Marketing Research and Analysis-II
(Application Oriented)

Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 38
MANOVA and ANCOVA in SPSS

Hello friends, welcome to the course Marketing Research and Analysis second part. So in the last lecture, we were discussing about the n-way ANOVA and we had started with MANOVA. So we said n-way ANOVA is the case where the researcher wants to examine the effect of any treatment with 2 or 3 levels. So if there is only 1, it is we say one treatment or one factor, then we say one-way ANOVA.

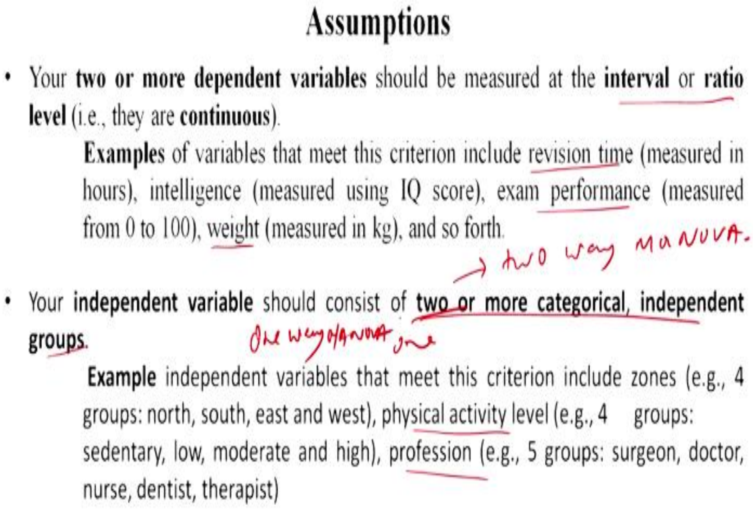
If there is more than one factor that means 2, 3 times of treatments with 2, 3 different levels, then we say it is a case of two or three factor ANOVA or that way we say it is a n-way ANOVA. There the major advantage is that when you have 2 treatments or 2 factors, the advantage is that it helps you to find out the interaction effect. So it not only tells you about absolute model, the overall model, but it also helps you to tell the main effects of the individual factors.

Suppose 2 factors are there, so the individual effect of the first factor, the individual effect of the second factor and then the interaction effect of the first and the second factor. So this interaction effect which is possible basically is a case of the factorial design which you have learned in the experimental research. So that was a case we solved 2 problems, one was in the primary method the other we solved through the formula.

Then we started with something a situation where the researcher is not having now 1 dependent variable but he has got more than 1 dependent variable and the number of independent variables can be 1 or 2. If it is 1, then it is okay, suppose 2 dependent variables are there and only 1 independent variable, then we say one-way MANOVAV or there are suppose 2 or more independent variables and 1 or more dependent variable, then we say it is a two-way MANOVA.

So MANOVA is a case where now the dependent variables have changed, are more than one and independent variables could be only 1 or more than 1 depending on whether it is a one-way MANOVA or two-way MANOVA.

(Refer Slide Time: 02:53)



Assumptions

- Your **two or more dependent variables** should be measured at the **interval** or **ratio level** (i.e., they are **continuous**).

  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.

  → two way MANOVA.

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

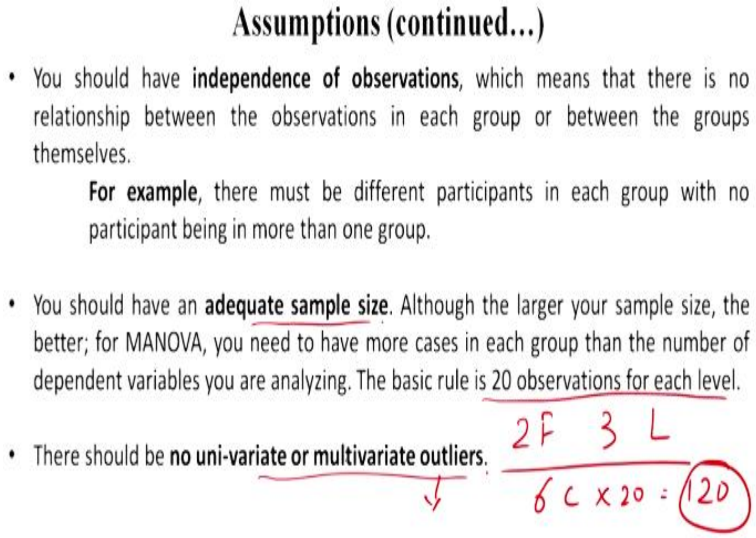
  One way MANOVA, one

  Example independent variables that meet this criterion include zones (e.g., 4 groups: north, south, east and west), physical activity level (e.g., 4 groups: sedentary, low, moderate and high), profession (e.g., 5 groups: surgeon, doctor, nurse, dentist, therapist)

