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

Lecture – 39
ANCOVA in SPSS

His friends, I welcome you all to the course Marketing Research and Analysis second part. In the last lecture, we were discussing about analysis of variance 1 factor and then we went ahead and discussed about n-way ANOVA where more than 1 factor is present, that means 1 factor is basically you can understand as 1 treatment, so the levels of 1 treatment or that means 1 independent variable which is categorical in nature and that independent variable or the treatment may have 2 or 3 levels, 2 levels and more than that.

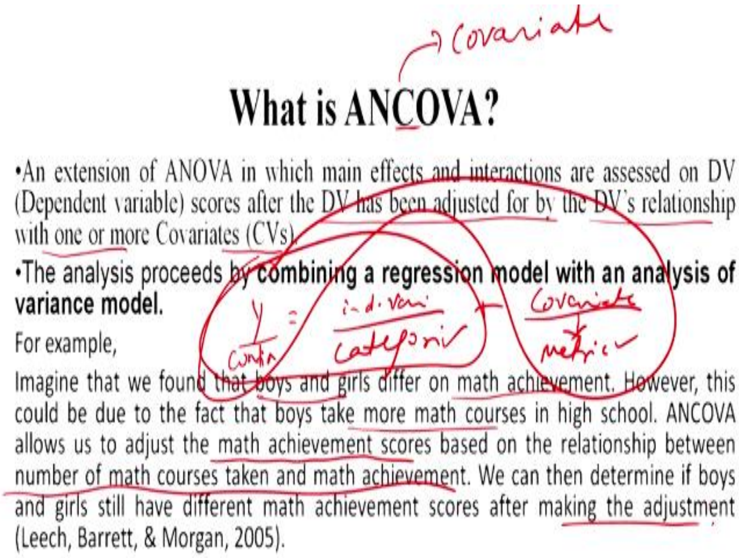
So in that condition we use a technique called the analysis of variance. When we have more than one factor that means 2 types of treatments given to see the effect on the dependent variable which is a metric variable continuous, measuring the ratio or interval scale, in that condition we use a two-way ANOVA or three-way ANOVA or four-way ANOVA, that is why it is called n-way ANOVA. As it goes on increasing the number of factors, independent variables category in nature as it increases, we say it is n-way ANOVA.

Then we discussed about the case of multivariate analysis of variance where we said what if there are more than one dependent variable, both measured in a continuous or metric manner. So in that condition, we said there are 2 possibilities that there is only 1 independent variable or 1 treatment with several levels or could be again 2 independent variables with 2 or 3 levels. So when we have 1 independent variable with 2 dependent variables, we said it is a one-way MANOVA.

But when we have more than 1 independent variable, all measured in categorical manner, then we said it is a two-way or more than that n-way MANOVA. So two-way MANOVA, three-way MANOVA again exactly, but what if in a condition now we want to control another important variable which is a confounding variable and its presence is affecting the dependent variable, the results that effects on the dependent variable.

So in such a condition when you have a case of a dependent variable and then an independent variable which is again categorical normally, but you have another variable which is measured in a continuous scale and this variable is now also part of the independent group and this effect the dependent variable, so when this happens, this is a case where we use a technique called analysis of covariance.

(Refer Slide Time: 03:30)



So I will explain what is each what is it mean basically. So let us start. ANCOVA as I said, ANCOVA is just nothing but an extension of the ANOVA, only the letter if you can see the letter C is attached here, which stands for a covariate. So what is this covariate? I will explain that. They are extension of an ANOVA in which the main effects and the interactions. Now I think by this time you must all be clear from my lectures what main effects means and what is an interaction effect.

So this main effect and interaction effects are assessed on the dependent variable scores after it has been adjusted for by the dependent variables relationship with 1 or more covariates. So let us now talk about only 1 covariate. So what is this covariate? A covariate is a part of the independent variable group, but this is not measured in a categorical scale, but it is measured in a continuous scale, that means it is measured in some interval or ratio scale. So the presence of this covariate makes an impact on the dependent variable.

This analysis proceeds by combining a regression model with an analysis of variance model, that means what? If you remember in the ANOVA what we said was Y my dependent variable is measured in a continuous manner or metric scale and my independent variable is categorical that is what we have said, but now since there is presence of a covariate, another independent variable which is a covariate, why it is called a covariate? Because it is measured in again a metric, a continuous manner. So this is categorical, this is continuous.

