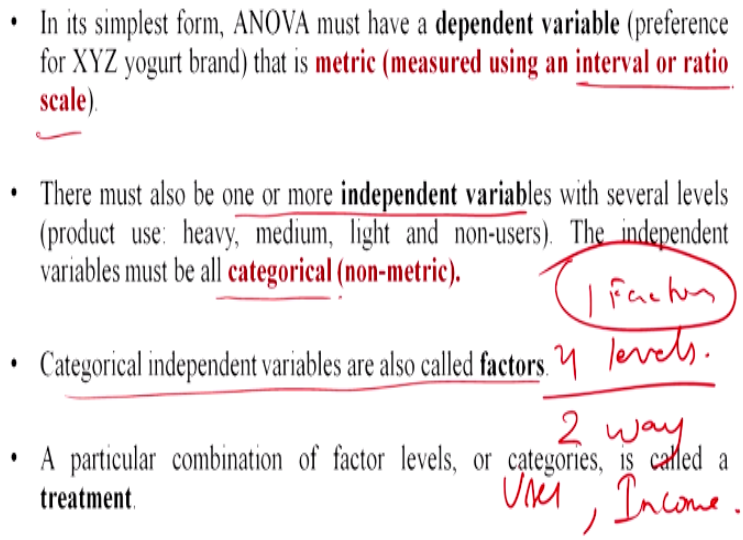
So that means there are 4 kinds of users. Somebody who consumes a lot of yogurt, somebody medium user, somebody light user, somebody who does not use at all. Is there any difference in the preference for the yogurt brand? Because there are several brands of yogurt, so do these users have a different taste of brand and does it depend on their usage. So it measured on let us say 9 point Likert scale.

So you know what is Likert scale, so it is an itemized scale where we are trying to study where each value has got some reference. So what it is saying? The null hypothesis that the 4 groups, in these 4 groups, were not different in having a preference for the XYZ brand and it

could be tested using ANOVA. So how does it look like? So let us say you have taken the score in a 1 to 9 scale. So let us say I am having 4 groups, so how I can make the table.

For example so I have 1, 2, 3, 4, there are 4 groups. So what is the group 1 given a score let us say 3, group 2 score 4, group 3 let us say 2, group 4, 5. So these scores that I gave and these are the brands, not the brands the users sorry. These are the users, so 4 types of users. So I am saying so particular yogurt brand you have taken XYZ and what score are these 4 groups of people, 4 users giving for this brand.

(Refer Slide Time: 07:49)



Now let us understand this. What is the simplest form it says, ANOVA must have one dependent variable. Now what is the dependent variable in this case? The taste or the preference for the yogurt brand. So the preference is the dependent variable and it is metric in nature, that means what it has to be measured in an interval or ratio scale, so a metric scale or a continuous scale it can be measured in an interval or a ratio. Now this is one thing.

Second thing is there must be one or more independent variables. Now what it says is basically when you want to see the effect on the dependent variable, so you have to have some independent variables, so at least you should have one independent variable or you could have more than one independent variable. So if you have one independent variable, we say it is a one-way ANOVA, but you have more than one independent variable, then we say it is a n-way ANOVA, two-way, three-way, n-way ANOVA.

So the independent variables must be all categorical. Now in this case how many groups we have, we have 4, now the heavy users, the middle, light users, the no users something we had right. So heavy, medium, light, and nonusers. So categorical independent variables are also called factors, so that is what I was saying. So how many factors do you have?:Let us say in this case you are only checking the user, user's interest, because it is only thing that you are testing that is the kind of user's effect on the brand.

It has only 1 factor is there which has 4 levels, so 1 factor 4 levels, so that is why it is called a one-way ANOVA. Had it been let us say the type of user and along with the income of the user we would have taken, then in that case we were trying to check not only the type of user's effect on the brands preference but also what income group he comes from, that also effects the preference for a particular brand.