So the assumption we had discussed here in the last lecture was that 2 or more dependent variable should be measured in an interval or ratio level. So these some are the revision time, exam performance, weight examples. Our independent variable should be 2 or more categorical independent groups. So independent variables could be for example I have said physical activity profession it could be.

It is not necessary 2 or more, 2 or more only when you are talking about two-way MANOVA. If it is only 1, let us say one categorical independent variable then with several levels, then it is a one-way MANOVA.

(Refer Slide Time: 03:37)



Assumptions (continued…)

- You should have **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.

- You should have an **adequate sample size**. Although the larger your sample size, the better; for MANOVA, you need to have more cases in each group than the number of dependent variables you are analyzing. The basic rule is 20 observations for each level.

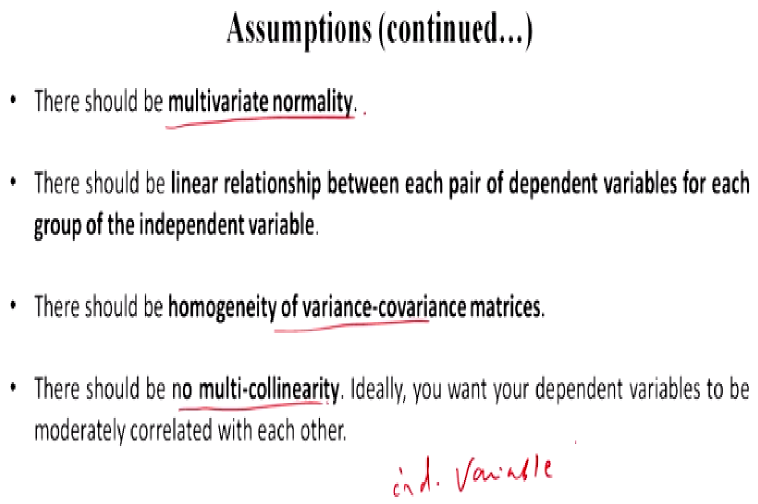
- There should be **no uni-variate or multivariate outliers**.

We said during the assumption that there should be independence of observations which means that there is no relationship between the observations in each group or between the groups that means different participation each group with no participation participants being in more than one group, that is a one of the assumptions. Sample size should be adequate. Now 20 observations for each level. Now what did I say or each cell you can say, understand it is not level, cell.

For example there are 2 factors and each factor has got let us say 3 levels, so many cells are formed? Now 6 cells are formed. So each cell multiplied by 20, so 120 should be your sample size. So this is one thing that you need to understand. There should be no univariate or multivariate outliers, now that is very important because you have understood from the very beginning that this is a test of means basically, so all kind of t test, z test, f test, MANOVA, ANOVA you talk about, they follow the means.

So when you follow the condition of mean, then if there is a presence of outlier, it will completely distort the whole statistical analysis. So you should ensure that there is no outlier.

(Refer Slide Time: 04:56)



Assumptions (continued...)

- There should be **multivariate normality**.
- There should be **linear relationship between each pair of dependent variables for each group of the independent variable**.
- There should be **homogeneity of variance-covariance matrices**.
- There should be **no multi-collinearity**. Ideally, you want your dependent variables to be moderately correlated with each other.