So when I am having 2 such independent variables and I am measuring effect on the dependent variable, so we can see that if we take only this much, it is like an ANOVA, is a case of ANOVA, but if I take this one along with this, if I take this Y and only the covariate which is again a continuous variable, then it is like a regression model, both are continuous, so both are metric, so that is what satisfies the case for a regression model. So what is it saying?

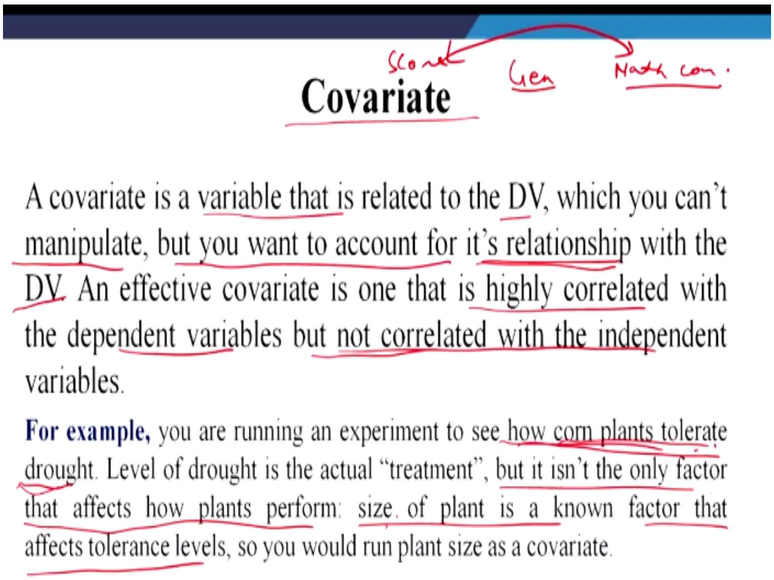
The analysis proceeds by combining a regression model with an analysis of variance model. For example imagine that we found that boys and girls differ on the math achievement or the score of math, how they perform in the math. However, this could be due to the fact that boys takes more math courses in high school, that means boys take more math courses due to pressure from family or maybe they are interested whatever, we have seen generally the boys are more interested into mathematics, the numbers, and girls into more creative side like arts, drama and all.

So may be boys take more math courses in comparison to girls. So ANCOVA allows us to adjust the math achievement scores based on the relationship between the number of math courses taken and math achievement. So if we do not take this for example let us say what would have happened? Had we taken the effect of gender on the math score, whatever you would have found might not have been justified or correct if you do not take this point how many number of courses that has a boy taken in comparison to the number of courses taken by a girl in the area of math.

So the more somebody has taken the courses, automatically we will infer that his understanding in the subject is higher than to the other boy or girl. So that is why the presence of this variable, the number of math courses taken, helps us to control and understand the effect, actually the right effect. So we can then determine if boys and girls still have different math achievement scores after making the adjustment for after the math courses that he has been taken.

So now we understand what is actually this ANCOVA. ANCOVA means we are finding an effect on the dependent variable with the help of the independent variable which is categorical, but we have a controlled variable which is the covariate basically measured in a continuous scale and by adjusting for this covariate we see what is the change in the dependent variable, that simple, that is exactly this is what we are saying.

(Refer Slide Time: 08:04)



So how do you define a covariate? It is a variable that is related to the dependent variable, it is related, so the math score achievement and the number of math courses he has taken are related which you cannot manipulate, you cannot manipulate, that is the data given to you, but you want to account for its relationship with the dependent variable, but although you cannot manipulate but you want to know whether this boy has taken how many courses and this girl has taken how many courses.