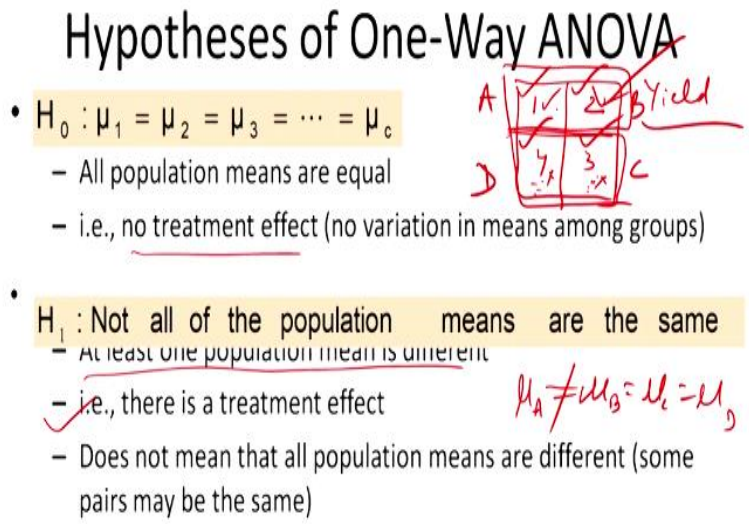
In that case, this would not have been a one-factor ANOVA, but it would have been a let us say a two-way ANOVA because there are 2 factors, one the user, the other is the income. A particular combination of factor levels or categories is called a treatment. So treatment is that what we are trying to check, what is the effect of the treatment? So when we give a treatment, if you remember when we did our experimental design.

The whole thing I got experimental design is about when you give a treatment to the variable what is the change in the dependent variable. For example let us say we are trying to take a bunch of students and trying to see whether a new method of teaching, may be through let us say some online mechanism or some kind of a new test that has emerged let us say or may be with the teaching done outside may be a in a garden, so if we you teach students in a garden does it have an effect?

So when we are trying to see so this is the treatment we are giving, what happens if we do not make any changes, so then it is a controlled group, but if we make a change, so what is the change? The change is that you are now removing the students from the classroom and you are taking to the outside into the open air and you are trying to teach them. So does this change have an effect on the student's ability to score? Now this is called a treatment.

So when you are giving a treatment, a treatment is for example again if you want to understand like when the farmer treats his soil with some fertilizer, so what is the effect of this fertilizer? So that is the treatment effect okay.

(Refer Slide Time: 11:55)



So hypothesis of a one-way ANOVA is that the null hypothesis, in this case there are 4 or let us say infinite, the $\mu 1 = \mu 2 = \mu 3 =$ goes on infinite. So the mean among all the levels there is no difference in the mean among the levels of the factor. So all population means are equal we are saying, there is no difference, so that is the null hypothesis, that is no treatment effect. That means suppose the farmer in this case, let us say this farmer has a small patch of land, now he breaks it up into 4 patches 1, 2, 3, 4.

So now he is trying to put a fertilizer here and let say here and tries to keep this as it is as. Now he wants to see whether the yield or the output that he derives from these 2 places and the output that he derives these 2 places where fertilizer is used, is there any difference or not? Now when he tries to do this study, so he says that if there is no difference, then the treatment effect is not there.

But if suppose here the yield increases or decreases whatever, then he can predict or he can say that because of the fertilizer the yield has changed. That means the treatment which is the fertilizer which I am giving in this case has got an effect on the final output which is the yield. What is the alternate hypothesis, not all of the populations means are the same. So at least one population mean is different.

That means if I would have divided this into 4 different patches and I would have put the fertilizers, so I am trying to see whether the fertilizer has got any effect or not. Now suppose I take the output from 4 different patches 1, 2, 3, 4. Now then I would say let us say there are 4 kind of fertilizers now I am using this case fertilizer A, fertilizer B, fertilizer C, fertilizer D. Now I am trying to see the effect of fertilizer on the output.

So my mean is my first null hypothesis is that the yield or the output that you derive by using fertilizer A = the yield from fertilizer B = the yield from fertilizer C = the yield from fertilizer D, that means the effect of the fertilizers is not there, all the fertilizers are giving the same output. So that is what it was saying, there was no treatment effect, but as a researcher you do not want to do that, you want actually to see that there is a difference.

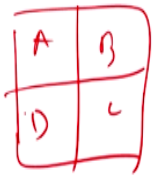
So if there is a difference what is the difference? Which one of them is giving the best yield or the poorest yield? So at least one population mean is different, so that means in this case μA, μB, μC in this case and μD, at least one of them should not be equal, at least, and that one which one you want to keep is up to the researcher, but he wants at least one of them should be different, then there is a treatment effect.

Then you can say well the kind of fertilizer has an effect. It does not mean that all population means are different, some pairs may be the same, that means may be in this case let us say μC that means C and D are giving the same result or even they may be equal to A also and only B is different or A and B are different C and D are the same, could be anything.

