Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 07

Lecture – 33

Analysis of Variance and Homogeneity of Variances with Univariate Data

Hello friends, welcome to the course multivariate procedure with R. So, you can recall that in the last 3 lectures, we have considered the test of hypothesis related to the mean from a normal population. So, first we consider the one sample test where there were two conditions that sigma square is known or sigma square is unknown. Then this one sample test was extended to two sample test. So, what was the difference? In the one sample test, we were interested only in the comparison of hypothetical value of the population mean. Then in the case of two sample test, we wanted to compare the population means of two populations on the basis of two samples of data.

Now can we extend it? We have done from 1 to 2. Now can you make it more than 2? In the last lecture, I have given you the examples that where would you like to use two sample test. Now in this lecture, we are going to extend it to more than two population that is we want to compare the population mean from various normal populations on the basis of samples of data means from every population we have got a sample. Now the next question is why should we do it, where it is going to be useful? Now in order to explain you, let me try to take a very simple example.

Think of a situation or think of a college where the students study from class 1 to say class 10. Now do you think if I want to understand what is the average age of the students who are studying in class 1, class 2, class 3, class 4, class 5, class 6, 7, 8, 9, 10, then what will happen? In case if I try to compare the ages of students studying in class 1 and class 10, do you think that are they going to be the same? Certainly not. There will be huge difference. So in case if I want to test H0: mu1 is equal to mu2 where mu1 is the age of the students in the class 1 population and mu2 is the age of the students in the class 10 population. They are not going to be the same.

They will be different. But on the other hand, if you try to compare the means of the age of the students in class 9 and class 10 or say class 1 and class 2, do you think that there is going to be really big difference in the sample values or in the population mean of these two populations? Maybe not. There is not much difference in the ages of the students studying in class 1 and class 2 or say class 9 or class 10. So now my question is this. If I want to compare the ages or the average ages of the populations which are coming from class 1, 2, 3, 4 up to 10.

So your null hypothesis which was earlier H0: mu1 is equal to mu2 will now be extended to H0: mu1 is equal to mu2 is equal to mu3 is equal to mu4 is equal to mu5 is equal to mu6 equal to mu7 equal to mu8 equal to mu9 and equal to mu10. So you are trying to compare the population means of 10 different population. So this is achieved by the test which is called as analysis of variance or briefly it is called as ANOVA, A-N-O-V-A ANOVA. And when we are trying to conduct such a test, then there is always an assumption about the population variance.

You can recall that in the last lecture when we considered the two sample test, we had assumed that the population variance of population 1 and 2, they are unknown but they are equal. Well, if they are not equal then you have a different procedure. So now in the case of analysis of variance, we are going to extend the two sample test to a more than two sample test. So the same assumption will continue here that we are interested in the null hypothesis H0 mu1 equal to mu2 up to mu10 but we are assuming that the variances of these population they are unknown but they are equal. So the question comes whenever you want to conduct such type of test of hypothesis on a real set of data, first you need to test the hypothesis about the equality of the variances or this is also called as we need to test the homogeneity of the variances that whether all the population variances they are the same or different and that is the same thing you have to do in a two sample test case also.

I mean first you have to test the hypothesis H0: sigma1 square equal to sigma2 square if that is accepted, then you go for the t test. If that is not accepted, then you have to go for some other procedure. So this is called here as a Bartlett test which conducts the test of hypothesis for the homogeneity of the variance or the equality of the variances. So in this lecture we are going to talk about two topics. One is analysis of variance and say another is Bartlett test.

I am not going to give you here much mathematical details but my main objective is to explain you the procedure and the main steps and after that I will try to illustrate it in the R software. So in order to understand the outcome of the R software, it is important for you to understand how they have been found and what are they indicating. So with this objective let us begin our lecture and try to understand analysis of variance and Bartlett test. Right, so in this lecture we are going to cover two topics analysis of variance and homogeneity of variances and yeah they are going to be handled only for the univariate data. So first we try to consider what is analysis of variance.