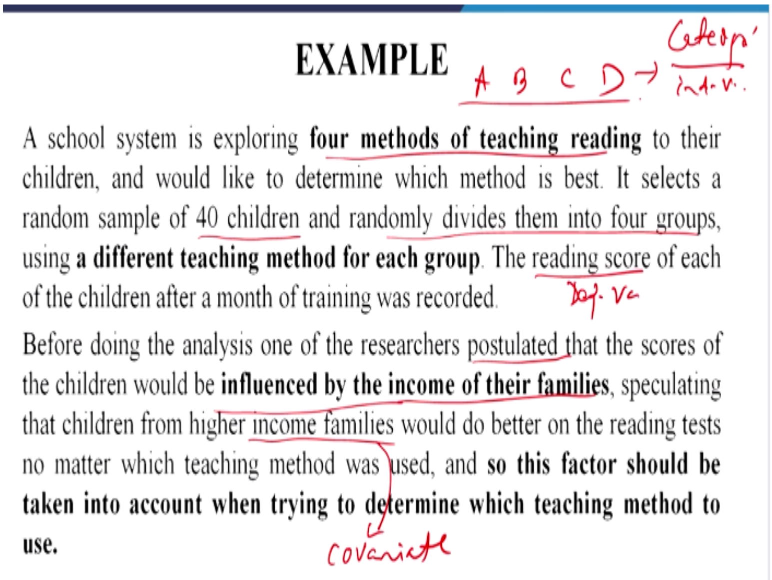
Because the number of courses they have taken that will have an effect or bearing on the dependent variable. So its relationship is important for us. An effective covariate is one that is highly correlated with the dependent variable but is not correlated with the independent variable, that means in this case the gender and the math course taken, logically they should not be related, but the score achieved in math and the number of course taken should have a strong correlation ship.

Let us take another example. You are running an experiment to see how corn plants tolerate drought? So we will take some dependent variable to measure that how the corn plants are tolerating a drought condition. Level of drought is the actual treatment. So level of drought suppose it is 1, not so much, 2 average, 3 severe drought, but it is not the only factor that affects how plants perform. So we say the scientists are thinking the level of drought is not only condition that will affect how the corn plants will tolerate the drought.

So then what is the other factor that affects? Let us see. So it says the size of the plant is a known factor that affects the tolerance level. So if a plant is only let us say few days old and if a plant is let us say a month old, so then the difference in the size in the plant in the same may be category will have an effect different effect on the dependent variable, that is how the plant tolerates the drought.

So here you see how the plant is tolerating the drought and the size of the plant will have high correlation, the dependent variable and the covariate will have high, but the level of drought and the size of plant that has got actually no correlation.

(Refer Slide Time: 10:42)



So now this is another example. A school is exploring 4 methods of teaching, so what is this 4 methods? Let us say A, B, C, and D to their children. So A, B, C, D are 4 methods, so what is this? This is a categorical variable. So this is my first independent variable, to their children and would like to determine which method is the best. It selects the random sample of 40 children and randomly divided them into 4 groups. So we have studied in factorial design that randomness is one of the most important criteria in any experimental design.

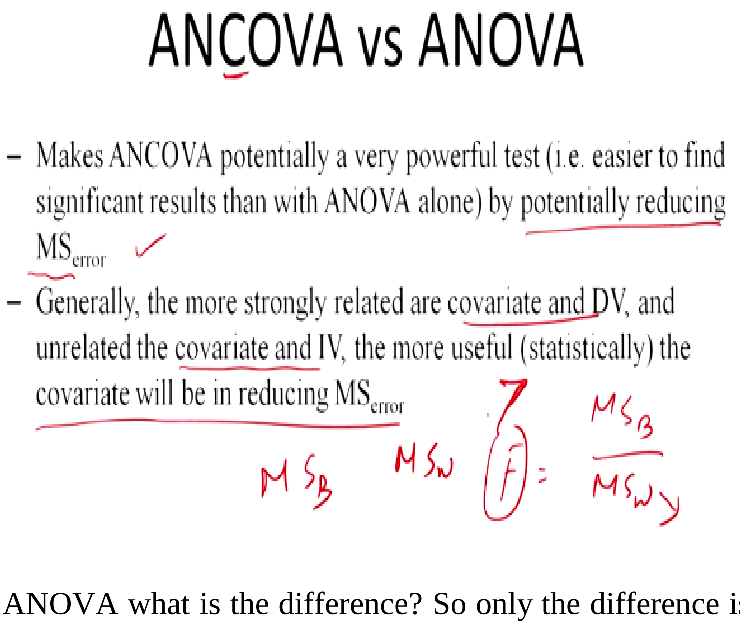
Using a different teaching method for each group, so 10, 10, 10, 10; 4 groups. The reading score of each of the children after 1 month of training was recorded. So first A, B, C, D is the

4 teaching methods of training and randomly 40 people were divided into this 4 groups, 1 month was spent on training and then the reading score was measured. Before doing the analysis, one of the researchers said or postulated thought that the scores of the children would be influenced by the income of their family, why?

How income of the families will affect? Now let us see. Speculating the children from higher income families would do better on the reading tests, no matter which teaching method was used, why because may be since the children from the higher income families have an access to more number of books and better educational facilities, so their reading ability, and their parents might be more educated than the rest, so their reading habits would be better in comparison to the ones who are coming from a poor income family.