ind. Variable

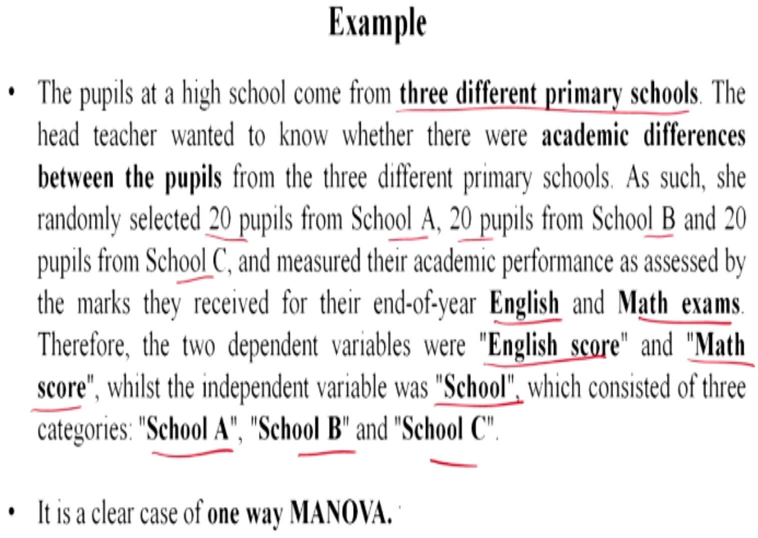
Next multivariate normality. Now normality is an assumption which is very essential for most of the statistical tests. If the data does not follow a normal distribution, then a parametric test is not possible and these are all parametric tests, the ANOVA, MANOVA and all. There should be linear relationship between each pair or dependent variable for each group of the independent variable, that means it says there should be some kind of correlation between the dependent and the independent variables.

So it says the linear relationship between each pair of dependent variables for each group of the independent variables, so there should be some relationship. Homogeneity of variance, now that has already we have said when there are several groups, these groups can only be compared when the groups have a similar kind of variance, otherwise they cannot be compared. Multicollinearity is another assumption which is required.

Ideally want your dependent variable to be moderately correlated with each other but no multicollinearity, that means multicollinearity happens when the independent variables are strongly correlated. So if they are very strongly correlated, sometimes the meaning happens that there is a redundancy of facts, unnecessary you have used more than suppose there are 3 independent variables out of which 2 are very, very highly correlated that means these 2 mean the same, so could have usually one of it and not 2.

Now how do we conduct this one-way MANOVA in SPSS? Because I had said earlier also MANOVA is very complicated and doing it in hand is beyond the class capacity at the moment, so it cannot be done. So whatever could be done you have shown it by hand, for example one way and all we have shown, two way, but MANOVA is very difficult, it is not possible. So we will show it how to do it by SPSS, one of the software packages. So this is the case, this is an example, let us read this example.

(Refer Slide Time: 07:08)



Example

- The pupils at a high school come from **three different primary schools**. The head teacher wanted to know whether there were **academic differences between the pupils** from the three different primary schools. As such, she randomly selected 20 pupils from School A, 20 pupils from School B and 20 pupils from School C, and measured their academic performance as assessed by the marks they received for their end-of-year **English** and **Math exams**. Therefore, the two dependent variables were "**English score**" and "**Math score**", whilst the independent variable was "**School**", which consisted of three categories: "**School A**", "**School B**" and "**School C**".
- It is a clear case of **one way MANOVA.**

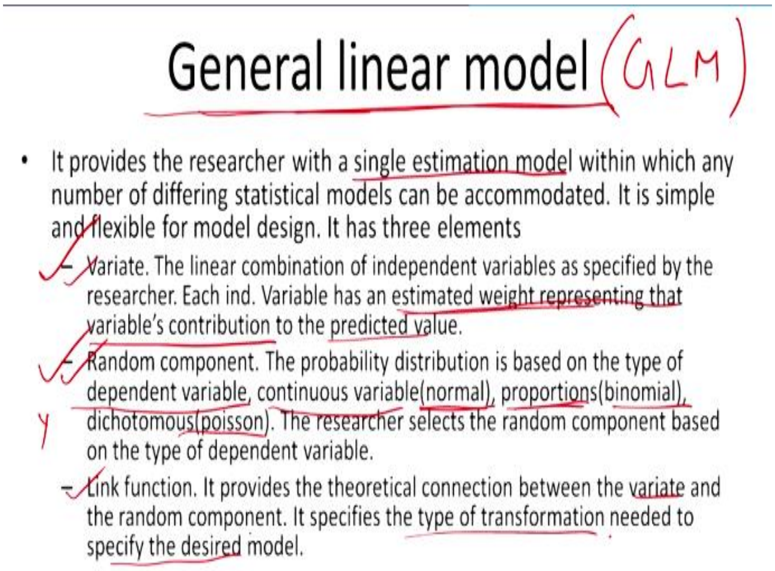
Students at a high school come from 3 different primary schools. The head teacher wanted to know whether there are academic differences between the pupils or the students form these 3 different primary schools. So that means there are 3 different primary schools, students have come from it passed out from it. Is there any academic significant difference between these students' results from passing out from these 3 schools.