So this is actually the test of hypothesis for the equality of means when we have more than two population means. Right, so suppose we have more than two samples say k samples which are drawn independently from each of the population. So we call them as k independent sample and suppose the sample 1 is of size n1, sample 2 is of size n2 and in general I can say here that each sample is of size ni which is drawn from a normal population with mean mu i which is unknown and a common variance sigma square that means sigma square is unknown but they are equal for all the population and the samples are independent and their respective population variance are unknown but equal and we want to examine, we want to test if population means mu1, mu2, mu3 are different or they are the same. So this is achieved by the analysis of variance. So essentially we want to conduct the test of hypothesis H0 mu1 equal to mu2 equal to mu k and the alternative hypothesis here is at least one mu i is different from mu j, right, where i is not equal to j equal to from 1 to k.

So if any pair of the mu i is mu j is not equal then the hypothesis will be rejected and the next step after this will be to find out that which of the pair is responsible for the rejection of the hypothesis. So now if you try to understand what is the idea behind this analysis of variance. So first we try to understand it that why it is called as analysis of variance. This earlier it was called as one sample test, two sample test or but here it is called as the analysis of variance. So we understand that whenever we are dealing with the statistical tool we are dealing with the random variation, right.

So whenever an experiment is conducted the variable will certainly enter into the recorded data. Now this variability can be caused by two types of variation. One variation which is due to some assignable causes that we know and this variation can be explained. No issues. I have no problem once I know that this is the reason for the variability.

And second factor is random causes. This is the part which is which cannot be explained and due to the unexplained variation. So now when we are trying to develop any good statistical procedure we always want that yes variation is going to be there but I should know the reason and the random variation should be as small as possible, right. So we expect that in any good statistical procedure the explained variation is as high as possible and the unexplained variation which is due to the random factor is as low as possible. So if I can partition the total variation into two parts which is due to the explained factors and another which is due to the unexplained factors then possibly we can get an idea about the decision, right, whether they are going to provide us the good decision or bad decision.

So the next question is how to partition the total variance in the data into two such components. This is the job of analysis of variance tool and this is achieved through the analysis of variance. So essentially the analysis of variance partition the variability in two types of causes due to assignable causes which is the explained variation and due to the random causes which is the unexplained variation. Now the question is how to partition this total variance in the data into two such components which I can say broadly as between sample and within sample. So the variability will be now divided into two types of variability between sample variability and within sample variability and this is the job of the analysis of variance, right.

So in the case of analysis of variance which is popularly called as ANOVA, A-N-O-V-A, this works by comparison by comparing the between sample variability and within sample variability. What is between sample variability? The between sample variability is the variability in the sample means say x1 bar, x2 bar, xk bar because now you have got here k independent samples for each of the sample you can find out the sample mean. So you will have here the k sample means, k sample means and if you try to find out the variability in the values of k sample mean that is the between sample variability. And then we have within sample variability that means in each of the k sample there are going to be n1, n2, nk number of units. So they will also contribute to the variability and we can find out the variability within the sample.

So for example in the first sample there are n1 units, so n1 units will contribute to the within sample variability. So I can say that when the between sample variability that is here in the 1 is much larger than the within sample variability which is here in the 2, what does this indicates? That between sample variability is much much higher than the within sample variability. This indicates that the population means are not equal. For example if you try to see the if I try to take the same example that we have suppose some student from class 1 and we try to record their ages and there are some students from say class 10 and we try to record their ages. Do you think that the ages within this class they will be small? The values of ages within the class 10 they will also be small, but the value of the ages between class 1 and class 10 this is going to be larger.

And this is going to indicate that the average ages of the students in class 1 and class 10 they are not the same. So then that is what I mean to say that indicates that the population means are not equal or I can say that H1 is true or H0 is rejected. So this approach depends on comparing the within and between group variability. So that is why it is termed as analysis of variance, right. Variance is for variability.