So this factor should be taken into account when trying to determine which teaching method to use. So now you see this income is my covariate okay. So you have a covariate, you have an independent variable and what is my dependent variable? My reading score is my dependent variable okay.

(Refer Slide Time: 12:53)



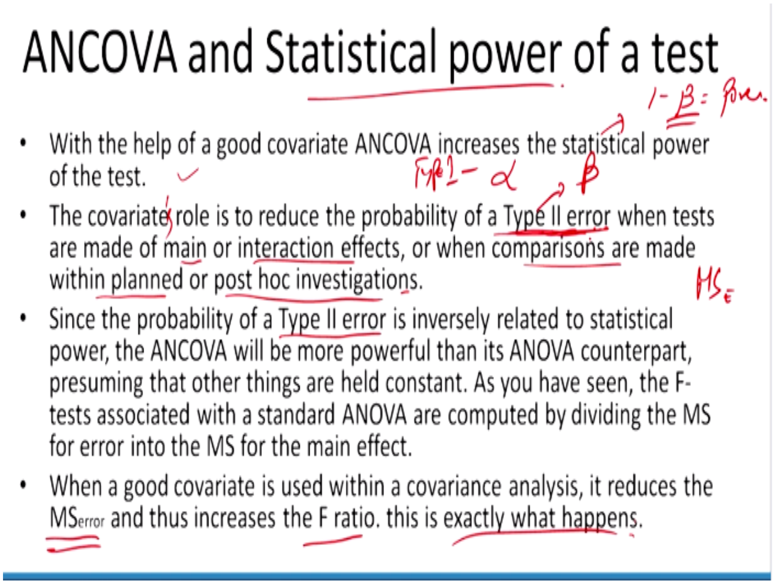
ANCOVA versus ANOVA what is the difference? So only the difference is we are saying C but let us see. Makes ANCOVA potentially a very powerful test. It is easier to find significant results than with ANOVA alone by potentially reducing the MS error. Now what is this MS error? I have already explained to you analysis of variance in which I have said to you 2 things, mean sum of square between the groups, mean sum of square within the groups or it is also called as error, mean sum of square error.

So it says ANCOVA if you take a proper covariate or you select your covariate properly, it helps you to potentially reduce the mean sum of square or within error in a significant manner. Generally, the more strongly related are the covariates and the dependent variable. Generally the covariate and the dependent variable are strongly correlated that is what I had explained earlier also and unrelated the covariate and the independent variable.

So the covariate and the independent variable are unrelated, not correlated, and the covariate and the dependent variable are but strongly correlated, so the more useful the covariate will be in reducing the MS error. So that means what in simple terms to understand if you select your covariate in the right way, then your statistical mean sum of square within or the internal within the group sum of square that error will reduce, and if the error reduces, you will only say what will have you think?

F ratio is nothing but M sum of square between/M sum of square within, the mean sum of square within. So if this is reducing, automatically my F value will go up.

(Refer Slide Time: 14:42)



So what is the relationship ANCOVA and statistical power of a test? With the help of a good covariate, ANCOVA increases the statistical power of the test, just I was explaining. The covariate’s role is to reduce the probability of a Type II error. Now do you remember Type II error? Type II and Type I error. Type I error is my alpha and Type II error is called beta. You remember that? Alpha is my Type I error, Type I error means when you have the chances of rejecting a true null hypothesis.

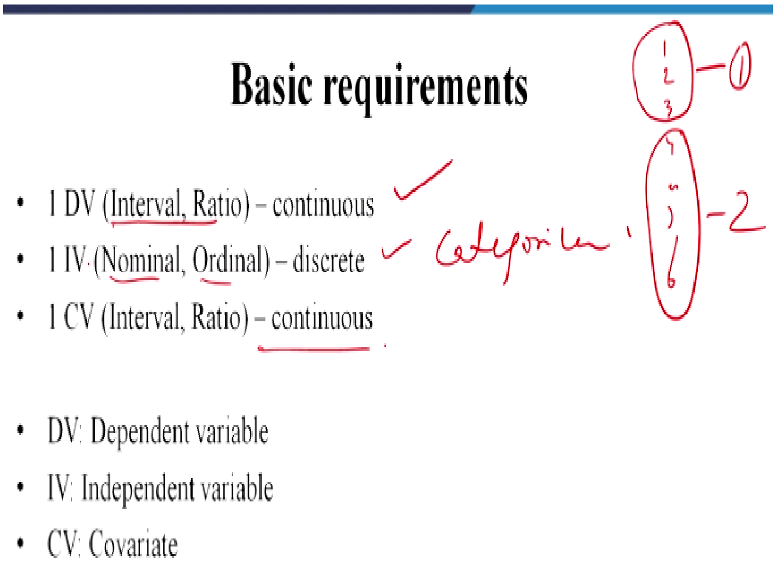
What is my beta, Type II? The chances of accepting a false null hypothesis. So that means a right covariate reduces the probability of a Type II error, that means it reduced the chances of accepting a false null hypothesis. So that means what? If you do this or this is called the power of a test. So 1-beta is called the power of a test. So if my beta is reducing, so what is happening? My power of the test is increasing, so the statistical power of a test increases.