(Refer Slide Time: 15:36)

One-way analysis of variance

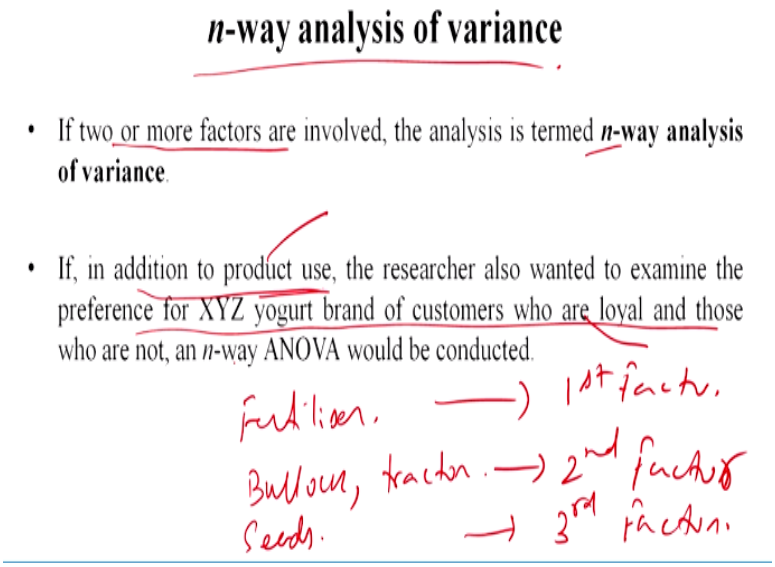
- **One-way analysis of variance** involves **only one categorical variable**, or a single factor. with several levels.
- The differences in preference of heavy users, medium users, light users and non-users would be examined by one-way ANOVA.
- In this, a treatment is the same as a factor level (medium users constitute a treatment).



So one-way analysis of variance involves only one categorical variable, what it says, only one categorical variable or factor with several levels, it should be several levels. The differences in preference of heavy users, medium users, light users, and nonusers would be examined by this one-way ANOVA. In this, a treatment is the same as a factor level, so in this fertilizer case what we said, there are 4 right A, B, C, and D. So there are 4 levels we are saying and we are trying to say that each has been given a different treatment, a different kind of fertilizer.

Now so we want to check the effect of this, whether they are giving the same result or they are giving different result okay.

(Refer Slide Time: 16:36)



Now what is the n-way analysis of variance? I explain. If 2 or more factors, for example now fertilizer you have taken in the last case, now you want to take also the along with the fertilizer you want to see the usage of technology, let us say whether the farmer is using just bullock or it is using some kind of a tractor, so what is he using. So the kind of technology being used also will have an impact on the yield, so this is the fertilizer this is first factor, and this is the second factor.

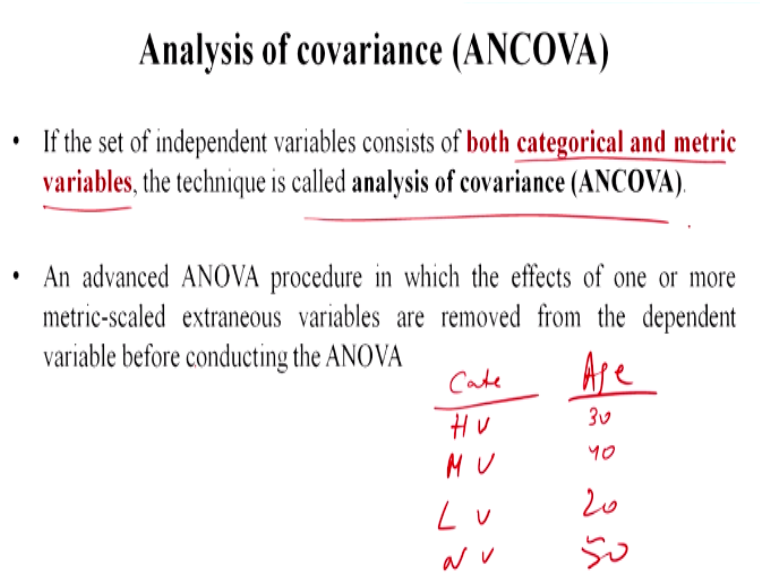
The second factor says that the kind of technology being used that will also have an effect on my dependent variable. So now in this condition, there are 2 things now the researcher is going to check, fertilizer and technology. So in this case since there are 2 factors, we say it is a two-way ANOVA right. Now suppose you would have the used the third case; fertilizer,

technology, and the seeds. Now he says the seeds are available from various brands, let us say now high quality, low quality, or medium quality.