Now how are you going to conduct this test? That is what I want to now explain you. So we would like to find out the total sample size. Sample sample size is the size total of say n1, n2, nk that is n1 + n2 + up to nk. And then for each of the sample which is based on the sample size ni we need to find out the sample mean x bar i and sample variance si square by these expressions, right. Sample mean is obtained by 1 upon ni summation xi and sample variance is obtained by 1 upon n - 1 summation xi - x bar i whole square.

And then we try to find out the mean of all the observation n1 + n2 + nk observation which is here n. So let x bar be the mean of all the observation from all the groups based on n observation. Now then we try to find out the between sample variability. So this is found as the variance of the k sample means weighted by sample size. And this express in terms of mean square treatment, right.

Or say this is because we are trying to use here the terminology between variability and within variability. So I am trying to indicate here as say by here MSBV, the mean square between variability. So it is just for the sake of understanding, right. And in general you will see that this is also called as mean square due to treatments. So it is the mean square due to between sample variability is defined by here SSBV divided by the degrees of freedom, right.

$$\boldsymbol{MSBV} = \frac{\boldsymbol{SSBV}}{\textbf{Degrees of freedom}(\boldsymbol{df1})} = \frac{\sum_{i=1}^{k} \boldsymbol{n_i(\bar{x}_i - \bar{x})^2}}{\boldsymbol{k-1}}$$

SSBV is computed by this expression which is the sum of squares due to between sample variability, right. So this is actually if you try to see this is x bar i - x bar. So it is trying to consider the variability in the sample means x bar 1, x bar 2, x bar here k. Next we try to find out the within sample variability. So this is found as the weighted mean of the sample variances which is expressed as the mean square error or shortly called as MSE.

$$\boldsymbol{SSBV = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}$$

So this MSE is defined as SSE divided by the degrees of freedom, right. So this SSE is actually here like this summation i goes from 1 to k and i - 1, Si square. This is the sum of square due to the random errors, right. And it gives us some idea about the variability in the observation due to the randomness in the values. And you can see here that the degrees of freedom here are here n - k.

$$MSE = \frac{SSE}{\textbf{Degrees of freedom } (df2)} = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k} \qquad \boldsymbol{SSE = \sum_{i=1}^{k}(n_i - 1)s_i^2}$$

I will try to show you that how these things actually come, right. So now after that we try to use some tools of statistical inference. Yes, I am not giving you here all the details and which is out of the view of this course also. But if you wish you can look into the books and you will get a complete idea. So now we have to find out the sum distribution of this MSBV and MSE based on that we are going to create a test statistics for conducting the test of hypothesis for H0, mu1 equal to mu2 equal to mu k, right.

So I am just giving you here the brief steps so that you can understand how different values are coming, right. So using the properties of normal distribution we try to compare the MSBV with MSE, right. And for that we take the ratio and that ratio will be called as F statistic, right. And in order to find out the F statistics we consider here the statistical properties that sum of square due to between variability divided by sigma square this follows a chi-square distribution with k - 1 degrees of freedom under H0. The sum of square due to error divided by sigma square follows chi-square distribution with n - k degrees of freedom under H0.

And both this SSBV and SSE are independently distributed. Now we try to take the ratio of this MSBV and MSE that is sum of squares divided by their respective degrees of freedom and this statistics will follow F distribution with k - 1 and n - k degrees of freedom under H0, right. And this degrees of freedom k - 1 this is indicated by df1 and n - k S here df2, right. So, the rule to get the hypothesis H0 mu1 equal to mu2 equal to mu k that is mu1 H0 is mu1 equal to mu2 equal to mu k is the same that reject H0 at alpha level of significance if p value is less than alpha, right. So now this MSBV is also known as the sum of square due to treatment and it is also indicated as say MSTR in various books.