So what it says is when it reduces the probability of Type II error when tests are made of main or interaction effects or when comparisons are made with planned or post hoc investigations. So understand that beta reduces with a right covariate. Since the probability of a Type II error is inversely related to the statistical power, the ANCOVA will be more powerful than its ANOVA counterpart presuming that other things are held constant.

So keeping all other things constant, the probability of Type II error is reducing, that means beta is reducing and the statistical power of a test is increasing. As you have seen the F test associated with a standard ANOVA are computed by dividing the MS for error, that is we say MS within error or within into the MS for the main effect. So when a good covariate is used within a covariance analysis, it reduces the mean sum of square error and thus increases the F ratio, this is exactly what happens.

So just understand this much when I am using ANCOVA, the ANCOVA is more powerful than ANOVA because it helps me in reducing my within error and thus it increases my F ration and my chance of significance or rejecting the null hypothesis increases.

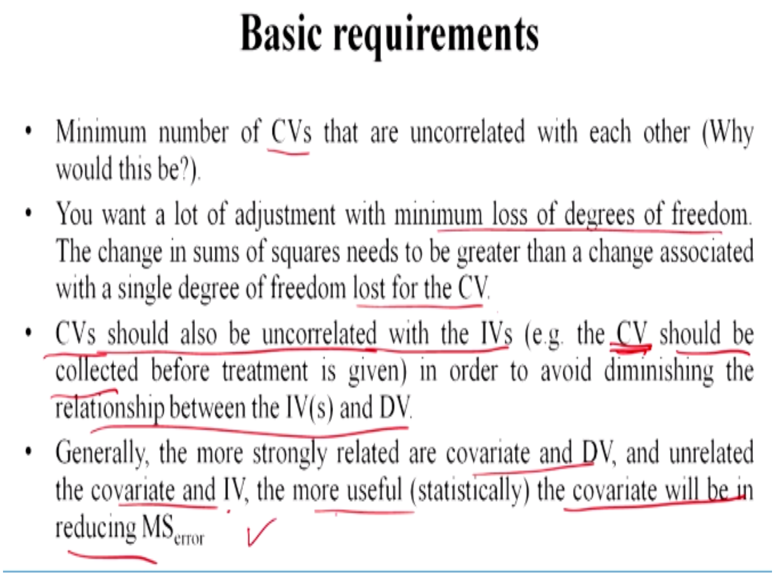
(Refer Slide Time: 17:31)



What are the basic requirements? I need one dependent variable which is measured in an interval or ratio, a continuous scale. I need one independent variable which is nominal or ordinal, so discrete or I say categorical. So if you do not have let us say independent variable which is let us say categorical in nature and you have a continuous scale also, you can convert that into a categorical scale. For example income you have the exact incomes of people so let us 1 lakh, 2 lakh, 3 lakh, 4 lakh, 5 lakh, 6 lakh.

So what I will do is I say this is one group 1, this is let us say another group. So now I have converted the same dependent continuous variable into a categorical variable, so that can also be done. One covariate I have which is again in a continuous. So this is just understanding what is dependent variable.

(Refer Slide Time: 18:29)



What is the basic requirement? The minimum number of covariates that are uncorrelated with each other. You should have minimum number of covariates and that should be uncorrelated, why because you want a lot of adjustment with the dependent variable with minimum loss of degrees of freedom. So the change in sum of square needs to be greater. The change in the sum of square needs to be greater than a change associated with the single degree of freedom lost for the covariate, as it is uncorrelated with the independent variable.