As such, she selected 20 people from school A, 20 from B, and 20 from C, so 20, 20, 20 and measured their academic performance. So in this case if you can see, what is the independent variable which is playing important role on the dependent variable? Now the dependent variable is my score, the student has scored, and which school is coming from is my independent variable.

So the category is we will say categorical variable A, B, C okay with 3 levels, so academic performance measured by the marks, marks received for the end of year English and Maths exam. So 2 exams are conducted English and Maths, and what is the scores students from the school A got for English and Maths, students from school B got for English and Maths and similarly for C, what did they get? How much is the score?

So the 2 dependent variables were what? English score and Maths score, so these are the my 2 dependent variables while my independent variable is only school, so I have only 1 independent variable. So if I have only 1 independent variable, then it is a case of one-way MANOVA. So it consist of 3 categories A, B, C, so 3 levels.

(Refer Slide Time: 09:01)



General linear model (GLM)

- It provides the researcher with a single estimation model within which any number of differing statistical models can be accommodated. It is simple and flexible for model design. It has three elements
  - Variate. The linear combination of independent variables as specified by the researcher. Each ind. Variable has an estimated weight representing that variable's contribution to the predicted value.
  - Random component. The probability distribution is based on the type of dependent variable, continuous variable(normal), proportions(binomial), dichotomous(poisson). The researcher selects the random component based on the type of dependent variable.
  - Link function. It provides the theoretical connection between the variate and the random component. It specifies the type of transformation needed to specify the desired model.

Now when I talk about ANOVA, only ANOVA, it is a comparison of means, but when I talk about MANOVA or two-way ANOVA, ANCOVA, so here we use a model called the general linear model. Now what is this general linear model and why it is so much in demand? Now a days, why every scientist is talking about the general linear model, researches also evolved it has improved with time.

In the earlier days when Fisher developed the techniques of ANOVA and all these things in 70-80 years back, at that time, they did not have this advantage with them, they did not know some of the facts because everything evolves with time. So this is a new model that has developed, which says is named as general linear model or if you see softwares you will see GLM. What is this GLM? I will explain.

It provides the researcher with a single, this is the important thing, single estimation model within which any number of differing statistical models can be accommodated, that means if a different statistical models for example regression, test of means, they can all be accommodated at the same place through a linear equation. It is simple and flexible for model design and it has 3 elements which are very important.

What are these 3 elements? So the first is a variate. What is a variate? The linear combination of the independent variables as specified by the researcher. Each independent variable has an estimated weight, please remember this, representing that variable's contribution to the predicted value. So the first element is the variate. Now each variate says has an estimated weight which represents how much it is contributing to my dependent variable.

Second random component, this says the probability distribution of the variables is based on the type of my dependent variable in the case of the general linear model. Suppose it is a continuous variable which in case of MANOVA or ANOVA you are doing, then you follow a normal distribution. Suppose you would have had a proportions, you had used proportion, the ratio and proportion right, proportions, then I will use my binominal distribution, the only distribution would be binomial in nature.

But suppose I am using a method like dichotomous which is yes no kind of, so then I am using a poisson distribution. The researcher selects the random component based on what kind of dependent variable he is using. So general linear model helps you in these 2 things with a third thing, what is the third? What is the link function? Link means connection. It provides the theoretical connection between the variate and the random component.

It specifies the type of transformation needed, so you must have seen earlier in my initial classes, we talked about transformation of variables. Now when the variables are of not following a particular distribution and we wanted to follow, then we try to transform the data. So we use a logarithm transformation, a square transformation, a cubical transformation, we use various kinds of transformation, inverse transformation.

So here we say this link function helps you to do that. It specifies what kind of transformation is needed to specify the desired model. So log transformation is, logit log and kind of identity, so these are the kind of transformations you generally talk about. So general linear model is helping you to create a single model, linear model, which can accommodate several statistical tools.

(Video Starts: 13:24) How do you do it? Just to show you, many people might not be aware, so I am showing it. When you go to SPSS, there is see analyze, go to general linear model and now had you been doing a two-way ANOVA, suppose this case, then you would have done a, this is for a two-way or a n-way ANOVAM, this is n-way ANOVA, but we are doing a multivariate now, why? Because we have more than one dependent variable.