$$\boldsymbol{F = \frac{MSBV}{MSE} \sim F(k-1, n-k)}$$

And this SSBV is also known as sum of square due to treatment and many books you will see that it is indicated by SSTR, sometime it is written as here like this also. And now this total sum of squares, total sum of squares which is here say here sum of all the observations or the total variability in the system that is divided into two parts, sum of square due to treatment and sum of square due to error or equivalently I can also write then total sum of squares is equal to say sum of square due to say BV + SSBV. So whatever you want to use they are the same thing. And these quantities are very conveniently displayed in the form of an ANOVA table, right. So you can see here, so you can see here this is a very simple table which is trying to compile all the information at one place.

So here the first column here is source of variation which is here between groups and within groups and within group is actually error and then we have here total. Then the next column is about degrees of freedom. So this SSBV upon sigma square that has got a chi-square distribution with k - 1 degrees of freedom. So this k - 1 is indicated here, right. SSE upon sigma square that has got a chi-square distribution with n - k degrees of freedom.

So this n - k is given here. Now similarly we can make a conclusion about the total also but then we need not to do it because these degrees of freedoms are additive. So if I try to add here k - 1 + n - k this will come out to be n - 1, right. So that is why I have not found here the different types of expression for the total. Now in the next column I am trying to give here the sum of square. So sum of squares due to between variation it is here indicated by SSBV or say we have also given the name as here SSTR.

Then we have obtained the sum of square due to error and now if you try to add here SSBV + SSE then you will get here total sum of square TSS. That is the reason that I have not given you here the expression for finding out the TSS, right. Now this additivity, this is due to the Fisher-Cochran theorem in statistics. So this partitioning of total sum of squares due to between variation and error it has been done orthogonally that means both these sum of squares are orthogonal to each other and then their chi-square distributions, their degrees of freedoms are also additive. So all these things they are ensured by the Fisher-Cockran theorem.

Well, I am not going into the details of these things. Now once you have obtained the say here sum of squares and the respective degrees of freedom then you can obtain here this mean squares due to between variation as SSBV divided by k - 1. So both these columns are used here. MSE is obtained here say SSE upon degrees of freedom n - k and now this both MSBV and MSE they are obtained here say MSBV upon MSE and this is here the f statistic, right, which is called here say f value. So this is here the analysis of variance table and you will see that in the software outcome also usually they try to indicate the outcome in the form of an analysis of variance table, right.

So let us try to take some examples and try to understand how to conduct this analysis of variance in the R software. Suppose we have a class of 20 students and based on that they have been graded and the grades are A, B and C which are given on the basis of the marks. For example, if there is a rule that somebody if somebody gets say more than 60 percent mark then the person is classified as first division. If between 45 and 60 percent mark then the person is classified as second division and so on. So similar in the class these 20 students marks have been converted into grades A, B and C.

So these marks are stored in the data vector here marks and the corresponding grades are stored in the data vector here grades which are obtained as factor, right. So factor is also a concept in the R software where we try to convert the numerical data into categorical variable which is popularly called as factor in the R software. So if you try to understand the data like this that the first student has got marks 20 and the person has got the grade C. Similarly the second person has got marks 10 and the second person also has got the grade C and so on, right. So now the question is that we want to know if the average marks in each of the group A, B, C are the same or different because you can see here that when we are trying to give the grades then there is certain range based on which we try to give the grades.

For example, here you can see here that both these students who have got marks 20 and 10 they have been given the grade C whereas if somebody has obtained the marks here 60 then it has got here the grade B, right. And another student who has got here the marks 40 that person has also got the grade here B, right. So this happens in all the class that the grading is done on the alphabets based on certain ranges of the mark. So now we want to understand whether the average marks in the group A, group B, group C or grade A, grade B, grade C they are the same or not. So in our software I will try to show you here two ways in which you can conduct the analysis of variant.