The covariate is uncorrelated with the independent variable, that means it says it should be, the example covariate should be collected before the treatment is given. If you collect the score, the covariate, what is the covariate in the math test? How many times the exams has been taken by the boys and how many times the exams have been taken by the girls, so if I collected this data beforehand, so then what it does?

It helps to avoid diminishing the relationship between the independent variable and the dependent variable. So that is it says CV should be also be uncorrelated with independent variables in order to avoid diminishing the relationship between the IV and the DV okay. Generally more strongly related are covariate and dependent variable. Covariate and the dependent variable the correlation is very high.

And unrelated covariate and the independent, the more useful the covariate will be in reducing MS error. It is more or less repeating the same thing which I have been telling the last time. So as you increase, as you select a proper covariate, the mean sum of square within or the within error reduces okay.

(Refer Slide Time: 20:20)

Assumptions

- Your **dependent variable** and **covariate variable(s)** should be measured on a **continuous** scale (i.e., they are measured at the **interval** or **ratio** level).

  Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

- Your **independent variable** should consist of two or more categorical, independent groups.

  Example: independent variables that meet this criterion include gender (e.g., two groups: male and female), ethnicity (e.g., three groups: Caucasian, African American and Hispanic), physical activity level (e.g., four groups: sedentary, low, moderate and high), profession (e.g., five groups: surgeon, doctor, nurse, dentist, therapist), and so forth.

Now coming to some of the assumptions. First of all when you do ANCOVA or ANOVA, for that all of these techniques, these are except the ANOVA, the others fall under GLM, general linear model methods because it encumbers a large number of statistical models into one linear equation. So what is it saying is the assumption is first of all there should be no missing data. If there is any missing data, you should correct for it.

So I have explained that in my earlier lectures how to correct for missing data, so you can make replacements. Then it says your dependent variable and covariate variable should be measured on a continuous scale, fine, no issues. Your independent variable should consist of 2 or more categorical independent groups, this is also fine, so there is nothing in this.

(Refer Slide Time: 21:14)

Assumptions (continued...)

- **Independence of observations,** which means that there is no relationship between the observations in each group or between the groups themselves.

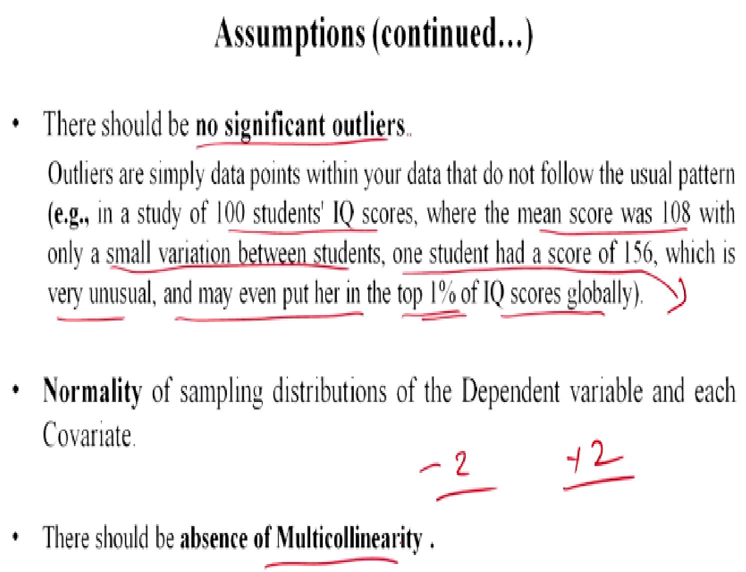
  For example, there must be different participants in each group with no participant being in more than one group. This is more of a study design issue than something you can test for, but it is an important assumption of a one-way ANCOVA.

  If your study fails this assumption, you will need to use another statistical test instead of a one-way ANCOVA (e.g., a repeated measures design).

Now coming to this point independence of observations. This was also applicable in ANOVA if you remember, which means that there is no relationship between the observations in each group or between the group themselves there is no relationship. So for example what it says? There must be different participants in each group with no participant being in more than one group, so this participant is not being repeated. If you want to repeat, then there is another kind of method which is called repeated measures ANOVA or repeated measures ANCOVA.

There is another method, this is more of a study design issue than something you can test for, but it is an important assumption of one-way ANCOVA. If your study fails this assumption, you will need to use another, this is what I was saying, statistical test instead of a one-way ANCOVA called the repeated measures design, repeated measures as the word suggests you are repeating the respondent, the sample.