Now when I am using this third factor, so the kind of seed, then I am saying it is a three-way ANOVA. So similarly you can go on increasing the number factors and that is why you term it as a n-way, so n stands for 1, 2, 3 whatever, number of factors. If in addition to product use, the researcher also wanted to examine the preference for XYZ brand of customers who are loyal, in this case you see that example, now along with the preference of the yogurt brand in addition to product usage, usage means heavy, light, medium that one.

If along with this, he is also interested to see the loyalty of the customer, in this way a n-way ANOVA could be conducted or in this particular it is a two-way ANOVA.

(Refer Slide Time: 18:57)



Now what is his analysis of covariance? Now many a times, people get confused and they get complexed, they get scared well what is this, it is very simple actually. Analysis of variance and covariance are more or less the same except that there is only one difference, what is the difference, you see. It says if the set of independent variables consist of both categorical and metric variables.

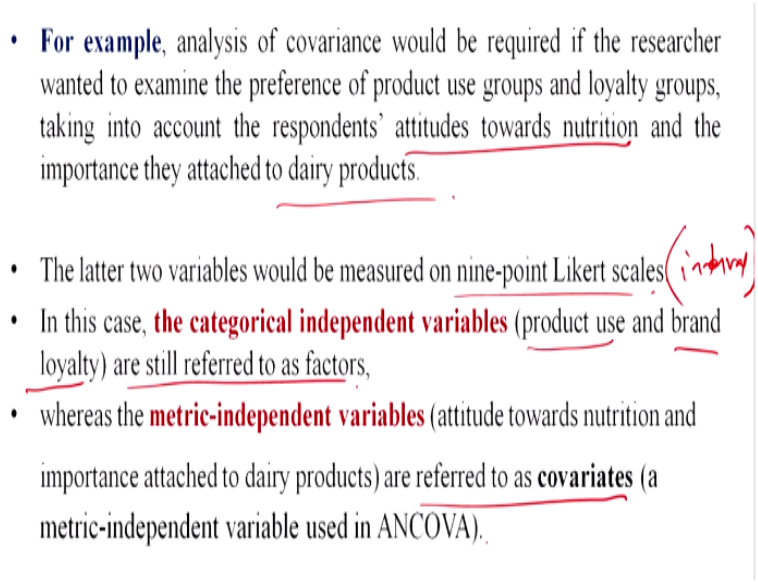
That means the number of factors that you are using the independent variables can be either both of them are categorical or one is categorical and other is metric, then this technique of analysis of covariance is used. That means what categorical variables we said let say

categorical in that case was heavy user, medium user, light user, no user and suppose you want to take another variable, let us say in that case the age.

Now age somebody is let us say 30, somebody 40, somebody 20, somebody 50. So when I am taking this age, this is a continuous variable, so there measured in a ratio scale, that means so in this condition we do not use a ANOVA, we will use analysis of covariance. An advanced ANOVA procedure in which the effects of one or more metric-scaled extraneous variable are removed from the dependent variable before conducing the ANOVA.

So it says that you try to, in simple if you understand that means there is one categorical variable and the other is a ratio-scaled continuous variable which is both taken together and we then find out the impact on the dependent variable.

(Refer Slide Time: 20:26)



Example of analysis of covariance. It would be required if the researcher wanted to examine the preference of product use groups and loyalty groups taking into account the respondents' attitude towards the nutrition and the importance they attached to dairy products. Now suppose you also measured what is the attitude of the respondents towards nutrition, how much value they gave to nutrition and thus how much importance they attach to the dairy products.

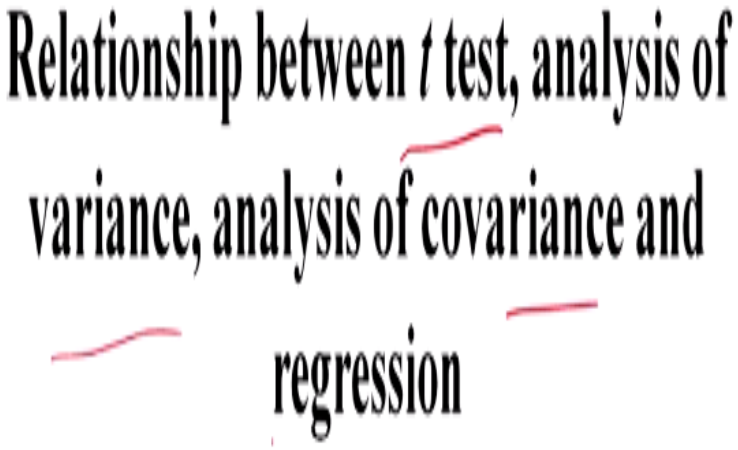
So this is something you have measured may be in a in a continuous scale, may be in an interval scale between 1 to 9 or the preference can be measured in several ways, so but it is a continuous scale or a let us say a metric scale. The latter 2 variables would be measured on 9-

point Likert scale, Likert scale is an interval scale, please remember. In this case, the categorical independent variable what are they, product use and brand loyalty, are still referred to as the factors.