So first command in the R base package is aov(), Analysis of Variants. Then you have to give here the formula. I mean formula you have to specify these two variables for example here marks and grades because they are related in certain way. Because I will try to explain you. Then whatsoever be the data, here you have to give here the data frame in which the variables are specified in the formula.

And then you have the option of here projection should be written or not. Then here QR here, QR is for the logical flag that should the QR decomposition be written or not. Then we have here contrast, right. Contrast are some linear parametric function where the sum of the coefficient is equal to 0 and so on. But anyway I am not going into that much of detail and you can yourself can see, but I want to take care of the minimum commands to make you understand that how to conduct the analysis of variant.

So now the first step what you can do that if you want to understand whether there is a difference in the average marks in the groups A, B, C. The first option is that you can create here a box plot, group box plot. And if you try to recall we have learned these things in the beginning and I had told you that as we move further we will be using these things. So if you try to see here these are the box plots of groups A, B and C.

This is for here A, this is for here B and this is for here C. And you can see here that the average that is the median, this middle value is in these boxes is trying to give you the second quartile which is the median and you can see here that they are quite different. So if all these box plots are very, very close to each other like this one then possibly I can say that the average marks are quite the same. But anyway we try to conclude the same thing which we have concluded from the graphic through this quantitative method also. So now I try to use here the command aov() and you can see here that how I am trying to write down here the formula marks tilde grades.

And then after that you get here this type of outcome. So you can see here that these are here grades and these are here residuals. Residuals is something like random errors in simple way what we have understood. Different software will have different meaning or say different terminologies. Now we are trying to find out here sum of squares. So I can say here this is the sum of squares due to grades or in your terminology this is SSBV or this is SS treatment or SS you have written as SSTR.

Now after this what is here this value 2144.952 this is the sum of squares due to residuals. So this is your here SSE. And then in the second row we have the degrees of freedom which is related to the sum of squares due to between variability BB or SS treatment and this is here 27 degrees of freedom related to SSE. So in case if you want to find out here the TSS total sum of squares then this can be obtained as a 16362.548 + 2144.952 and the degrees of freedom of TSS will be 2 + 27. And now it is giving you here in this row here residual standard error which is 8.913061. So this is the value of here square root of say S square which is or the value of standard error that is value of the estimated value of the sigma. Now there is another line here estimated effects may be unbalanced. While we are not going into that much detail about the concepts of this analysis of variance so who can safely ignore this line.

And you can see here this is here the screenshot of the same thing. And now if you want to make it here more information here for example if you want to get about the value of statistics etc. then you can use here the command here summary and the same command here aov() masked tilde grades. So this command will give you this type of outcome here where you have here the results about the statistical inference for this different sum of squares. So you can see here this is here the degrees of freedom which is here like this. So where it is a line but I can show you it on the just here the screenshot.

I think later you can consider this screenshot which has a nice outcome. Then in the second column we have sum of squares. So this value here is the sum of squares due to grades. This is something like SSBV and this 2145 is the SSE. Now this now in the next column we have here mean square which is 8181 is obtained by here 16363 upon here 2 that is the SSBV divided by the corresponding degrees of freedom and then the 79 this is obtained as 2145 upon 27.

And then finally we have here f value. This is here 103. And after that we have here see here p value. This is your here p value. So now it is trying to give you here different option that okay these stars are giving significance code. But if you try to see here if you the p value here is something like 2.32 into 10 power of here - 13 very close to 0 and if you try to take here alpha is equal to 0.055 percent level of significance then you can very easily say that p value is less than alpha and your rules is that reject H0 when p value is less than alpha. So reject H0 means you are trying to have here mu A is equal to mu B is equal to mu C. That means that the means of the grades of group A, B and C they are not the same. This is rejected.