(Refer Slide Time: 22:07)



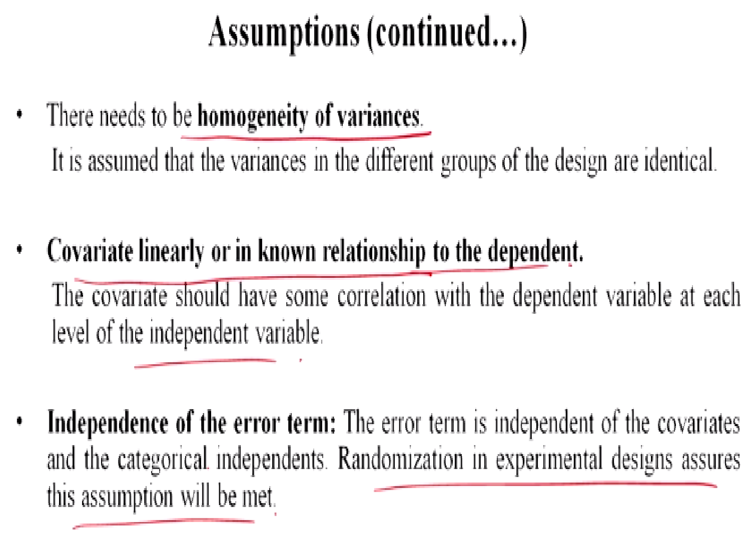
There should be no significant outliers. A slight difference is okay, but a significant outlier or a larger outlier, larger or too small whatever, can effect because you remember that these techniques are basically a test of means and proportions, so they get affected if there is presence of an outlier. So it says an example, outliers are simply data points within your data that do not follow the usual pattern.

In a study of 100 students IQ scores, the mean score was 108 with only a small variation between the students, 1 student had a score of 156 which is very unusual and may even put her in the top 1% of IQ scores globally. Now this is an outlier. So if this outlier is justified if really somebody has got it, he or she is a genius, so no issues, but generally such outliers will distort the whole statistics. Then normality, so we should ensure that our variables, the normality of the sampling distribution of the dependent variable and each covariate.

The dependent variable and the covariate should be following a normal distribution and you know how to check for normal distribution. You can do it through a plot, through a normal histogram also or a PP plot or you can measure it by calculating in SPSS you can go to explore and find out the Mardia's coefficient. So you can calculate the Mardia's coefficient, I have already shown, you can go back to my lectures earlier and you can check that and I said it should range between -2 to +2, so you can check back that.

There should be an absence of multicollinearity. If you have let us say more than 2 covariates, these 2 covariates or 3 covariates whatever, then there should be an absence of multicollinearity. If there is multicollinearity, the serious strong correlation, heavy correlation between the covariates, then it is again a problem.

(Refer Slide Time: 25:03)

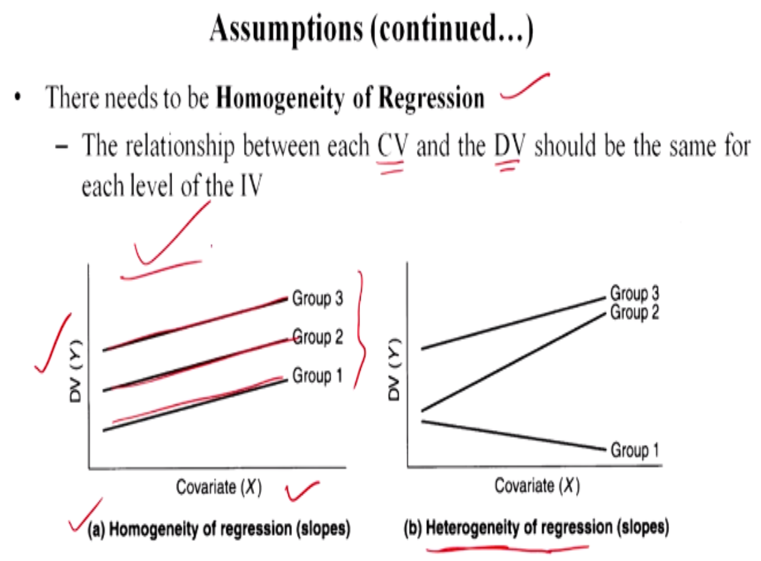


There needs to be homogeneity of variance. By now, I think you must have heard n number of times what homogeneity of variance means. Homogeneity of variance means that there has to be all the groups need to have an equal variance among within themselves, that means the variance in the group should be equal. That means if there are 5 groups, all the groups should have equal variance, that is what it says homogeneity of variance. If that is there, then only we should go ahead and compare the groups.