In this case what was happening, you have only one dependent variable and more than one independent variable. So more than one independent variable and only one dependent variable, so 2 or more you can say independent variable, but in this case we have more than 2 or more dependent variable and 1 or more independent variable, so this case is a case of general linear model multivariate condition. So next what we will do in this case?

In our case example, school was my, what was school independent variable, so school is my independent variable which comes into this category of fixed factor. Now English score and Maths score are my dependent variables, so I will take it here. We do not have a case of a covariate, I will explain what is a covariate. I had explained in one of the classes when I said about ANCOVA, we will do it later on in the next lecture we have ANCOVA we will explain that.

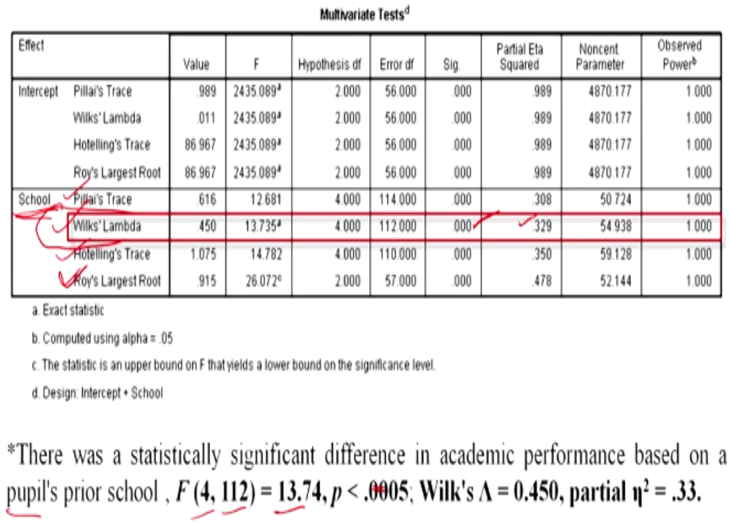
Then once you have done it, then you go to the next. So now we will do profile plots, you check the profile plots. Now first take the school. School is my independent variable, so I am taking it in the horizontal axis and I am continuing. So after I have done with it, then I am checking for a post hoc test. Now what is the post hoc? There are 2 things in this research. See whenever you have more than 2 groups let us say, your null hypothesis says there is muA = muB = muC.

Well suppose this hypothesis is my null hypothesis, what is my alternate, at least one of them is different correct. Now suppose my alternate has been accepted, my null has been rejected, so that means what? At least one of them is different. The question that comes to you is which one is different? So to do that, we do a post hoc test, we compares the means individually across each other.

So post hoc tests out of there are several methods, the most likeable method is a Tukey method, Bonferroni, LSD, so there are few methods but Tukey is the most applicable method. So there are 2 methods that i said, one you have the post hoc, the other is you have also a priori method, which is not there here. So the priori method also tells you to compare before the research is done. So let us talk about the post hoc only.

So once you have done all these, then you proceed with the steps. Now you see so we have marked Tukey and continue again. Now we will see the means, for what? The school. Why we are checking the means? you want to compare it and you want to see the main effects, but here since you have only one independent variable school, so what is the question of checking main and what is the question of checking the interaction, no interaction is possible, only one school is there.

So had there been school let us say the type of income group they are coming from, so 2 factors, then an interaction was required. So here we will only go for descriptive you see, estimates of size, observed power and this one homogeneity. So now again we click on the continue button **(Video Ends: 17:05).**
(Refer Slide Time: 17:10)



Multivariate Tests[d]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Intercept	Pillai's Trace	.989	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
	Wilks' Lambda	.011	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
	Hotelling's Trace	86.967	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
	Roy's Largest Root	86.967	2435.089[a]	2.000	56.000	.000	.989	4870.177	1.000
School	Pillai's Trace	.616	12.681	4.000	114.000	.000	.308	50.724	1.000
	Wilks' Lambda	.450	13.735[a]	4.000	112.000	.000	.329	54.938	1.000
	Hotelling's Trace	1.075	14.782	4.000	110.000	.000	.350	59.128	1.000
	Roy's Largest Root	.915	26.072[c]	2.000	57.000	.000	.478	52.144	1.000

a. Exact statistic

b. Computed using alpha = .05

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

d. Design: Intercept + School

*There was a statistically significant difference in academic performance based on a pupil's prior school , F **(4, 112) = 13.74,** p **< .0005**; **Wilk's Λ = 0.450, partial η² = .33.**