And it was also clear from the box plot also that they are very different. So let me try to show you this thing on the R console also so that you can get here more confidence. Right. So let me try to first enter this data in the R console and then now if I try to take here this command here same. So if I try to show you here the box plot so you can see here let me try to make this one go smaller. You can see here this is here the box plot and if I try to make it here this analysis of variance so you can see here you are going to get here the same outcome here like this.

And if you want to have the summary of these things so you can see here this is here like this summary of this analysis of variance. It is here like this. So you can see here it is not very difficult. The main thing here is that you need to know how to interpret it.

That is more important. The interpretation is more important than doing this analysis. That is a straightforward command. So now let me come back to our slides and now I would try to give you the idea about another command to conduct this analysis of variance. Well you can use any one of them. So another command to conduct this one way ANOVA, I will try to give you idea of what is one way.

It is oneway.test and so this is going to do the same thing what we have done in the command anova that it will try to test H0: mu1 equal to mu2 equal to mu k. And in this case the variances are not must heavily assumed to be equal. So that is the main difference between the two and the command here is oneway.test then you have to give the formula and data then subset then na.action, variance.equal to false or true. So this option is here. Then formula is as I told you that we have used the command here marks tilde grid like this.

So this LHS tilde RHS LHS gives the sample values and RHS gives the corresponding groups. And data is the optional matrix of the data frame in which you have to give the data. Subset is another optional vector which is trying to specify the subset of observation to be used and na.action is a function which indicate that what should happen when the data has missing values NA and var.equal that is the variance is equal or not that command can be used by specifying var. equal. This is a logical variable which can take the value true or false.

So now if you try to see on the same dataset I try to conduct this test. So we have a command here oneway.test marks tilde grid that is the most simple option. So it will give you here this type of option. So you can see here it is giving you here the value of statistics then the degrees of freedom of the numerator, degrees of freedom of the denominator and then here p values here right. So and then you can see here this is here the screenshot.

And here I have intentionally taken that variance.equal is equal to false. Well if you try to take it here variance.equal is equal to true then you will get the same result what you have got earlier. But I wanted to show you that what really happens when you try to make this option that the variance of different population are not the same that is based on the different type of approach. But that has some issues. One issue which you can observe here that the degrees of freedom here in the denominator should use here this here is coming out to be 14.

865 and the degrees of freedom which is 2.000. It is coming in the fraction. So we have certain procedure by which we try to approximate the degrees of freedom when the variances are not the same. But if you try to use here var.equal is equal to true then you will get the same outcome. But my objective was to show you that what really happens. So before I try to explain you about the butt list test let me try to show you these things on the R console also. So let me try to copy this command here and you can see here that it will give you here like this.

And if you try to use here the quickest command where you have an option var.equal is equal to false then you can see here this is the default. Both the options are coming out to be the same. But on the other hand if you try to use here the option here true then you can see here this won't happen and you are getting here the same option which you have obtained earlier. So that is why I have shown you this option over here so that you can yeah you should know how to conduct these things because in real life you will be facing different types of a scavenger and an issue. Now in the last let me try to very quickly give you information about the butt list test which is used for conducting the test of hypothesis for the equality of different variances.

For example, you have assumed in the analysis of variance that you have got k samples which are drawn independently from k different normal population which are having different means mu1, mu2, mu k but they are having variances like sigma1 square, sigma2 square, sigma k square. So in the usual analysis of variance you assume that all sigma1 square, sigma2 square, sigma k square they have got the same value but first you need to test. So that is why when you try to conduct the analysis of variance first you must conduct that test of hypothesis about the homogeneity of variance. If that is accepted then you try to conduct the analysis of variance with an option for example like var.equal is equal to true otherwise use the option var.equal is equal to false. So a basic assumption in this analysis of variance is that all the samples are come from a k normal population which are independent and they have got different means but the same variance.