Another important thing is covariate linearly or in known relationship to the dependent, here there is something called covariate should have thus some correlation with the dependent variable at each level of the independent variable. So it says there has to be some correlation at least between the dependent and the covariate. Independence of the error term, the error term is independent of the covariates and the categorical independents.

Randomization in experimental assures this assumption will be met. So the error term is independent of the covariates and the categorical independent variables okay.

(Refer Slide Time: 25:32)



Next is, there needs to be homogeneity of regression. Now what is homogeneity of regression? The relationship between each covariant and the dependent variable should be the same for each level of IV. Now look at this. So this is dependent variable, this is my covariate, there are 3 levels. So if you see these 3 lines you can understand they are not meeting each other, they are not interacting at any point, so there is a homogeneity of regression.

But look at this, same condition, group 3, group 2, and group 1. So now this is not showing a homogeneity, rather this is a case of a heterogeneity of regression. So you should always be trying to see that the homogeneity of regression assumption is met. How to do that? I will explain.

(Refer Slide Time: 26:15)

Checking the assumptions for ANCOVA

Assumptions	How to check	What to do if the assumption is not met
Covariates should not be highly correlated (if using more than 1)	Check correlation before performing analysis. Use Analys e → Correlate → Bivariate and check that none of the covariates have high correlation values (r>0.8)	If there are some highly correlated covariates, one must select which covariates are of most importance and use those in the model.
Residuals should be normally distributed	Use the Save menu within GLM to request the standardised residuals for each subject to be added to the dataset and then use Analyze → Descriptive Statistics → Explore to produce histograms/ QQ plot / Shapiro Wilk tests of residuals.	If the residuals are very skewed, the results of the ANOVA are less reliable. One possible method of solving this issue is transformation of the dependent variable which may help with this assumption violation.
Homogeneity (equality) of variance: The variances (SD squared) should be similar for all the groups.	The Levene's test is carried out if the Homogeneity of variance test option is selected in the Options menu. If $p > 0.05$, equal variances can be assumed. If $p < 0.05$	The Levene's test is carried out if the Homogeneity of variance test option is selected in the Options menu. If $p > 0.05$, equal variances can be assumed. If $p < 0.05$, the results of the ANOVA are less reliable. One possibility it to transform the data.

This is how you check the assumption for an ANCOVA. So there are 3 assumptions. Covariate should not be highly correlated. So how to do that? Check the correlation before performing the analysis, check the correlation among the covariates. Go to SPSS, analyze, then correlate, bivariate correlation and check that none of the covariates have a correlation more than 0.81. What to do if the assumption is not met?

If there are some highly correlated covariates, one must select which covariates are of most importance and use only those in the model. Suppose you have 2-3 covariates and there is a strong correlation between let us say 1 and 3, then you may decide which one to keep and which one to avoid. Because if there is strong correlation, they actually mean the same thing, they have meaning close to the same thing.

Residuals should be normally distributed or the errors should be normally distributed. What does it mean? How to check? First use the save menu within GLM to request the standardized residuals for each subject. So when you go to the GLM model, there is an option for converting into standardize, so for each subject to be added to the dataset. So when you do a standardized, you standardize the scores, so the standardized values will be added to your dataset.

So simply if you want to do it, do it like this. Go to analyze, then go to descriptive statistics, go to explore, and in explore, you go for there is there is a provision for standardize course. I think descriptive also it is there, I will show you. So explore to produce histograms and there you can check whether the data is normal or not. If the residuals are very skewed, the results of the ANOVA are less reliable. One possible method of solving this is transformation of the data and that also we have covered in earlier classes.

So transformation, you can convert into inverse transformation, log transformation, a square transformation, a cubical transformation, any of these. Last, homogeneity of variance, the variance should be similar for all the groups. So Levene's test is the test which is carried out to check the homogeneity of variance. Now if your homogeneity of variance is violated, then the results of the ANOVA are less reliable. So in such a condition, the best thing is to transform the data. I hope this is all for the day.

So we will stop here, and in the coming class, we will further move ahead with ANCOVA and then we will also try to cover MANCOVA and I will also help you to explain how to do that with an example in the SPSS. So that will not only give you a theoretical clarity, but it will also help you to understand and utilize the techniques in your own research work. So for today, thank you so much, all the best.