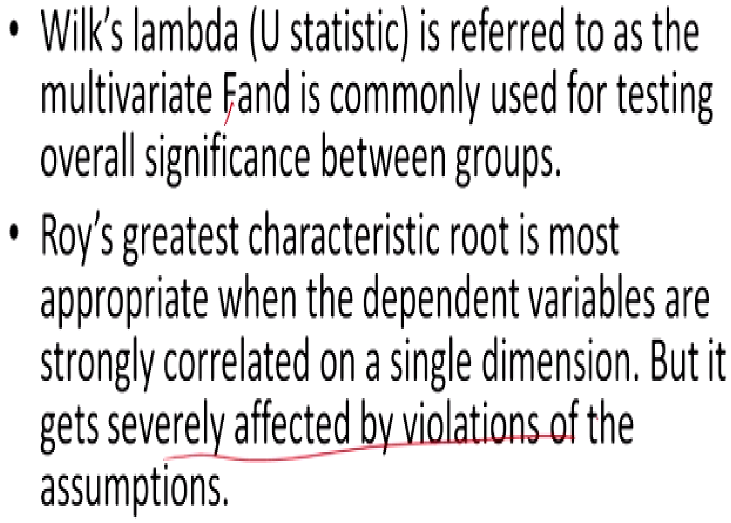
Now how do you report? Let us see. So this is very important. After this, I will show you also a problem on SPSS. How do you report? Now this is how you will get a table and here if you see the school, this is the independent variable we are talking about. So there are something called Pillai's Trace, Wilk's Lambda, Hotelling's Trace, Roy's Largest Root, out of which this is the one which we generally report in any research publication.

Roy's Largest Root is also good, but it is highly sensitive to normality, your basic assumptions of MANOVA, but Wilk's Lambda is more robust. So how do you write? Now if you look at it, it is significant. The eta square, now what is this eta square? What does it tell you? The eta square tells you the effect of size, the effect basically. So the higher it is, it is 0 to 1, so 1 means it has a large effect and 0 means absolutely no effect at all.

So when you write the research report, you should write like this. There was a significance in this case, significant difference in academic performance based on the students' or the pupil's prior school. How did I know? Now F = 4 and 112. Now 4 is my numerator 112 is my denominator. So these are my degrees of freedom is equal to this much and at p was checked at 0.0005, this is point 0.5 it should be, not 0005, so just strike, 0.05.

Wilk's Lambda is equal is 0.45, and eta square is point 0.33. So this is how you write the report for a MANOVA. Why I am telling this is many people are confused, they do the test but how do you report it? How do you write it? This they are not very clear.

(Refer Slide Time: 19:16)



- Wilk's lambda (U statistic) is referred to as the multivariate F and is commonly used for testing overall significance between groups.
- Roy's greatest characteristic root is most appropriate when the dependent variables are strongly correlated on a single dimension. But it gets severely affected by violations of the assumptions.

So Wilk's Lambda is referred to as the multivariate F and is commonly used for testing overall significance between groups. Roy's greatest characteristic root is also appropriate when the dependent variables are strongly correlated. If the dependent variables are very strongly or moderately correlated or highly correlated on a single dimension, but it gets severely affected by violations of assumptions, the problem is here, otherwise if your dependent variables are strongly correlated, then you should go for a Roy's greatest characteristics root.

Now what we will do is solve the problem on SPSS. **(Video starts: 19:51)** So two-way MANOVA. So this is the case we were talking about. So these are 2 scores and there are 3 schools. So let us first see and may be again we will go back to the slide. So first analyze, then I go next to what? I go to General Linear Model. So what will I go? Will I go to univariate or multivariate, I will go for multivariate, why? Because we have a case of MANOVA, where there more than 1 dependent variable.

So which is my fixed factor or my independent variable? Do you remember this, fixed factor means we are talking about the factors, the treatments basically. So what treatment is there? The school is the treatment in this case. Maths is my dependent variable, English is my dependent variable. So the score of Maths and English depends on which school does the student come from. Now I will go to plots as I showed you there and I add, continue.

Post hoc, I want to compare obviously between the 3 scores A, B, C and which one is the best out of it okay. Then I go to options, I go to schools, display means for school. See this is not coming, why? Because the interaction effect is not required here, only the main effect of school will be shown to you, nothing else right. So what you do is you take the estimates of effect size, the eta square, which is what you say the eta here, homogeneity, and observed power of the test.