Whereas the metric independent variable, which one the metric independent variable is the attitude towards the nutrition and importance to the dairy products. How much attachment they have for nutrition, how much value they give to nutrition, and how much do they love the dairy products are referred to as covariates. So this presence of a covariate actually is responsible for the development of the technical analysis of covariance.

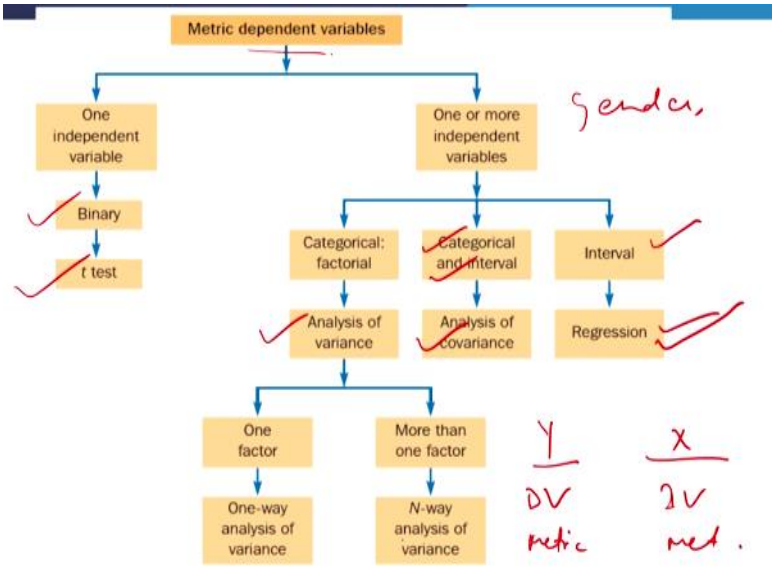
So a metric independent variable used in the ANCOVA, I think you are clear. So in this case, you have minimum one categorical independent variable and one metric independent variable. So presence of the metric and a categorical mix gives rise to the need of a ANCOVA technique.

(Refer Slide Time: 22:19)



Now let us understand the relationship between a t test, ANOVA, analysis of covariance and regression. So these all related, how they are related, we will see.

(Refer Slide Time: 22:33)



Now if you see when you want to measure the metric dependent variable, the dependent variable is metric that means a continuous or something, and you have one independent variable and that independent variable is in a form of let us say 2 levels which is categorical, say male female, high income low income, so when you take loyal not loyal, measured as 1 and 0. So when this is measured in a binary.

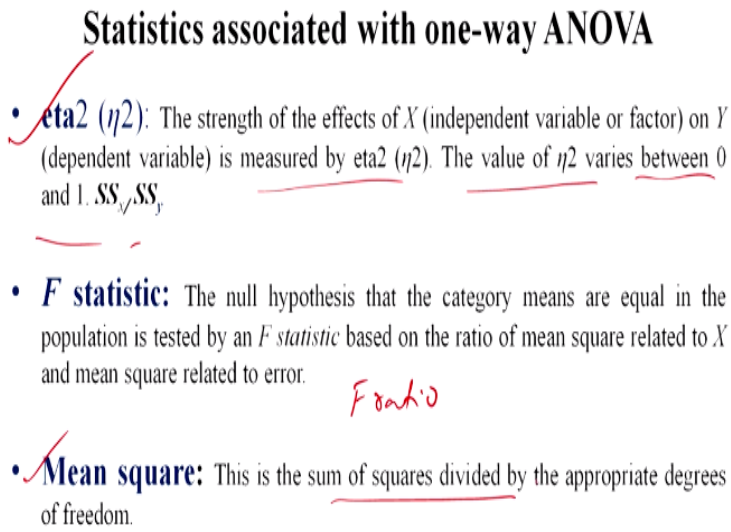
So when the independent variable is male or female, now I want to see the effect of or the impact of love for movies by male and female students, now how much do the male and female students have a love for movies. Now love for movies is my dependent variable and independent variable is whether it is a male or a female, the gender. So in such condition we said we use a t test, but when you have 1 or more independent variables, so now have not 1 independent variable, earlier you had the gender but now along with gender you have something else.