So in order to check whether they have got the same variance or not we use the Bartlett test and it tests that if k samples have equal variances across the sample and this is called as homogeneity of variance. So basically we want to test the null hypothesis H0sigma1 square equal to sigma2 square equal to … up to sigma k square and my alternative is that at least one sigma i square is different from sigma j square for j0 equal to i, i goes from 1 to k. So now in this case the test statistics is here like this. Well it is only for your information. We are going to use the software and the pooled variance that we used in the case of two samples if you recall this can be computed in this case by SP square which is here like this.

$$T = \frac{(n-k)\ln s_p^2 - \sum_{i=1}^{k}(n_i - 1)\ln s_i^2}{1 + \left(\frac{1}{3(k-1)}\right)\left[\left\{\sum_{i=1}^{k}\frac{1}{n_i - 1}\right\} - \frac{1}{n-k}\right]}$$

The same definition has been extended and the pooled variance is defined as the weighted average of the group variances. So now we have here k group. So each of the group has got the variance ni – 1, si square. So that is now taken the average. So in order to conduct this Bartlett test on the data in the R software we have a command here bartlett.test all in lower case alphabets and then the formula here is given by the same as we did in the case of analysis of variance marks tilde grades.

And once you do it actually this is available in the base package of R. Once you try to do it, it will give you here the value of the Bartlett statistics here like this 1.1382 which has got 2 degrees of freedom and its p value here is 0.566. So now based on this p value you can take appropriate decision and then you can conduct the test of hypothesis with the assumption that whether your variance should be equal or not.

Now I am sure that you know about it and let me try to show you this Bartlett test in the R software. So you can see here which is here coming like this. So this is the same thing which I shown you on the slide. So now we come to an end to this lecture and you can see here that we have considered here the analysis of variance and Bartlett test for the for testing the homogeneity of the variances. Now you can see here the analysis of variance what I have considered here.

This is actually the one way analysis of variance. Well analysis of variance itself is a whole subject. So when we have only one factor which is affecting the outcome then we consider one way analysis of variance. When we have two factors then we try to consider two way analysis of variance etc. But definitely we are not going into the part of analysis of variance but we wanted to use the tool of analysis of variance for comparing the population from say more than two different populations. And now you have seen that we have considered earlier one sample test, two sample test for the equality of the for testing the equality of the population mean.

But now we have extended it to more than two populations. The next question comes here if your null hypothesis is rejected that means all mu1, mu2, mu k they are not the same. Then the next aspect is that how to find out that which of the this mu1, mu2, mu k they are same and which are different. So this is achieved by the multiple comparison test. So I am not going to cover here but this is for your information only and most of this multiple comparison test they are available in the R software. So I cannot cover here each and everything otherwise the whole this course will be based only for this analysis of variance but my very humble request to you all is that you please try to look into the books.

You have a concept of one way analysis of variance, two way analysis of variance, different types of multiple comparison test etc. And you simply have to just see what they are trying to do and you have to simply find out the command by which you can execute them in the R software. Now you can see the conducting of test of hypothesis is not difficult in the R software. You simply have to look into the value of the p that is p value but the more important part is which of the test statistics is going to be used whether it is Z test, P test or something else that is the point which you have to learn. So there are many many parameters different types of hypothesis which can be tested using different types of statistics but surely my scope here is limited and now my objective is how I can extend it to a multivariate setup.

So you please try to revise at least one sample test, two sample test and I have not covered here the confidence interval estimation because in the I have explained you how to construct the confidence interval estimation in the case of one sample, in the two sample also it can be done and in the analysis of variance also some of the multiple comparison test like Chiffre test, 2K test etc. they are based on the confidence interval. So but now I believe that you have sufficient background so that you can learn yourself also and I am sure that it would not be difficult for you and you will do a very good job. So try to take different data sets, try to see what is happening, try to expose them to the R software and try to analyse the outcome and try to understand how can you interpret it.

The better you can interpret it the better you are. So you try to practice it and I will see you in the next lecture till then goodbye. Thank you.