What is the significance level? You may change it if you want, but here it is only 5%, now let us run it. So when I am doing it, I have three scores 15, 15, 15 candidates each. The question is should we have 15? It should have more right because there are 3 levels and I only said beforehand there should be 20 per each cell, so there are 3 levels, so it should be at least 60, but I have only 15.

So this is not adequate, but just to understand this is for a class this is to run this SPSS tutorial I am showing you, but when you do it, ensure that for each level you have 20, 20 cases, so there are 60 cases at least should be there. So Maths you see 1, 2, 3. So what is the school 1 the Maths average 74, 64, 64, overall average 67; English 67, 69, 65. Now what is this box's test of equality of covariance?

This as it says if you can see test the null hypothesis that have observed covariance matrices of the dependent variables equal across groups, so basically the box test tells you about the equality of the groups, whether there is significant difference or the groups are similar in nature. So it is basically significant means the null hypothesis is rejected, that means there is a significant difference among the groups, this is a very powerful test, this always should be reported also in your research paper.

Now coming to the effect multivariate test. Now we are interested for this part right. So let us go to the Wilk's Lambda, what is the value? 0.503. What is my F value? 8.404. What is my degree of freedom? 4 and what is my significance? The significance is 0.000. What is my effect size? 0.291. What is my observed power? 0.998 and it has been tested at 0.05. So first of all let us see the 2 dependent variables, Maths and English.

Actually in this case somehow the data what we have used because I have already said the adequate data was also not there, it was only 45, it could have been 60, so it is coming if you see both are significant, but actually our hypothesis says that Levene's test which talks about the homogeneity of variance across the groups should be actually the null hypothesis should be accepted, well all the groups have equal variance, but in this case that hypothesis is getting rejected.

So when you write you kindly write this in your report that in this, in our study, the error variance, the variance across the groups was not same or equal. Now coming to the test between subjects. Now if you look at the school, Maths, English. Type III sum of square is basically the Pearson sum of squares. So what is the finding, the significant school Maths score, there is a significant difference in the Maths score among the 3 schools.

There is also a significant difference in the English score among the 3 schools, but if you see the effect size is more in case for Maths than in comparison for English. The observed power that means the power of the test is much higher for Maths, that means the difference in the Maths score is more clearly evident than in comparison to the score for English among these 3 schools, the marks of the students of the 3 schools. So basically this is what we want to mention and then we will see a multiple comparison.

So let us see, is there any difference, yeah school 1 versus school 2, the mean difference 9.67, is it significant, yes, it is only for Maths; 1 and 3, yes there is significant difference; so and 1 and 2 and 1 and 3 is done. So, 2 and 3 is left, 2 and 3 is different? No. So the Maths score for the students of school 2 and school 3, we cannot say it is different. Now coming to the

English score, English score between school first and school 2, is there any significant difference? No; 1 and 3, no; 2 and 3, yes.

So if you see, this study says well in this case if you look at the mean also, so i-j, that means what i-j, that i is this one j, so 1-2 is 9.67, that means school 1 is higher than school 2, school 1 is higher than school 3 also. So school 1 has got the highest score and it is significantly different from the rest. But 2 and 3, if you see the difference between 2 and 3, so what is the difference? Only 0.47, where 3 is higher than 2 and that is not significant obviously.

So this applies same to 1 and 2 and in this case in English the school difference of 1 and 2 is, 2 has higher marks in English than school 1, and 1 and 3 if you compare, yes, 1 is higher than 3. So 2 is the highest, then I think 1, then 3. So this is all what you can find out from the case of MANOVA **(Video Ends: 27:25)**. So today what we have done is we have understood the multiple analysis of variance right in which I have explained what a MANOVA means and when it is to be used.

MANOVA is a technique which is the part of experimental design and very highly used by researchers both in engineering, medical, pure science, basic science, or even in the psychological labs. So here, the researcher can determine the effect of an independent variable one or more on several dependent variables and then can see the effect of these, which earlier we could not do it in other tests. So that is what MANOVA has a great advantage.

So I hope this class must have cleared, must have given you some insight, and even I explained how to write the report, the findings of the research work. So I am sure this class must have been of some value to you, you must have been slightly clear what MANOVA is and how MANOVA is to be written and how it is to be utilized, and the more you use it, I think you will get clear and clearer. So this is a basic understanding that I have tried to give you. Thank you very much and all the best.