So when you have more than 1 independent variable what happens is, suppose it is a categorical one, then you have analysis of variance. Suppose you have more independent variables but they are all categorical, then you have a two-way ANOVA, three-way ANOVA whatever, but suppose it is a categorical and then an interval the next one, then analysis of covariance.

Suppose you have the independent variable measured in an interval scale or a continuous scale, then the technique is use a regression. That means Y is my dependent variable in a continuous measured in a metric and X suppose is also independent variable is also measured

in metric, then it is a case of a regression. Now coming to analysis of variance, one factor one-way analysis of variance, more than one factor we have n-way.

(Refer Slide Time: 24:49)

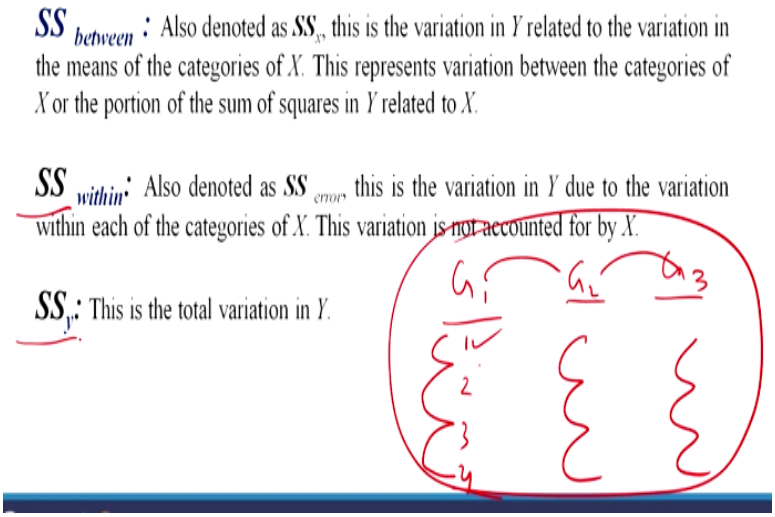


Now let us get into the one-way ANOVA. So some more statistics associated with one-way ANOVA. The first one is called eta square (η2). Now what is this eta square? The strength of the effect of X independent variables of X there is an independent variable or factor or Y is measured by this eta square. The value of eta square varies between 0 and 1. So it tells you the strength of the effect of X on Y, so how much is the strength of the independent variable on the dependent variable is measured by this term called η2.

F statistic or we say F ration f ratio, it is the null hypothesis that the category means are equal in the population is tested by an F ratio, what does it mean? So if my statistic is significant or not significant, then on basis of this ratio, I can say whether my means that means the μ1, μ2, μ3, μ4 in that earlier case which we discussed are either same or not same, at least one of them is different, if it is not significant, then we will say well there is some difference among the means.

But so if the F statistics is significant that means there is some difference among the means of the population, but if is not significant that means you cannot say that there is a difference and you would accept the null hypothesis. Mean square, this is another term that is used, it is sum of squares divided by the degrees of freedom, now what it is I will slowly explain.

(Refer Slide Time: 26:25)



Another term is sum of square between. Now there are 2 things in this when we talk about the ANOVA. So let us say there is a group 1, there is a group 2, there is a group 3. So what is the difference in the sum of squares, that means the among the values in the corresponding groups is called by sum of square between the groups. Now what is the sum of square within? The sum of square of within is the variables within.

So the difference in the sum of squares within the variables, within the group is called my sum of square within and is denoted as sum of square error also. This is the variation in Y due to the variation within each of the categories of X and this variation is not accounted for by X, that means it is not accounted by the independent variable, it happens within the variables, inside the independent variable.

Similarly sum of square of Y, that means the overall, the total sum of square of each value, each cell from value from the overall mean is called my sum of square of error. So these 3 things are very important before we understand and we go deep into the analysis of variance. So just I hope you have understood today that analysis of variance is a technique or analysis of covariance for that when you have more than one independent variable.

And this independent variables are measured either in categorical or a categorical plus non categorical or a continuous method. So in this case if it is both categorical or the independent variables are measured in categorical, then you would say it is an analysis of variance test, but if the independent variable one is in categorical the other is non-categorical, then you use a technique called ANCOVA.

So now we are continuing with the analysis of variance technique, and in the next lecture, we will solve some problems and go understand it deep how do you solve such a problem when you face it in the real life. So thank you for today. We will meet in the next class.